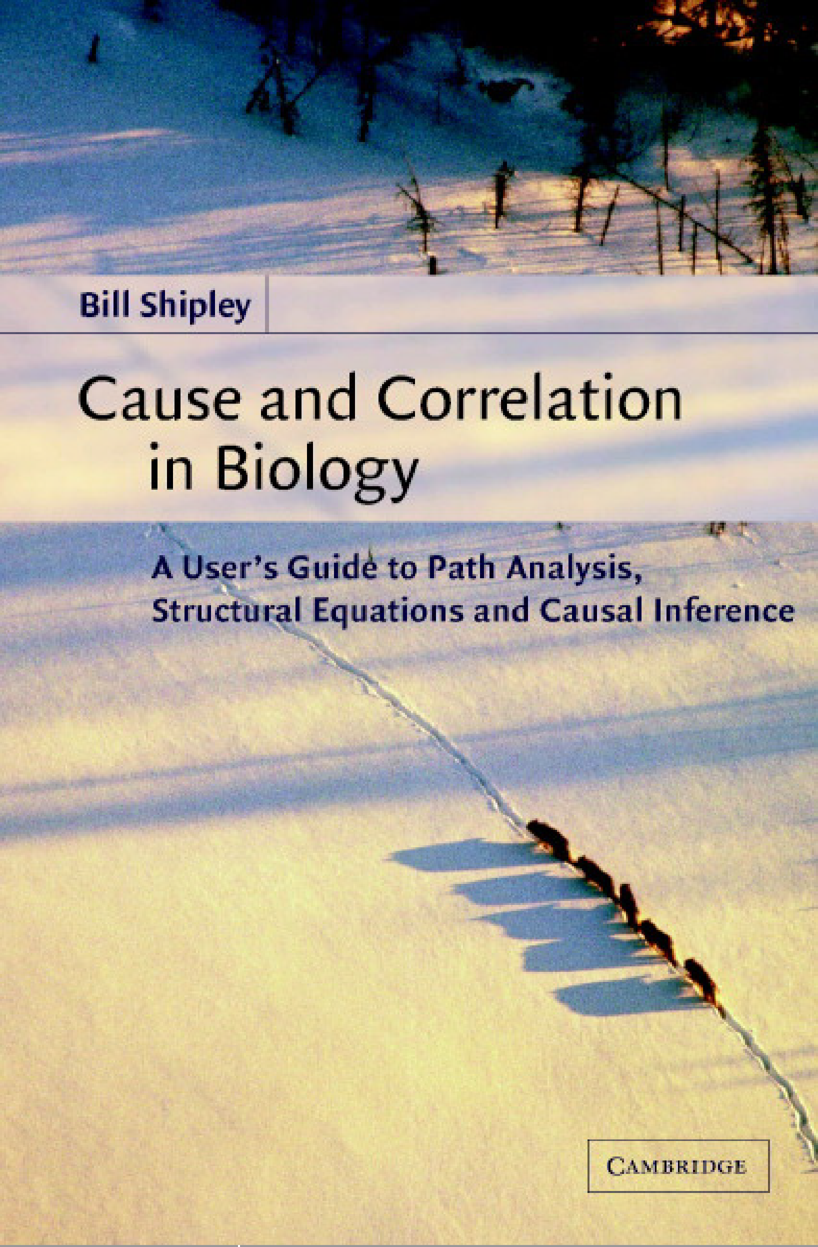


Bill Shipley

Cause and Correlation in Biology

A User's Guide to Path Analysis, Structural Equations and Causal Inference

CAMBRIDGE

CAMBRIDGE

more information - www.cambridge.org/0521791537

This page intentionally left blank

Cause and Correlation in Biology

A User's Guide to Path Analysis, Structural Equations and Causal Inference

This book goes beyond the truism that 'correlation does not imply causation' and explores the logical and methodological relationships between correlation and causation. It presents a series of statistical methods that can test, and potentially discover, cause–effect relationships between variables in situations in which it is not possible to conduct randomised or experimentally controlled experiments. Many of these methods are quite new and most are generally unknown to biologists. In addition to describing how to conduct these statistical tests, the book also puts the methods into historical context and explains when they can and cannot justifiably be used to test or discover causal claims. Written in a conversational style that minimises technical jargon, the book is aimed at practising biologists and advanced students, and assumes only a very basic knowledge of introductory statistics.

BILL SHIPLEY teaches plant ecology and biometry in the Department of Biology at the Université de Sherbrooke, Quebec, Canada. His present ecological research concentrates on comparative ecophysiology and the ways in which plant attributes interact to produce ecological outcomes. He has also contributed significantly to research in topics including plant competition, species richness and plant community ecology. His statistical research is equally diverse, covering such areas as permutation and bootstrap methods, path analysis, dynamic game theory and non-parametric regression smoothers. This rare combination of practical experience in both experimental science and statistical research makes him well positioned to communicate statistical methods to practising biologists in a meaningful way.

Cause and Correlation in Biology

A User's Guide to Path Analysis, Structural Equations and Causal Inference

BILL SHIPLEY

Université de Sherbrooke,
Sherbrooke (Qc) Canada

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published in printed format 2000

ISBN 0-511-01772-3 eBook (netLibrary)
ISBN 0-521-79153-7 hardback
ISBN 0-521-52921-2 paperback

À ma petite Rhinanthe, David et Élyse.

Contents

Preface

This book describes a series of statistical methods for testing causal hypotheses using observational data – but it is not a statistics book. It describes a series of algorithms, derived from research in Artificial Intelligence, that can discover causal relationships from observational data – but it is not a book about Artificial Intelligence. It describes the logical and philosophical relationships between causality and probability distributions – but it is certainly not a book about the philosophy of statistics. Rather it is a *user's guide*, written for biologists, whose purpose is to allow the practising biologist to make use of these important new developments when causal questions can't be answered with randomised experiments.

I have written the book assuming that you have no previous training in these methods. If you have taken an introductory statistics course – even if it was longer ago than you want to acknowledge – and have managed to hold on to some of the basic notions of sampling and hypothesis testing using statistics, then you should be able to understand the material in this book. I recommend that you read each chapter through in its entirety, even if you don't feel that you have mastered all of the notions. This will at least give you a general feeling for the goals and vocabulary of each chapter. You can then go back and pay closer attention to the details.

The book is addressed to biologists, mostly because I am a practising biologist myself, but I hope that it will also be of interest to statisticians, scientists in other fields and even philosophers of science. I have not written the book as a textbook simply because the discipline to which the material in this book naturally belongs does not yet exist. Whatever the name eventually given to this new discipline, I firmly believe that it will exist, and be generally recognised as a distinct discipline, in the future. The questions that this new discipline addresses, and the elegance of its results, are too important. None the less, the chapters follow a logical progression that would be well suited to an upper level undergraduate, or graduate, course. I have used the manuscript of this book for such a purpose and every one of my students is still alive.

It is a pleasure and an honour to acknowledge the many people who have contributed to this project. First, Jim and Marg Shipley started everything. Robert van Hulst supplied much of the initial impulse through our conversations about science and causality while I was still an undergraduate. He has also read every one of the manuscript chapters and suggested many useful changes. Paul Keddy kept my interest burning during my Ph.D. studies and also commented on the first two chapters. As usual, his comments went to the heart of the matter.

The late Robert Peters had a large impact on my thoughts about causality and even convinced me, for a number of years, that ecologists are best to give up on the concept – not because he viewed the notion of causality as meaningless (he never believed this despite his empiricist reputation) but because it was simply too slippery a notion to demonstrate without randomised experiments. His constant prodding must have caused me to stop while wandering through the library one day when, almost subconsciously, I saw a book with the following provocative title: *Discovering causal structure. Artificial intelligence, philosophy of science, and statistical modeling* (Glymour *et al.* 1987). That book was my introduction to a more sophisticated understanding of causality. Rob Peters was much too young when he passed away and I am sorry that he never read the book that you are about to begin. I am not sure that he would have approved of everything in it but I know that he would have appreciated the effort.

Martin Lechowicz introduced me to the notion of path analysis at a time when this method had been mostly forgotten by biologists. He and I have collaborated for a number of years on this topic and he read the entire manuscript of this book, providing many insightful comments. Steve Coté and Jim Grace also read parts of this book. Jim, in particular, provided some important counterpoint to my thoughts on latent variable models. Marco Festa-Bianchet provided the unpublished data that is reported in Chapter 5. I must also acknowledge my graduate students, Margaret McKenna, Driss Meziane, Jarceline Almeida-Cortez, Luc St-Pierre and Muhaymina Sari, as well as the many members of the SEMNET Internet discussion group.

Finally, I want to thank Judea Pearl for kindly responding to my many emails about d-separation and basis sets and to Clark Glymour, Richard Scheines and Peter Spirtes of Carnegie–Mellon University for their generosity in extending an invitation to visit with them and for patiently answering my many questions about their discovery algorithms. Clark Glymour read and commented on some of the manuscript chapters.

I hope that you find this book to be useful, interesting and readable. I welcome your comments and feedback. Especially, if you don't agree with me.

Bill Shipley

1 Preliminaries

1.1 The shadow's cause

The *Wayang Kulit* is an ancient theatrical art, practised in Malaysia and throughout much of the Orient. The stories are often about battles between good and evil, as told in the great Hindu epics. What the audience actually sees are not actors, nor even puppets, but rather the shadows of puppets projected onto a canvas screen. Behind the screen is a light. The puppet master creates the action by manipulating the puppets and props so that they will intercept the light and cast shadows. As these shadows dance across the screen the audience must deduce the story from these two-dimensional projections of the hidden three-dimensional objects. Shadows, however, can be ambiguous. In order to infer the three-dimensional action, the shadows must be detailed, with sharp contours, and they must be placed in context.

Biologists are unwitting participants in nature's Shadow Play. These shadows are cast when the causal processes in nature are intercepted by our measurements. Like the audience at the *Wayang Kulit*, the biologist cannot simply peek behind the screen and directly observe the actual causal processes. All that can be directly observed are the consequences of these processes in the form of complicated patterns of association and independence in the data. As with shadows, these correlational patterns are incomplete – and potentially ambiguous – projections of the original causal processes. As with shadows, we can infer much about the underlying causal processes if we can learn to study their details, sharpen their contours, and especially if we can study them in context.

Unfortunately, unlike the Puppet Master in a *Wayang Kulit*, who takes care to cast informative shadows, nature is indifferent to the correlational shadows that it casts. This is the main reason why researchers go to such extraordinary lengths to randomise treatment allocations and to control variables. These methods, when they can be properly done, simplify the correlational shadows to manageable patterns that can be more easily mapped to the underlying causal processes.

It is uncomfortably true, although rarely admitted in statistics texts, that many important areas of science are stubbornly impervious to experimental designs based on randomisation of treatments to experimental units. Historically, the response to this embarrassing problem has been to either ignore it or to banish the very notion of causality from the language and to claim that the shadows dancing on the screen are all that exists. Ignoring a problem doesn't make it go away and defining a problem out of existence doesn't make it so. We need to know what we can safely infer about causes from their observational shadows, what we can't infer, and the degree of ambiguity that remains.

I wrote this book to introduce biologists to some very recent, and intellectually elegant, methods that help in the difficult task of inferring causes from observational data. Some of these methods, for instance structural equations modelling (SEM), are well known to researchers in other fields, although largely unknown to biologists. Other methods, for instance those based on causal graphs, are unknown to almost everyone but a small community of researchers. These methods help both to test pre-specified causal hypotheses and to discover potentially useful hypotheses concerning causal structures.

This book has three objectives. First, it was written to convince biologists that inferring causes without randomised experiments is possible. If you are a typical reader then you are already more than a little sceptical. For this reason I devote the first two chapters to explaining why these methods are justified. The second objective is to produce a user's guide, devoid of as much jargon as possible, that explains how to use and interpret these methods. The third objective is to exemplify these methods using biological examples, taken mostly from my own research and from that of my students. Since I am an organismal biologist whose research deals primarily with plant physiological ecology, most of the examples will be from this area, but the extensions to other fields of biology should be obvious.

I came to these ideas unwillingly. In fact, I find myself in the embarrassing position of having publicly claimed that inferring causes without randomisation and experimental control is probably impossible and, if possible, is not to be recommended (Shipley and Peters 1990). I had expressed such an opinion in the context of determining how the different traits of an organism interact as a causal system. I will return to this theme repeatedly in this book because it is so basic to biology[1] and yet is completely unamen-

[1] This is also the problem that inspired Sewall Wright, one the most influential evolutionary biologists of the twentieth century, the inventor of path analysis, and the intellectual grandparent of the methods described in this book. The history of path analysis is explored in more detail in Chapter 3.

able to the one method that most modern biologists and statisticians would accept as providing convincing evidence of a causal relationship: the randomised experiment. However, even as I advanced the arguments in Shipley and Peters (1990), I was dissatisfied with the consequences that such arguments entailed. I was also uncomfortably aware of the logical weakness of such arguments; the fact that I did not know of any provably correct way of inferring causation without the randomised experiment does not mean that such a method can't exist. In my defence, I could point out that I was saying nothing original; such an opinion was (and still is) the position of most statisticians and biologists. This view is summed up in the mantra that is learnt by almost every student who has ever taken an elementary course in statistics: *correlation does not imply causation*.

In fact, with few exceptions[2], correlation does imply causation. If we observe a systematic relationship between two variables, and we have ruled out the likelihood that this is simply due to a random coincidence, then *something* must be causing this relationship. When the audience at a Malay shadow theatre sees a solid round shadow on the screen they know that some three-dimensional object has cast it, although they may not know whether the object is a ball or a rice bowl in profile. A more accurate sound bite for introductory statistics would be that a simple correlation implies an *unresolved* causal structure, since we cannot know which is the cause, which is the effect, or even if both are common effects of some third, unmeasured variable.

Although correlation implies an unresolved causal structure, the reverse is not true: causation implies a completely resolved correlational structure. By this I mean that once a causal structure has been proposed, the complete pattern of correlation and partial correlation is fixed unambiguously. This point is developed more precisely in Chapter 2 but is so central to this book that it deserves repeating: the causal relationships between objects or variables determine the correlational relationships between them. Just as the shape of an object fixes the shape of its shadow, the patterns of direct and indirect causation fix the correlational 'shadows' that we observe in observational data. The causal processes generating our observed data impose constraints on the patterns of correlation that such data display.

The term 'correlation' evokes the notion of a probabilistic association between random variables. One reason why statisticians rarely speak of

[2] It could be argued that variables that covary because they are time-ordered have no causal basis. For instance, Monday unfortunately always follows Sunday and day always follows night. However, the first is simply a naming convention and there is a causal basis for the second: the earth's rotation about its axis in conjunction with its rotation around the sun. A more convincing example would be the correlation between the sizes of unrelated children, as they age, who are born at the same time.

causation, except to distance themselves from it, is because there did not exist, until very recently, any rigorous translation between the language of causality (however defined) and the language of probability distributions (Pearl 1988). It is therefore necessary to link causation to probability distributions in a very precise way. Such rigorous links are now being forged. It is now possible to give mathematical proofs that specify the correlational pattern that must exist given a causal structure. These proofs also allow us to specify the class of causal structures that must include the causal structure that generates a given correlational pattern. The methods described in this book are justified by these proofs. Since my objective is to describe these methods and show how they can help biologists in practical applications, I won't present these proofs but will direct the interested reader to the relevant primary literature as each proof is needed.

Another reason why some prefer to speak of associations rather than causes is perhaps because causation is seen as a metaphysical notion that is best left to philosophers. In fact, even philosophers of science can't agree on what constitutes a 'cause'. I have no formal training in the philosophy of science and am neither able nor inclined to advance such a debate. This is not to say that philosophers of science have nothing useful to contribute. Where directly relevant I will outline the development of philosophical investigations into the notion of 'causality' and place these ideas into the context of the methods that I will describe. However, I won't insist on any formal definition of 'cause' and will even admit that I have never seen anything in the life sciences that resembles the 'necessary and sufficient' conditions for causation that are so beloved of logicians.

You probably already have your own intuitive understanding of the term 'cause'. I won't take it away from you, although, I hope, it will be more refined after reading this book. When I first came across the idea that one can study causes without defining them, I almost stopped reading the book (Spirtes, Glymour and Scheines 1993). I can advance three reasons why you should not follow through on this same impulse. First, and most important, the methods described here are not logically dependent on any particular definition of causality. The most basic assumption that these methods require is that causal relationships exist in relation to the phenomena that are studied by biologists[3].

The second reason why you should continue reading even if you are sceptical is more practical and, admittedly, rhetorical: scientists commonly deal with notions whose meaning is somewhat ambiguous. Biologists

[3] Perhaps quantum physics does not need such an assumption. I will leave this question to people better qualified than I. The world of biology does not operate at the quantum level.

are even more promiscuous than most with one notion that can still raise the blood pressure of philosophers and statisticians. This notion is 'probability', for which there are frequentist, objective Bayesian and subjective Bayesian definitions. In the 1920s von Mises is reported to have said: 'today, probability theory is not a mathematical science' (Rao 1984). Mayo (1996) gave the following description of the present degree of consensus concerning the meaning of 'probability': 'Not only was there the controversy raging between the Bayesians and the error [i.e. frequentist] statisticians, but philosophers of statistics of all stripes were full of criticisms of Neyman–Pearson error [i.e. frequentist-based] statistics . . .'. Needless to say, the fact that those best in a position to define 'probability' cannot agree on one does not prevent biologists from effectively using probabilities, significance levels, confidence intervals, and the other paraphernalia of modern statistics[4]. In fact, insisting on such an agreement would mean that modern statistics could not even have begun.

The third reason why you should continue reading, even if you are sceptical, is eminently practical. Although the randomised experiment is inferentially superior to the methods described in this book, when randomisation can be properly applied, it can't be properly applied to many (perhaps most) research questions asked by biologists. Unless you are willing simply to deny that causality is a meaningful concept then you will need some way of studying causal relationships when randomised experiments cannot be performed. Maintain your scepticism if you wish, but grant me the benefit of your doubt. A healthy scepticism while in a car dealership will keep you from buying a 'lemon'. An unhealthy scepticism might prevent you from obtaining a reliable means of transport.

I said that the methods in this book are not logically dependent on any particular definition of causality. Rather than *defining* causality, the approach is to *axiomise* causality (Spirtes, Glymour and Scheines 1993). In other words, one begins by determining those attributes that scientists view as necessary for a relationship to be considered 'causal' and then develop a formal mathematical language that is based on such attributes. First, these relationships must be *transitive*: if A causes B and B causes C, then it must also be true that A causes C. Second, such relationships must be 'local'; the technical term for this is that the relationships must obey the *Markov condition*, of which there are local and global versions. This is described in more detail in Chapter 2 but can be intuitively understood to mean that events are caused only by their proximate causes. Thus, if event A causes event C

[4] The perceptive reader will note that I have now compounded my problems. Not only do I propose to deal with one imperfectly defined notion – causality – but I will do it with reference to another imperfectly defined notion: a probability distribution.

only through its effect of an intermediate event B ($A \rightarrow B \rightarrow C$), then the causal influence of A on C is blocked if event B is prevented from responding to A. Third, these relationships must be *irreflexive*: an event cannot cause itself. This is not to say that every event must be causally explained; to argue in this way would lead us directly into the paradox of infinite regress. Every causal explanation in science includes events that are accepted (measured, observed . . .) without being derived from previous events[5]. Finally, these relationships must be *asymmetric*: if A is a cause of B, then B cannot simultaneously be a cause of A[6]. In my experience, scientists generally accept these four properties. In fact, so long as I avoid asking for definitions, I find that there is a large degree of agreement between scientists on whether any particular relationship should be considered causal or not. It might be of some comfort to empirically trained biologists that the methods described in this book are based on an almost empirical approach to causality. This is because deductive definitions of philosophers are replaced with attributes that working scientists have historically judged to be necessary for a relationship to be causal. However, this change of emphasis is, by itself, of little use.

Next, we require a new mathematical language that is able to express and manipulate these causal relationships. This mathematical language is that of directed graphs[7] (Pearl 1988; Spirtes, Glymour and Scheines 1993). Even this new mathematical language is not enough to be of practical use. Since, in the end, we wish to infer causal relationships from correlational data, we need a logically rigorous way of translating between the causal relationships encoded in directed graphs and the correlational relationships encoded in probability theory. Each of these requirements can now be fulfilled.

[5] The paradox of infinite regress is sometimes 'solved' by simply declaring a First Cause: that which causes but which has no cause. This trick is hardly convincing because, if we are allowed to invent such things by fiat, then we can declare them anywhere in the causal chain. The antiquity of this paradox can been seen in the first sentence of the first verse of Genesis: 'In the beginning God created the heavens and the earth.' According to the Confraternity Text of the Holy Bible, the Hebrew word that has been translated as 'created' was used only with reference to divine creation and meant 'to create out of nothing'.

[6] This does not exclude feedback loops so long as we understand these to be dynamic in nature: A causes B at time t, B causes A at time $t+\Delta t$, and so on. This is discussed more fully in Chapter 2.

[7] Biologists will find it ironic that this graphical language was actually proposed by Wright (1921), one of the most influential evolutionary biologists of the twentieth century, but his insight was largely ignored. This history is explored in Chapters 3 and 4.

1.2 Fisher's genius and the randomised experiment

Since this book deals with causal inference from observational data, we should first look more closely at how biologists infer causes from experimental data. What is it about these experimental methods that allows scientists to comfortably speak about causes? What is it about inferring causality from non-experimental data that make them squirm in their chairs? I will distinguish between two basic types of experiment: controlled and randomised. Although the controlled experiment takes historical precedence, the randomised experiment takes precedence in the strength of its causal inferences.

Fisher[8] described the principles of the randomised experiment in his classic *The design of experiments* (Fisher 1926). Since he developed many of his statistical methods in the context of agronomy, let's consider a typical randomised experiment designed to determine whether the addition of a nitrogen-based fertiliser can cause an increase in the seed yield of a particular variety of wheat. A field is divided into 30 plots of soil (50 cm × 50 cm) and the seed is sown. The treatment variable consists of the fertiliser, which is applied at either 0 or 20 kg/hectare. For each plot we place a small piece of paper in a hat. One half of the pieces of paper have a '0' and the other half have a '20' written on them. After thoroughly mixing the pieces of paper, we randomly draw one for each plot to determine the treatment level that each plot is to receive. After applying the appropriate level of fertiliser independently to each plot, we make no further manipulations until harvest day, at which time we weigh the seed that is harvested from each plot.

The seed weight per plot is normally distributed within each treatment group. Those plots receiving no fertiliser produce 55 g of seed with a standard error of 6. Those plots receiving 20 kg/hectare of fertiliser produce 80 g of seed with a standard error of 6. Excluding the possibility that a very rare random event has occurred (with a probability of approximately 5×10^{-8}), we have very good evidence that there is a positive *association* between the addition of the fertiliser and the increased yield of the wheat. Here we see the first advantage of randomisation. By randomising the treatment allocation, we generate a sampling distribution that allows us to calculate the probability of observing a given result by chance if, in reality, there is no effect of the treatment. This helps us to distinguish between chance associations and systematic ones. Since one error that a researcher can make is to confuse a real difference with a difference due to sampling

[8] Sir Ronald A. Fisher (1890–1962) was chief statistician at the Rothamsted Agricultural Station, (now IACR – Rothamsted), Hertfordshire. He was later Galton Professor at the University of London and Professor of Genetics at the University of Cambridge.

fluctuations, the sampling distribution allows us to calculate the probability of committing such an error[9]. Yet Fisher and many other statisticians[10] since (Kempthorpe 1979; Kendall and Stuart 1983) claim further that the process of randomisation allows us to differentiate between associations due to causal effects of the treatment and associations due to some variable that is a common cause both of the treatment and response variables. What allows us to move so confidently from this conclusion about an *association* (a 'co-relation') between fertiliser addition and increased seed yield to the claim that the added fertiliser actually *causes* the increased yield?

Given that two variables (X and Y) are associated, there can be only three elementary, but not mutually exclusive, causal explanations: X causes Y, Y causes X, or there are some other causes that are common to both X and Y. Here, I am making no distinctions between 'direct' and 'indirect' causes; I argue in Chapter 2 that such terms have no meaning except relative to the other variables in the causal explanation. Remembering that transitivity is a property of causes, to say that X causes Y does not exclude the possibility that there are intervening variables ($X \rightarrow Z_1 \rightarrow Z_2 \rightarrow \ldots \rightarrow Y$) in the causal chain between them. We can confidently exclude the possibility that the seed produced by the wheat caused the amount of fertiliser that was added. First, we already know the only cause of the amount of fertiliser to be added to any given plot: the number that the experimenter saw written on the piece of paper attributed to that plot. Second, the fertiliser was added before the wheat plants began to produce seed[11]. What allows us to exclude the possibility that the observed association between fertiliser addition and seed yield is due to some unrecognised common cause of both? This was Fisher's genius; the treatments were randomly assigned to the experimental units (i.e. the plots with their associated wheat plants). By definition, such a random process ensures that the order in which the pieces of paper are chosen (and therefore the order in which the plots receive the treatment) is causally independent of any attributes of the plot, its soil, or the plant at the moment of randomisation.

[9] It is for this reason that Mayo (1996) called such frequency-based statistical tests 'error probes'.

[10] 'Only when the treatments in the experiment are applied by the experimenter using the full randomisation procedure is the chain of inductive inference sound; it is only under these circumstances that the experimenter can attribute whatever effect he observes to the treatment and to the treatment only' (Kempthorpe 1979).

[11] Unless your meaning of 'cause' is very peculiar, you will not have objected to the notion that causal relationships cannot travel backwards in time. Despite some ambiguity in its formal definition, scientists would agree on a number of attributes associated with causal relationships. Like pornography, we have difficulty defining it but we all seem to know it when we see it.

Let's retrace the logical steps. We began by asserting that, if there was a causal relationship between fertiliser addition and seed yield, then there would also be a systematic relationship between these two variables in our data: *causation implies correlation*. When we observe a systematic relationship that can't reasonably be attributed to sampling fluctuations, we conclude that there was some causal mechanism responsible for this association. Correlation does not necessarily imply a causal relationship from the fertiliser addition to the seed yield, but it does imply *some* causal relationship that is responsible for this association. There are only three such elementary causal relationships and the process of randomisation has excluded two of them. We are left with the overwhelming likelihood that the fertiliser addition caused the increased seed yield. We cannot categorically exclude the two alternative causal explanations, since it is always possible that we were incredibly unlucky. Perhaps the random allocations resulted, by chance, in those plots that received the 20 kg of fertiliser per hectare having soil with a higher moisture-holding capacity or some other attribute that actually caused the increased seed yield? In any empirical investigation, experimental or observational, we can only advance an argument that is beyond reasonable doubt, not a logical certainty.

The key role played by the process of randomisation seems to be to ensure, up to a probability that can be calculated from the sampling distribution produced by the randomisation, that no uncontrolled common cause of both the treatment and the response variables could produce a spurious association. Fisher said as much himself when he stated that randomisation 'relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed'. Is this strictly true? Consider again the possibility that soil moisture content affects seed yield. By randomly assigning the fertiliser to plots we ensure that, *on average*, the treatment and control plots have soil with the same moisture content, therefore removing any chance correlation between the treatment received by the plot and its soil moisture[12]. But the number of attributes of the experimental units (i.e. the plots with their attendant soil and plants) is limited only by our imagination. Let's say that there are 20 different attributes of the experimental units that could cause a difference in seed yield. What is the probability that at least one of these was sufficiently concentrated, by chance, in the treatment plots to produce a significant difference in seed yield even if the fertiliser had no causal effect? If this probability is not large enough for you, then I can easily posit 50 or 100 different

[12] More specifically, these two variables, being causally independent, are also probabilistically independent in the statistical population. This is not necessarily true in the sample, owing to sampling fluctuations.

attributes that could cause a difference in seed yield. Since there is a large number of potential causes of seed yield, then the likelihood that at least one of them was concentrated, by chance, in the treatment plots is not negligible, even if we had used many more than the 30 plots.

Randomisation therefore serves two purposes in causal inference. First, it ensures that there is no causal effect coming from the experimental units to the treatment variable or from a common cause of both. Second, it helps to reduce the likelihood in the sample of a chance correlation between the treatment variable and some other cause of the treatment, but doesn't completely remove it. To cite Howson and Urbach (1989):

> Whatever the size of the sample, two treatment groups are *absolutely certain* to differ in some respect, indeed, in infinitely many respects, any of which might, unknown to us, be causally implicated in the trial outcome. So randomisation cannot possibly guarantee that the groups will be free from bias by unknown nuisance factors [i.e. variables correlated with the treatment]. And since one obviously doesn't know what those unknown factors are, one is in no position to calculate the probability of such a bias developing either.

This should not be interpreted as a severe weakness of the randomised experiment in any practical sense, but does emphasise that even the randomised experiment does not provide any automatic assurance of causal inference, free from subjective assumptions.

Equally important is what is not required by the randomised experiment. The logic of experimentation up to Fisher's time was that of the controlled experiment, in which it was crucial that all other variables be experimentally fixed to constant values[13] (see, for example, Feiblman 1972, page 149). R. A. Fisher (1970) explicitly rejected this as an inferior method, pointing out that it is logically impossible to know whether 'all other variables' have been accounted for. This is not to say that Fisher did not advocate physically controlling for other causes in addition to randomisation. In fact, he explicitly recommended that the researcher do this whenever possible. For instance, in discussing the comparison of plant yields of different varieties, he advised that they be planted in soil 'that appears to be uniform'. In the context of pot experiments he recommended that the soil be thor-

[13] Clearly, this cannot be literally true. Consider a case in which the causal process is: $A \rightarrow B \rightarrow C$ and we want to experimentally test whether A causes C. If we hold variable B constant then we would incorrectly surmise that A has no causal effect on C. It is crucial that common causes of A and C be held constant in order to exclude the possibility of a spurious relationship. It is also a good idea, although not crucial for the causal inference, that causes of C that are independent of A also be held constant in order to reduce the residual variation of C.

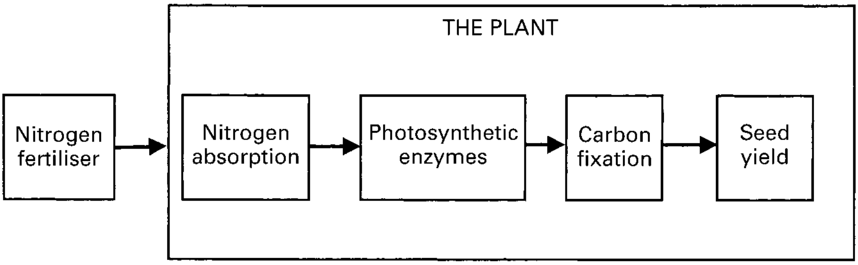


Figure 1.1. An hypothetical causal scenario that is not amenable to a randomised experiment.

oughly mixed before putting it in the pots, that the watering be equalised, that they receive the same amount of light and so on. The strength of the randomised experiment is in the fact that we do not have to physically control – or even be aware of – other causally relevant variables in order to reduce (but not logically exclude) the possibility that the observed association is due to some unmeasured common cause in our sample.

Yet strength is not the same as omnipotence. Some readers will have noticed that the logic of the randomised experiment has, hidden within it, a weakness not yet discussed that severely restricts its usefulness to biologists; a weakness that is not removed even with an infinite sample size. In order to work, one must be able to randomly assign values of the hypothesised 'cause' to the experimental units independently of any attributes of these units. This assignment must be direct and not mediated by other attributes of the experimental units. Yet, a large proportion of biological studies involves relationships between different attributes of such experimental units.

In the experiment described above, the experimental units are the plots of ground with their wheat plants. The attributes of these units include those of the soil, the surrounding environment and the plants. Imagine that the researcher wants to test the following causal scenario: the added fertiliser increases the amount of nitrogen absorbed by the plant. This increases the amount of nitrogen-based photosynthetic enzymes in the leaves and therefore the net photosynthetic rate. The increased carbon fixation due to photosynthesis causes the increased seed yield (Figure 1.1).

The first part of this scenario is perfectly amenable to the randomised experiment since the nitrogen absorption is an attribute of the plant (the experimental unit), while the amount of fertiliser added is controlled completely by the researcher independently of any attribute of the plot or its wheat plants. The rest of the hypothesis is impervious to the randomised experiment. For instance, both the rate of nitrogen absorption and the concentration of photosynthetic enzymes are attributes of the plant (the experimental unit). It is impossible to randomly assign rates of nitrogen absorption to each plant independently of any of its other attributes. Yet this is the crucial step in the randomised experiment that allows us to distinguish correlation from causation. It is true that the researcher can induce a *change* both in the rate of nitrogen absorption by the plant and in the concentration of photosynthetic enzymes in its leaves but in each case these changes are due to the addition of the fertiliser. After observing an association between the increased nitrogen absorption and the increased enzyme concentration the randomisation of fertiliser addition does not exclude different causal scenarios, only some of which are shown in Figure 1.2.

While reading books about experimental design one's eyes often skim across the words 'experimental unit' without pausing to consider what these words mean. The experimental unit is the 'thing' to which the treatment levels are randomly assigned. The experimental unit is also an experimental *unit.* The causal relationships, if they exist, are between the external treatment variable and each of the attributes of the experimental unit that show a response. In biology the experimental units (for instance plants, leaves or cells) are integrated wholes whose parts cannot be disassembled without affecting the other parts. It is often not possible to randomly 'assign' values of one attribute of an experimental unit independently of the behaviour of its other attributes[14]. When such random assignments can't be done then one can't infer causality from a random experiment. A moment's reflection will show that this problem is very common in biology. Organismal, cell and molecular biology are rife with it. Physiology is hopelessly entangled. Evolution and ecology, dependent as they are on physiology and morphology, are often beyond its reach. If we accept that one can't study causal relationships without the randomised experiment, then a large proportion of biological research will have been gutted of any demonstrable causal content.

The usefulness of the randomised experiment is also severely reduced because of practical constraints. Remember that the inference is from the randomised treatment allocation to the experimental unit. The experimental unit must be the one that is relevant to the scientific hypothesis of interest. If the hypothesis refers to large-scale units (populations, ecosystems, landscapes) then the experimental unit must consist of such units. Someone wishing to know whether increased carbon dioxide (CO_2) con-

[14] This is not to say that it is always impossible. For instance, one can randomly add levels of insulin to the blood because the only cause of these changes (given proper controls) is the random numbers assigned to the animal. One can't randomly add different numbers of functioning chloroplasts to a leaf.

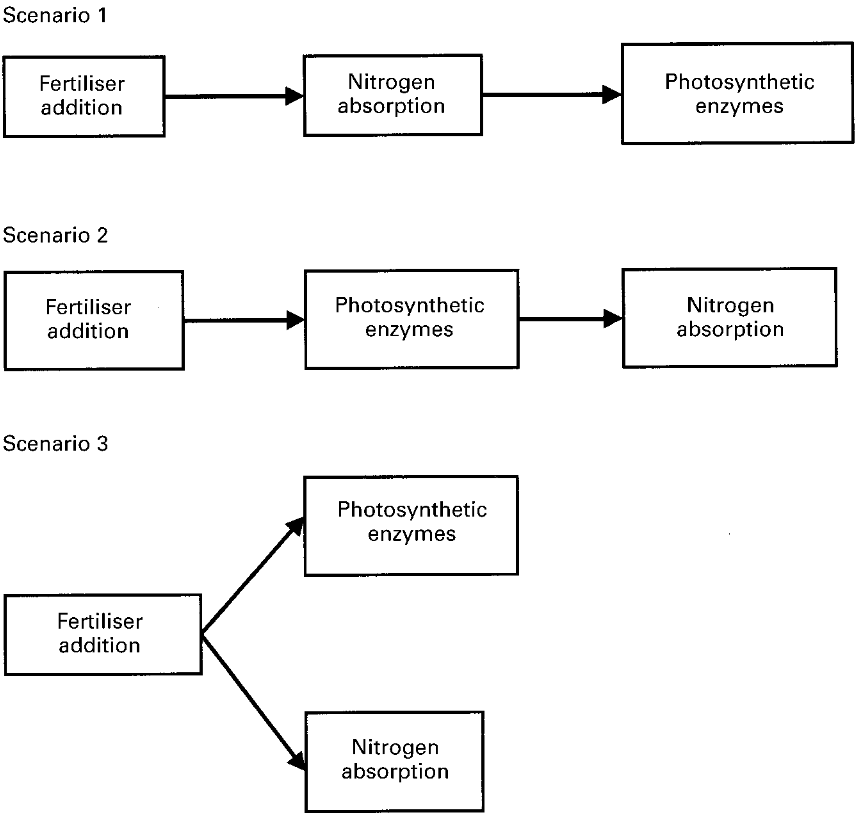


Figure 1.2. Three different causal scenarios that could generate an association between increased nitrogen absorption and increased enzyme concentration in the plant following the addition of fertiliser in a randomised experiment.

centrations will change the community structure of forests will have to use entire forests as the experimental units. Such experiments are never done and there is nothing in the inferential logic of randomised experiments that allows one to scale up from different (small-scale) experimental units. Even when proper randomised experiments can be done in principle, they sometimes can't be done in practice, owing to financial or ethical constraints.

The biologist who wishes to study causal relationships using the randomised experiment is therefore severely limited in the questions that can be posed. The philosophically inclined scientist who insists that a positive response from a randomised experiment is an operational *definition* of a causal relationship would have to conclude that causality is irrelevant to much of science.

1.3 The controlled experiment

The currently prevalent notion that scientists cannot convincingly study causal relationships without the randomised experiment would seem incomprehensible to scientists before the twentieth century. Certainly biologists *thought* that they were demonstrating causal relationships long before the invention of the randomised experiment. A wonderful example of this can be found in *An introduction to the study of experimental medicine* by the great nineteenth century physiologist, Claude Bernard[15]. I will cite a particularly interesting passage (Rapport and Wright 1963), and I ask that you pay special attention to the ways in which he tries to control variables. I will then develop the connection between the controlled experiment and the statistical methods described in this book.

> In investigating how the blood, leaving the kidney, eliminated substances that I had injected, I chanced to observe that the blood in the renal vein was crimson, while the blood in the neighboring veins was dark like ordinary venous blood. This unexpected peculiarity struck me, and I thus made observation of a fresh fact begotten by the experiment, but foreign to the experimental aim pursued at the moment. I therefore gave up my unverified original idea, and directed my attention to the singular coloring of the venous renal blood; and when I had noted it well and assured myself that there was no source of error in my observation, I naturally asked myself what could be its cause. As I examined the urine flowing through the urethra and reflected about it, it occurred to me that the red coloring of the venous blood might well be connected with the secreting or active state of the kidney. On this hypothesis, if the renal secretion was stopped, the venous blood should become dark: that is what happened; when the renal secretion was re-established, the venous blood should become crimson again; this I also succeeded in verifying whenever I excited the secretion of urine. I thus secured experimental proof that there is a connection between the secretion of urine and the coloring of blood in the renal vein.

Our knowledge of human physiology has progressed far from the experiments of Claude Bernard (physiologists might find it strange that he spoke of renal 'secretions'); yet his use of the controlled experiment would be immediately recognisable and accepted by modern physiologists. Fisher was correct in describing the controlled experiment as an inferior way of obtaining causal inferences, but the truth is that the randomised experiment is unsuited to much of biological research. The controlled experi-

[15] Rapport and Wright (1963) describe Claude Bernard (1813–1878) as an experimental genius and 'a master of the controlled experiment'.

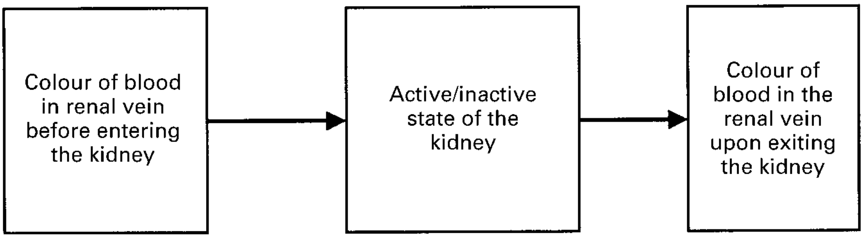

Colour of blood in renal vein before entering the kidney
Active/inactive state of the kidney
Colour of blood in the renal vein upon exiting the kidney

Figure 1.3. The hypothetical causal explanation invoked by Claude Bernard.

ment consists of proposing a hypothetical structure of cause–effect relationships, deducing what would happen if particular variables are controlled, or 'fixed' in a particular state, and then comparing the observed result with its predicted outcome. In the experiment described by Claude Bernard, the hypothetical causal structure could be conceptualised as shown in Figure 1.3.

The key notion in Bernard's experiment was the realisation that, if his causal explanation were true, then the type of *association* between the colour of the blood in the renal vein as it enters and leaves the kidney would change, depending on the state of the hypothesised cause, i.e. whether the kidney was secreting or not. It is worth returning to his words: 'On this hypothesis, if the renal secretion was stopped, the venous blood should become dark: that is what happened; when the renal secretion was re-established, the venous blood should become crimson again; this I also succeeded in verifying whenever I excited the secretion of urine. I thus secured experimental proof that there is a connection between the secretion of urine and the coloring of blood in the renal vein.' Since he explicitly stated earlier in the quote that he was inquiring into the 'cause' of the phenomenon, it is clear that he viewed the result of his experiments as establishing a *causal connection* between the secretion of urine and the colouring of blood in the renal vein.

Although the controlled experiment is an inferior method of making causal inferences relative to the randomised experiment, it is actually responsible for most of the causal knowledge that science has produced. The method involves two basic parts. First, one must propose an hypothesis stating how the measured variables are linked in the causal process. Second, one must deduce how the associations between the observations must change once particular combinations of variables are controlled so that they can no longer vary naturally, i.e. once particular combinations of variables are 'blocked'. The final step is to compare the patterns of association, after such controls are established, with the deductions. Historically, variables have been blocked by physically manipulating them. However (this is an important point that will be more fully developed and justified in Chapter 2), it is the control of variables, not how they are controlled, that is the crucial step. The weakness of the method, as Fisher pointed out, is that one can never be sure that all relevant variables have been identified and properly controlled. One can never be sure that, in manipulating one variable, one has not also changed some other, unknown variable. In any field of study, as Bernard documents in his book, the first causal hypotheses are generally wrong and the process of testing, rejecting, and revising them is what leads to progress in the field.

1.4 Physical controls and observational controls

It is the control of variables, not how they are controlled, that is the crucial step in the controlled experiment. What does it mean to 'control' a variable? Can such control be obtained in more than one way? In particular, can one control variables on the basis of observational, rather than experimental, observations? The link between a physical control through an experimental manipulation and a statistical control through conditioning will be developed in the next chapter, but it is useful to provide an informal demonstration here using an example that should present no metaphysical problems to most biologists.

Body size in large mammals seems to be important in determining much of their ecology. In populations of Bighorn Sheep in the Rocky Mountains, it has been observed that the probability of survival of an individual through the winter is related to the size of the animal in the autumn. However, this species has a strong sexual dimorphism, males being up to 60% larger than females. Perhaps the association between body size and survival is simply due to the fact that males have a better probability of survival than females and this is unrelated to their body size. In observing these populations over many years, perhaps the observed association arises because those years showing better survival also have a larger proportion of males. Figure 1.4 shows these two alternative causal hypotheses. I have included boxes labelled 'other causes' to emphasise that we are not assuming the chosen variables to be the only causes of body size or of survival.

Notice the similarity to Claude Bernard's question concerning the cause of blood colour in the renal vein. The difference between the two alternative causal explanations in Figure 1.4 is that the second assumes that the association between spring survival and autumn body size is due only to

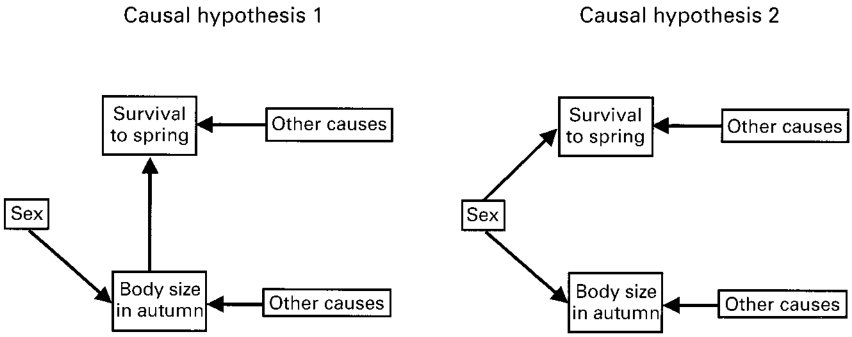


Figure 1.4. Two alternative causal explanations for the relationship between sex, body size of Bighorn Sheep in the autumn and the probability of survival until the spring.

the sex ratio of the population. Thus, if the sex ratio could be held constant, then the association would disappear. Since adult males and females of this species live in separate groups, it would be possible to physically separate them in their range and, in this way, physically control the sex ratio of the population. However, it is much easier to simply sort the data according to sex and then look for an association within each homogeneous group. The act of separating the data into two groups such that the variable in question – the sex ratio – is constant within each group represents a *statistical control*. We could imagine a situation in which we instruct one set of researchers to physically separate the original population into two groups based on sex, after which they test for the association within each of their experimental groups, and then ask them to combine the data and give them to a second team of researchers. The second team would analyse the data using the statistical control. Both groups would come to identical conclusions[16]. In fact, using statistical controls might even be preferable in this situation. Simply observing the population over many years and then statistically controlling for the sex ratio on paper does not introduce any physical changes in the field population. It is certainly conceivable that the act of physically separating the sexes in the field might introduce some unwanted, and potentially uncontrolled, change in the behavioural ecology of the animals that might bias the survival rates during the winter quite independently of body size.

Let's further extend this example to look at a case in which it is not as easy to separate the data into groups that are homogeneous with respect

[16] It is not true that statistical and physical controls will always give the same conclusion. This is discussed in Chapter 2.

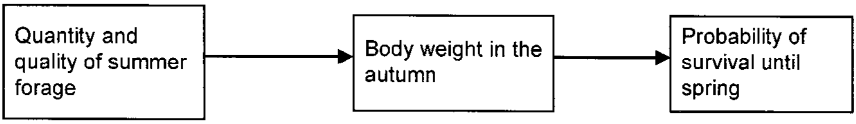

Quantity and quality of summer forage
Body weight in the autumn
Probability of survival until spring

Figure 1.5. A hypothetical causal explanation for the relationship between the quality and quantity of summer forage, the body weight of the Bighorn Sheep in the autumn and the probability of survival until the spring.

to the control variable. Perhaps the researchers have also noticed an association between the amount and quality of the rangeland vegetation during the early summer and the probability of sheep survival during the next winter. They hypothesise that this pattern is caused by the animals being able to eat more during the summer, which increases their body size in the autumn, which then increases their chances of survival during the winter (Figure 1.5).

The logic of the controlled experiment requires that we be able to compare the relationship between forage quality and winter survival after physically preventing body weight from changing, which we can't do[17]. Since 'body weight' is a continuous variable, we can't simply sort the data and then divide it into groups that are homogeneous for this variable. This is because each animal will have a different body weight. Nonetheless, there is a way of comparing the relationship between forage quality and winter survival while controlling for the body weight of the animals during the comparison. This involves the concept of statistical conditioning, which will be more rigorously developed in Chapters 2 and 3. An intuitive understanding can be had with reference to a simple linear regression (Figure 1.6).

The formula for a linear regression is: $Y_i = \alpha + \beta X_i + \mathrm{N}(0,\sigma)$. Here, the notation '$\mathrm{N}(0,\sigma)$' means 'a normally distributed random variable with a population mean of zero and a population standard deviation of σ'. As the formula makes clear, the observed value of Y consists of two parts: one part that depends on X and one part that doesn't. If we let '$\mathrm{E}(Y|X)$' represent the expected value of Y given X, then we can write:

[17] It is actually possible, in principle if not in practice, to conduct a randomised experiment in this case, so long as we are interested only in knowing whether summer forage quality causes a change in winter survival. This is because the hypothetical cause (vegetation quality and quantity) is not an attribute of the unit possessing the hypothetical effect (winter survival). Again, it is impossible to use a randomised experiment to determine whether body size in the autumn is a cause of increased survival during the winter.

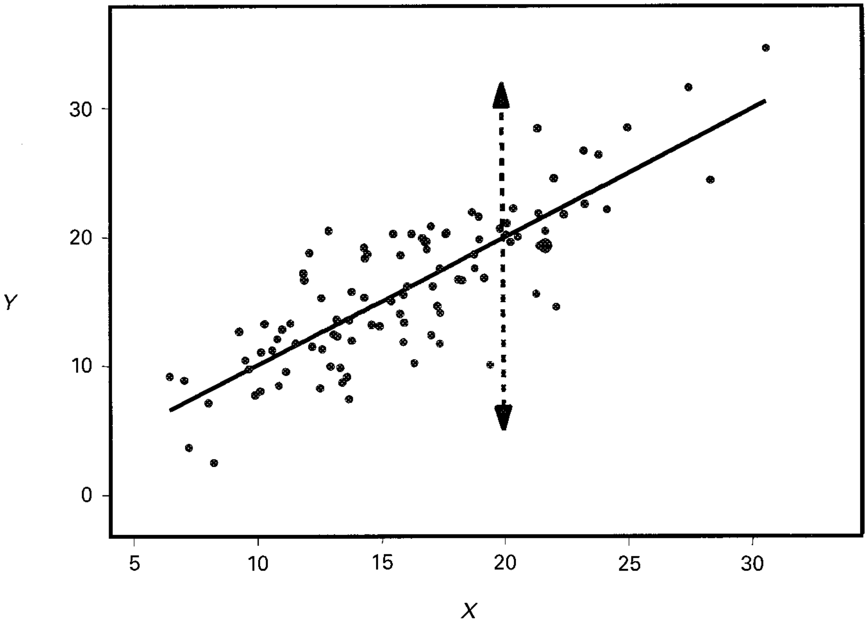


Figure 1.6. A simple bivariate regression. The solid line shows the expected value of Y_i given the value of X_i (E[$Y_i|X_i$]). The dotted line shows the possible values of Y_i that are independent of X_i (the residuals).

$$\mathrm{E}(Y|X_i) = \alpha + \beta X_i$$

$$Y_i = \mathrm{E}(Y|X_i) + \mathrm{N}(0,\sigma)$$

$$Y_i - \mathrm{E}(Y|X_i)) = \mathrm{N}(0,\sigma).$$

Thus, if we subtract the expected value of each *Y*, given *X*, from the value itself, then we get the variation in *Y* that is independent of *X*. This new variable is called the *residual* of *Y* given *X*. These are the values of *Y* that exist for a constant value of *X*. For instance, the vertical arrow in Figure 1.6 shows the values of *Y* when *X*=20.

If we want to compare the relationship between forage quality and winter survival while controlling for the body weight of the animals during the comparison, then we have to remove the effect of body weight on each of the other two variables. We do this by taking each variable in turn, subtracting the expected value of its given body weight, and then see whether there is still a relationship between the two sets of residuals. In this way, we can hold constant the effect of body weight in a way similar to experimentally holding constant the effect of some variable. The analogy is not exact.

There are situations in which statistically holding constant a variable will produce patterns of association different from those that would occur when one is physically holding constant the same variable. To understand when statistical controls cast the same correlational shadows as experimental controls, and when they differ, we need a way of rigorously translating from the language of causality to the language of probability distributions. This is the topic of the next chapter.

2 From cause to correlation and back

2.1 Translating from causal to statistical models

The official language of statistics is the probability calculus, based on the notion of a probability distribution. For instance, if you conduct an analysis of variance (ANOVA) then the key piece of information is the probability of observing a particular value of Fisher's *F* statistic in a random sample of data, given a particular hypothesis or model. To obtain this crucial piece of information, you (or your computer) must know the probability density function of the *F* statistic. Certain other (mathematical) languages are tolerated within statistics but, in the end, one must link one's ideas to a probability distribution in order to be understood. If we wish to study causal relationships using statistics, it is necessary that we translate, without error, from the language of causality to the only language that statistics can understand: probability theory.

Such a rigorous translation device did not exist until recently (Pearl 1988). It is no wonder that statisticians have virtually banished the word 'cause' from statistics – it has no equivalent in their language[1]. Within the world of statistics the scientific notion of causality has, until recently, been a stranger in a strange land. Posing causal questions in the language of probability calculus is like a unilingual Englishman asking for directions to the Louvre in Paris from a Frenchman who can't speak English. The Frenchman might understand that directions are being requested, and the Englishman might see fingers pointing in particular directions, but it is not at all sure that works of art will be found. Imperfect translations between the language of causality and the language of probability theory are equally disorienting.

Mistakes in translation come in all kinds. The most dangerous ones are the subtle errors in which a slight change in inflection or context of a word can change the meaning in disastrous ways. Because the French word *demande* both sounds like the English word 'demand' and has roughly the

[1] Fisherian statistics does deal with causal hypotheses, but the causal inferences come from the experimental design, not from the mathematical details; see Chapter 1.

same meaning (it simply means 'to ask for', without any connotation of obligation), I have seen French-speaking people come up to a store clerk and, while speaking English, 'demand service'. They think that they are politely asking for help while the clerk thinks they are issuing an ultimatum. I once came close to being beaten by an enraged boyfriend simply because (I thought) I was complimenting his girlfriend on her long hair, which was drawn in a ponytail. The word for 'tail' in French is *queue*, which takes a feminine gender. There is another word in colloquial Canadian French, *cul* (the 'l' is silent), that sounds almost the same. It takes a masculine gender, is pronounced only slightly differently, and can be roughly translated as a person's rear end; the correctly translated word rhymes with 'pass' but the reader will understand if I don't give the literal translation. So, while trying to make conversation with the boyfriend I told him that his girlfriend had a nice *cul* instead of a nice *queue*. I immediately knew, from the look of rage on his face, that I had chosen the wrong word.

The same subtle mistakes of translation can occur when translating between the language of causality and the mathematical language of probability distributions. I began the first chapter by comparing causes and correlations to three-dimensional objects and their two-dimensional shadows. Clearly, there is a close relationship between the object and its shadow. Just as clearly, they are not the same thing. The goal of this chapter is to describe the relationship between variables involved in a causal process and the probability distribution of these variables that the causal process generates. Causal processes cast probability shadows but 'causes' and 'probability distributions' are not the same thing either. It is important to understand exactly how the translation is made between causal processes and probability distributions in order to avoid the scientific equivalent of a punch in the nose from an enraged boyfriend.

I will make the distinction between a causal model, an observational model and a statistical model. Since every child knows that rain causes mud[2], I will illustrate the difference between these three types of model with this analogy. The statement 'rain causes mud' implies an asymmetric relationship: the rain will create mud, but the mud will not create rain. I will use the symbol '→' when I want to refer to such causal relationships. This leads naturally to the sort of 'box and arrow' diagrams with which most biologists are familiar (Figure 2.1).

To complete the description it is necessary to add the convention that, unless a causal relationship is explicitly included, it is understood not

[2] My children seem to have mastered this metaphysical concept well before age 5. This is another example of how deeply ingrained is the notion of causality.

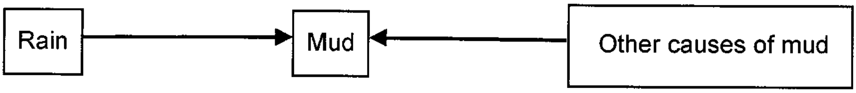


Figure 2.1. The causal relationships between rain, mud and other causes of mud.

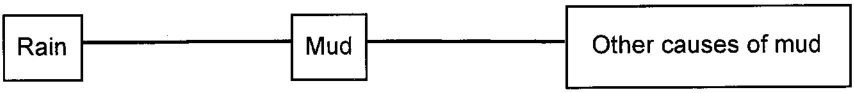


Figure 2.2. The observational relationships between rain, mud and other causes of mud.

to exist. So, in Figure 2.1, the fact that there are no arrows between 'rain' and 'other causes of mud' means that there is no direct causal relationship between them; in fact, there is no causal relationship of any kind in this example, since the two are causally independent.

The observational model that is related to this causal model is the statement that 'having observed rain will give us information about what we will observe concerning mud'. Notice that this observational statement deals with information, not causes, and is not asymmetric. If we learn that it has rained, then we will have added information concerning the presence of mud in our yard, but observing mud in our yard will also give us information about whether or not it has rained. I will use the symbol '—' when I refer to such observational relationships. This leads to the model in Figure 2.2.

Notice that, although rain and other causes of mud are causally independent, they are not observationally independent given the state of mud; knowing that it has not rained but that there is mud in the front yard gives you information on the existence of other causes of mud.

The statistical model differs only in degree, not in kind, from the observational model. The statistical model (Figure 2.3) specifies the mathematical relationship between the variables as well as the probability distributions of the variables. Now we can use the equivalence operator of algebra ('='), since we are stating a quantitative equivalence.

This mathematical statement says that the value obtained by measuring the depth of the mud, in centimetres, is the same as (is 'equivalent to') the value that is obtained by measuring the amount of rain that falls, in centimetres, multiplying this value by 0.1, and adding another value (in centimetres) obtained from a random value taken from a normal distribution whose population mean is zero and whose population standard deviation is 0.1.

$$\text{Mud (cm)} = 0.1\text{Rain (cm)} + N(0, 0.1)$$

Figure 2.3. A statistical model relating rain and mud.

$$\text{Rain (cm)} = 10\text{Mud (cm)} + N(0,1)$$

Figure 2.4. Another statistical model relating rain and mud.

What is the point of all this? According to Pearl (1997) a century of confusion between correlation and causation can be traced, in part, to a mistranslation of the word ‘cause’. When scientists and statisticians attempt to express notions of causality using mathematics they mistranslate ‘cause’, a word having connotations of asymmetry and all of the other properties discussed in Chapter 1, as the algebraic notion ‘=’ used in the language of probability theory. The symbols ‘→’ and ‘=’ do not mean the same thing. It is perfectly correct to rearrange the equation in Figure 2.3 in order to imply that the amount of rain can be predicted from the amount of mud (Figure 2.4) even though any 5 year old child would recognise this as causally nonsensical.

This mistake is the scientific equivalent of telling a boyfriend that his girlfriend has ‘un beau cul’ rather than ‘une belle queue’. The conceptual error occurs because we have replaced ‘→’ with ‘=’. After translating from the language of causality to the language of observations, we have used the syntax of this observational language to produce a perfectly reasonable statement for this observational language, but then we have performed a literal translation back into the language of causality without recognising the difference in syntax. There are computer programs that attempt to translate between human languages and those that use literal word-by-word translations run into the same problems. A newspaper headline like ‘Bill Gates worth $1 000 000 000’, after being literally translated (word for word) into a different language and then re-translated back into English, might come up with a phrase like ‘payment request for door in the fence costs $1 000 000 000’!

In the next few sections I develop a translation device to move between causal models and observational (statistical) models. To do this we require the necessary and sufficient conditions needed to specify a joint probability distribution that must exist given a causal process. Put another way, we require the necessary and sufficient conditions needed to specify the correlational shadow that will be cast by a causal process. This provides the key to translating between causal and statistical models. These sections require more effort to understand but in each case I will also provide a more intuitive description and some worked examples.

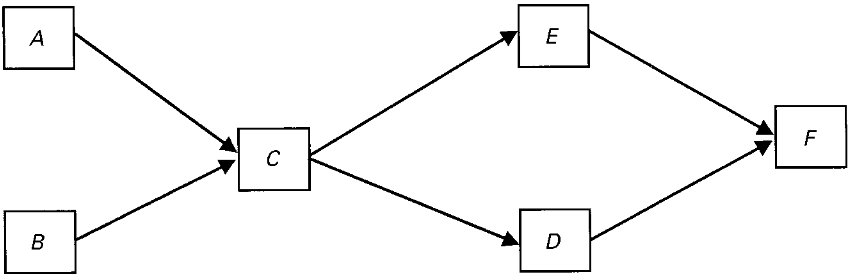


Figure 2.5. A directed graph describing the causal relationships between five variables or vertices (*A* to *F*).

The strategy for translation from the physical world, in which the notion of causation is applicable, to the mathematical world of probability theory, in which the abstract notion of algebraic equivalence is applicable, involves two steps. First, since algebra cannot express the sorts of relationship that we term 'causal', we need a new mathematical language that can; this language is that of directed graphs. Second, we need a translation device that can unambiguously convert the statements expressed in such directed graphs into statements concerning conditional independence of random variables obeying a particular probability distribution. This translation device is called 'd-separation' (short for *directed* separation).

2.2 Directed graphs

It is now time to introduce some terminology concerning directed (sometimes called causal) graphs. These terms, although unfamiliar to most biologists, are quite easy to grasp and use. These terms will be defined using the causal graph shown in Figure 2.5.

Here is a *partial* verbal (as opposed to mathematical) description of what Figure 2.5 means. Two of the six variables (*A* and *B*) are *causally independent*, meaning that changes in either will not affect the value of the other. Each of the four other variables (*C*, *D*, *E* and *F*) are *causally dependent* on *A* and *B*, either directly (*C*) or indirectly (*D*, *E* and *F*). By 'causally dependent' I mean that changes in either *A* or *B* will provoke changes in each of *C*, *D*, *E* and *F* but changes in any of these will not provoke changes in either *A* or *B*. *A* and *B* are *direct* causes of *C* because changes in *A* or *B* will provoke changes in *C* irrespective of the behaviour of either *D*, *E* or *F*. *A* and *B* are *indirect causes* of *D*, *E* and *F* because changes in *A* or *B* will only provoke changes in these variables by causing changes in *C*; if *C* is prevented from changing then *A* and *B* will no longer cause changes in these three other variables. *C* is a direct *common* cause of *D* and *E* and an indirect cause of *F* through its effects on *D* and *E*. Finally, both *D* and *E* are direct causes of *F*, although they are not themselves causally independent.

It is clear that this directed graph is a very economic way of expressing even the previous incomplete verbal description of this causal system. This economy of description is a major reason why researchers in artificial intelligence adopted directed graphs as a way of economically programming causal knowledge (Pearl 1988)[3]. In order to better use and interpret directed graphs, a few definitions are needed.

In graph theory a directed graph is a set of vertices, represented by letters enclosed in boxes in Figure 2.5, and a set of edges, represented by lines; these lines can have either no arrowheads, single or double arrowheads. The arrowheads denote the direction of the functional relationship between the vertices at either end of the line[4]. Since biologists will use directed graphs to represent causal relationships between variables, you can replace the abstract term 'vertex' with the more familiar word 'variable' and the abstract term 'edge' with the more familiar word 'effect'. The symbols at the ends of the lines can be either an arrowhead or a 'missing' mark. Thus, the notation '$X \rightarrow Y$' means that *X* is a direct cause of *Y*. The notation '$X \leftarrow Y$' means that *Y* is a direct cause of *X*. Finally, the notation '$X \leftrightarrow Y$' means that neither *X* nor *Y* are causes of the other but both share common unknown causes represented by some unknown vertex not included in the causal graph. This last notation is needed later when we use incomplete causal graphs with unspecified latent vertices.

A *direct cause* is a causal relationship between two vertices that exists independently of any other vertex in the causal explanation. This denoted by an arrow (→) whose tail is at the cause and whose head is pointing to its direct effect. For instance, both *A* and *B* are direct causes of *C* in Figure 2.5. Furthermore, *A* and *B* are the *causal parents* of *C*, and *C* is their *causal child*. A cause is only direct in relation to the other vertices in the causal explanation. This point is important because a common error is to incorrectly equate a 'direct' cause relative to others in the causal graph with the more

[3] More accurately, directed graphs can economically store the conditional independence constraints implied by a causal system of an arbitrary joint probability distribution. This is explained in more detail below.

[4] In the jargon of graph theory, an undirected graph consists of a set of vertices {*A*, *B*, *C*, . . .} and a binary set denoting the presence or absence of edges (lines) between each pair of vertices. The graph becomes directed when we include a set of symbols for each edge showing direction. It is also possible to construct partially directed graphs. A graph is acyclic if there are no paths that lead a vertex back onto itself, otherwise it is cyclic. The causal graph in Figure 2.5 is therefore a directed acyclic graph, or DAG.

fundamental claim that the cause is somehow 'direct' with respect to any other variable that might exist. Whenever you read the words 'direct cause' you should mentally add the words 'relative to the other variables that are explicitly invoked in the causal explanation'.

An *indirect cause* is a causal relationship between two vertices that is conditional on the behaviour of other vertices in the causal explanation. Again, a cause is only indirect in relation to the other vertices in the causal explanation. For instance, in Figure 2.5 the vertex A is an indirect cause of vertex D ($A \to C \to D$) because its causal effect is conditional on the behaviour of vertex C. Furthermore, A and B are *causal ancestors* of D in Figure 2.5 and D is a *causal descendant* of both A and B.

Perhaps an example would help at this point. If we wish to give a causal description of the murder of a victim by a gunman and this explanation involves only these two 'variables' then we would say that the gunman's actions were the direct cause of the victim's death and write 'Gunman's actions→Murder of victim'. On the other hand, if we also include the presence of the bullet penetrating the victim's heart in our causal explanation then we would say that the bullet was the direct cause of death, the gunman was an indirect cause, and write 'Gunman's actions→Bullet→Murder of victim'. If we wish to go into more gruesome physiological detail then we would describe how the bullet interrupts the heart and the bullet would no longer be a direct cause of the victim's death. Virtually any causal mechanism can be further decomposed into a more detailed causal mechanism and so describing a cause as 'direct' or 'indirect' can be meaningful only in relative terms in the context of the other variables that make up the causal explanation. This is simply the reductionist method common in science and the trick is always to choose a level of causal complexity that is sufficiently detailed that it meets the goals of the study while remaining applicable in practice.

A *directed path* between two vertices in a causal graph exists if it is possible to trace an ordered sequence of vertices that must be traversed, when following the direction of the edges (head to tail), in order to travel from the first to the second. If no such directed path exists, then the two vertices *are causally independent*; causal conditional independence is defined below. It is possible for there to be more than one directed path linking two vertices. In Figure 2.5 there are two different directed paths between A and F: $A \to C \to D \to F$ and $A \to C \to E \to F$.

An *undirected path* between two vertices in a causal graph exists if it is possible to trace an ordered sequence of vertices that must be traversed, *ignoring* the direction of the edges (head to tail), in order to travel from the first to the second. An undirected path can also be a directed path, but this is not necessarily the case. For instance, there is an undirected path between *A* and *B* in Figure 2.5 (*A*→*C*←*B*) that is not also a directed path.

A *collider* vertex on a path is a vertex with arrows pointing into it from both directions. Thus the vertex *F* in the undirected path *D*→*F*←*E* in Figure 2.5 is a collider. It is possible for the same vertex to be a collider along one path and a non-collider along another path. A vertex that is a collider along an undirected path is *inactive* in its normal (unconditioned) state. This means that, in its normal (unconditioned) state, a collider blocks (prevents) the transmission of causal effects along such a path. The contrary of a collider is a *non-collider*. The vertex *C* in the path *A*→*C*→*D* in Figure 2.5 is a non-collider. A vertex that is a non-collider along a path is said to be active in its normal (unconditioned) state. This means that, in its normal (unconditioned) state, a non-collider permits the transmission of causal effects along such a path. It is sometimes easier to imagine a path as an electrical circuit and the variables (vertices) along the path as switches. A variable along a path that is a collider is like a switch that is normally OFF and a variable along a path that is a non-collider is like a switch that is normally ON.

An *unshielded collider* vertex is a set of three vertices *A*→*B*←*C* along a path such that *B* is a collider and, additionally, there is no edge between *A* and *C*. In Figure 2.5 the vertex *F* in the undirected path *D*→*F*←*E* is not only a collider but also an unshielded collider, since there is no edge between *D* and *E*. The contrary of an unshielded collider is a *shielded collider*.

2.3 Causal conditioning

I have been referring to the letters in the causal graph as 'vertices'. Once we include the notion of a probability distribution that is generated by the causal graph, these vertices will also represent random variables. These vertices can be conceived to exist in one of two binary states along a given path: active or inactive. As stated above, the natural state of a non-collider is the active (ON) state and the natural state of a collider is the inactive (OFF) state. Again, it is possible for a vertex to be active along one path and inactive along another. Intuitively, one can think of the arrows as pointing out the direction of causal influence. Thus a vertex that is both an effect and a cause (a non-collider), for example vertex *C* along the path *A*→*C*→*D*, is active because it allows the causal influence of *A* to be transmitted to *D*. In the same way, a vertex that is an effect of two vertices and therefore a cause to neither (a collider) is inactive because it blocks the causal influence from being transmitted along the path. An example is the vertex *F* along the path *D*→*F*←*E* in Figure 2.5. *Conditioning* on a vertex in a causal graph means to change its state; if it was active, then conditioning inactivates it but, if it was inactive, then conditioning activates it. So, since vertex C along the path $A \rightarrow C \rightarrow D$ is naturally active (ON), conditioning on it changes its state to inactive (OFF), thus blocking any indirect causal influence of A on D.

2.4 d-separation

Remembering that we are still not discussing probability distributions or statistical models, and are still concerned only with properties of directed acyclic graphs, we can now define what is meant by 'independence' of vertices, or of groups of vertices, in a causal graph upon conditioning on some other set of vertices. This property is called *d-separation* ('directed separation': Verma and Pearl 1988; Pearl 1988; Geiger, Verma and Pearl 1990). The definition of d-separation uses the definitions above and, although it is awkward to define in words, it is very easy to understand when looking at a causal graph. The formal definition is given in Box 2.1. I then give a more informal definition, and finally I illustrate it using figures.

Box 2.1. Formal definition of *d-separation*[5]

Given a causal graph G, if X and Y are two different vertices in G and $\mathbf{Q}$ is a set of vertices in G that does not contain X or Y, then X and Y are d-separated given $\mathbf{Q}$ in G if and only if there exists no undirected path U between X and Y, such that (i) every collider on U is either in $\mathbf{Q}$ or else has a descendant in $\mathbf{Q}$ and (ii) no other vertex on U is in $\mathbf{Q}$.

Informally, d-separation gives the necessary and sufficient conditions for two vertices in a directed acyclic (causal) graph to be observationally (probabilistically) independent upon conditioning on some other set of vertices. d-separation is the translation device between the language of causality and the language of probability distributions. To know whether two vertices (X, Y) are d-separated given some set of other vertices in the causal graph, which we will call $\mathbf{Q}$, do the following:

1. List every undirected path between X and Y.
2. For every such undirected path between X and Y (which is an ordered sequence of vertices that must be traversed, ignoring the directions of the arrows), see whether *any* non-colliding vertices in

[5] d-separation can also be extended to determining causal independence of two sets of vertices $\boldsymbol{A}$ and $\boldsymbol{B}$, upon conditioning on a third set $\mathbf{Q}$.

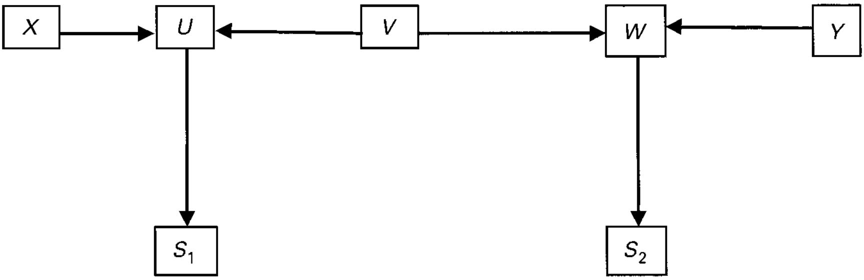


Figure 2.6. A directed graph used to illustrate the notion of d-separation.

this path are in the conditioning set **Q**. If so, then the path is blocked and there is no causal influence between *X* and *Y* along this path. Remembering that conditioning on a non-collider changes its state to inactive, then at least one of the vertices in **Q** blocks any causal influence between *X* and *Y* along this undirected path.

3. For every such undirected path between *X* and *Y*, see whether *every* collider vertex along this path is either a member of the conditioning set **Q** or else has a causal descendant that is a member of the conditioning set **Q**. If not, then the path is blocked and there is no causal influence between *X* and *Y* along this path. Remembering that conditioning on a collider changes its state from inactive to active, then there is at least one collider along this undirected path that remains inactive and so this path cannot transmit causal influence between *X* and *Y*.
4. *X* and *Y* are d-separated given **Q** if every undirected path between them is blocked.

The use of d-separation to deduce probabilistic independence upon conditioning from a causal system is best understood using a diagram (Figure 2.6) from Spirtes, Glymour and Scheines (1993). Table 2.1 lists some of the d-separation statements that can be obtained from Figure 2.6. I will use the notation 'I(*X*,**Q**,*Y*)' to mean 'vertices *X* and *Y* are independent given the conditioning set **Q**'. The negation '~I(*X*,**Q**,*Y*)' means that 'vertices *X* and *Y* are *not* independent given the conditioning set **Q**'. The set **Q** can include the null set ϕ, denoting unconditional causal independence.

The causal inferences listed in Table 2.1 are not exhaustive. After a few minutes of practice it is easy to simply read off the conditional independence relations from such a causal graph. d-separation leads to a wealth of very useful results involving causal inference, many of which will be described in later chapters. However, until d-separation is related to probability distributions, it provides no way of inferring causal relationships from

Table 2.1. V*arious probabilistic independence relationships of the directed graph in Figure 2.6 that can de deduced using d-separation*

Independence relation	Explanation
I(X,$\boldsymbol{\phi}$,V). X and V unconditionally independent	There are no directed paths between X and V
~I(X,U,V). X and V not independent, conditioned on U	Since $X \rightarrow U \leftarrow V$ collides at U, conditioning on U activates this path
~I(X,S_1,V). X and V not independent, conditioned on S_1	Since S$_1$ is a causal descendant of U, conditioning on S_1 activates U along path $X \rightarrow U \leftarrow V$
~I(U,$\boldsymbol{\phi}$,W). U and W are not unconditionally independent	The path $U \leftarrow V \rightarrow W$ is naturally active. U and W share a common cause (V) and V is not in the conditioning set $\{\phi\}$.
I(U,V,W). U and W are independent, conditioned on V	There is only one naturally active path between U and W: $U \leftarrow V \rightarrow W$. Conditioning on V inactivates V, blocking this path
I(X,$\boldsymbol{\phi}$,Y). X and Y are unconditionally independent	The only undirected path between X and Y is naturally blocked by both U and W
~I(X,{U,W},Y). X is not independent of Y, conditioned simultaneously on U and W	The only undirected path between X and Y has two colliders, and both are in the conditioning set. This activates the undirected path
~I(X,{S_1,S_2},Y). X is not independent of Y, conditioned simultaneously on S_1 and S_2	The only undirected path between X and Y has two colliders, and the causal descendants of both are in the conditioning set. This activates the undirected path
I(X,{U,W,V},Y). X is independent of Y, conditioned simultaneously on U, W and V	Although conditioning on both U and W activates these two colliders, conditioning on V disactivates this non-collider

observational data. Before making this link explicit, we first need some notions from probability theory.

2.5 Probability distributions

The vertices of a causal graph represent attributes in a causal system, for instance the nitrogen concentration in a leaf or the body mass of a sheep. When we randomly sample observational units (leaves, sheep) possessing these attributes (nitrogen concentration, body mass) from some statistical population which is governed by this causal system, then the vertices of the causal graph are also random variables that obey a probability distribution. Since causal relationships involve at least two such random variables, we must deal with joint probability distributions.

As I have already briefly mentioned, the notion of 'probability' differs depending on whether one subscribes to a frequentist, objective Bayesian or subjective Bayesian school of statistics. Since almost all statistical methods familiar to biologists derive from a frequentist perspective, I will use this definition. One begins with a hypothetical statistical population (say, all Wheat plants grown in Europe) that contains all of the observational units (individual plants) of interest. Each observational unit has a variable (say, the protein content of a seed) that can take different values (1.2 mg, 3.1 mg . . .). The proportion of observational units (individual plants) in the statistical population (Wheat grown in Europe) taking different values of the variable of interest (seed protein content) is the probability of this variable in this statistical population. Another way of saying this is that the probability of a random variable (X) taking a value $X=x_i$ (or having a value within an infinitesimal interval around x_i) in a statistical population of size N is the limiting frequency of $X=x_i$ in a random sample of size n as n approaches N.

A probability distribution is the distribution of the limiting (relative) frequencies of $X=x_1$, x_2, . . . in such a statistical population. Happily, it is an empirical fact that the distribution of many variables, when randomly sampled, can be closely approximated by various mathematical functions. Many of these functions are well known to biologists (normal distribution, Poisson distribution, binomial distribution, Fisher's F distribution, chi-squared distribution) and there are many less well-known functions that can be used as well. It is always an empirical question whether or not one of these mathematical distributions is a sufficiently close approximation of one's data to be acceptable. For instance, the relative frequency of the seed protein content per plant is likely to follow a normal distribution. The formula for the normal distribution is:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

When only one variable is measured on each observational unit, then one obtains a *univariate* distribution. When one measures more than one variable on each observational unit (say, both the protein content and the average seed weight per plant) then one obtains a *multivariate* distribution[6]. If one obtains the relative frequencies of values of each unique set of multivariate observations, then one has a multivariate probability distribution. Again, there are many multivariate mathematical functions that approximate such multivariate probability distributions. Figure 2.7 shows two versions of a bivariate normal distribution.

2.6 Probabilistic independence

By definition, two random variables (X, Y) are (unconditionally) independent if the joint probability density of X and Y is the product of the probability density of X and the probability density of Y. Thus:

$$\text{If } \mathrm{I}(X,\boldsymbol{\phi},Y) \text{ then } P(X,Y)=P(X)\times P(Y)$$

For instance, if X and Y are each distributed as a standard normal distribution and they are also independent (Figure 2.7A), then the joint probability distribution can be obtained as follows:

$$f(X;0,1)=\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{\frac{-(X)^2}{2}}$$

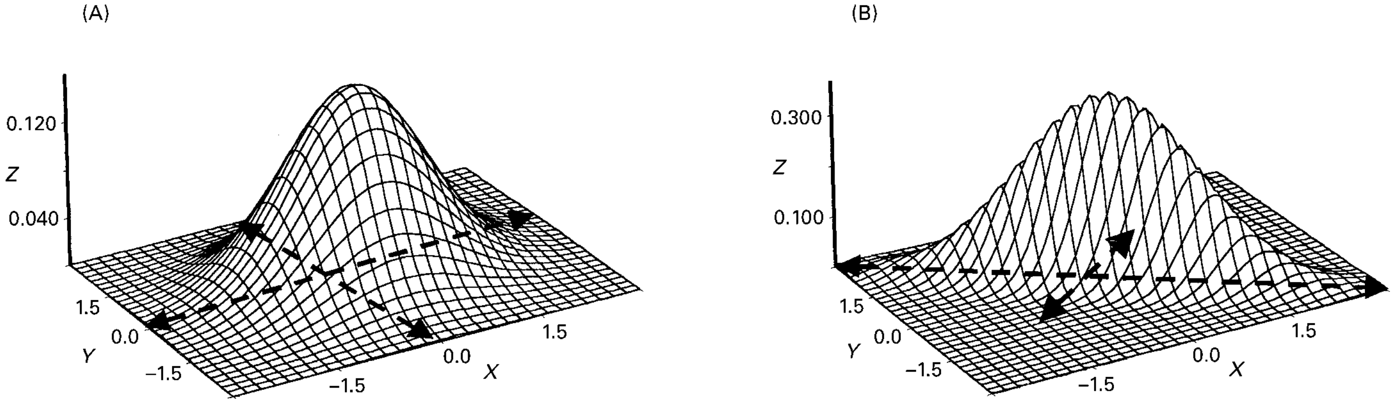
$$f(Y;0,1)=\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{\frac{-(Y)^2}{2}}$$

$$f(X;Y)=f(X;0,1)\times f(Y;0,1)=\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{\frac{-(X^2+Y^2)}{2}}$$

If two random variables (X, Y) are not (unconditionally) independent then the joint probability density of X and Y is not the product of the two univariate probability densities. If the variables are dependent then one can't simply multiply one univariate probability density by the other because we have to take into consideration the interaction between the two (Figure 2.7B).

Figure 2.7A shows the bivariate normal density function of two independent variables. Note that the mean value of Y is the same (0) no matter what the value of X, and vice versa; the value of one variable doesn't

[6] In this case, a bivariate normal distribution.



$Z = f(X,Y) = f(X)f(Y)$

$Z = f(X,Y) \neq f(X)f(Y)$

Figure 2.7. Two different versions of a bivariate normal probability distribution. (A) The joint distribution of two independent, normally distributed random variables. (B) The joint distribution of two normally distributed random variables that are not independent.

change the average value (expected value) of the other variable. Figure 2.7B shows the bivariate normal density function of two dependent variables. Here, the mean value of Y is not independent of the value of X.

Similarly, X and Y are independent, conditional on ('given') a set of other variables $\boldsymbol{Z}$, if the joint probability density of X and Y given $\boldsymbol{Z}$ equals the product of the probability density of X given $\boldsymbol{Z}$ and the probability density of Y given $\boldsymbol{Z}$ for all values of X, Y and $\boldsymbol{Z}$ for which the probability density of $\boldsymbol{Z}$ is not equal to zero[7]. The notion of conditional independence will be explained in more detail in Chapter 3. Thus:

$$\text{If } \mathrm{I}(X,\boldsymbol{Z},Y) \text{ then } P(X,Y \mid Z) = P(X \mid Z) \times P(Y \mid Z)$$

2.7 Markov condition

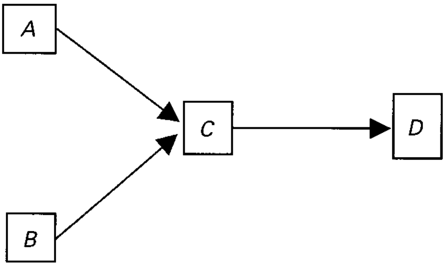
Many ecologists, especially those who study vegetation dynamics, are familiar with Markov chain models (Van Hulst 1979). These models predict vegetation dynamics based on a 'transition matrix'. The transition matrix gives the probability that a location that is occupied by a species s_i at time t will be replaced by species s_j at time $t+1$. The model is 'Markovian' because of the assumption that changes in the vegetation at time $t+1$ depend at most on the state of the vegetation at time t, but not on states of the vegetation at earlier times. Stated another way, these models are Markovian because they assume that the more distant past $(t-1)$ affects the immediate future $(t+1)$ only indirectly through the present (t), thus: $(t-1)\rightarrow(t)\rightarrow(t+1)$.

In the context of causal models, the Markov condition is a property both of a directed acyclic (causal) graph and the joint probability distribution that is generated by the graph. The condition is satisfied if, given a vertex v_i in the graph, or a random variable v_i in the probability distribution, v_i is independent of all ancestral causes given its causal parents[8]. In the context of a causal model, this assumption is simply the reasonable claim that, once we know the direct causes of an event, then knowledge of more distant (indirect) causes provides no new information. To use a previous example[9], assume that the only cause of an increased concentration of photosynthetic enzymes in a leaf is the added fertiliser that was put on the ground, and that the only cause of an increased photosynthetic rate is the increased concentration of photosynthetic enzymes. Then, knowing how

[7] This can be generalized to joint distributions of sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ conditional on another set $\boldsymbol{Z}$.

[8] $\mathrm{P}(v_i) = \Pi \mathrm{P}(v_i \mid \text{parents}(v_i))$.

[9] Fertiliser→photosynthetic enzymes→photosynthetic rate.



$$P(A, B, C, D) = P(A) \times P(B) \times P(C|\{A, B\}) \times P(D|C)$$

Figure 2.8. A causal graph involving four variables and the joint probability distribution that is generated by it.

much fertiliser was added gives us no new information about the photosynthetic rate once we already know the concentration of photosynthetic enzymes in the leaf.

An important property of probability distributions that obey the Markov condition is that they can be decomposed into conditional probabilities involving only variables and their causal parents. For example, Figure 2.8 shows a causal graph and the joint probability distribution that is generated by it. This decomposition states that to know the probability distribution of D, we need only know the value of C; i.e. $P(D|C)$. To know the probability distribution of C we need only know the values of A and B; i.e. $P(C|\{A,B\})$. A and B are independent and so to know the joint probability distribution of A and B we need only know the marginal distributions of A and B; i.e. $P(A)P(B)$.

2.8 The translation from causal models to observational models

Although causal models and observational models are not the same thing, there is a remarkable relationship between the two. Consider first the case of causal graphs that do not have feedback relationships; that is, directed paths from some vertex that do not lead back to the same vertex. Theorem 10 of Pearl (1988) states that for any causal graph without feedback loops (a directed acyclic graph, or DAG), every d-separation statement obtained from the graph implies an independence relation in the joint probability distribution of the random variables represented by its vertices.

This central insight has been a long time in coming, and I imagine that many readers will wonder whether the effort was worth the return, so let me rephrase it:

> *Once we have specified the acyclic causal graph, then every d-separation relation that exists in our causal graph must be mirrored in an equivalent statistical independency in the observational data if the causal model is correct.*

The above statement is incredibly general; it does not depend on any distributional assumptions of the random variables or on the functional form of the causal relationships. In the same way, if even one statistical independency in the data disagrees with what d-separations of the causal graph predict, then the causal model must be wrong. This is the translation device that we needed in order to properly translate the causal claims represented in the directed graph into the 'official' language of probability theory used by statisticians to express observational models. After wading through the jargon developed above, I hope that the reader will recognise the elegant simplicity of this strategy (Figure 2.9). First, express one's causal hypothesis in a mathematical language (directed graphs) that can properly express the asymmetric types of relationship that scientists imply when they use the language of causality. Second, use the translation device (d-separation) to translate from this directed graph into the well-known mathematical language (probability theory) that is used in statistics to express notions of association. Finally, determine the types of (conditional) independence relationship that must occur in the resulting joint probability distribution. Continuing with the analogy of a correlation as being an observational shadow of the underlying causal process, the translation device (d-separation) is the method by which one can predict these shadows. The shadows are in the form of conditional independence relationships that the joint probability distribution (and therefore the observational model) must possess if the data are really generated by the hypothesised directed graph.

2.9 Counterintuitive consequences and limitations of d-separation: conditioning on a causal child

Although d-separation can also be used to obtain predictions concerning how a causal system will respond following an external manipulation[10], d-separation is really only a mathematical operation that gives the correlational consequences of conditioning on a variable in a causal system. One non-intuitive consequence is that two causally independent variables will be correlated if one conditions on any of their common children. This is because conditioning on a collider vertex along a path between vertices X

[10] This is explained later in this chapter.

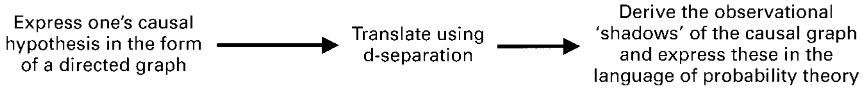

Express one's causal hypothesis in the form of a directed graph
Translate using d-separation
Derive the observational 'shadows' of the causal graph and express these in the language of probability theory

Figure 2.9. The strategy used to translate from a causal model to an observational model.

and *Y* means that *X* and *Y* are not d-separated. This has important consequences for applied regression analysis and shows how such a method can give very misleading results if these are interpreted as giving information about causal relationships.

Consider a causal system in which two causally independent variables (*X* and *Y*) jointly cause variable *Z*: $X \rightarrow Z \leftarrow Y$. To be more specific, let's assume that the nitrogen content (*X*) and the stomatal density (*Y*) of the leaves of individuals of a particular species jointly cause the observed net photosynthetic rate (*Z*). Further, assume that leaf nitrogen content and stomatal density are causally independent. So, the causal graph is: leaf nitrogen→net photosynthetic rate←stomatal density. Let the functional relationships between these variables be as follows:

$$\text{leaf nitrogen} = N(0,1)$$

$$\text{stomatal density} = N(0,1)$$

$$\text{net photosynthesis} = 0.5 \text{ leaf nitrogen} + 0.5 \text{ stomatal density} + N(0,0.707)$$

These three equations can be used to conduct numerical simulations[11] that can demonstrate the consequences of conditioning on a common causal child (net photosynthetic rate). Since I use this method repeatedly in this book, I will explain how it is done in some detail. The first equation states that the leaf nitrogen concentration of a particular plant has causes not included in the model. Since the plant is chosen at random, the leaf nitrogen concentration is simulated by choosing at random from a normal distribution whose population mean is zero and whose population standard deviation is 1. The second equation states that the stomatal density of the same leaf of this individual also has causes not included in the model (not the same unknown causes, since otherwise it would not be causally independent) and its value is simulated by choosing another (independent) number from the same probability distribution. The third equation states that the net photosynthetic rate of this same leaf is jointly caused by the two

[11] Such simulations are often called Monte Carlo simulations, after the famous gambling city, because they make use of random number generators to simulate a random process.

previous variables. The quantitative effect of these two causes on the net photosynthetic rate is obtained by adding 0.5 times the leaf nitrogen concentration plus 0.5 times the stomatal density plus a new (independent) random number taken from a normal distribution whose population mean is zero, whose population variance is $1-2(0.5^2)$, and whose population standard deviation is therefore the square root of this value; this third random variable represents all those other causes of net photosynthetic rate other than leaf nitrogen and stomatal density and these other unspecified causes are not causally connected to either of the specified causes. By repeating this process a large number of times, one obtains a random 'sample' of 'observations' that agree with the generating process specified by the equations[12]. As is described in Chapter 3, this model is actually a very simply path model. After generating 1000 independent 'observations' that agree with these equations, and respecting the causal relationships specified by our causal system, here are the regression equations that are obtained:

$$\text{leaf nitrogen} = N(0.035, 1.006)$$

$$\text{stomatal density} = N(-0.031, 1.017)$$

$$\text{net photosynthesis} = 0.003 + 0.527 \text{ leaf nitrogen} + 0.498 \text{ stomatal density} + N(0, 0.693)$$

Happily, the partial regression coefficients as well as the means and standard deviations of the random variables are what we should find, given sampling variation with a sample size of 1000. What happens if we give these data to a friend who mistakenly thinks that leaf nitrogen concentration is actually caused by net photosynthetic rate and stomatal density? That is, she mistakenly thinks that the causal graph is: net photosynthetic rate→leaf nitrogen←stomatal density. We know, because we generated the numbers, that leaf nitrogen and stomatal density are actually independent (the Pearson correlation coefficient between them is −0.037) but this is the set of regression equations that results from this incorrect causal hypothesis:

$$\text{net photosynthesis} = N(0.001, 0.994)$$

$$\text{stomatal density} = N(-0.031, 1.017)$$

$$\text{leaf nitrogen} = 0.023 + 0.70 \text{ net photosynthesis} - 0.366 \text{ stomatal density} + N(0, 0.799)$$

[12] Many commercial statistical packages can generate random numbers from specified probability distributions. A good reference, along with FORTRAN subroutines, is Press *et al.* (1986).

Tests of significance for the two partial regression coefficients show that each is significantly different from zero at a probability of less than 1×10^{-6}. Why would the multiple regression mistakenly report a highly significant 'effect' of stomatal density on leaf nitrogen when we know that they are both statistically and causally independent (because we made them that way in the simulation)? There is no 'mistake' in the statistics; rather it is our friend's interpretation that is mistaken. The regression equation is an observational model and it is simply telling us that knowing something about the net photosynthetic rate gives us extra information about (or helps to predict) the amount of nitrogen in the leaf, *when we compare leaves with the same stomatal density*[13]. This is exactly what d-separation, applied to the correct causal graph, tells us will happen: leaf nitrogen and stomatal density, while unconditionally d-separated, are not d-separated (therefore observationally associated) upon conditioning on their causal child (net photosynthetic rate).

This counterintuitive claim is easier to understand with an everyday example. Consider again the simple causal world consisting only of rain, watering pails and mud, related as: rain→mud←watering pails. Now, in this world there are no causal links between watering pails and rain. Knowing that no one has dumped water from the watering pail tells us nothing about whether or not it is raining; we can predict nothing about the occurrence of rain by knowing something about the watering pail. On the other hand, if we see that there is mud (the causal child of the two independent causes), *and* we know that no one has dumped water from the watering pail (i.e. conditional on this variable) then we can predict that it has rained. Conditioning on a common child of the two causally independent variables (rain and watering pails) renders them observationally dependent. This is because information, unlike causality, is symmetrical.

Many researchers believe that the more variables that can be statistically controlled in a multiple regression, the less biased and the more reliable the resulting model. The above example shows this to be wrong and warns against such methods as stepwise multiple regression if the resulting model is to be interpreted as something more than simply a prediction device[14]. This point is almost never mentioned in most statistics texts.

[13] Remember that a partial regression coefficient is a function of the partial correlation coefficient. The partial correlation coefficient measures the degree of linear association between two variables upon conditioning on some other set of variables; see Chapter 3.

[14] Even as a prediction device, such models are only valid if no manipulations are done to the population.

2.10 Counterintuitive consequences and limitations of d-separation: conditioning due to selection bias

There is also an interesting consequence of d-separation that might occur in experiments using artificial selection. 'Body condition' is a somewhat vague concept that is sometimes used to refer to the general health and vigour of an animal. It is occasionally operationalized as an index based on a weighting of such things as the amount of subcutaneous fat, the parasite load, or other variables judged relevant to the health of the species. Imagine a wildlife manager who wants to select for an improved body condition of Bighorn Sheep. His measure of body condition is obtained by adding together the thickness of subcutaneous fat in the autumn (in centimetres) and a score for parasite load (0 = none, 1 = average load, 2 = above-average load) as follows: body condition = 0.5 fat + parasite load. These two components of body condition are causally unrelated. He decides to protect all individuals whose body condition is greater than 3 and removes all others from the population by allowing hunters to kill them. The causal graph of this process is: fat thickness→body condition←parasite load. If someone were to then measure the fat thickness and parasite load in the remaining population after the selective hunt, she would find that these two variables were correlated, even though there is, in reality, no causal link between the two[15]. This occurs because the selection process has removed all those individuals not meeting the selection criterion and this effectively results in conditioning on body condition.

We can simulate this with the following generating equations[16].

fat thickness = Gamma(shape = 2)

parasite load = Multinomial(p = 1/3,1/3,1/3)

body condition = 0.5 fat thickness + parasite load

After generating 1000 independent 'sheep' following this process we find the Spearman non-parametric correlation coefficient between fat thickness and parasite load in the original population before artificial selection to be −0.018, consistent with independence. There were 493 'sheep'

[15] On the other hand, if this process were to be repeated for a number of generations and the two attributes were heritable, then there would develop a causal link, since the average values of the attributes in the next generation would depend on who survives, and this is caused by the same attributes in the previous generation.

[16] Gamma(shape = 2) is the incomplete Gamma distribution which gives values greater than zero with a right-tailed skew. Multinomial(1/3,1/3,1/3) means a multinomial distribution with equal probability of values being 0, 1 or 2.

whose body condition was at least 3, and so these are kept to represent the post-selection population, the rest being killed. The Spearman non-parametric correlation coefficient between fat thickness and parasite load for this post-selection population was -0.593. This occurs even though these two variables are causally independent.

2.11 Counterintuitive consequences and limitations of d-separation: feedback loops and cyclic causal graphs

The relationship between d-separation in an acyclic causal model (a directed acyclic graph) and independencies in a probability distribution is therefore very general. What happens if there are feedback loops in the causal model? We don't know for sure, although this is an area of active research (Richardson 1996b). Spirtes (1995) has shown that d-separation in a cyclic causal model still implies independence in the joint probability distribution that it generates, but only if the relationships are linear. Pearl and Dechter (1996) have also shown that the relationship between d-separation and probabilistic independence also holds if all variables are discrete without any restriction on the functional form of the relationships. Unfortunately, Spirtes (1995) has also shown, by a counter-example, that d-separation does not always imply probabilistic independence when the functional relationships are non-linear and the variables are continuous. There are some grammatical constructs in the language of causality for which no one has yet found a good translation.

There are other curious properties of causal models with feedback loops. Consider Figure 2.10. Such a causal model seems to violate many properties of causes. The relationship is no longer asymmetrical, since X causes Z (indirectly through Y) and Z also causes X. The relationship is no longer irreflexive, since X seems to cause itself through its effects on Y and Z.

These counterintuitive aspects of feedback loops can be resolved if we remember that causality is a process that must follow time's arrow but causal graphs do not explicitly include this time dimension. Causal graphs with feedback loops represent either a 'time slice' of an ongoing dynamic process or a description of this dynamic process at equilibrium, an interpretation that appears to have been first proposed by F. M. Fisher (1970). Richardson's very interesting Ph.D. thesis (Richardson 1996b) provides a history of the use and interpretation of such cyclic, or 'feedback' models[17]

[17] In the literature of structural equation modelling, cyclic or feedback models are called 'non-recursive'. This whole subject area is replete with confusing and intimidating jargon.

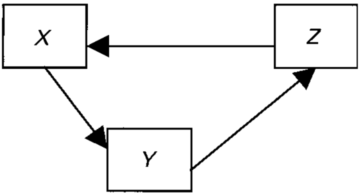


Figure 2.10. A cyclic causal graph that seemingly violates many of the properties of 'causal' relationships.

in economics. A more complete causal description of the process shown in Figure 2.10 is given in Figure 2.11; the subscripts on the vertices index the state of that vertex at a given time. From Figure 2.11 we see that, once the explicit time dimension is included in the directed graph, the apparent paradoxes disappear. Rather than circles, when we ignore the time dimension (as in Figure 2.10) we have spirals that never close on themselves when the time dimension is included. Just as the 20 year old Bill Shipley is not the same individual as I am as I write these words, the 'X' that causes Y at time $t=1$ will not be that same 'X' that is caused by Z at time $t=4$ in Figure 2.11.

Conceived in this way, both acyclic and cyclic causal models represent 'time slices' of some causal process. Samuel Mason, described by Heise (1975), provided a general treatment of feedback loops in causal graphs over 40 years ago for the case of linear relationships between variables. None the less, trying to model causal processes with feedback using directed graphs that ignore this time dimension is more complicated and requires that we make assumptions about the linearity of the functional relationships.

2.12 Counterintuitive consequences and limitations of d-separation: imposed conservation relationships

Relationships derived from imposed (as opposed to dynamic) conservation constraints are superficially similar to cyclic relationships, but they are conceptually quite different. By 'conservation' I mean variables that are constrained to maintain some conserved property. For instance, if I purchase fruits and vegetables in a shop and then count the total amount of money that I have spent, I can represent this as: money spent on fruits→total money spent←money spent on vegetables. If the total amount of money that I can spend is not fixed, then the amount that I spend on fruits and the amount that I spend on vegetables are causally independent. However, if the total amount of money is fixed, or *conserved*, due to some influence outside of the causal system then every dollar that I spend on fruit causes a decrease in the amount of money that I spend on vegetables. There is now a causal link

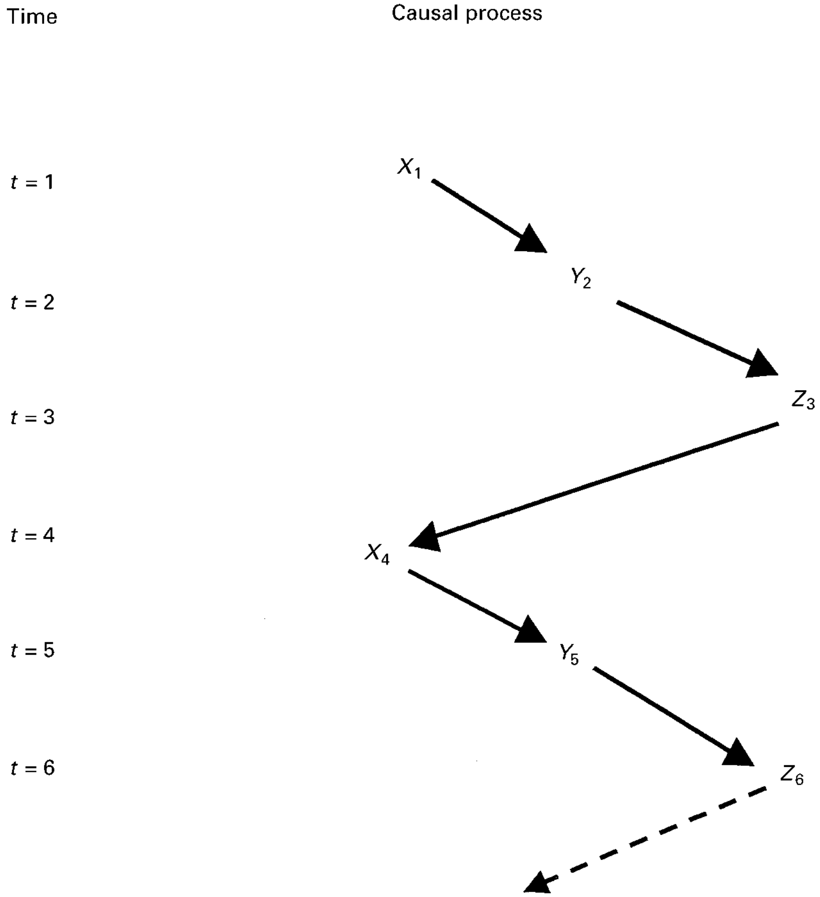


Figure 2.11. The causal relationships between *X*, *Y* and *Z* from Figure 2.10 when the time dimension is included in the causal graph.

between the amount of money spent on fruits and on vegetables due only to the requirement that the total amount of money be conserved.

There is no obvious way to express such relationships in a causal graph. One might be tempted to modify our original acyclic graph by adding a cyclic path between 'fruits' and 'vegetables' but, if we do this, then we can't interpret such a cyclic graph as a static graph of a dynamic process; the conservation constraint is imposed from outside and is not due to a dynamic equilibrium that results from the prior interaction of 'money spent on fruits' and 'money spent on vegetables'. In other words, it is not as if spending one dollar more on fruits at time $t=1$ causes me to spend one dollar less on vegetables at time $t=2$, which then causes me to spend one dollar less on fruits at time $t=3$, and so on until some dynamic equilibrium is attained. The conservation of the total amount of money spent is imposed from outside the causal system.

One might also be tempted to interpret the conservation requirement as equivalent to physically fixing the total amount of money at a constant value. If this were true, then one could maintain the causal graph 'money spent on fruits→total money spent←money spent on vegetables' but with the variable 'total money spent' being fixed due to the imposed conservation requirement. Because 'total money spent' is now viewed as being fixed rather than being allowed to vary randomly, then 'money spent on fruits' would not be d-separated from 'money spent on vegetables' (remember d-separation); this is because 'total money spent' is the causal child of each of 'money spent on fruits' and 'money spent on vegetables'. This would indeed imply a correlation between 'fruits' and 'vegetables'. Unfortunately, our causal system does not imply simply that the money spent on fruits is *correlated* with the money spent on vegetables, but that there is actually a causal connection between them that exists only when the conservation requirement is in place. d-separation upon conditioning on a common causal child does not imply that any new causal connections form between the causal parents. Perhaps the best causal representation is to consider that the causal graph 'money spent on fruits→total money spent←money spent on vegetables' is actually replaced by the causal graph 'money spent on fruits←total money spent→money spent on vegetables' with the convention that 'total money spent' is not random.

Systems that contain imposed conservation laws (conservation of energy, mass, volume, number, etc.) cannot yet be properly expressed using directed graphs and d-separation. In fact, such 'causal' relationships resemble Plato's notion of 'formal causes' rather than the 'efficient causes' with which scientists are used to working. It is important to keep in mind, however, that this does not apply to conservation relationships that are due to a dynamic equilibrium, for which cyclic graphs can be used, but rather to conservation relationships that are imposed independently of the casual parents of the conserved variable.

2.13 Counterintuitive consequences and limitations of d-separation: unfaithfulness

Let's go back to the relationship between d-separation and probabilistic independence. We now know that once we have specified the acyclic causal model, then every d-separation relation that exists in our causal model must be mirrored in an equivalent statistical independency in the observational data if the causal model is correct. This does not depend on any distributional assumptions of the random variables or on the functional form of the causal relationships. Is the contrary also true? Can there be independencies in the data that are not predicted by the d-separation criterion?

Yes, but only as limiting cases. For instance, this can occur if the quantitative causal effect of two variables along different directed paths exactly cancel each other out. Two examples are shown in Figure 2.12. In these causal models we see that no vertex is unconditionally d-separated from any other vertex. Assume that the joint probability distribution over the three vertices is multivariate normal and that the functional relationships between the variables are linear. Under these conditions, we can use Pearson's partial correlation to measure probabilistic independence[18]. By definition, the partial correlation between X and Z, conditioned on Y,[19] is given by:

$$\rho_{XZ.Y}=\frac{\rho_{XZ}-\rho_{XY}\rho_{ZY}}{\sqrt{(1-\rho_{XY}^2)(1-\rho_{ZY}^2)}}$$

It can happen that $\rho_{XZ.Y}=0$ (i.e. $\rho_{XZ}=\rho_{XY}\rho_{ZY}$) even though X and Z are not d-separated given Y, if the correlations between each pair of variables exactly cancel each other. Using the rules of path analysis (Chapter 4), this will happen only if Y is perfectly correlated with X in the first model in Figure 2.12, or if the indirect effect of X on Z is exactly equal in strength but opposite in sign to the direct effect of X on Z.

When this occurs, we say that the probability distribution is *unfaithful* to the causal graph (Pearl 1988; Spirtes, Glymour and Scheines 1993). I will call such probabilistic independencies that are not predicted by d-separation, and that depend on a particular combination of quantitative effects, *balancing independencies*, to emphasise that such independencies require a very peculiar balancing of the positive and negative effects between the variables along different paths. Clearly, this can occur only under very special conditions, and anyone who wanted to link a causal model with such an unfaithful probability distribution would require strong external evidence to support such a delicate balance of causal effects. This is not to say that these things are impossible. It sometimes occurs that an organism attempts to maintain some constant set-point value by balancing different causal effects; an example is the control of the internal CO_2 concentration of a leaf, as described in Chapter 3. Essentially, in proposing such a claim we are saying that nature is conspiring to give the impression of independence by exactly balancing the positive and negative effects.

[18] Pearson partial correlations are explained more fully in Chapter 3.

[19] See p. 84 for a more detailed explanation of the notation $XZ.Y$.

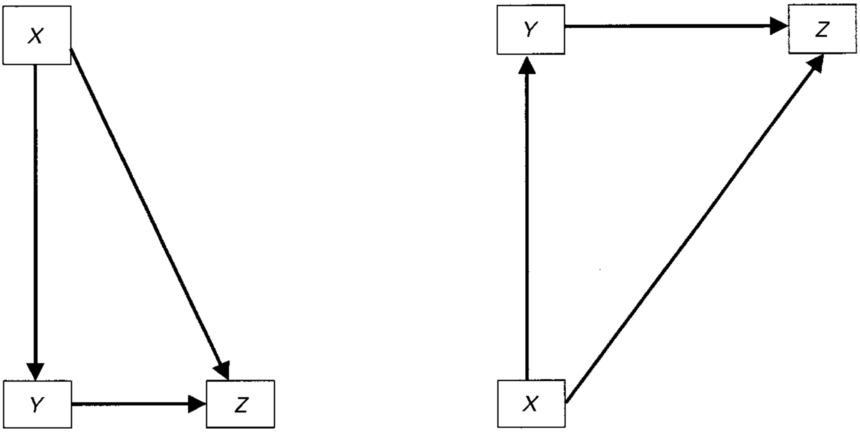


Figure 2.12. Two causal graphs for which special combinations of causal strengths can result in unfaithful probability distributions.

2.14 Counterintuitive consequences and limitations of d-separation: context-sensitive independence

Another way in which independencies can occur in the joint probability distribution without being mirrored in the d-separation criterion is due to *context-sensitive* independence. An example of this in biology is enzyme induction[20]. Imagine a case in which the number (G) of functional copies of a gene determines the rate (E) at which some enzyme is produced. If there are no functional copies of the gene then the enzyme is never produced. However, the rate at which these genes are transcribed is determined by the amount (I) of some environmental inducer. If the environment completely lacks the inducer, then no genes are transcribed and the enzyme is still never produced. It is possible to arrange an experimental set-up in which the number (G) of functional genes is causally independent of the concentration (I) of the inducer in the environment[21]. Both the number of functional genes and the concentration of the inducer are causes of enzyme production. We can construct a causal graph of this process (Figure 2.13).

Now, applying d-separation to the causal graph in Figure 2.13 predicts that G is independent of I, but that E is dependent on both G and I.

[20] A classic example is the *lac* operon of *Escherichia coli,* whose transcription in the presence of lactose induces the production of β-galactosidase, lac permease and transacetylase, thus converting lactose into galactose and glucose (De Robertis and De Robertis 1980).

[21] Whether this would be true in the biological population is an empirical question. Perhaps the presence of a functional gene was selected based on the presence of the inducer. In this case, the inducer would be a cause of the presence (and perhaps the number of copies) of the gene.

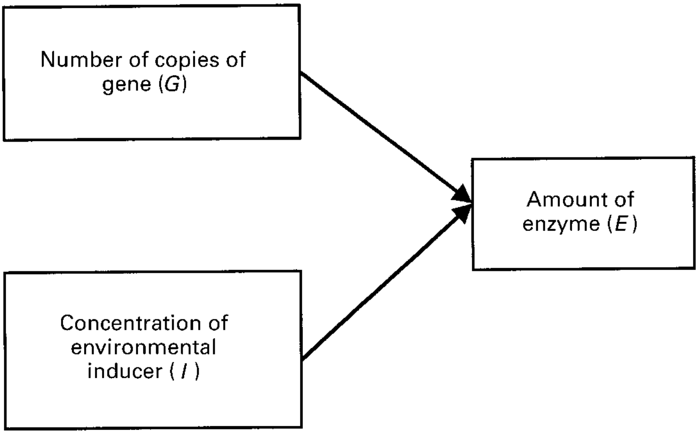

Number of copies of gene (*G*)
Amount of enzyme (*E*)
Concentration of environmental inducer (*I*)

Figure 2.13. A biological example of a causal process that can potentially result in context-sensitive independence.

However, if there are no copies of *G* (i.e. $G=0$) then the concentration of the inducer will be independent of the amount of enzyme that is produced (which will be zero). Similarly, if there is no inducer (i.e. $I=0$) then the number of copies of the gene will be independent of the amount of enzyme that is produced (which will be zero). In other words, for the special cases of $G=0$ and/or $I=0$ d-separation predicts a dependence when, in fact, there is independence. Note that the d-separation theorem still holds; d-separation does not predict any *independence* relations that do not exist. So long as the experiment involves experimental units, at least some of which include $G\neq 0$ and $I\neq 0$, the d-separation criterion still predicts both probabilistic independence and dependence. Similarly, if both *G* and I were true random variables (i.e. in which the experimenter did not fix their values), then any reasonably large random sample would include such cases.

2.15 The logic of causal inference

Now that we have our translation device and are aware of some of the counterintuitive results and limitations that can occur with d-separation, we have to be able to infer causal consequences from observational data by using this translation device. The details of how to carry out such inferences will occupy most of Chapters 3 to 7. Before looking at the statistical details, however, we must first consider the logic of causal and statistical inferences.

Since we are talking about the logic of inferences from empirical experience, it is useful to briefly look at what philosophers of science have had to say about valid inference. Logical positivism, itself being rooted in the British empiricism of the last century that so influenced people like Karl Pearson[22], was dominant in this century up to the mid 1960s. This philosophical school was based on the verifiability theory of meaning; to be meaningful, a statement had to be of a kind that could be shown to be either true or false. For logical positivism, there were only two kinds of meaningful statement. The first kind was composed of *analytical* statements (tautologies, mathematical or logical statements) whose truth could be determined by deducing them from axioms or definitions. The second kind was composed of *empirical* statements that were either self-evident observations ('the water is 23 °C') or could be logically deduced from combinations of basic observations whose truth was self-evident[23]. Thus logical positivists emphasised the hypothetico-deductive method: a hypothesis was formulated to explain some phenomenon by showing that it followed deductively from the hypothesis. The scientist attempted to validate the hypothesis by deducing logical consequences of the hypothesis that were not involved in its formulation and testing these against additional observations. A simplified version of the argument goes like this:

- If my hypothesis is true, then consequence *C* must also be true.
- Consequence *C* is true.
- Therefore my hypothesis is true.

Readers will immediately recognise that such an argument commits the logical fallacy of affirming the consequent. It is possible for the consequence to be true even though the hypothesis that deduced it is false, since there can always be other reasons for the truth of *C*.

Popper (1980) pointed out that, although we cannot use such an argument to verify hypotheses, we can use it to reject them without committing any logical fallacy:

- If my hypothesis is true, then consequence *C* must also be true.
- Consequence *C* is false.
- Therefore my hypothesis is false.

Practising scientists would quickly recognise that this argument, although logically acceptable, has important shortcomings when applied to empirical studies. It was recognised as long ago as the turn of the century (Duhem 1914) that no hypothesis is tested in isolation. Every time that we draw a conclusion from some empirical observation we rely on a whole set

[22] This is explored in more detail in Chapter 3.

[23] That even such simple observational or experiential statements cannot be considered objectively self-evident was shown at the beginning of the twentieth century by Duhem (1914).

of auxiliary hypotheses (A_1, A_2 . . .) as well. Some of these have been repeatedly tested so many times and in so many situations that we scarcely doubt their truth. Other auxiliary assumptions may be less well established. These auxiliary assumptions will typically include those concerning the experimental or observational background, the statistical properties of the data, and so on. Did the experimental control really prevent the variable from changing? Were the data really normally distributed, as the statistical test assumes? Such auxiliary assumptions are legion in every empirical study, including the randomised experiment, the controlled experiment or the methods described in this book involving statistical controls. A large part of every empirical investigation involves checking, as best one can, such auxiliary assumptions so that, once the result is obtained, blame or praise can be directed at the main hypothesis rather than at the auxiliary assumptions.

So, Popper's process of inference might be simplistically paraphrased[24] as:

- If auxiliary hypotheses A_1, A_2, . . . A_n are true, and
- if my hypothesis is true, then consequence C must be true.
- Consequence C is false.
- Therefore, my hypothesis is false.

Unfortunately, to argue in such a manner is also logically fallacious. Consequence C might be false, not because the hypothesis is false, but rather because one or more of the auxiliary hypotheses are false. The empirical researcher is now back where he started: there is no way of determining either the truth of falsity of his or her hypothesis in any absolute sense from logical deduction. This conclusion applies just as well to the randomised experiment, the controlled experiment or the methods described in this book. Yet, most biologists would recognise the falsifiability criterion as important to science and would probably modify the simplistic paraphrase of Popper's inference by attempting to judge which, the auxiliary hypotheses and background conditions, or the hypothesis under scrutiny, is on firmer empirical ground. If the auxiliary assumptions seem more likely to be true than the hypothesis under scrutiny, yet the data do not accord with the predicted consequences, then the hypothesis would be tentatively rejected. If there are no reasoned arguments to suggest that the auxiliary assumptions are false, and the data also accord with the predictions of the hypothesis under scrutiny, then the hypothesis would be tentatively accepted.

Pollack (1986) called such reasoning *defeasible* reasoning[25]. Reveal-

[24] *Simplistic* because it is wrong. Popper did not make such a claim.

[25] *Defeasible* because it can be *defeated* with subsequent evidence.

ingly, practising scientists have explicitly described their inferences in such terms for a long time. At the turn of the century T. H. Huxley likened the decision to accept or reject a scientific hypothesis to a criminal trial in a court of law (reproduced in Rapport and Wright 1963) in which guilt must be demonstrated beyond reasonable doubt.

Let's apply this reasoning to the examples in Chapter 1 involving the randomised and the controlled experiments. Later, I will apply the same reasoning to the methods involving statistical control.

Here is the logic of causal inference with respect to the randomised experiment to test the hypothesis that fertiliser addition increases seed yield:

- If the randomisation procedure was properly done so that the alternative causal explanations were excluded;
- if the experimental treatment was properly applied;
- if the observational data do not violate the assumptions of the statistical test;
- if the observed degree of association was not due to sampling fluctuations;
- then by the causal hypothesis the amount of seed produced will be associated with the presence of the fertiliser.
- There is/is not an association between the two variables.
- Therefore, the fertiliser addition might have caused/did not cause the increased seed yield.

This list of auxiliary assumptions is only partial. In particular, we still have to make the basic assumption linking causality to observational associations, as described in Chapter 1. At this stage we must either reject one of the auxiliary assumptions or tentatively accept the conclusion concerning the causal hypothesis. If the probability associated with the test for the association is sufficiently large[26], traditionally above 0.05, then we are willing to reject one of the auxiliary assumptions (the observed measure of

[26] See Cowles and Davis (1982b) for a history of the 5% significance level. The first edition of Fisher's (1925) classic book states: 'It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant'. The words 'convenient' and 'formal' emphasise the somewhat arbitrary nature of this value. In fact, this level can be traced back even further to the use of three times the probable error (about 2/3 of a standard deviation). Strictly speaking, twice the standard deviation of a normal distribution gives a probability level of 0.0456; perhaps Fisher simply rounded this up to 0.05 for his tables. E. S. Pearson and Kendall (1970) record Karl Pearson's reasons at the turn of the century: $p=0.5586$ 'thus we may consider the fit remarkably good'; $p=0.28$ 'fairly represented'; $p=0.10$ 'not very improbable'; $p=0.01$ 'this very improbable result'. Note that some doubt began at 0.1 and Pearson was quite convinced at $p=0.01$. The midpoint

association was not due to sampling fluctuations) rather than accept the causal hypothesis. Thus we reject our causal hypothesis. This rejection must remain tentative. This is because another of the auxiliary assumptions (not listed above) is that the sample size is large enough to permit the statistical test to differentiate between sampling fluctuations and systematic differences. Note, however, that it is not enough to propose any old reason to reject one of the auxiliary assumptions; we must propose a reason that has empirical support. We must produce *reasonable* doubt – in the context of the assumption concerning sampling fluctuations scientists generally require a probability above 0.05. Here it is useful to cite from the first edition of Fisher's (1925) influential *Statistical methods for research workers*: 'Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.' It is clear that Fisher was demanding reasonable doubt concerning the null hypothesis, since he asks only that a result 'rarely fail' to reject it. What if the probability of the statistical test was sufficiently small, say 0.01, that we do not have reasonable grounds to reject our auxiliary assumption concerning sampling fluctuations? What if we do not have reasonable grounds to reject the other auxiliary assumptions? What if the sampling variation was small compared with a reasonable effect size? Then we must tentatively accept the causal hypothesis. Again, this acceptance must remain tentative, since new empirical data might provide such reasonable doubt. Is there any automatic way of measuring the relative support for or against each of the auxiliary assumptions and of the principal causal hypothesis? No. Although the support (in terms of objective probabilities) for some assumptions can be obtained – for instance, those concerning normality or linearity of the data – there are many other assumptions that deal with experimental procedure or lack of confounding variables for which no objective probability can be calculated. This is one reason why so many contemporary philosophers of science prefer Bayesian methods to frequency-based interpretations of probabilistic

footnote 26 (*cont.*)
between 0.1 and 0.01 is 0.05. Cowles and Davis (1982a) conducted a small psychological experiment by fooling students into believing that they were participating in a real betting game (with money) that was, in reality, fixed. The object was to see how unlikely a result people would accept before they began to doubt the fairness of the game. They found that 'on average, people do have doubts about the operation of chance when the odds reach about 9 to 1 [i.e. 0.09], and are pretty well convinced when the odds are 99 to 1 [i.e. ~0.0101] . . . If these data are accepted, the 5% level would appear to have the appealing merit of having some grounding in common sense'.

inference (see, for example, Howson and Urbach 1989). Such Bayesian methods suffer from their own set of conceptual problems (Mayo 1996). In the end, even the randomised experiment requires subjective decisions on the part of the researcher. This is why the independent replication of experiments in different locations, using slightly different environmental or experimental conditions and therefore having different sets of auxiliary assumptions, is so important. As the causal hypothesis continues to be accepted in these new experiments, it becomes less and less reasonable to suppose that incorrect auxiliary assumptions are conspiring to give the illusion of a correct causal hypothesis.

Here is the logic of our inferences with respect to the controlled experiment to test the hypothesis that renal activity causes the change in the colour of the renal vein blood, as described in Chapter 1:

- If the activity of the kidney was effectively controlled;
- if the colour of the blood was accurately determined;
- if the experimental manipulation did not change some other uncontrolled attribute besides kidney function that is a common cause of the colour of blood in the renal vein before entering, and after leaving the kidney;
- if there was not some unknown (and therefore uncontrolled) common cause of the colour of blood in the renal vein before entering, and after leaving the kidney;
- if a rare random event did not occur;
- then by the causal hypothesis, blood will change colour only when the kidney is active.
- The blood did change colour in relation to kidney activity.
- Therefore, kidney activity does cause the change in the colour of blood leaving the renal vein.

Again, this list of auxiliary assumptions is only partial. Again, one must either produce reasonable evidence that one or more of the auxiliary assumptions is false or tentatively accept the hypothesis. In particular, more of these auxiliary assumptions concern properties of the experiment or of the experimental units for which we cannot calculate any objective probability concerning their veracity. This was one of the primary reasons why Fisher rejected the controlled experiment as inferior. In the controlled experiment these auxiliary assumptions are more substantial but it is still not enough to raise any doubt; there must be some empirical evidence to support the decision to reject one of these assumptions. Since we want the data to cast doubt or praise on the principal causal hypothesis and not on the auxiliary assumptions, we will ask only for evidence that casts reasonable doubt. It is not enough to reject the causal hypothesis simply because 'experimental manipulation *might have* changed some other uncontrolled attribute besides kidney function that is a common cause of the colour of blood in the renal vein before entering, and after leaving the kidney'. We must advance *some* evidence to support the idea that such an uncontrolled factor actually exists. For instance, a critic might reasonably point out that some other attribute is also known to be correlated with blood colour and that the experimental manipulation was known to have changed this attribute. Although such evidence would certainly not be sufficient to demonstrate that this other attribute definitely was the cause, it might be enough to cast doubt on the veracity of the principal hypothesis. This is the same criterion as we used before to choose a significance level in our statistical test. Rejecting a statistical hypothesis because the probability associated with it was, say, 0.5 would not be reasonable. Certainly, this gives some doubt about the truth of the hypothesis but our doubt is not sufficiently strong that we would have a clear preference for the contrary hypothesis. It is the same defeasible argument that might be raised in a murder trial. If the prosecution has demonstrated that the accused had a strong motive, if it produced a number of reliable eyewitnesses and if it produced physical evidence implicating the accused, then it would not be enough for the defence to claim simply that 'maybe someone else did it'. If, however, the defence could produce some contrary empirical evidence implicating someone else, then reasonable doubt would be cast on the prosecution's argument. In fact, I think that the analogy between testing a scientific hypothesis and testing the innocence of the accused in a criminal trial can be stretched even further. There is no objective definition of reasonable doubt in a criminal trial; what is reasonable is decided by the jury in the context of legal precedence. In the same way, there is no objective definition of reasonable doubt in a scientific claim. In the first instance reasonable doubt is decided by the peer reviewers of the scientific article and, ultimately, reasonable doubt is decided by the entire scientific community. One should not conclude from this that such decisions are purely subjective acts and that scientific claims are therefore simply relativistic stories whose truth is decided by fiat by a power elite. Judgements concerning reasonable doubt and statistical significance are constrained in that they must deliver predictive agreement with the natural world in the long run.

Now let's look at the process of inference with respect to causal graphs.

- If the data were generated according to the causal model;
- if the causal process generating the data does not include non-linear feedback relationships;

- if the statistical test used to test the independence relationships is appropriate for the data;
- if a rare sampling fluctuation did not occur;
- then each d-separation statement will be mirrored by a probabilistic independence in the data.
- At least one predicted probabilistic independence did not exist;
- therefore, the causal model is wrong.

By now, you should have recognised the similarity of these inferences. We can prove by logical deduction that d-separation implies probabilistic independence in such directed acyclic graphs. We can prove that, barring the case of non-linear feedback with non-normal data (an auxiliary assumption), every d-separation statement obtained from any directed graph must be mirrored by a probabilistic independence in any data that were generated according to the causal process that was coded by this directed graph. We can prove that, barring a non-faithful probability distribution (another auxiliary assumption, but one that is only relevant if the causal hypothesis is accepted, not if it is rejected), there can be no independence relation in the data that is not mirrored by d-separation. So, if we have used a statistical test that is appropriate for our data and have obtained a probability that is sufficiently low to reasonably exclude a rare sampling event, then we must tentatively reject our causal model. As in the case of the controlled experiment, if we are led to tentatively accept our causal model, then this will require that we can't reasonably propose an alternative causal explanation that also fits our data as well. As always, it is not sufficient to simply claim that '*maybe* there is such an alternative causal explanation'. One must be able to propose an alternative causal explanation that has at least enough empirical support to cast reasonable doubt on the proposed explanation.

2.16 Statistical control is not always the same as physical control

We have now seen how to translate from a causal hypothesis into a statistical hypothesis. First, transcribe the causal hypothesis into a causal graph showing how each variable is causally linked to other variables in the form of direct and indirect effects. Second, use the d-separation criterion to predict what types of probabilistic independence relationship must exist when we observe a random sample of units that obey such a causal process. In Chapter 1 I alluded to the fact that the key to a controlled experiment is *control* over variables, not how the control is produced. It is time to look at this more carefully. The relationship between control through external (experimental) manipulation and probability distributions is given by the

Manipulation Theorem (Spirtes, Glymour and Scheines 1993). Let me introduce another definition in Box 2.2.

Box 2.2. Definition of a backdoor path

Given two variables, *X* and *Y*, and a variable *F* that is a causal ancestor of both *X* and *Y*, a backdoor path goes from *F* to each of *X* and *Y*. Thus

$$X \leftarrow \quad \leftarrow \leftarrow F \rightarrow \rightarrow \quad \rightarrow Y$$

Whenever someone directly physically controls some set of variables through experimental manipulation, he or she is changing the causal process that is generating the data. Whenever someone physically fixes some variable at a given level the variable stops being random[27] and is then under the complete control of the experimenter. In other words, whatever causes might have determined the 'random' values of the variable *before* the manipulation have been removed *by* the manipulation. The only direct cause of the controlled variable after the manipulation has been performed is the will of the experimenter.

Imagine that someone has randomly sampled herbaceous plants growing in the understorey of an open stand of trees. The measured variables are the light intensities experienced by the herbaceous plants, their photosynthetic rates and the concentration of anthocyanins (red-coloured pigments) in their leaves. Each of these three are random variables, since they are outside the control of the researcher. One cause of variation in light intensity at ground level is the presence of trees. The researcher proposes two alternative causal explanations for the data (Figure 2.14).

To test between these two explanations, the researcher experimentally manipulates light intensity by installing a neutral-shade cloth between the trees and the herbs, and then adds an artificial source of lighting. Remembering that this is a controlled experiment, the researcher would want to take precautions to ensure that other environmental variables (temperature, humidity and so on) are not changed by this manipulation. The Manipulation Theorem, in graphical terms[28], states that the probability distribution of this new causal system can be described by taking the original (unmanipulated) causal graphs, removing any arrows leading into

[27] The notion of 'randomness' is another example of a concept that is regularly invoked in science even though it is extraordinarily difficult to define.

[28] The Manipulation Theorem also predicts how the joint probability distribution in the new manipulated causal system differs, if at all, from the original distribution before the manipulation.

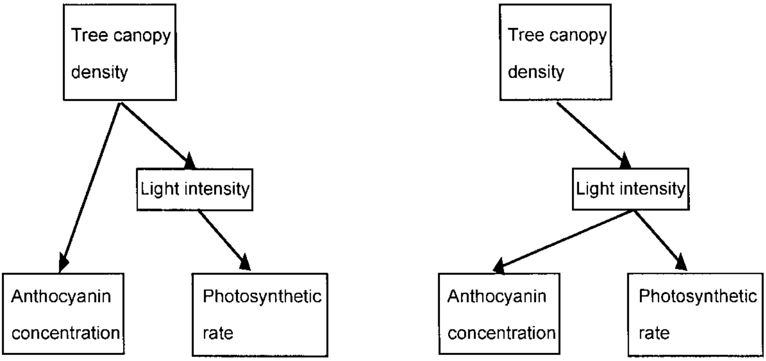


Figure 2.14. Two different causal scenarios linking the same four variables.

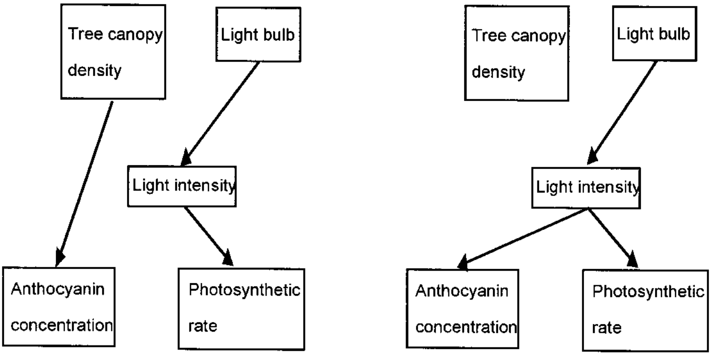


Figure 2.15. Experimental manipulation of the causal systems that are shown in Figure 2.14.

the manipulated variable (light intensity) and adding a new variable representing the new causes of the manipulated variable (Figure 2.15).

d-separation will predict the pattern of probabilistic independencies in this new causal system. Notice that anthocyanin concentration is d-separated from photosynthetic rate according to the first hypothesis in both the manipulated system (Figure 2.15), when light intensity is experimentally fixed, and in the unmanipulated system (Figure 2.14), when light intensity is statistically fixed by conditioning. The same d-connection relationships between anthocyanin concentration and photosynthetic rate hold in the second scenario whether based on physically or on statistically controlling light intensity. In other words, statistical and experimental controls are alternative ways of doing the same thing: predicting how the associations between variables will change once other sets of variables are 'held con-

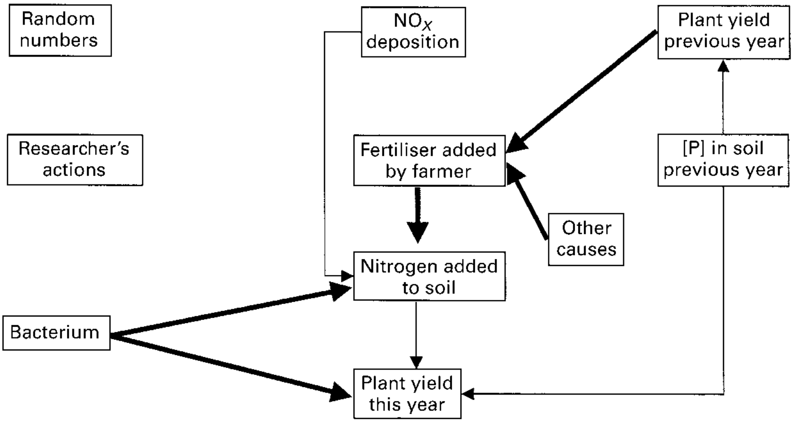


Figure 2.16. A hypothetical causal system before experimental manipulation.

stant'. This does not mean that the two types of control always predict the same types of observational independency in our data; remember the example of d-separation upon conditioning on a causal child, described previously. Once we have a way of measuring how closely the predictions agree with the observations, then we have a way of testing, and potentially falsifying, causal hypotheses even in cases in which we cannot physically control the variables of interest.

With these notions we can now go back and look again at the randomised experiment in Chapter 1. Let's consider an example involving an agricultural researcher who is interested in determining whether, and how, the addition of a nitrate fertiliser can increase plant yield. To be more specific, imagine that the plant is Alfalfa, which contains a bacterium in its roots that is capable of directly fixing atmospheric nitrogen (N_2). The researcher meets a farmer who tells him that adding such a nitrate fertiliser in the past had increased the yield of Alfalfa. After further questioning, the researcher learns that the farmer had tended to add more fertiliser to those parts of the field that, in previous years, had produced the lowest yields. The researcher knows that other things can also affect the amount of fertiliser that a farmer will add to different parts of a field. For instance, parts of the field that cause the farmer to slow down the speed of his tractor will therefore tend to receive more fertiliser, and so on[29]. Imagine that, unknown to the researcher, the actual causal processes are as shown in Figure 2.16. There are only three sources of nitrogen: the nitrate that is added to the soil by the fertiliser, by

[29] Readers with experience with tractors will have to assume that the governor is not functioning!

NO_X deposition, and from N_2 fixation by the bacterium. The amount of fertiliser added by the farmer in different parts of the field is determined by the yield of plants the previous year as well as the contours of the field. In reality, all the sources of nitrogen and the soil phosphate level [P] are causes of yield.

Before experimenting with this system, the researcher has previous causal knowledge of only part of it, shown by the thicker arrows in Figure 2.16. He knows that the bacterium will increase Alfalfa yield. He knows that the bacterium will increase the nitrate concentration in the soil. He knows that the yield of Alfalfa in previous years has affected the amount of nitrate fertiliser that the farmer had added, and he knows that the amount of added nitrate fertiliser is *associated* with increased yields. What he doesn't know is whether or not nitrogen added to the soil is the cause of the subsequent plant yield.

Since the experiment has not yet begun, the 'random numbers' in Figure 2.16 do not affect any actions by the researcher and the researcher has no causal effect on any variable in the system. The 'random numbers' and the 'researcher's actions' are therefore causally independent of each other and of every other variable in the system.

Based only on the *partial* knowledge shown by the thick arrows, can the researcher use d-separation and statistical control to confidently infer that the added nitrate fertiliser causes an increase in plant yield? No. He knows that the yields of previous years were a cause of the farmer's fertiliser addition and not *vice versa*; therefore he knows that he can block any possible backdoor path between the amount of fertiliser added and plant yield that passes through the variable 'plant yield the previous year'. Unfortunately, he also knows that this was not the *only* possible cause of the amount of fertiliser added by the farmer to different parts of the field. Therefore, he can't exclude the possibility that there is some backdoor path that does not include the variable 'plant yield the previous year' and that is generating the association between present plant yield and the amount of fertiliser added by the farmer. Remember that, to invoke such a possibility, one must be able to present some empirical evidence that such a backdoor path might exist, but this would be easy to do. For instance, if the tractor slows down as it begins to go up a slope (and therefore deposits more fertiliser), and if water (which is known to increase plant yield) tends to accumulate at the bottom of the slope then we have a possible backdoor path (fertiliser added ← tractor slowed down ← hill → water accumulation → plant yield).

The researcher knows that it is possible to randomly assign different levels of nitrate fertiliser to plots of ground in a way that is not caused by

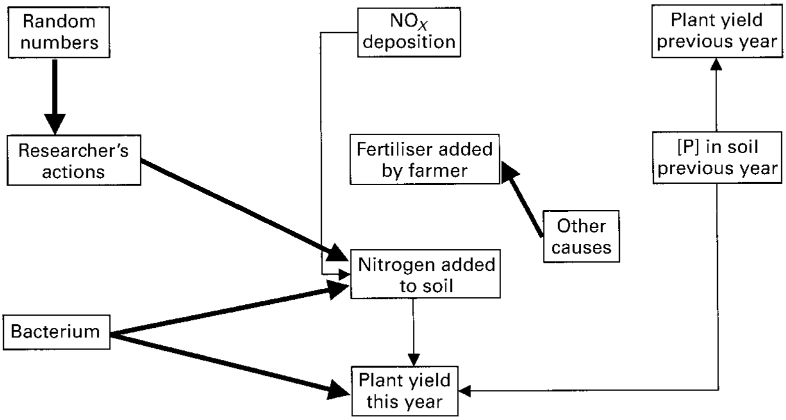


Figure 2.17. Experimental manipulation of the causal system shown in Figure 2.16 based on a randomised experiment.

any attribute of these plots. He convinces the farmer not to add any fertiliser. The previous cause of the amount of fertiliser added has been erased in this new context and so the arrow from 'plant yield previous year' to 'fertiliser added by farmer' is removed from the causal graph. Since the farmer has agreed not to add any fertiliser, the value of this variable is fixed at zero, and so all arrows coming out of this variable are also erased. The researcher decides to add nitrate fertiliser to different plots at either 0 or 20 kg/hectare, based only on the value of randomly chosen numbers. Therefore we add an arrow from 'random numbers' to 'researcher's actions' and also an arrow from 'researcher's actions' to 'nitrate added to soil'. Remember that an arrow signifies a *direct* cause, i.e. a causal effect that is not mediated through other variables in the causal explanation. Therefore we can't add an arrow from 'researcher's actions' to 'plant yield this year' unless we believe that the researcher's actions do cause a change in plant yield this year and that this cause is not completely mediated by some other set of variables in the causal system. Therefore, the causal structure that exists after the experimental manipulation is shown in Figure 2.17.

Given this new causal scenario, we can now use d-separation to determine whether there is a causal relationship between the amount of nitrate fertiliser added by the researcher and the plant yield that year. If one can trace a directed path beginning at 'researcher's actions' and passing through 'plant yield this year' by following the direction of the arrows, then the two are not d-separated. This necessarily implies that there will be a statistical association between the two variables. If no such directed path exists, then the addition of nitrate fertiliser by the researcher does not cause a change in plant yield this year. In fact, these two variables are not d-separated in this causal graph and so such a randomised experiment would detect an effect of fertiliser addition on plant yield. In Chapter 1 I said that if there is a statistical association between two variables, X and Y, then there can be only three elementary (but not mutually exclusive) causal explanations: X causes Y (shown by a directed path leading from X and passing into Y), or Y causes X (shown by a directed path leading from Y and passing into X), or there is some other variable (F) that is a cause of both X and Y (shown by a backdoor path from F and into both X and Y). Because the researcher has agreed to act completely in accordance with the results of the randomisation process, we know that no arrows point into 'researcher's actions' except the one coming from 'random numbers'. The random numbers are not caused by any attribute of the system. Therefore the researcher knows that there can be no backdoor paths confounding the results because he knows that there are no arrows pointing into 'researcher's actions' except for the one coming from 'random numbers'. If there is a statistical association between 'researcher's actions' and 'plant yield this year' that can't reasonably be attributed to random sampling fluctuations then the researcher knows that the association must be due to a directed path coming from 'researcher's actions' and passing through 'plant yield this year'. This is why such a randomised experiment, in conjunction with a way of calculating the probability of observing such a random event, can provide a strong inference concerning a causal effect. The reader should note that even the randomisation process might not allow the researcher to conclude that 'nitrate added to the soil' is a *direct* cause of increased plant yield. In Figure 2.17 the researcher has already concluded that there is a backdoor path from these two variables emanating from the presence of the nitrogen-fixing bacterium, and so to make such a claim he would have to provide evidence beyond a reasonable doubt that his actions did not somehow affect the abundance or activity of these bacteria.

Now, let's modify the causal scenario a bit. Imagine that the farmer has agreed to let the researcher conduct an experiment and promises not to add any fertiliser while the experiment is in progress, but insists that the parts of the field that had produced the lowest plant yield last year must absolutely receive more fertiliser this year. The researcher decides to allocate the fertiliser treatment in the following way: after choosing the random numbers as before, he also adds 5 kg/h to those plots whose previous yields were below the median value. Figure 2.18 shows this causal scenario. By doing so he is no longer conducting a true randomised experiment.

Now, using d-separation we see that there would be an association between 'researcher's actions' and 'plant yield this year' even if there were no causal effect of the amount of nitrate fertiliser added and the plant yield

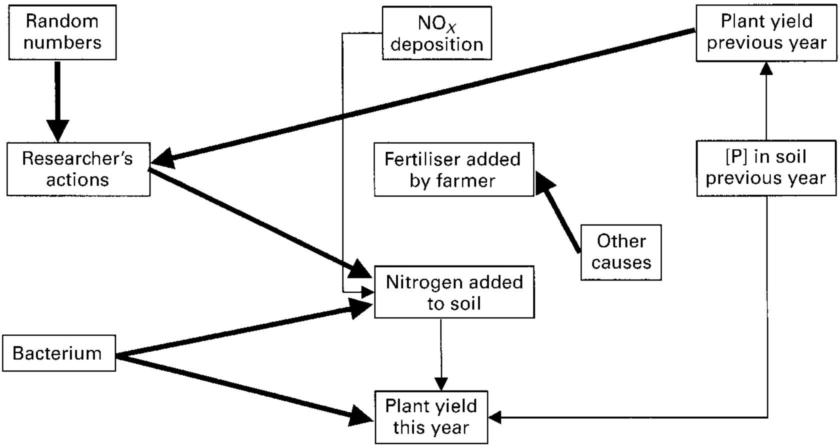


Figure 2.18. Experimental manipulation of the causal system shown in Figure 2.16 that is not based on a randomised experiment.

that follows. The reason is because there is now a backdoor path linking the two variables through the common cause '[P] in soil the previous year'. This path has been created by allowing 'plant yield the previous year' to be a cause of the researcher's actions. Yet all is not lost. He systematically assigned fertiliser levels based *only* on the yield data of the previous year plus the random numbers. This means that he knows that there are only two independent causes determining how much fertiliser each plot received. He also knows, because of d-separation, that any causal signal passing from any unknown variable into 'researcher's actions' through 'plant yield the previous year' is blocked if he statistically controls for 'plant yield the previous year'. He can make this causal inference without knowing anything else about the causal system. Therefore he knows that once he statistically conditions on 'plant yield the previous year' then any remaining statistical association, if it exists, must be due to a causal signal coming from 'researcher's actions' and following a directed path into 'plant yield this year'. This causal inference is just as solid as in the previous example in which treatment allocation was due only to random numbers. What allows him to do this in this controlled, but not strictly randomised, experiment but not in the original non-manipulated system in which the farmer applied the fertiliser based on previous yield data? If you compare Figures 2.16 (non-manipulated) and 2.18 (controlled, non-randomised manipulation) you will see that in Figure 2.16 there were other causes, besides yield, that influenced the farmer's actions. These other causes were both unknown and unmeasured, thus preventing the researcher from statistically controlling for them, and this left open the possibility of other backdoor paths that would confound the causal inference. In Figure 2.18 the experimental design ensured that the only cause (i.e. previous yields) was already known and measured.

Using either randomised experiments or this controlled approach, the researcher could conclude[30] that his action of adding nitrate fertiliser does cause a change in Alfalfa yield and in the amount of nitrate in the soil.

Under what conditions could he infer that the *soil* nitrate levels (as opposed to nitrate fertiliser *addition*) causes the change in Alfalfa yield? That is, what would allow him to infer that the fertiliser addition increased soil nitrate concentration, which, in turn, increased Alfalfa yield? Although he was able to randomise and to exert experimental control over the amount of fertiliser added to the soil, this is not the same as randomly assigning values of soil nitrate to the plots and he has not exerted *direct* experimental control over soil nitrate levels. Because of this he cannot unambiguously claim that the experiment has demonstrated that soil nitrate levels cause an increase in plant yield. In other words, there might be a backdoor path from the fertiliser addition to each of soil nitrate and plant yield even though soil nitrate levels may have had no direct effect on plant yield. For instance, perhaps the fertiliser addition reduced the population level of some soil pathogen whose presence was reducing plant growth?

He can test the hypothesis that the association between soil nitrate levels and plant yield is due only to a backdoor path emanating from the amount of added fertiliser by measuring soil nitrate levels and then statistically controlling for this variable. d-separation predicts that, if this new causal hypothesis is true, then the effect of fertiliser addition will still exist. If the effect of fertiliser addition was due only to its effect on soil nitrate levels, then d-separation predicts that the effect of fertiliser addition on plant yield will disappear once the soil nitrate level is statistically controlled. Since he knows, from previous biological knowledge, that there is at least one backdoor path linking soil nitrate and plant yield (due to the effect of the nitrogen-fixing bacteria in the root nodules) then he can determine whether there is some other common cause generating a backdoor path if he can measure and then control for the amount of this bacterium.

2.17 A taste of things to come

Up to now, we have been inferring the properties of the observational model (the joint probability distribution) given the causal model that generates it. Can we also do the contrary? If we know the entire pattern of

[30] Given the typical assumptions of the statistical test used, and assuming that he is not in the presence of an unusual event.

statistical independencies and conditional independencies in our observational model, can we specify the causal structure that must have generated it? No. It is possible for different causal structures to generate the same set of d-separation statements and, therefore, the same pattern of independencies. None the less, it is possible to specify a *set* of causal models that all predict the same pattern of independencies that we find in the probability distribution; these are called *equivalent* models, and these are described in Chapter 8. By extension, we can exclude a vast group of causal models that could not have generated the observational data. There are two important consequences of this.

First, after proposing a causal model and finding that our observational data are consistent with it (i.e. that the data do not contradict any of the d-separation statements of our causal model), we can determine which other causal models would also be consistent with our data[31]. By definition, our data can't distinguish between such equivalent causal models and so we will have to devise other sorts of observation to differentiate between them.

Second, we can exploit the independencies in our observational data to generate such equivalent models even if we do not yet have a causal model that is consistent with our data. This leads to the topic of exploratory methods, which is also discussed in Chapter 8. Such exploratory methods are very useful when theory is not sufficiently well developed to allow us to propose a causal explanation – a condition that occurs often in organismal biology.

However, before delving into these topics, we must first look at the mechanics of fitting such observational models, generating their correlational 'shadows', and comparing the observed shadows (the patterns of correlation and partial correlation) with the predicted shadows. This leads into the topic of path models and, more generally, structural equations. Chapters 3 to 7 deal with these topics.

[31] This statement must be tempered due to practical problems involving statistical power.

3 Sewall Wright, path analysis and d-separation

3.1 A bit of history

> The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated with well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.
>
> The present paper is an attempt to present a method of measuring the direct influence along each separate path in such a system and thus of finding the degree to which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain the method can be used to find the logical consequences of any particular hypothesis in regard to them.

So begins Sewall Wright's 1921 paper in which he describes his 'method of path coefficients'. In fact, he invented this method while still in graduate school (Provine 1986) and had even used it, without presenting its formal description, in a paper published the previous year (Wright 1920). The 1920 paper used his new method to describe and measure the direct and indirect causal relationships that he had proposed to explain the patterns of inheritance of different colour patterns in Guinea Pigs. The paper came complete with a path diagram (i.e. a causal graph) in which actual drawings of the colour patterns of Guinea Pig coats were used instead of variable names.

Wright was one of the most influential evolutionary biologists of the twentieth century, being one of the founders of population genetics and intimately involved in the modern synthesis of evolutionary theory and genetics. Despite these other impressive accomplishments Wright viewed path analysis as one of his more important scientific contributions and continued to publish on the subject right up to his death (Wright 1984). The method was described by his biographer (Provine 1986) as 'the quantitative backbone of his work in evolutionary theory'. His method of path coefficients is the intellectual predecessor of all of the methods described in this book. It is therefore especially ironic that path analysis – the 'backbone' of his work in evolutionary theory – has been almost completely ignored by biologists.

This chapter has three goals. First, I want to explore why, despite such an illustrious family pedigree, path analysis and causal modelling have been largely ignored by biologists. To do this I have to delve into the history of biometry at the turn of the century but it is important to understand why path analysis was ignored in order to appreciate why its modern incarnation does not deserve such a fate. Next I want to introduce a new inferential test that allows one to test the causal claims of the path model rather than only 'measuring the direct influence along each separate path in such a system'. The inferential method described in this chapter is not the first such test. Another inferential test was developed quite independently by sociologists in the early 1970s, based on a statistical technique called maximum likelihood estimation. Since that method forms the basis of modern structural equation modelling, I postpone its explanation until the next chapter. Finally, I present some published biological examples of path analysis and apply the new inferential test to them.

3.2 Why Wright's method of path analysis was ignored

I suspect that scientists largely ignored Wright's work on path analysis for two reasons. First, it ran counter to the philosophical and methodological underpinnings of the two main contending schools of statistics at the turn of the twentieth century. Second, it was methodologically incomplete in comparison with Fisher's (1925) statistical methods, based on the analysis of variance combined with the randomised experiment, which had appeared at about the same time.

Francis Galton invented the method of correlation. Karl Pearson transformed correlation from a formula into a concept of great scientific importance and championed it as a replacement for the 'primitive' notion of causality. Despite Pearson's long-term programme to provide 'mathematical contributions to the theory of evolution' (Aldrich 1995), he had little training in biology, especially in its experimental form. He was educated as a mathematician and became interested in the philosophy of science early in his career (Norton 1975). Presumably his interest in heredity and genetics came from his interest in Galton's work on regression, which was itself applied to heredity and eugenics[1]. In 1892 Pearson published a book entitled *The grammar of science* (Pearson 1892). In his chapter entitled 'Cause and effect' he gave the following definition: 'Whenever a sequence of perceptions D, E, F, G is invariably preceded by the perception C . . ., C is said to be the *cause* of D, E, F, G.' As will become apparent later, his use of the word 'perceptions' rather than 'events' or 'variables' or 'observations' was an important part of his phenomenalist philosophy of science. He viewed the relatively new concept of correlation as having immense importance to science and the old notion of causality as so much metaphysical nonsense. In the third edition of his book (Pearson 1911) he even included a section entitled 'The category of association, as replacing causation'. In the third edition he had this to say:

> The newer and I think truer, view of the universe is that all existences are associated with a corresponding variation among the existences in a second class. Science has to measure the degree of stringency, or looseness of these concomitant variations. Absolute independence is the conceptual limit at one end to the looseness of the link, absolute dependence is the conceptual limit at the other end to the stringency of the link. The old view of cause and effect tried to subsume the universe under these two conceptual limits to experience – and it could only fail; things are not in our experience either independent or causative. All classes of phenomena are linked together, and the problem in each case is how close is the degree of association.

These words may seem curious to many readers because they express ideas that have mostly disappeared from modern biology. None the less, these ideas dominated the philosophy of science at the beginning of the twentieth century and were at least partially accepted by such eminent scientists as Albert Einstein. Pearson was a convinced phenomenalist and logical positivist[2]. This view of science was expressed by people such as Gustav Kirchhoff, who held that science can only discover new connections between phenomena, not discover the 'underlying reasons'. Ernst Mach, who dedicated one of his books to Pearson, viewed the only proper goal of

[1] Galton published his *Hereditary genius* in 1869 in which he studied the 'natural ability' of men (women were presumably not worth discussing). He was interested in 'those qualities of intellect and disposition, which urge and qualify a man to perform acts that lead to reputation . . .'. He concluded that '[those] men who achieve eminence, and those who are naturally capable, are, to a large extent, identical'. Lest we judge Galton and Pearson too harshly, remember that such views were considered almost self-evident at the time. Charles Darwin is reputed to have said of Galton's book: 'I do not think I ever in my life read anything more interesting and original . . . a memorable work' (Forrest 1974).

[2] It is more accurate to say that his ideas were a forerunner to logical positivism.

science as providing economical descriptions of experience by describing a large number of diverse experiences in the form of mathematical formulae (Mach 1883). To go beyond this and invoke unobserved entities such as 'atoms' or 'causes' or 'genes' was not science and such terms must be removed from its vocabulary. So, Mach (and Pearson) held that a mature science would express its conclusions as functional – i.e. mathematical – relationships that can summarise and predict direct experience, not as causal links that can explain phenomena (Passmore 1966).

Pearson had thought long and hard about the notion of causality and had concluded, in accord with British empiricist tradition and the people cited above, that association was all that there was. Causality was an outdated and useless concept. The proper goal of science was simply to measure direct experiences (phenomena) and to economically describe them in the form of mathematical functions. If a scientist could predict the likely values of variable *Y* after observing the values of variable *X*, then he would have done his job. The more simply and accurately he could do it, the better his science. If we go back to Chapter 2, Pearson did not view the equivalence operator of algebra ('=') as an imperfect *translation* of a causal relationship because he did not recognise 'causality' as anything but correlation in the limit[3]. By the time that Wright published his method of path analysis, Pearson's British school of biometry was dominant. One of its fundamental tenets was that 'it is this conception of correlation between two occurrences embracing all relationships from absolute independence to complete dependence, which is the wider category by which we have to replace the old idea of causation' (Pearson 1911).

Given these strong philosophical views, imagine what happened when Wright proposed using the biometrists' tools of correlation and regression . . . to peek beneath direct observation and deduce systems of causation from systems of correlation! In such an intellectual atmosphere Wright's paper on path analysis was seen as a direct challenge to the Biometrists. One has only to read the title ('Correlation and causation') and the introduction of Wright's (1921) paper, cited at the beginning of this chapter, to see how infuriating it must have seemed to the Pearson school.

The pagan had entered the temple and, like the Macabees, someone had to purify it. The reply came the very next year (Niles 1922). Said H. E. Niles: 'We therefore conclude that philosophically the basis of the method of path coefficients is faulty, while practically the results of applying it where it can be checked prove to be wholly unreliable'. Although he found fault

[3] And yet, citing the philosopher David Hume, Pearson did accept that associations could be time-ordered from past to future. Nowhere in his writings have I found him express unease that such asymmetries could not be expressed by the equivalence operator.

in some of Wright's formulae (which were, in fact, correct) the bulk of Niles' scathing criticism was openly philosophical: '"Causation" has been popularly used to express the condition of association, when applied to natural phenomena. There is no philosophical basis for giving it a wider meaning than partial or absolute association. In no case has it been proved that there is an inherent necessity in the laws of nature. Causation is correlation . . .' (Niles 1922).

Any Mendelian geneticist during that time – of whom Wright was one – would have accepted as self-evident that a mere correlation between parent and offspring told nothing about the mechanisms of inheritance. Therefore, concluded these biologists, a series of correlations between traits of an organism told nothing of how these traits interacted biologically or evolutionarily[4]. The Biometricians could never have disentangled the genetic rules determining colour inheritance in Guinea Pigs, which Wright was working on at the time, simply by using correlations or regressions. Even if distinguishing causation from correlation appeared philosophically 'faulty' to the Biometricians, Wright and the other Mendelian geneticists were experimentalists for whom statements such as 'causation is correlation' would have seemed equally absurd. For Wright, his method of path analysis was not a statistical *test* based on standard formulae such as correlation or regression. Rather, his path coefficients were interpretative parameters for measuring direct and indirect causal effects based on a causal system that had already been determined. His method was a statistical translation, a mathematical analogue, of a biological system obeying asymmetrical causal relationships.

As the fates would have it, path analysis soon found itself embroiled in a second heresy. Three years after Wright's 'Correlation and causation' paper, Fisher published his *Statistical methods for research workers* (1925). Fisher certainly viewed correlation as distinct from causation. For him the distinction was so profound that he developed an entire theory of experimental design to separate the two. He viewed randomisation and experimental control as the only reliable way of obtaining causal knowledge. Later in his life Fisher wrote another book criticising the research that identified tobacco smoking as a cause of cancer on the basis that such evidence was not based on randomised trials[5] (Fisher 1959). I have already described the

[4] Pearson was strongly opposed to Mendelism and, according to Norton (1975), this opposition was based on his philosophy of science; Mendelians insisted on using unobserved entities ('genes') and forces ('causation').

[5] I don't know whether Fisher was a smoker. If he was, I wonder what he would have thought if, because of a random number, he was assigned to the 'non-smoker' group in a clinical trial?

assumptions linking causality and probability distributions, unstated by Fisher but needed to infer causation from a randomised experiment, as well as the limitations of these assumptions, when one is studying different attributes of organisms. Despite these limitations, Fisher's methods had one important advantage over Wright's path analysis: they allowed one to rigorously test causal hypotheses while path analysis could only estimate the direct and indirect causal effects *assuming* that the causal relationships were correct.

Mulaik (1986) has described these two dominant schools of statistics in the twentieth century. His phenomenalist and empiricist school starts with Pearson. Examples of the statistical methods of this school were correlation, regression[6], common-factor and principal component analyses. The purpose of these methods was primarily, as Mach directed, to provide an economical description of experience by economically describing a large number of diverse experiences in the form of mathematical formulae. The second school was the Realist school begun by Fisher. It emphasised the analysis of variance, experimental design based on the randomised experiment and the hypothetico-deductive method. These Fisherian methods were not designed to provide functional relationships but rather to ensure conditions under which causal relationships could be reliably distinguished from non-causal relationships.

With hindsight then, it seems that path analysis simply appeared at the wrong time. It did not fit into either of the two dominant schools of statistics and it contained elements that were objectionable to each. The Phenomenalist school of Pearson disliked Wright's notion that one *should* distinguish 'causes' from correlations. The Realist school of Fisher disliked Wright's notion that one *could* study causes by looking at correlations. Professional statisticians therefore ignored it. Biologists found Fisher's methods, complete with inferential tests of significance, more useful and conceptually easier to grasp and so biologists ignored path analysis too. A statistical method, viewed as central to the work of one of the most influential evolutionary biologists of the twentieth century, was largely ignored by biologists.

[6] Regression based on least squares was, of course, developed well before Pearson by people like Carl Friedrich Gauss and had been based on a more explicit causal assumption that the independent variable plus independent measurement errors were the causes of the dependent variable. This distinction lives on under the guise of Type I and Type II regression.

3.3 d-sep tests

Wright's method of path analysis was so completely ignored by biologists that most biometry texts do not even mention it. Those that do (Li 1975; Sokal and Rohlf 1981) described it as Wright originally presented it, without even mentioning that it was reformulated by others, primarily economists and social scientists, such that it permitted inferential tests of the causal hypothesis and allowed one to include unmeasured (or 'latent') variables. The main weakness of Wright's method – that it required one to assume the causal structure rather that being able to test it – had been corrected by 1970 (Jöreskog 1970) but biologists are mostly unaware of this.

Two different ways of testing causal models will be presented in this book. The most common method is called structural equations modelling (SEM) and is based on maximum likelihood techniques. This method is described in Chapters 4 to 7 and it does have a number of advantages when testing models that include variables that cannot be directly observed and measured (so-called *latent* variables) and for which one must rely on observed indicator variables that contain measurement errors. SEM also has some statistical drawbacks. The inferential tests are asymptotic and can therefore require rather large sample sizes. The functional relationships must be linear. Data that are not multivariate normal are difficult to treat.

These drawbacks led me to develop an alternative set of methods that can be used for small sample sizes, non-normally distributed data or non-linear functional relationships (Shipley 2000). Since these methods are derived directly from the notion of d-separation that was described in Chapter 2, I will call these *d-sep* tests. The main disadvantage of d-sep tests is that they are not applicable to causal models that include latent (unmeasured) variables.

The link between causal conditional independence, given by d-separation, and probabilistic independence suggests an intuitive way of testing a causal model: simply list all of the d-separation statements that are implied by the causal model and then test each of these using an appropriate test of conditional independence. There are a number of problems with this naïve approach. First, even models with a small number of variables can include a large number of d-separation statements. Second, we need some way of combining all of these tests of independence into a single composite test. For instance, if we had a model that implied 100 independent d-separation statements and tested each independently at the traditional 5% significance level we would expect, on average, that five of these tests would reach significance simply as a result of random sampling fluctuations. Even worse, the d-separation statements in a causal model are almost never

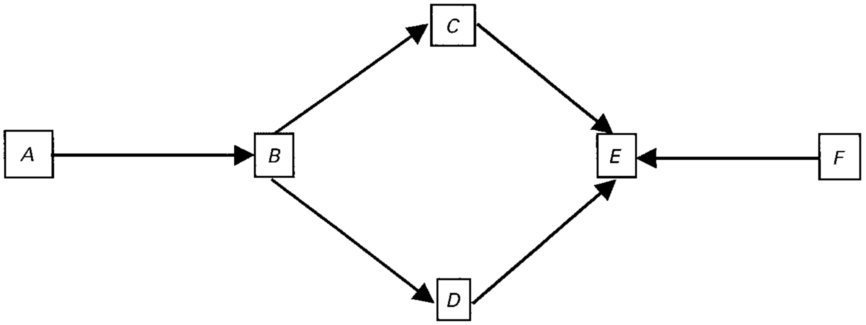


Figure 3.1. A directed acyclic graph (DAG) involving six variables.

completely independent and so we would not even know what the true overall significance level would be. Each of these problems can be solved.

3.4 Independence of d-separation statements

Given an acyclic[7] causal graph, we can use the d-separation criterion to predict a set of conditional probabilistic independencies that must be true if the causal model is true. However, many of these d-separation statements can be themselves predicted from other d-separation statements and are therefore not independent. Happily, Pearl (1988) described a simple method of obtaining the minimum number of d-separation statements needed to completely specify the causal graph and proved that this minimum list of d-separation statements is sufficient to predict the entire set of d-separation statements. This minimum set of d-separation statements is called a *basis set*[8]. The basis set is not unique. This method is illustrated in Figure 3.1.

To obtain the basis set, the first step is to list each unique pair of non-adjacent vertices. That is, list each pair of variables in the causal model that do not have an arrow between them. So, in Figure 3.1 the list is: $\{(A,C), (A,D), (A,E), (A,F), (B,E), (B,F), (C,D), (C,F), (D,F)\}$. Pearl's (1988) basis set is given by d-separation statements consisting of each such pair of vertices conditioned on the parents of the vertex having higher causal order. The number of pairs of variables that don't have an arrow between

[7] This restriction will be partly removed later. Remember that d-separation also implies probabilistic independence in cyclic causal models in which all variables are discrete and in cyclic causal models in which functional relationships are linear.

[8] Let $\boldsymbol{S}$ be the set of d-separation facts (and therefore the set of conditional independence relationships) that are implied by a directed acyclic graph. A basis set $\boldsymbol{B}$ for $\boldsymbol{S}$ is a set of d-separation facts that (i) implies, using the laws of probability, all other elements of $\boldsymbol{S}$, and (ii) no proper subset of $\boldsymbol{B}$ sustains such implications.

Table 3.1. *A basis set for the DAG shown in Figure 3.1 along with the implied d-separation statements*

Non-adjacent variables	Parent variables of either non-adjacent variable	d-separation statement
A, C	B	$A \perp\!\!\!\perp C \mid B$
A, D	B	$A \perp\!\!\!\perp D \mid B$
A, E	C, D	$A \perp\!\!\!\perp E \mid CD$
A, F	None	$A \perp\!\!\!\perp F$
B, E	A, C, D	$B \perp\!\!\!\perp E \mid ACD$
B, F	A	$B \perp\!\!\!\perp F \mid A$
C, D	B	$C \perp\!\!\!\perp D \mid B$
C, F	B	$C \perp\!\!\!\perp F \mid B$
D, F	B	$D \perp\!\!\!\perp F \mid B$

them is always equal to the total number of pairs minus the number of arrows in the causal graph. In general, if there are V variables and A arrows in the causal graph, then the number of elements in the basis set will be:

$$\frac{V!}{2(V-2)!} - A$$

Unfortunately the conditional independencies derived from such a basis set are not necessarily mutually independent in finite samples (Shipley 2000). A basis set that does have this property is given by the set of unique pairs of non-adjacent vertices, of which each pair is conditioned on the set of causal parents of both (Shipley 2000). Remember that an exogenous variable has no parents, so the set of 'parents' of such a variable is empty (such an empty set is written '$\{\phi\}$'or $\boldsymbol{\phi}$). The second step in getting the basis set that will be used in the inferential test is to list all causal parents of each vertex in the pair. Using Figure 3.1 and the notation for d-separation introduced in Chapter 2[9], Table 3.1 summarises the d-separation statements that make up the basis set.

Each of the d-separation statements in Table 3.1 predicts a (conditional) probabilistic independence. How you test each predicted conditional independence depends on the nature of the variables. For instance, if the two variables involved in the independence statement are normally and linearly distributed, you could test the hypothesis that the Pearson partial correlation coefficient is zero. Other tests of conditional independence are

[9] In other words, $X \perp\!\!\!\perp Y \mid \mathbf{Q}$ means that vertex X is d-separated from vertex Y, given the set of vertices $\mathbf{Q}$.

described below. At this point, assume that you have used tests of independence that are appropriate for the variables involved in each d-separation statement and that you have obtained the exact probability level assuming such independence. By 'exact' probability levels, I mean that you can't simply look at a statistical table and find that the probability is ≤ 0.05; rather, you must obtain the actual probability level – say, $p = 0.036$.

Because the conditional independence tests implied by the basis set are mutually independent, we can obtain a composite probability for the entire set using Fisher's test. Since this test seems not to have a name, I have called it Fisher's C (for 'combined') test. If there are a total of k independence tests in the basis set, and p_i is the exact probability of the ith test assuming independence, then the test statistic is:

$$C = -2\sum_{i=1}^{k} \ln(p_i)$$

If all k independence relationships are true, then this statistic will follow a chi-squared distribution with $2k$ degrees of freedom. This is not an asymptotic test unless you use asymptotic tests for some of the individual independence hypotheses. Furthermore, you can use different statistical tests for different individual independence hypotheses. In this sense, it is a very general test.

3.5 Testing for probabilistic independence

In this section, I want to be more explicit concerning what 'independence' and 'conditional independence' mean and the different ways that one can test such hypotheses given empirical data. Let's first start with the simplest case: that of unconditional independence.

The difference between the value of a random quantity X_i and its expected value μ is $(X_i - \mu)$. Since these differences can be either negative or positive, and we want to know simply the deviation around the expected value, not the direction of the deviation, we can take the square of the difference: $(X_i - \mu)^2$. The expected value of this squared difference[10] is the variance: $\mathrm{E}[(X_i - \mu_X)^2] = \mathrm{E}[(X_i - \mu_X)(X_i - \mu_X)]$.

The covariance is simply a generalisation of the variance. If we have two different random variables (X, Y) measured on the same observational units, then the covariance between these two variables is defined as: $\mathrm{E}[(X_i - \mu_X)(Y_i - \mu_Y)]$. If X and Y behave independently of each other, then large positive deviations of X from its mean (μ_X) will be just as likely to be

[10] The formula to estimate this in a sample is given in Box 3.1.

paired with large or small, negative or positive, deviations of *Y* from its mean (μ_Y). These will cancel each other out in the long run (remember, we are envisaging a complete statistical population) and the expected value of the product of these two deviations, $E[(X_i - \mu_X)(Y_i - \mu_Y)]$, will be zero. So, probabilistic independence of *X* and *Y* implies a population zero covariance[11]. If *X* and *Y* tend to behave similarly, increasing or decreasing together, then large positive values of *X* will often be paired with large positive values of *Y* and large negative values of *X* will often be paired with large negative values of *Y*. In such cases, the covariance will be large and positive. If *X* and *Y* tend to behave in opposite ways, then the covariance between them will be negative.

A Pearson correlation coefficient is simply a standardised covariance. Neither a variance nor a covariance have any upper or lower bounds. Changing the units of measurement (say, from metres to millimetres) will change both the variance and the covariance. If we divide the covariance between two variables by the product of their variances (taking the square root of this product in order to ensure that the range goes from +1 to −1), then we obtain a Pearson correlation coefficient. Box 3.1 summarises these points.

Box 3.1. Variance, covariance and correlation

Population variance (sigma^2, σ^2) of a random variable *X*: $E[(X - \mu_X)^2]$
Variance (s^2) of a random variable *X* from a sample of size *n*:

$$\frac{\sum_i (X_i - \overline{X})^2}{n-1}$$

Population covariance (sigma$_{XY}$, σ_{XY}) between two random variables *X*, *Y*:

$$E[(X - \mu_X)(Y - \mu_Y)].$$

Covariance (s_{XY}) between two random variables *X*, *Y* from a sample of size *n*:

$$\frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Population Pearson correlation (rho$_{XY}$, ρ_{XY}) between two random variables, *X*,*Y*:

$$\frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]E[Y - \mu_Y)^2]}} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

[11] But not the converse!

Pearson correlation coefficient (r_{XY}) between two random variables, X,Y from a sample of size n:

$$\frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}$$

The formulae in Box 3.1 are valid so long as both X and Y are random variables. If we want to conduct an inferential test of independence using these formulae, we have to pay attention to the probability distributions of X and Y and the form of the relationship between them in case they are not independent. Different assumptions concerning these points require different statistical methods.

Case 1: *X* and *Y* are both normally distributed and any relationship between them is linear

Tests of the independence of X and Y involving this set of assumptions are treated in any introductory statistics book. First, one can transform the Pearson correlation coefficient so that it follows Student's t-distribution. If X and Y, sampled randomly and measured on n units, are independent (so the null hypothesis is that $\rho = 0$) then the following transformation will follow a Student's t-distribution[12] with $n-2$ degrees of freedom:

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

This test is exact. So long as you have at least three independent observations then you can test for the independence of X and Y[13].

It is also possible to transform a Pearson correlation coefficient so that it asymptotically follows a standard normal distribution (i.e. a normal distribution with a mean of zero and a variance of 1). For sample sizes of at least 50 (and approximately even for sample sizes as low as 25) one can use Fisher's z-transform:

$$z = 0.5\sqrt{n-3}\,\ln\left(\frac{1+r}{1-r}\right)$$

[12] For partial correlations, described below, one simply replaces r with the value of the partial correlation coefficient, and the numerator $(n-2)$ becomes $(n-2-p)$ where p is the number of conditioning variables.

[13] Of course, with so few observations you would have so little statistical power that only very strong associations would be detected.

If X and Y are independent then the probability of z can be obtained from a standard normal distribution. Finally, one can use Hotelling's (1953) transformation[14], which is acceptable for sample sizes as low as 10:

$$z = \sqrt{(n-1)}\left[0.5\ln\left(\frac{1+r}{1-r}\right) - \frac{1.5\ln\left(\frac{1+r}{1-r}\right)+r}{4(n-1)}\right]$$

Case 2: X and Y are continuous but not normally distributed and any relationship between them is only monotonic

If X or Y are not normally distributed and any relationship between them is not linear but is monotonic[15], then we can use Spearman's correlation coefficient. Although there exist statistical tables giving probability levels for Spearman's correlation coefficient, one can use exactly the same formulae as for Pearson's correlation coefficient so long as the sample size is greater than 10 (Sokal and Rohlf 1981).

The first step is to convert X and Y to their ranks. In other words, sort each value of X from smallest to largest and replace the actual value of each X by its order in the rank; the smallest number becomes 1, the second smallest number becomes 2, and so on. Do the same thing for Y. Now that you have converted each X and each Y to its rank, you can simply put these numbers into the formula for a Pearson's correlation coefficient and test as before.

One complication is when there are ties. Spearman's coefficient assumes that the underlying values of X and Y are continuous, not discrete. Given such an assumption then equal values of X (or Y) will only occur due to limitations in measurement. To correct for such ties, first sort the values ignoring ties, and then replace the ranks of tied values by the mean rank of these tied values. Box 3.2 gives an example of the calculation of a Spearman rank correlation coefficient.

[14] Both Fisher's and Hotelling's transformations can be used to test null hypotheses in which ρ equals a value different from zero. This useful property allows one to compute confidence intervals around the Pearson correlation coefficient.

[15] A non-monotonic relationship is one in which X increases with increasing Y over part of the range and decreases with increasing Y over another part of the range. If you think that a graph of X and Y has hills and valleys, then the relationship is non-monotonic.

Box 3.2. Spearman's rank correlation coefficient

Here are 10 simulated pairs of values and the accompanying scatterplot (Figure 3.2). The X values were drawn from a uniform distribution and rounded to the nearest unit. The Y values were drawn from the following equation: $Y_i = X_i^{0.2} + \beta(5,1)$ where the random component is drawn from a β distribution with shape parameters of 5 and 1.

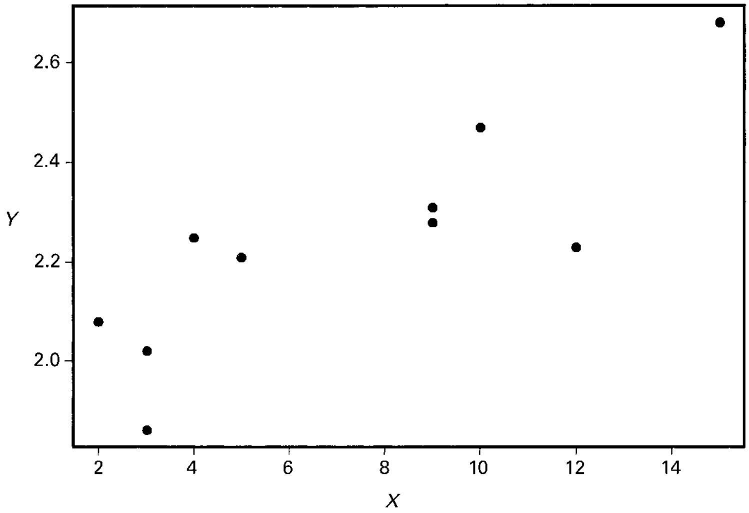


Figure 3.2. A scatterplot of randomly generated pairs of values from a bivariate non-normal distribution and possessing a non-linear monotonic relationship.

Values of X, Y and their ranks

X	Y	Rank	Rank	Rank	Rank
		X	Y	X	Y
2	2.08	1	3	1	3
3	2.02	2	2	2.5	2
15	2.68	10	10	10	10
10	2.47	8	6	8	6
5	2.21	5	4	5	4
12	2.23	9	5	9	5
3	1.86	3	1	2.5	1
4	2.25	4	7	4	7
9	2.31	6	9	6.5	9
9	2.28	7	8	6.5	8

In the above table, X, Y are the original values. Columns 3 and 4 of the table are the ranks of X and Y before correcting for ties (the underlined values). Columns 5 and 6 are the ranks after correcting for the two pairs of ties values of X (there were two values of 3 and two values of 9). To calculate the Spearman rank correlation coefficient of X and Y, simply use the values in columns 5 and 6 and enter them into the formula for the Pearson's correlation coefficient. In the above example, the Spearman rank correlation coefficient is 0.726. Assuming that X and Y are independent in the statistical population, we can convert this to a standard normal variate using Hotelling's z-transform, giving a value of 2.47. This value has a probability under the null hypothesis of 0.014.

Case 3: *X* and *Y* are continuous and any relationship between them is not even monotonic

This case applies when the relationship between X and Y might have a very complicated form, with X and Y being positively related in some parts of the range and negatively related in other parts, and therefore when neither a Pearson nor a Spearman correlation can be applied. This situation requires more computationally demanding methods, including form-free regression and permutation tests. Each of these topics is dealt with much more fully in other publications but will be introduced intuitively here because these notions are needed for the analogous case in conditional independence. Form-free regression is a vast topic, which includes kernel smoothers, cubic-spline smoothers (Wahba 1991) and local (loess) smoothers (Cleveland and Devlin 1988; Cleveland, Devlin and Grosse 1988; Cleveland, Grosse and Shyu 1992). Collectively, these methods form the basis of generalised additive models (Hastie and Tibshirani 1990). Permutation tests for association are described by Good (1993, 1994).

3.6 Permutation tests of independence

To begin, consider a simple linear regression of Y on X, where both are random variables. The correlation between X and Y is the same as the correlation between the observed value of Y and the predicted value of Y given X, that is: $E[Y|X]$. To test for an association between X and Y in this regression context we need to do three things. First, we have to estimate the predicted values of Y for each value of X. For linear regression we simply obtain the slope and intercept to get these values and in the general case we would use form-free regression methods. Second, we need to calculate a measure of the association between the observed and predicted values of Y; we can use a Pearson correlation coefficient, a Spearman correlation coefficient, or any of a large number of other measures that can be found in the statistical literature. Finally, we need to know the probability of having observed such a value when, in fact, *X* and *Y* really are independent. This is where a permutation test comes in handy.

Remembering the definition of probabilistic independence given in Chapter 2, we know that if *X* and *Y* are independent then the probability of observing any particular value of *Y* is the same whether or not we know the value of *X*. In other words, any value of *X* is just as likely to be paired with any other value of *Y* as with the particular *Y* that we happen to observe. The permutation test works by making this true in our data. After calculating our measure of association in our data, we randomly rearrange the values of *X* and/or *Y* using a random number generator. In this new randomly mixed 'data set' the values of *X* and *Y* really are independent because we forced them to be so; we have literally forced our null hypothesis of independence to be true and the value of the association between *X* and *Y* is due only to chance. We do this a very large number of times until we have generated an empirical frequency distribution of our measure of association[16]. The exact number of times that we randomly permute our data will depend on the true probability level of our actual data and the accuracy that we want to obtain in our probability estimate. Manly (1997) showed how to determine this number, but it is typically between 1000 and 10000 times. On modern computers this will take only a few seconds. The last step is to count the proportion of times that we observe at least as large a value of association within the permuted data sets, or its absolute value for a 2-tailed test, as we actually observed in our original data. Box 3.3 gives an example of this permutation procedure.

3.7 Form-free regression

Box 3.3. Loess regression and permutation tests

The following three graphs (Figure 3.3) show a simulated data set generated from a complicated non-linear function (solid line of the first graph) along with a loess regression (broken line) using a local quadratic fit and a neighbourhood size of one half the range of *X*. The middle graph shows the same complicated non-linear function in the range 1 to 3 of the *X* values and the graph to the right shows this in the range 1.5 to 2.5 of the *X* values.

[16] For small samples one can generate all unique permutations of the data. The use of random permutations, described here, is generally applicable and the estimated probabilities converge on the true probabilities as the number of random permutations increase.

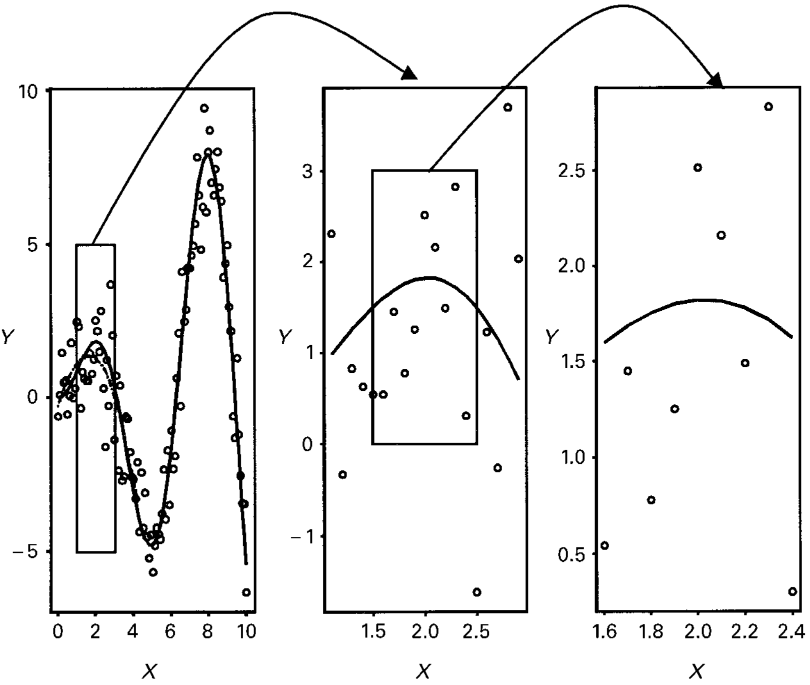


Figure 3.3. The graph on the left shows a highly non-linear function (the solid line) between *X* and *Y* and the loess fit (dotted line, mostly superimposed on the solid line). The small rectangle is reproduced in the middle graph and the small rectangle in this middle graph is reproduced in the graph on the right.

The loess regression (the dotted line in the left graph) doesn't actually give a parametric function linking *Y* to *X*, but does give the predicted value of *Y* for each unique value of *X*; i.e. it gives the estimate of E[$Y|X$]; the solid and broken lines in the left-most graph completely overlap except in the range of $X=2$. To estimate a permutation probability of the non-linear correlation of *X* and *Y*, we can first calculate the Pearson correlation coefficient between the observed *Y* values (the circles in the figure) and the predicted values of *Y* given *X* (the loess estimates). In this example, $r=0.956$. If we don't want to assume any particular probability distribution for the residuals, then we can generate a permutation frequency distribution for the correlation coefficient. To do this, we randomly permute the order of the observed *Y* values (or the predicted values, it doesn't matter which) to get a 'new' set of $Y^\star$ values and recalculate the Pearson correlation coefficient between $Y^\star$ and E[$Y^\star|X$]. The following histogram (Figure 3.4) shows the relative frequency of the Pearson correlation coefficient in 5000 such permutations; the arrow indicates the value of the observed Pearson correlation coefficient. None of the 5000 permutation data sets had a Pearson correlation whose absolute value was at least 0.956. Since the residuals were actually generated from a unit normal distribution, we can calculate the probability of observing a value of 0.956 with 101 observations. It is approximately 1×10^{-39}.

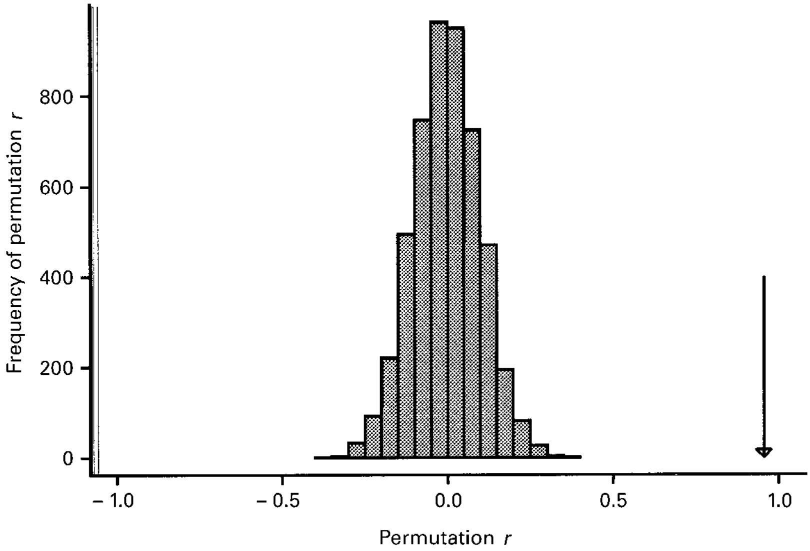


Figure 3.4. The frequency distribution of the Pearson correlation coefficient in 5000 random permutations of the simulated data set involving the observed *Y* values and the predicted loess values. The arrow shows the observed Pearson correlation in the original simulated data set.

The first graph in Box 3.3 (Figure 3.3) shows a highly non-linear relationship between *X* and *Y* and it is unlikely that we would be able to deduce the actual function that generated these data[17]. On the other hand, if we concentrate on smaller and smaller sections of the graph, the relationship becomes simpler and simpler. The basic insight of form-free regression methods is that even complicated functions can be quite well approximated by simple linear, quadratic or cubic functions in the neighbourhood of a given value of *X*. Within such a neighbourhood, shown by the boxes in the graphs of Box 3.3, we can use these simpler functions to calculate the expected value of *Y* at that particular value of *X*. We then go on to the next value of *X*, move the neighbourhood so that it is centred around this new value of *X*, and calculate the expected value of the new *Y*, and so on. In

[17] The actual function was: $Y = X \sin(X) + \varepsilon$, where the error term comes from a unit normal distribution.

this way, we do not actually estimate a parametric function predicting *Y* over the entire range of *X* but we do get very good estimates of the predicted values of *Y* given each unique value of *X*. To obtain the predicted values of *Y* given *X*, we use weighted regression (linear, quadratic or cubic) where each (X, Y) pair in the data set is weighted according to its distance from the value of *X* around which the neighbourhood is centred. In local, or loess[18], regression the neighbourhood size can be chosen according to different criteria such as minimising the residual sum of squares and the weights are chosen based on the tricube weight function. Shipley and Hunt (1996) described this in more detail in the context of plant growth rates[19].

3.8 Conditional independence

So far we have been talking about unconditional independence; that is, the independence of two variables without regard to the behaviour of any other variables. Such unconditional independence is implied by two variables in a causal graph that are d-separated without conditioning on any other variable. d-separation upon conditioning implies *conditional* independence. The notion of conditional independence seems paradoxical to many people. How can two variables be dependent, even highly correlated, and still be independent upon conditioning on some other set of variables?

Consider the following causal graph: $\varepsilon 1 \rightarrow X \leftarrow Z \rightarrow Y \leftarrow \varepsilon 2$. Does it seem equally paradoxical if I say that *X* and *Y* will behave similarly owing to the common causal effect of *Z*, but that they will no longer behave similarly if I prevent *Z* from changing? If *Z* doesn't change, then the only changes in *X* and *Y* will come from the changes in $\varepsilon 1$ and $\varepsilon 2$, and these two variables are d-separated and therefore unconditionally independent. A moment's reflection will convince you that if *Z* is allowed to change (vary) then both *X* and *Y* will change as well in a systematic fashion, since they are both responding to *Z*. If the variables in the causal graph are random then the correlation between *X* and *Y* will be due to the fact that both share common variance due to *Z*. If we restrict the variance in *Z* more and more, then *X* and *Y* will share a smaller and smaller amount of common variance. In the limit, if we prevent *Z* from changing at all, then *X* and *Y* will no longer share any common variance; the only variation in *X* and *Y* will come from the independent error variables $\varepsilon 1$ and $\varepsilon 2$ and so *X* and *Y* will then

[18] The word 'loess' comes from the geological term 'loess' which is a deposit of fine clay or silt along a river valley. I suppose that this evokes the image of a very wavy surface that traces the form of the underlying geological formation. At least some statisticians have a sense of the poetic.

[19] The S-PLUS program performs multivariate form-free regression (StatSci 1995).

be independent. In such a case we would be comparing values of X and Y when Z is constant. This is the intuitive meaning of conditional independence. To illustrate, I generated 10000 independent sets of $\varepsilon1$, X, Z, Y and $\varepsilon2$ according to the following generating equations:

$$\varepsilon1 = N(0,1-0.9^2)$$

$$\varepsilon2 = N(0,1-0.9^2)$$

$$Z = N(0,1)$$

$$Y = 0.9Z + \varepsilon1$$

$$X = 0.9Z + \varepsilon2$$

Since X, Y and Z are all unit normal variables, the population correlations are $\rho_{X,Z}=0.9$, $\rho_{Y,Z}=0.9$ and $\rho_{X,Y}=0.81$. Notice that X and Y are highly correlated even though neither X nor Y is a cause of the other. Figure 3.5 shows three scatterplots. The plot on the left shows the relationship between X and Y when no restrictions are placed on the variance of Z. The sample correlation between X and Y in this graph is 0.8016, compared with the population value of 0.81. The graph in the middle plots only those values of X and Y for which the value of Z is between -2 and $+2$, thus restricting the variance of Z a little bit. The sample correlation between X and Y has been decreased slightly to 0.7591. The graph on the right plots those values of X and Y for which the value of Z is between -0.5 and $+0.5$, thus restricting the variance of Z much more. The sample correlation between X and Y is now only 0.2294. Clearly, the degree of association between X and Y is decreasing as Z is prevented more and more from varying.

If we calculate the correlation between X and Y as we restrict the variation in Z more and more, we can get an idea of what happens to the correlation between X and Y in the limit when the variance of Z is zero. This limit is the correlation between X and Y when Z is fixed (or 'conditioned') to a constant value; this is called the *partial* correlation between X and Y, conditional on Z and it is written '$\rho_{XY.Z}$' or '$\rho_{XY|Z}$'. Figure 3.6 plots the sample correlation between X and Y as Z is progressively restricted in its variance.

As expected, as the range of Z around its mean (zero) becomes smaller and smaller, the correlation between X and Y also becomes smaller and approaches zero. Given the causal graph that governed these data, we know that X and Y are not unconditionally d-separated and therefore are not unconditionally independent. However, X and Y are d-separated given Z and therefore X and Y are independent conditional on Z.

If we remember that a regression of X on Z gives the expected value

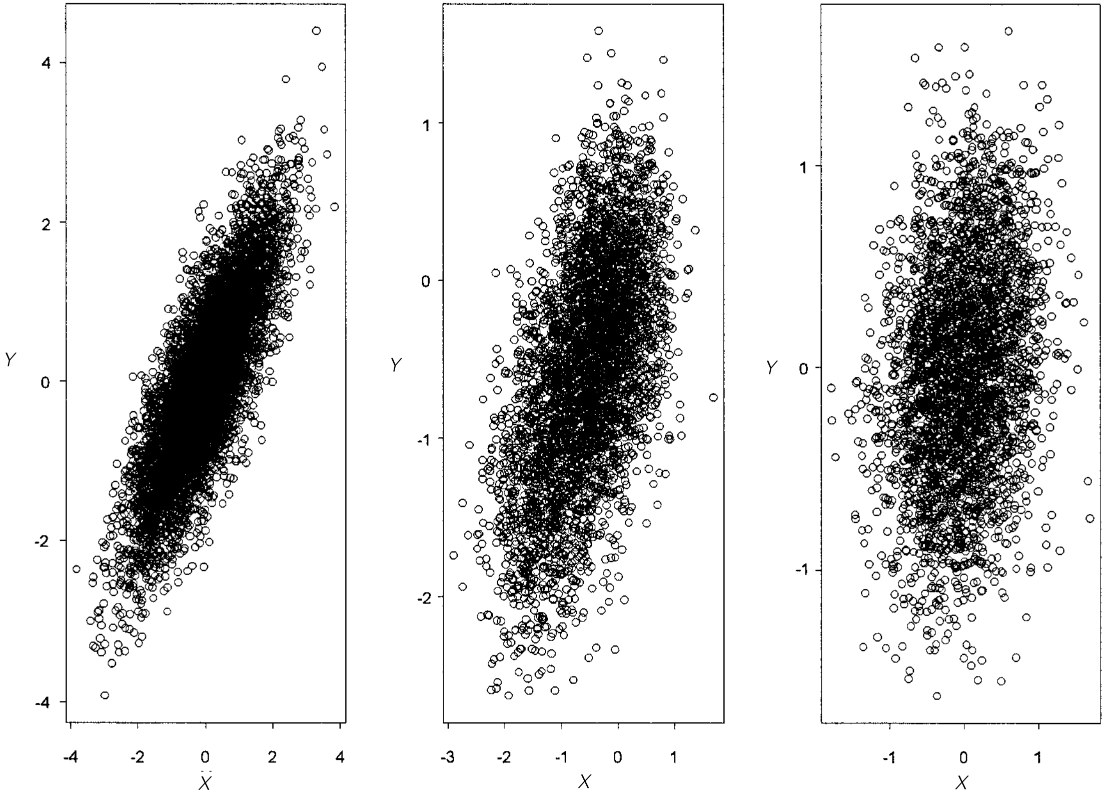


Figure 3.5. The graph on the left shows10 000 observations of X and Y that were generated from the causal graph: $\varepsilon_1 \to X \leftarrow Z \to Y \leftarrow \varepsilon_2$ and parameterised as given in the text. The middle graph shows only those (X, Y) observations for which $|Z|$ is less than 2. The graph on the right shows only those (X, Y) observations for which $|Z|$ is less than 0.5.

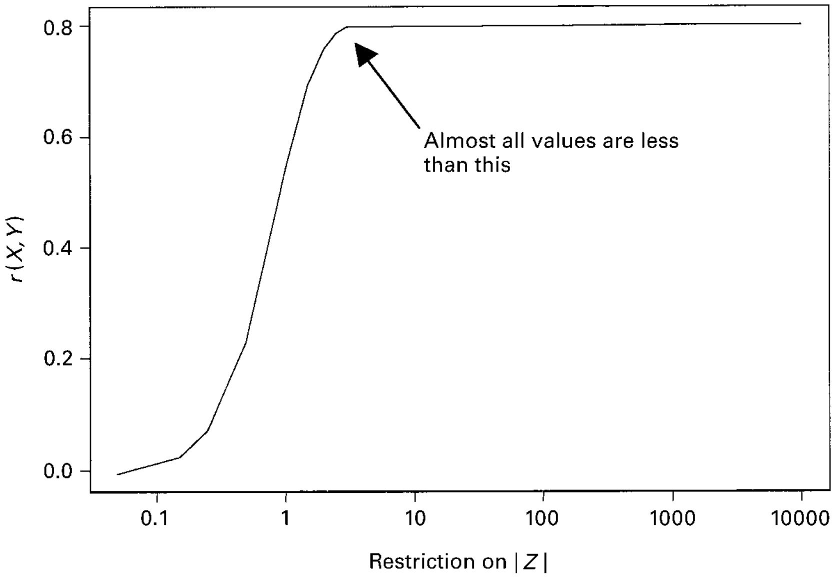


Figure 3.6. The Pearson correlation coefficient between *X* and *Y* in the data shown in Figure 3.5 (left) when the absolute value of *Z* is restricted to various degrees. The limiting value of the correlation coefficient when |*Z*| is restricted to a constant value is the partial correlation between *X* and *Y*.

of *X* conditional on *Z*, then the residuals around this regression are the values of *X* for fixed values of *Z*. This gives us another way of visualising the partial correlation of *X* and *Y* conditional on *Z*: it is the correlation between the residuals of *X*, conditional on *Z*, and the residuals of *Y*, conditional on *Z*. If I regress, in turn, each of *X* and *Y* on *Z* in the above example and calculate the correlation coefficient between the residuals of these two regressions, I get a value of -0.0060.

This view of a conditional independence provides us with a very general method of testing for it. If *X* and *Y* are predicted to be d-separated given some other set of variables $\mathbf{Q}=\{A, B, C, \ldots\}$ then regress (perhaps using form-free regression) each of *X* and *Y* on the set **Q** and then test for independence of the residuals using, if you want, any of the methods of testing unconditional independence described above. If the residuals are normally distributed and linearly related then you can use the test for Pearson correlations. If the residuals appear, at most, to have a monotonic relationship then you can use the test for a Spearman correlation. If the residuals have a more complicated pattern then you can use one of the non-parametric smoothing techniques available, followed by a permutation test. The only difference is that you have to reduce the degrees of freedom in the tests by the number of variables in the conditioning set.

Most of these tests can be performed using standard statistical programs[20]. If your statistical program can invert a matrix, then there are faster ways of calculating partial Pearson or Spearman correlations. These are explained in Box 3.4.

Box 3.4. Calculating partial covariances and partial correlations

Given a sample covariance matrix $\boldsymbol{S}$, the inverse of this matrix is called the *concentration* matrix, $\boldsymbol{C}$. The negative of the off-diagonal elements c_{ij} give the partial covariance between variables i and j, conditional on (holding constant) all of the other variables included in the matrix. This gives an easy way of estimating partial covariances and partial correlations of any order. To get the partial covariance between variables X and Y conditional on a set of other variables $\mathbf{Q}$, simply create a covariance matrix in which the only variables are X, Y, and the remaining variables in $\mathbf{Q}$. After inverting this matrix, this partial covariance is the negative of the element in the row pertaining to X and the column pertaining to Y, i.e. $-c_{XY}$. The partial correlation between X and Y is given by:

$$r_{X,Y|\mathbf{Q}} = \frac{-c_{XY}}{\sqrt{c_{XX} \times c_{YY}}}$$

The partial correlation between two variables conditioned on n other variables is said to be a *partial correlation of order n*. The unconditional correlation coefficient is simply a partial correlation of order zero. Some texts give recursion formulae for partial correlations of various orders, although partials of higher orders are very tedious to calculate by such means. For instance, the formula for a partial correlation of order 1 between X and Y, conditional on Z, is:

$$\rho_{X,Y|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{(1-\rho_{XZ}^2)(1-\rho_{YZ}^2)}}$$

As an example, consider the following causal graph: $W \rightarrow X \rightarrow Z \rightarrow Y$. 100 independent (W,X,Y,Z) observations were generated according to structural equations with all path coefficients equal to 0.5 and the variances of all four variables equal to 1.0. Here is the sample covariance matrix:

[20] My Toolbox (Appendix) contains a program to calculate partial correlations of various orders.

	W	*X*	*Y*	*Z*
W	1.43347870	−0.75265627	−0.06269845	0.10179918
X	−0.75265627	1.52762094	−0.53911722	−0.03777874
Y	−0.06269845	−0.53911722	1.71116716	−0.90033856
Z	0.10179918	−0.03777874	−0.90033856	1.73196991

The inverse of the matrix (rounded to the nearest 100th) obtained by extracting only the elements of the covariance matrix pertaining to *W*, *X* and *Y* is:

	W	*X*	*Y*
W	1.43	−0.75	−0.01
X	−0.75	1.53	−0.56
Y	−0.01	−0.56	1.24

The partial correlation between *W* and *Y*, conditional on *X*, is:

$$r_{WY|X} = \frac{-0.01(-1)}{\sqrt{1.43 \times 1.24}} = 0.0075$$

The same method can be used to obtain partial Spearman partial correlations, by simply ranking the variables as described in Box 3.2 and then proceeding in the same way as for Pearson partial correlations.

3.9 Spearman partial correlations

This next section presents some Monte Carlo results to explore the degree to which the sampling distribution of Spearman partial correlations, after appropriate transformation, follows either a standard normal or a Student's t-distribution. This section is not necessary to understand the application of d-sep tests for path models, only to justify the use of Spearman partial correlations in testing for conditional independence.

There has been remarkably little published in the primary literature concerning inferential tests related to non-parametric conditional independence[21]. It is known that the expected values of first-order partial Kendall or Spearman partial correlations need not be strictly zero even when two variables are conditionally independent given the third (Shirahata 1980; Korn 1984). On the other hand, Conover and Iman (1981) recommended

[21] Kendall and Gibbons (1990) briefly discuss Spearman and Kendall partial correlations and provide a table of significance values for first-order Kendall partial correlations for small sample sizes.

the use of partial Spearman correlations for most practical cases in which the relationships between the variables are at least monotonic. A Spearman partial correlation is simply a Pearson partial correlation applied to the ranks of the variables in question. Therefore the conditional independence of non-normally distributed variables with non-linear, but monotonic, functional relationships between the variables can be tested with Spearman's partial rank correlation coefficient simply by ranking each variable (and correcting for ties as described in Box 3.2) and then applying the same inferential tests as for Pearson partial correlations. For instance, if one accepts Conover and Iman's (1981) recommendations, then a Spearman partial rank correlation will be approximately distributed as a standard normal variate when *z*-transformed.

How robust is this recommendation? To explore this question, Table 3.2 presents the results of some Monte Carlo simulations to determine the effects of sample size, the distributional form of the variables, and the effect of non-linearity on the sampling distribution of the *z*-transformed Spearman partial correlation coefficient. The random components of the generating equations (ε_i) were drawn from four different probability distributions: normal, gamma, beta or binomial. I chose the shape parameters of the gamma and beta distributions to produce different degrees of skew and kurtosis. Gamma($\lambda=1$) is a negative exponential distribution. Gamma($\lambda=5$)[22] is an asymmetrical distribution with a long right tail. Beta(1,1) is a uniform distribution, beta(1,5) is a highly asymmetrical distribution with a long right tail and beta(5,1) is a highly asymmetrical distribution with a long left tail. The final (discrete) probability distribution was symmetrical with an expected value of 2 and had ordered states of $X=0$, 1, 2, 3 or 4; these were generated from a binomial distribution of the form $C(5,X)0.5^X0.5^{1-X}$. Random numbers were generated using the random number generators given by Press *et al.* (1986). The generating equations were of the form:

$$X_1=\varepsilon_1$$

$$X_i=\alpha_i X_{(i-1)}^{\beta i}+\varepsilon_i;\ \ i>1$$

These generating equations are based on a causal chain ($X_1\rightarrow X_2\rightarrow X_3\rightarrow$ . . .) with sufficient variables (3, 4 or 5) to produce zero partial associations of orders 1 to 3. When β_i equals 1.0 the relationships between the variables are linear and when β_i is different from 1.0 then the relationships between the variables are non-linear but monotonic. The

[22] λ is a constant affecting the shape of the distribution and is sometimes referred to as a waiting time for the event in a Poisson random process of unit mean.

results in Table 3.2 are based on models with $\beta_i = 1$ (linear) and 0.5 (non-linear) but other values give similar results. All the simulation results in Table 3.2 are based on 1000 independent simulated data sets. In interpreting Table 3.2, remember that the *z*-transformed Spearman partial correlations should be approximately distributed as a standard normal variate whose population mean is zero, whose population standard deviation is 1.0, and whose 2-tailed 95% limit is $|1.96|$.

Generally, the sampling distribution of the *z*-transformed Spearman rank partial correlations is a very good approximation of a standard normal distribution. In fact, the only significant deviation from a standard normal distribution (based on a Kolmogorov–Smirnov test) was observed for the ranks of normally distributed variables, for which one would not normally use a Spearman partial correlation. The empirical standard deviations were always close to 1.0 and the empirical means only once differed significantly, but very slightly, from zero at high levels of replication. Approximate 95% confidence intervals for the empirical 0.05 significance level (i.e. the 2-tailed 95% quantiles), based on 1000 simulations, are 0.037 to 0.064 (Manly 1997).

The results of this simulation study support the recommendations of Conover and Iman (1981). These results are also consistent with the theoretical values given by Korn (1984) for the special case of a Spearman first-order partial based on trivariate normal and trivariate log-normal distributions, where the limiting values of the Spearman partial correlation are less than, or equal to, an absolute value of 0.012, thus giving an expected absolute z-score of ≤ 0.024. Korn (1984) gave a pathological example in which the above procedure will not work even after ranking the data because there is a non-monotonic relationship between the variables; he recommended that one first check[23] to see whether the relationships between the ranks are approximately linear before using Spearman partial correlations.

3.10 Seed production in St Lucie's Cherry

St Lucie's Cherry (*Prunus mahaleb*) is a small species of tree that is found in the Mediterranean region and relies on birds for the dispersal of its seeds. As in most plants, seedlings from seeds that are dispersed some distance from the adult are more likely to survive, since they will not be shaded by their own parent or eaten by granivores that are attracted to the parent tree. For species whose seeds can survive the passage through the digestive tract of the dispersing animal, it is also evolutionarily and ecologically advantageous

[23] This can be done by simply plotting the scatterplots of the ranked data.

Table 3.2. *Results of a Monte Carlo study of the distribution of z-transformed Spearman partial correlations. Four different distributional types were simulated for the random components. Sample size was the number of observations per simulated data set. Linear (L) and non-linear (NL) functional relationships were used. The empirical mean, standard deviation and the 2-tailed 95% limits of 1000 simulated data sets are shown*

Distribution of ε_i	Sample size	Order of partial	Linear/non-linear	Mean of z	Standard deviation of z	2-tailed 95% quantile	Theoretical probability
Normal	25	1	L	0.08	1.03	2.04	0.04
Normal	50	1	L	0.08	0.97	2.01	0.05
Normal	400	1	L	0.08	1.04	2.16	0.03
Normal	50	2	L	0.03	0.99	1.86	0.06
Normal	50	3	L	−0.07	1.00	1.85	0.06
Gamma(1)	25	3	L	0.01	1.05	2.09	0.04
Gamma(1)	50	3	L	−0.02	0.96	1.82	0.07
Gamma(1)	50	3	NL	0.03	0.96	2.02	0.04
Gamma(5)	50	3	NL	−0.07	0.99	1.93	0.05
Beta(1,1)	50	3	L	−0.02	0.99	2.00	0.05
Beta(1,1)	50	3	NL	0.03	1.02	2.08	0.04
Beta(1,5)	50	3	NL	0.03	1.02	2.08	0.04
Beta(5,1)	50	3	NL	−0.05	0.99	1.78	0.07
Beta(5,1)	400	3	NL	0.00	1.02	2.01	0.04
Binomial	50	3	NL	0.01	0.99	1.95	0.05

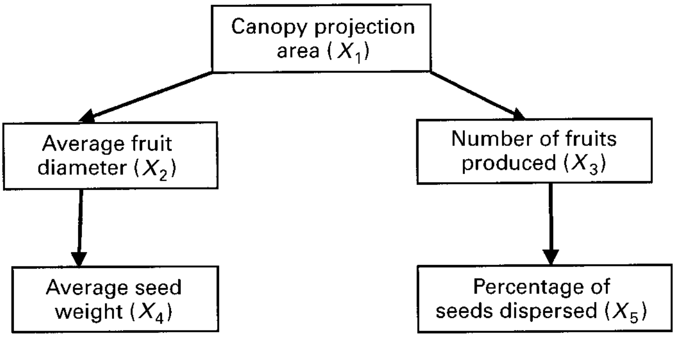

Canopy projection area (X_1)
Average fruit diameter (X_2)
Number of fruits produced (X_3)
Average seed weight (X_4)
Percentage of seeds dispersed (X_5)

Figure 3.7. Proposed causal relationships between five variables related to seed dispersal in St Lucie's Cherry.

for the fruit to be eaten by the animal, since the seed will be deposited with its own supply of fertiliser. Not all frugivores of St Lucie's Cherry are useful fruit dispersers. Some birds just consume the pulp while either leaving the naked seed attached to the tree or simply dropping the seed to the ground directly beneath the parent. In order to estimate selection gradients, Jordano (1995) measured six traits of 60 individuals of this species: the canopy projection area (a measure of photosynthetic biomass), average fruit diameter, the number of fruits produced, average seed weight, the number of fruits consumed by birds and the percentage of these consumed fruits that were properly dispersed away from the parent by passage through the gut. Based on five of these variables for which I had data (I was lacking the total number of fruits consumed by birds) I proposed the path model shown in Figure 3.7 (Shipley 1997), using the exploratory path models described in Chapter 8.

We can use this model to illustrate the d-sep test. The first step is to obtain the d-separation statements in the basis set that are implied by the causal graph in Figure 3.7. There are six such statements, since there are five variables and four arrows. Table 3.3 lists these d-separation statements.

We next have to decide how to test the independencies that are implied by these six d-separation statements. The original data showed heterogeneity of variance, as often happens with size-related variables, but transforming each variable to its natural logarithm stabilises the variance. Figure 3.8 shows the scatterplot matrix of these ln-transformed data.

Since the relationships appear to be linear and histograms of each variable did not show any obvious deviations from normality, we can test the predicted independencies using Pearson partial correlations. The results are shown in Table 3.3. Fisher's C statistic is 7.73, with 12 degrees of freedom (df), for an overall probability of 0.806. The difference between the observed and predicted (partial) correlations would occur in about an 80%

Table 3.3. *Shown are the d-separation statements in the basis set of the causal graph shown in Figure 3.7, along with the Pearson and Spearman partial correlations that are implied by the d-separation statements. The probabilities, assuming that the population partial correlations are zero, are listed as well*

	Pearson partial correlations		Spearman partial correlations	
d-separation statement	Estimate	Probability assuming independence	Estimate	Probability assuming independence
$X_4 \perp\!\!\!\perp X_1 \mid X_2$	−0.066	0.617	−0.063	0.635
$X_4 \perp\!\!\!\perp X_3 \mid X_2 X_1$	0.142	0.289	0.144	0.279
$X_4 \perp\!\!\!\perp X_5 \mid X_2 X_3$	0.004	0.976	0.075	0.574
$X_2 \perp\!\!\!\perp X_3 \mid X_1$	0.021	0.873	0.059	0.655
$X_2 \perp\!\!\!\perp X_5 \mid X_3 X_1$	−0.155	0.244	−0.160	0.229
$X_1 \perp\!\!\!\perp X_5 \mid X_3$	0.076	0.565	0.102	0.443

of data sets (in the long run) even if the data really were produced by the causal structure in Figure 3.7. This doesn't mean that the data were produced by such a causal structure but it does mean that we have no reason to reject it on the basis of the statistical test. If we want to reject it anyway, then we will need to produce reasonable doubt. Perhaps the assumption of normality, upon which the test of the Pearson partial correlations is based, was producing incorrect probability estimates. Table 3.3 also lists the Spearman partial correlations. The overall probability of the model ($\chi^2 = 9.99$, 12 df), based on the individual probability levels of these Spearman partial correlations, was 0.616. On the other hand, there are equivalent models that also produce non-significant probability estimates (Shipley 1997) and if any of these equally well-fitting other models do not contradict what is known of the biology of these trees, then they might constitute reasonable doubt[24].

The original data of Jordano (1995) was fit to a latent variable model using maximum likelihood methods[25]. Neither the model chi-squared statistic nor the model degrees of freedom were given. It is therefore not possible to judge the fit of that original model[26], but it is possible to extract those

[24] The model was actually developed using the exploratory methods of Chapter 8. This, too, should give us reason to question the model until independent data can be tested against it. At this point, all we can reasonably say is that the data are consistent with the model and so deserve further study.

[25] These methods are described in Chapters 4 to 7.

[26] One measured variable (total number of seeds dispersed) was not provided to me, so I can't fit his original model.

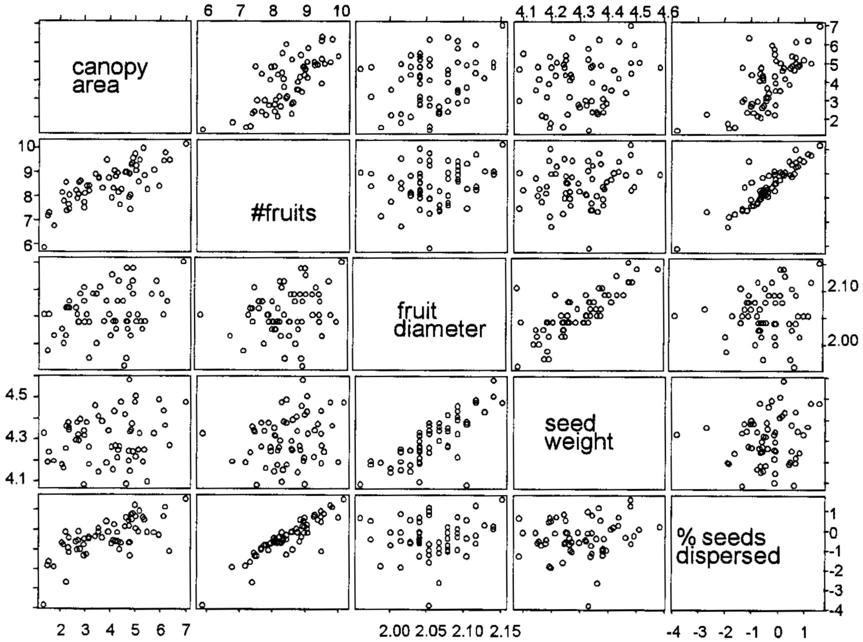


Figure 3.8. Scatterplot matrix of the empirical observations (all variables transformed to their natural logarithms).

d-separation statements involving only the measured variables available to myself from the original latent-variable model. Jordano's published model implies four d-separation statements in the basis set that can be tested: {(canopy projection area $\perp\!\!\!\perp$ average fruit diameter), (canopy projection area $\perp\!\!\!\perp$ average seed weight), (number of fruits produced $\perp\!\!\!\perp$ average fruit diameter), (number of fruits produced $\perp\!\!\!\perp$ average seed weight)}. The d-sep test, based on Pearson correlations, gives a probability of 0.005 ($\chi^2=21.85$, 8 df). Using Spearman correlations the probability is 0.019 ($\chi^2=18.24$). These low probabilities, based on a subset of the original measured variables, provide reasonable doubt concerning Jordano's (1995) model.

3.11 Specific leaf area and leaf gas exchange

The leaves of most flowering plants are photosynthetic organs. Since carbon fixation is so central to the survival of plants, one might expect that there is a tight integration of leaf form and physiology to provide for this necessary function. However, land plants face a dilemma. They need to keep their tissues turgid but these humid tissues find themselves surrounded by air or soil that is not saturated with water. The leaves (and other tissues) are protected by a cuticle to prevent dehydration. Unfortunately, this severely restricts not only the diffusion of water vapour, but also other gases, especially CO_2 that is required for photosynthesis, from diffusing into the leaves. The production of stomates is the evolutionary solution to this problem. Stomates are small openings on the surface of the leaves through which gases can diffuse and the size of the stomatal openings is controlled by guard cells.

As soon as the stomates begin to open, CO_2 begins to diffuse from the outside air into the intercellular spaces of the leaf through a process of passive diffusion. Since the leaf is photosynthesising, CO_2 is being removed from the intercellular spaces, creating a diffusion gradient. However, the air inside the leaf is always saturated with water vapour. As soon as the stomates begin to open, this water vapour also begins to diffuse out of the leaf, since the outside air is not saturated with water. In essence, the leaf has to accept a loss (water) in order to effect a gain (carbon). Cowan and Farquhar (1977) proposed a theoretical model of stomatal regulation to predict how the leaf should control its stomates in order to maximise carbon gain relative to water loss. The basic insight of this model was that the leaf should restrict carbon fixation below the maximum level because when the internal CO_2 level in the leaf reaches a certain level the main carboxylating enzyme (ribulose bisphosphate carboxylase (Rubisco)) becomes saturated, and further increases in carbon fixation require the regeneration of ATP from the light reaction of photosynthesis. The second stage results in a greatly reduced rate of increase of carbon fixation per increase in the internal CO_2 concentration, but the rate of water loss continues at its former rate. Thus Cowan and Farquhar's principal insight was that the leaf should maintain the intercellular CO_2 concentration at the break-point between Rubisco limitation and ribulose bisphosphate regeneration limitation so that the carboxylating capacity and the capacity to regenerate Rubisco are co-limiting.

On the basis of these theoretical notions, Martin Lechowicz and I (Shipley and Lechowicz 2000) proposed a path model based on five variables: specific leaf mass (SLM: leaf dry mass divided by leaf area, g/m^2), leaf organic nitrogen concentration ($mmol/m^2$), stomatal conductance to water ($mmol/m^2$ per s), net photosynthetic rate ($\mu mol/m^2$ per s) and internal CO_2 concentration ($\mu l/l$). The proposed model is shown in Figure 3.9. Our data were the mean values from 40 herbaceous species typical of wetland environments.

There are five outliers in the data in relation to the internal CO_2 concentration. These are 'C_4' species. The other 35 species are C_3 species. C_4 species have an additional metabolic pathway in which atmospheric carbon is first fixed by phosphoenol pyruvate carboxylase in the mesophyll cells to form malate or aspartate. This molecule, a 4-carbon acid, is then

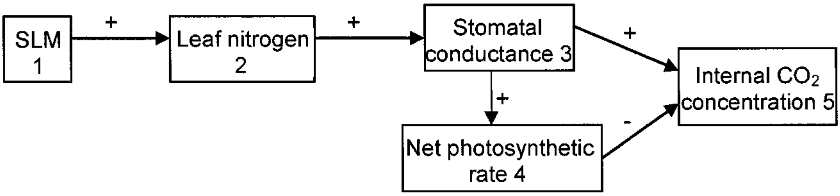


Figure 3.9. Proposed causal relationships between five variables related to interspecific leaf morphology and gas exchange. SLM, specific leaf mass.

transferred into bundle-sheath cells deeper in the leaf. Here these C_4 acids are decarboxylated and the freed CO_2 enters the normal Calvin cycle of the dark reaction of photosynthesis. An advantage of C_4 photosynthesis is that plants exhibiting it are able to absorb CO_2 strongly from a lower concentration of CO_2 within the leaf. They can do this without Rubisco acting as an oxygenase, rather than a carboxylase, under conditions of low CO_2 and high O_2. This means that C_4 plants do not exhibit the wasteful process of photorespiration under conditions of high illumination and low availability of water. Because of this, they are able to maintain high rates of photosynthesis even when the stomates are nearly closed. The basis set implied by the model in Figure 3.9, along with the relevant statistics, is summarised in Table 3.4.

There is no strong evidence for any deviation of the data from the predicted correlational shadow, as given by the d-separation statements. However, a reasonable alternative model would be that the leaf nitrogen content, which is due primarily to enzymes related to photosynthesis, directly causes the net photosynthetic rate. In other words, what if Cowan and Farquhar's (1977) model of stomatal regulation is wrong, and the leaf is regulating its stomates to maximise the net rate of CO_2 fixation independently of water loss? In this case, the observed rate of stomatal conductance would be a consequence of the net photosynthetic rate rather than its cause and the net photosynthetic rate would be directly caused by leaf nitrogen content. We can test this alternative model too and Table 3.5 summarises the results.

This alternative model is clearly rejected when both the C_3 and C_4 species are analysed together, since there are only about 2 out of 10000 chances of observing such a large difference at random. This lack of fit is coming from the predicted independence between leaf nitrogen level (2) and stomatal conductance (3), conditioned jointly on specific leaf mass (1) and net photosynthetic rate (4). This, of course, is the critical distinction between the path model in Figure 3.9 and the alternative model. When

Table 3.4. *The d-separation statements in the basis implied by the model in Figure 3.9, along with the Pearson and Spearman partial correlations and their 2-tailed probabilities. Results are shown for the full data set of 40 species and for the 35 species of C_3 species only. Numbers refer to the variables shown in Figure 3.9*

	Both C_3 and C_4 species				Only C_3 species			
	Pearson		Spearman		Pearson		Spearman	
d-sep	*r*	*p*(*r*)	*r*	*p*(*r*)	*r*	*p*(*r*)	*r*	*p*(*r*)
$1 \perp\!\!\!\perp 3 \mid 2$	−0.286	0.0777	−0.234	0.1523	−0.298	0.0871	−0.226	0.1986
$1 \perp\!\!\!\perp 4 \mid 3$	0.165	0.3163	0.217	0.1841	0.109	0.5392	0.188	0.2860
$1 \perp\!\!\!\perp 5 \mid 3,4$	0.035	0.8328	0.099	0.5560	0.043	0.8139	0.215	0.2303
$2 \perp\!\!\!\perp 4 \mid 1,3$	−0.092	0.5837	−0.069	0.6809	0.160	0.3743	0.156	0.3870
$2 \perp\!\!\!\perp 5 \mid 1,3,4$	0.262	0.1169	0.058	0.7327	−0.079	0.6678	−0.006	0.9758
Fisher's *C*:	$\chi^2 = 13.15$, 10 df, $p = 0.216$		$\chi^2 = 9.713$, 10 df, $p = 0.466$		$\chi^2 = 9.301$, 10 df, $p = 0.503$		$\chi^2 = 10.621$, 10 df, $p = 0.388$	

Table 3.5. *The d-separation statements implied by an alternative model in which the model in Figure 3.9 is changed to make leaf nitrogen cause net photosynthetic rate which then causes stomatal conductance, along with the Pearson and Spearman partial correlations and their 2-tailed probabilities. Results are shown for the full data set of 40 species and for the 35 species of C_3 species only*

	Both C_3 and C_4 species				Only C_3 species			
	Pearson		Spearman		Pearson		Spearman	
d-sep	*r*	*p*(*r*)	*r*	*p*(*r*)	*r*	*p*(*r*)	*r*	*p*(*r*)
$1 \perp\!\!\!\perp 3 \mid 2$	−0.286	0.0777	−0.234	0.1523	−0.298	0.0871	−0.226	0.1986
$1 \perp\!\!\!\perp 3 \mid 4$	0.286	0.0777	0.279	0.0853	0.221	0.2092	0.1723	0.3298
$1 \perp\!\!\!\perp 5 \mid 3,4$	0.035	0.8328	0.099	0.5560	0.043	0.8139	0.215	0.2303
$2 \perp\!\!\!\perp 3 \mid 1,4$	0.599	7×10^{-5}	0.569	2×10^{-4}	0.371	0.0338	0.339	0.0541
$2 \perp\!\!\!\perp 5 \mid 1,3,4$	0.262	0.1169	0.058	0.7327	−0.079	0.6678	−0.006	0.9758
Fisher's *C*:	$\chi^2 = 33.96$, 10 df, $p = 0.0002$		$\chi^2 = 27.59$, 10 df, $p = 0.0021$		$\chi^2 = 16.00$, 10 df, $p = 0.1000$		$\chi^2 = 14.27$, 10 df, $p = 0.161$	

looking only at the C_3 species, the alternative model does not have a large degree of lack of fit although the critical prediction still shows a reasonably large lack of fit ($r_{23|14}=0.371$, $p=0.0338$) and is always poorer that that provided by the structure shown in Figure 3.9.

Because of such results, and other reasons described in the original reference, I prefer the causal structure shown in Figure 3.9. Such a conclusion must remain tentative. After all, the conclusion is based on only 40 species and a larger sample size might detect some more subtle lack of fit that was too small to be found in the present data set.

Given the model in Figure 3.9, and given that we have not been able to reject it, we can now fit the path equations. Although Wright's original method was based on standardised variables, I prefer to use the original variables because the variables each have well-established units of measurement. The least squares regression equations, using only the C_3 species, are shown below. The residual variation is indicated by $N(0,\sigma)$.

$$\ln(\%\ \text{nitrogen})=0.78+0.90\ \ln(\text{SLM})+N(0,0.243),\quad R=0.85$$

$$\ln(\text{conductance})=-6.60+1.15\ \ln(\%\ \text{nitrogen})+N(0,0.56),\quad R=0.69$$

$$\ln(\text{photo})=3.08+0.55\ \ln(\text{conductance})+N(0,0.31),\quad R=0.81$$

$$\ln(CO_2\ \text{internal})=6.42+0.14\ \ln(\text{conductance})-0.1\ \ln(\text{photo})+N(0,0.04),\quad R=0.77$$

Each of the slopes is significant at a level below 10^{-4} and the sign of each is in the predicted direction. With these path equations we can begin to simulate how the entire suite of leaf traits would change if we change the specific leaf mass (the exogenous variable in this model) or if we observe species with different SLMs. We get the functional relationships by back-converting the variables in the equations from their natural logarithms. Of course, each of these variables may also change with changing environmental conditions. By including these environmental variables we could generate the response surfaces across which the suite of leaf traits would move as the environment changes.

4 Path analysis and maximum likelihood

James Burke (1996), in his fascinating book, *The pinball effect*, demonstrates the curious and unexpected paths of influence leading to most scientific discoveries. People often speak of the 'marriage of ideas'. If so then the most prolific intellectual offspring come, not from the arranged marriages preferred by research administrators, but from chance meetings and even illicit unions. The popular view of scientific discoveries as being linear causal chains from idea to solution is profoundly wrong; a better image would be a tangled web with many dead ends and broken strands. If most present knowledge depends on unlikely chains of events and personalities, then what paths of discovery have been deflected because the right people did not come together at the right time? Which historical developments in science have been changed because two people, each with half of the solution, were prevented from communicating due to linguistic or disciplinary boundaries? The second stage in the development of modern structural equation modelling is a case study in such historical contingencies and interdisciplinary incomprehension.

During the First World War, and in connection with the American war effort, Sewall Wright was on a committee allocating pork production to various US states on the basis of the availability of corn[1]. He was confronted with a problem that had a familiar feel. Given a whole series of variables related to corn availability and pork production, how do all these variables interact to determine the relationship between supply and demand, and the fluctuations between these two? It occurred to him that his new method of path analysis might help. He calculated the correlation coefficients between each pair of variables for five years, giving 510 separate correlations. After much trial and error he developed a model involving only four variables (corn price, summer hog price, winter hog price and hog breeding) and only 14 paths that still gave a 'good match' between observed and predicted correlations. He described his results in a manuscript that was submitted as a bulletin of the US Bureau of Animal Industry. It was

[1] This next section is based on Wright's biography (Provine 1986).

promptly rejected because officials at the Bureau of Agricultural Economics considered it to be an intrusion onto their turf. Happily for Wright, he had also shown it to the son of the secretary of agriculture (Henry A. Wallace) who was interested in animal breeding and quantitative modelling. Wallace, using his political influence, intervened to have the manuscript published as a US Department of Agriculture bulletin (Wright 1925).

Although economists later developed methods that were very similar to path analysis, Wright's foray into economics does not seem to have been very influential. During the Second World War, Wright presented a seminar on path analysis to the Cowles Commission, where economists were developing methods that were the forerunner of SEM. Neither Wright nor the economists recognised the link between the two approaches or the usefulness of such a marriage (Epstein 1987). None the less, some economists were independently trying to express causal processes in functional form[2] (Haavelmo 1943). In economics, constraints on the covariance matrix (for example, zero partial correlations due to d-separation) were called '*overidentifying constraints*'. Since most work in this area was in parameter estimation, not theory testing, such constraints were mostly avoided because they made consistent estimation difficult.

In the 1950s the political scientist Herbert Simon began to derive the causal claims of a statistical model[3]. This led some social scientists to think about expressing causal processes as statistical models that implied certain structural, or *overidentifying*, constraints. One such person was H. M. Blalock, who began deriving overidentifying constraints, in the form of zero partial correlations, that were implied by the structure of the causal process (Blalock 1961, 1964). Wright's method of path analysis had been largely rediscovered by social scientists, with the important difference that the emphasis shifted from being an *a posteriori* description of an assumed causal process – as Wright viewed his method – to being a (tentative) test of an assumed causal process. The late 1960s and early 1970s saw many applications of path analysis in sociology, political science and related social science disciplines.

The most important next step was the work of people such as Jöreskog (1967, 1969, 1970, 1973) and Keesling (1972), who developed ways of combining confirmatory factor analysis (see Chapter 5) and path analysis using maximum likelihood estimation techniques. The advance was not simply in using a new method of estimating the path coefficients. More importantly, the use of maximum likelihood allowed the resulting series of

[2] Some economists referred to Wright's work in passing (Goldberger 1972; Griliches 1974) but only for historical completeness.

[3] Summarized by Simon (1977).

equations describing the hypothesised causal process (a series of *structural equations*) to be tested against data in order to see whether the overidentifying constraints (the zero partial correlation coefficients and other types of constraint) agreed with the observations. This advance solved the main weakness of Wright's original method of path analysis, since one did not simply have to *assume* the causal structure, as Wright did. Now, one could test the statistical consequences of the causal structure and therefore potentially falsify the hypothesised causal structure[4]. Unfortunately, by the 1970s most biologists had forgotten about Wright's method of path analysis and disciplinary boundaries prevented the new SEM approach from penetrating into biology.

Wright's method was essentially the application of multiple regression based on standardised variables in the order specified by the path diagram (the causal graph). This, along with ANOVA and most other familiar statistical methods, consists of modelling the individual observations. In other words, the path coefficients were obtained using least square techniques by minimising the squared differences between the observed and predicted values of the individual observations, as is usual in multiple regression. Structural equations models, of which modern path analysis is a specialised version[5], concentrate instead on the pattern of covariation between the variables and minimise the difference between the observed and predicted pattern of covariation among them. The basic steps are:

1. Specify the hypothesised causal structure of the relationships between the variables.
2. Translate the causal model into an observational model. Write down the set of linear equations that follow this structure and specify which parameters (slopes, variances, covariances) are to be estimated from the data (i.e. that are *free*) and which are *fixed* (i.e. are not to be changed to accommodate the data) based on the causal hypothesis.
3. Derive the predicted variance and the covariance between each pair of variables in the model using covariance algebra. Covariance algebra gives the rules of path analysis that Wright had already derived.

[4] The logical and axiomatic relationships between probability distributions and causal properties had not yet been developed. This led to much confusion concerning the causal interpretation of structural equations models (Pearl 1997). One reason why I discuss these points in detail is to prevent the same sterile debates from recurring among biologists.

[5] There are a number of different names for this general class of models: structural equations models, LISREL models, covariance structure models.

4. Estimate these free parameters using maximum likelihood or related methods, while respecting the values of the fixed parameters. This estimation is done by minimising the difference between the observed covariances of the variables in the data and the covariances of the variables that are predicted by the causal model.
5. Calculate the probability of having observed the measured minimum difference between the observed and predicted covariances, assuming that the observed and predicted covariances are identical except for random sampling variation.
6. If the calculated probability that the remaining differences between observed and predicted covariances are due only to sampling variation is sufficiently small (say below 0.05) then one concludes that the observed data were not generated by the causal process specified by the hypothesis and that the proposed model should be rejected. If, on the contrary, the probability is sufficiently large (say above 0.05) then one concludes that the data are consistent with such a causal process.

4.1 Testing path models using maximum likelihood

Step 1: Translate the hypothetical causal system into a path diagram

This first step should be almost second nature by now, but there are a few notational conventions that must be introduced. Path diagrams contain three different types of variable. Variables that have been directly observed and measured are enclosed in squares; these variables are called *manifest* variables in SEM jargon. Variables that are hypothesised to have a causal role in the model, but which have not been directly observed or measured, are enclosed in circles; these variables are called *latent* variables in SEM jargon[6]. The third type of variable is the residual *error* variable[7] and it is not enclosed at all. This type of variable represents all other unmodelled causes of the variable into which it points. It is also generally defined as a normally distributed random variable with a mean of zero and a variance of 1, although it

[6] By convention, a path model is simply a structural equations model that does not involve unmeasured, or latent, variables.

[7] A common error is to assume that the residual error in a structural equations model is the same as the residual error in a regression model. The two are not necessarily the same. In a regression model the residuals are always uncorrelated (or orthogonal) with the predictors. The residuals in SEM need not be uncorrelated with the predictors or with each other.

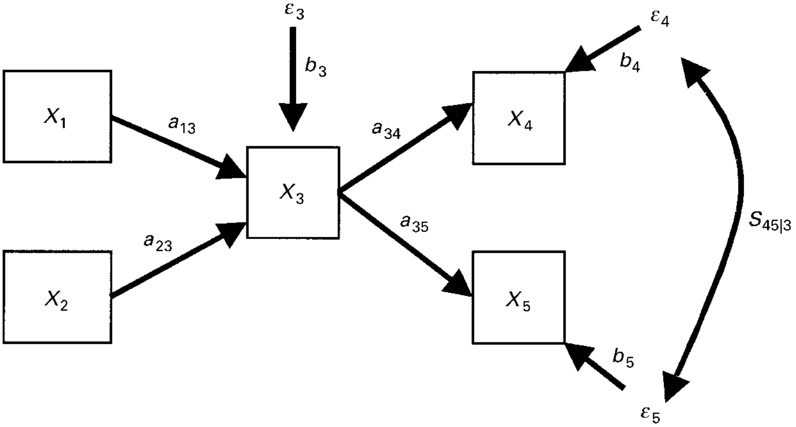


Figure 4.1. A path model involving five variables; $S_{45|3}$ is a free covariance between ε_4 and ε_5.

is possible to model it with a variance different from 1. A second common classification is between a variable that has no causal parents in the model, called *exogenous*, and a variable that is caused by some other variable in the model, called *endogenous*[8]. Finally, there are two types of arrow. A straight arrow indicates a causal relationship between two variables just as it does in the directed graphs of previous chapters. A curved, double-headed arrow indicates an unknown causal relationship linking the two variables. This means that there is a free covariance in the structural equations. These conventions are shown in Figure 4.1.

Step 2: Translate the causal model into an observational model in the form of a set of structural equations

As the arrows in Figure 4.1 suggest, the hypothesised relationships are asymmetrical causal ones. When we translate the causal model into mathematical equations we obtain an observational (statistical) model. Because we are now dealing with a statistical model, we must make assumptions concerning the form of the functional relationships and the sampling distribution of the random variables. Contrary to the path diagram, this new model is not strictly a causal model because it is expressed in the language of algebra using the equivalence operator ('='). It is an imperfect translation of the causal model and we must not forget this and begin manipulating these algebraic equations in ways contrary to the original asymmetrical causal relations

[8] Yes, I know; I didn't invent these terms. I will try to limit the jargon to a minimum but you will have to be aware of these terms in order to read the literature.

expressed in the path diagram. In almost all structural equations models, the relationships are assumed to be additively linear. In most structural equations models, the random variables are assumed to be multivariate normal. 'Multivariate normal' means that the variables follow a multivariate normal distribution, which is a somewhat stronger assumption than assuming simply that each variable is normally distributed. Different ways of assessing this assumption, and the degree of non-normality that can be tolerated, are described in Chapter 6.

If your causal model is sufficiently detailed that you are willing to hypothesise the numerical values of some parameters (path coefficients, variances or covariances) then you can include this information in the model by specifying the parameter to be fixed. If you are not able or willing to make such an assumption (except, of course, that the parameter is not zero) then the parameter is estimated from the data and is therefore *free*. Specifying that a variable is not a direct cause of another (i.e. that there is not an arrow from the one to the other in the path diagram) is the same as specifying that the path coefficient of this 'missing' arrow is fixed at zero. Each parameter that is fixed adds one degree of freedom to the inferential test.

In such models the interest is in the relationships between the variables, not the mean values of the variables themselves[9]. For this reason, all variables are 'centred' by subtracting the mean value of each variable from each observation. For instance, if the mean of X_1 in Figure 4.1 were 6, then we would replace each value of X_1 by $(X_1 - 6)$. This trick ensures that the mean of each transformed variable is zero and therefore that the intercepts are zero. Assuming that all of our variables are already centred, these are the structural equations corresponding to Figure 4.1, where $\mathrm{Cov}(X_1,X_2)$ means the population covariance between X_1 and X_2:

$$X_1 = N(0,\sigma_1) \qquad \varepsilon_3 = N(0,\sigma_3)$$

$$X_2 = N(0,\sigma_2) \qquad \varepsilon_4 = N(0,\sigma_4)$$

$$X_3 = a_{13}X_1 + a_{23}X_2 + b_3\varepsilon_3 \quad \varepsilon_5 = N(0,\sigma_5)$$

$$X_4 = a_{34}X_3 + b_4\varepsilon_4$$

$$X_5 = a_{35}X_3 + b_5\varepsilon_5$$

$$\mathrm{Cov}(X_1,X_2) = \mathrm{Cov}(X_1,\varepsilon_3) = \mathrm{Cov}(X_1,\varepsilon_4) = \mathrm{Cov}(X_1,\varepsilon_5) = \mathrm{Cov}(X_2,\varepsilon_3) = \mathrm{Cov}(X_2,\varepsilon_4) = \mathrm{Cov}(X_2,\varepsilon_5) = \mathrm{Cov}(\varepsilon_3,\varepsilon_4) = \mathrm{Cov}(\varepsilon_3,\varepsilon_5) = 0$$

$$\mathrm{Cov}(\varepsilon_4,\varepsilon_5) = \sigma_{45}$$

[9] Means can also be modelled but this requires a little bit more work.

Happily, most commercial SEM programs do most of this translation work for you; you have only to specify which parameters are free and which variables are direct causes of which other variables. In fact, all you have to do is draw the path model in the latest versions of most commercial SEM programs[10]. Notice that some parameters (σ_i,a_{ij}) in the above equations do not have numerical values and therefore have to be estimated; before looking at the data they are 'free' to take on any value.

Let's go through these equations more slowly to understand exactly how the causal model in Figure 4.1 has been represented in equation form. First, variables X_1 and X_2 are exogenous in the model; we don't know, or are not interested in explicitly modelling, the causal parents of these two variables. In the equations I have specified that X_1 and X_2 are each normally distributed random variables whose mean is zero and whose standard deviation is unknown. Therefore, these two standard deviations are free and must be estimated from the data[11]. Next, X_3 is written as a linear function of both X_1 and X_2 in accordance with Figure 4.1. Since I don't know the numerical strength of the direct causal effects of these two variables, the path coefficients (a_{13} and a_{23}) are also free and must be estimated from the data. If my causal hypothesis had been sufficiently well developed that I could specify what the values of these path coefficients were then I would have entered the predicted values rather than having to estimate them from the data. In addition, the combined direct effect of the other unknown causes of X_3 are not known either and so b_3, the path coefficient from the error variable (ε_3), is also free and must be estimated. Remember that all of the error variables (ε) are unit normal variables, i.e. with a zero mean and a standard deviation of 1. Multiplying a unit normal variable by a constant (b_3 in this case) makes its variance equal to the constant. Therefore, the part of the variance of X_3 that is not accounted for by X_1 and X_2 is b_3. In this particular equation the residual is exactly analogous to the residuals of a multiple regression, since it is made to be uncorrelated with either X_1 or X_2 but this is not always the case. Next, each of X_4 and X_5 are also written as linear functions of X_3 with the accompanying free path coefficients.

Since there are five variables in the model, there are 10 different pairs of variables and therefore 10 different covariances between the unique pairs of variables. Since X_1 and X_2 are causally independent, the covariance between these two variables must be zero (remember d-separation). X_1, X_2, and the unknown other causes of X_3 (i.e. ε_3) are also independent of each other and of the unknown other causes of X_4 and X_5 and so each of these

[10] The Appendix lists some common SEM programs.

[11] Fixing the error variance at 1.0 and freely estimating the path coefficient associated with the error variable, or fixing the path coefficient to 1.0 and freely estimating the error variance, are two equivalent ways of doing the same thing.

pairs of covariances must also be zero. Finally, the causal model in Figure 4.1 states that there is some causal influence linking X_4 and X_5 but the researcher does not know what it is. Perhaps X_4 causes X_5? Perhaps X_5 causes X_4? Perhaps there is a reciprocal causal relationship? Perhaps there is some unknown common cause of both X_4 and X_5? Adding a free covariance (the translation of a curved double-headed arrow) is an admission of ignorance as to the causal origin of the covariance. Each of the above causal relationships would generate a non-zero covariance between X_4 and X_5 even after controlling for X_3. Therefore, we allow ε_4 and ε_5 to have a non-zero covariance and the numerical value of this non-zero covariance must also be estimated from the data.

This completes the best translation of the causal model into the observational (statistical) model as can be obtained consistent with the statistical assumptions that are needed to estimate the free parameters. It will be important to evaluate these assumptions when judging whether the results of the analysis can be trusted, as is true of any statistical method.

Step 3: Derive the predicted variance and the covariance between each pair of variables in the model using covariance algebra

Box 4.1. Basic rules of covariance algebra

The notation E(X) means the expected value of X. So, the population covariance between two variables – symbolized here as Cov(X_1,X_2) – is defined as:

$$\mathrm{Cov}(X_1,X_2)=\mathrm{E}[(X_1-\mathrm{E}(X_1))(X_2-\mathrm{E}(X_2))]=\mathrm{E}(X_1X_2)-\mathrm{E}(X_1)\mathrm{E}(X_2)$$

If the variables are centred about their expected values, this reduces to:

$$\mathrm{Cov}(X_1,X_2)=\mathrm{E}(X_1,X_2)$$

Since a variance is simply the covariance of a variable with itself, we can write the population variance (Var) as:

$$\mathrm{Var}(X_1)=\mathrm{Cov}(X_1,X_1)$$

If k is a constant and X_1, X_2, X_3 are random variables then we can also state the following useful rules:

(1) $\mathrm{Cov}(k,X_1)=0$

(2) $\mathrm{Cov}(kX_1,X_2)=k\mathrm{Cov}(X_1,X_2)$

(3) $\mathrm{Cov}(k_1X_1,k_2X_2)=k_1k_2\mathrm{Cov}(X_1,X_2)$

(4) $\mathrm{Cov}(X_1+X_2,X_3)=\mathrm{Cov}(X_1,X_2)+\mathrm{Cov}(X_2,X_3)$

The set of structural equations allows us to derive the predicted values for the covariances between each pair of variables. Since a Pearson correlation coefficient is simply a covariance that has been standardised, one can also derive the predicted values for the correlations between each pair of variables. This step uses the rules of path analysis that were derived originally by Wright. Box 4.1 summarises a few basic rules of covariance algebra that will be useful in discussing this section.

From Chapter 2 we know that two vertices in the path diagram that are d-separated correspond to two random variables that are independently distributed, meaning that the population covariance between them must be zero. If the two vertices are not d-separated, then the corresponding random variables are not independently distributed and so (given the assumption of linearity made by SEM) the covariance between them can't be zero. This justifies the list of zero covariances in the structural equations given above. For those vertices that are not unconditionally d-separated (and are therefore correlated in some way), we can use the rules of covariance algebra to obtain formulae giving their covariances. Take, for instance, variables X_1 and X_3 in Figure 4.1. We can write:

$$\text{Cov}(X_1,X_3) = \text{Cov}(X_1, (a_{13}X_1 + a_{23}X_2 + b_3\varepsilon_3)) = \text{a}_{13}\text{Cov}(X_1,X_1) + a_{23}\text{Cov}(X_1,X_2) + b_3\text{Cov}(X_1,\varepsilon_3).$$

Looking at the path diagram and applying the d-separation operation, we see that X_1 is independent of both X_2 and ε_3 and therefore the population covariances involving X_1 and these two variables are zero. Therefore, the population covariance between X_1 and X_3 is simply $a_{13}\text{Cov}(X_1,X_1)$ or $a_{13}\text{Var}(X_1)$. In this way we can obtain formulae for the expected value for each pair of variables in the model. These are shown in Table 4.1.

This must seem like a lot of work. Most commercial SEM programs do all of this work for you and the important point at this stage is only that you have an intuitive understanding of why we can express the covariances between each pair of variables as a function of path coefficients, variances and covariances. For those who are used to working with matrix algebra, Box 4.2 gives a more formal derivation of the predicted covariance matrix based on the Bentler–Weeks model (for a concise description, see Bentler 1995).

If we go back to the analogy of correlations being the shadows that are cast by causal processes, then Table 4.1 is a description of the 'shape' of the shadow that is cast by the hypothesised causal process shown in Figure 4.1. Imagine that we were describing the shadow cast by a solid square whose size was unknown to us (i.e. the length of whose sides are *free* param-

Table 4.1. *Predicted population variances and covariances for the observed variables shown in Figure 4.1. Since this is a symmetrical matrix, the values on the lower triangle of this matrix are the same as those on the upper triangle*

	X_1	X_2	X_3	X_4	X_5
X_1	$\mathrm{Var}(X_1)$	0	$a_{13}\mathrm{Var}(X_1)$	$a_{13}a_{34}\mathrm{Var}(X_1)$	$a_{23}a_{35}\mathrm{Var}(X_1)$
X_2		$\mathrm{Var}(X_2)$	$a_{23}\mathrm{Var}(X_2)$	$a_{23}a_{34}\mathrm{Var}(X_2)$	$a_{23}a_{35}\mathrm{Var}(X_2)$
X_3			$\mathrm{Var}(X_3)$		$a_{35}\mathrm{Var}(X_3)$
X_4				$\mathrm{Var}(X_4)$	
					$a_{34}a_{35}\mathrm{Var}(X_3)+b_4b_5\mathrm{Cov}(\varepsilon_4\varepsilon_5)$
X_5					$\mathrm{Var}(X_5)$

eters). We would describe the shadow as having four equal sides of unknown length (the first constraint) with four sides that meet in such a way that they make four corners having 90° angles (the second constraint). The general *shape* of the shadow is fixed (a square) but the numerical *values* (the lengths of the sides) are free parameters and can be estimated by measuring the real shadow.

Box 4.2. The Bentler–Weeks model

Definitions: Let the endogenous (i.e. dependent) variables in the model be written in a column vector called η and let the exogenous (i.e. independent variables, including the error variables) be written in a column vector called ε. Let the coefficients of the effects of dependent causes to dependent effects be a matrix called $\boldsymbol{\beta}$ (rows are dependent effects and columns are dependent causes) and let the coefficients of the effects of independent causes to dependent effects be a matrix called $\boldsymbol{\tau}$ (rows are dependent effects and columns are independent causes). Then the system of structural equations can be written as: $\eta = \boldsymbol{\beta}\eta + \boldsymbol{\gamma}\varepsilon$.

For instance, the path model in Figure 4.1 would be written:

$$\begin{bmatrix} x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ a_{34} & 0 & 0 \\ a_{35} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} a_{13} & a_{23} & b_3 & 0 & 0 \\ 0 & 0 & 0 & b_4 & 0 \\ 0 & 0 & 0 & 0 & b_5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

In reduced form the equation is: $\boldsymbol{\eta}=(\boldsymbol{I}-\boldsymbol{\beta})^{-1}\boldsymbol{\gamma}\varepsilon$, where $\boldsymbol{I}$ is the identity matrix. Predicted covariances between exogenous variables: $\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']=\boldsymbol{\varsigma}$. Predicted covariances between endogenous and exogenous variables: $\mathrm{E}[\boldsymbol{\eta}\boldsymbol{\varepsilon}']=(\boldsymbol{I}-\boldsymbol{\beta})^{-1}\boldsymbol{\gamma}\boldsymbol{\varsigma}$. Predicted covariances between endogenous variables: $\mathrm{E}[\boldsymbol{\eta}\boldsymbol{\eta}']=(\boldsymbol{I}-\boldsymbol{\beta})^{-1}\boldsymbol{\gamma}\boldsymbol{\varsigma}\boldsymbol{\gamma}\,(\boldsymbol{I}-\boldsymbol{\beta})^{-1'}$.

Step 4: Estimate the free parameters by minimising the difference between the observed and predicted variances and covariances

The hypothesised object was the solid square and from this we have predicted the shape of the shadow that it would cast. Is our hypothesis correct? To decide, we look at the actual shadow, choose values for the length of the sides of our hypothesised square that make it as numerically close to the observed shadow as possible while respecting the constraints, and then measure the remaining lack-of-fit. This is the same basic logic used to fit and test a structural equations model. We first choose values for the free parameters in our predicted covariance matrix that make it as numerically close as possible to the observed covariance matrix, while respecting the constraints applied to the predicted covariance matrix. How this is done depends on the assumptions that have been made concerning the distributional form of the random variables; in SEM the usual assumption is that the random variables follow a multivariate normal distribution.

The general strategy for obtaining the best values for the free parameters is easy enough to grasp: choose values of the free parameters that make the numerical values of the predicted covariance matrix (i.e. the values in Table 4.1 after replacing the variables by their numerical values) as close as possible to the actual covariances measured in the data. This is usually done using maximum likelihood estimation (Fisher 1950). Eliason (1993) and Bollen (1989) described the mechanics of this technique and Box 4.3 gives a brief introduction for those who are interested in this topic. In essence, the numerical algorithm used to maximise the likelihood is a bit like playing the game of 20 questions (Is it alive? Is it a mammal? Is it a carnivore? Does it live in Africa? . . .). Box 4.3 gives a more precise definition of the likelihood function but this function can be intuitively understood to measure the discrepancy between the observed data and the sort of data that would have been observed had the free parameter been equal to our chosen value. We start with an initial guess of the values of the free parameters and calculate the likelihood of the data given the current parameter values. We then see whether we can modify our guess of the values of the free parameters in such a way as to improve the likelihood. We continue with this process until we find values such that any change to them will do worse than the present values.

Another analogy for maximising a likelihood function might be a person who is blindfolded and finds herself in a landscape with various hills and valleys. Her job is to walk down to the valley floor without peeking. She begins by taking an initial step in a direction based on her best guess. If she sees that she has moved down-slope then she continues in the same direction with a second step in the same direction. If not, she changes direction and tries again. She continues with this process until she finds herself at a position on the landscape in which every possible change in direction results in movement up-slope. She therefore knows that she is in a valley. Unfortunately, if the landscape is very complicated she may have found herself in a small depression rather than on the true valley floor. The only way to find out would be to start over at a different initial position and see whether she again ends up in the same place.

Box 4.3. Maximum likelihood estimation

The probability of occurrence of a random variable, X_i, is given by a probability function (for discrete variables) or a probability density function (for continuous variables). For instance, the probability density function of a univariate normal random variable is:

$$f(X;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

The notation means that x is the random variable and μ and σ are population parameters that are fixed. Now, if we take a series of N independent observations of the random variable, then the joint probability density function for these N observations is: $f(X_1;\mu,\sigma)f(X_2;\mu,\sigma)f(X_3;\mu,\sigma)\ldots f(X_N;\mu,\sigma)$. The objective of maximising the likelihood of a parameter is to find a value of this parameter that maximises this joint probability density function. In other words, find a value for the parameter (for example μ) that maximises the likelihood of having observed the series of observations. This objective turns the probability density function on its head. Now the observed values (X_i) are fixed and we view the population parameters as variables. We are envisaging a whole series of different normal distributions and we want to choose the most likely one given our data. So, the likelihood function of the univariate normal distribution is:

$$\mathrm{L}(\mu,\sigma;X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

and the joint likelihood function of the entire set of data is: $L(\mu,\sigma;X_1)L(\mu,\sigma;X_2)L(\mu,\sigma;X_3)\ldots L(\mu,\sigma;X_N)$.

The natural logarithm of a series of positive numbers is an increasing function of these numbers. Because it is difficult to maximise a product but easier to maximise a sum, we use the logarithm of the likelihood function. For instance, imagine that we have observed eight values (1.20, 0.08, 0.34, 0.57, 0.46, 0.48, 0.56, 1.01) from a normal distribution whose population variance (σ^2) is 1 and we want to find the maximum likelihood value for the population mean (μ). Figure 4.2 shows a graph of the log-likelihood function over the range $\mu = -4$ to $+4$.

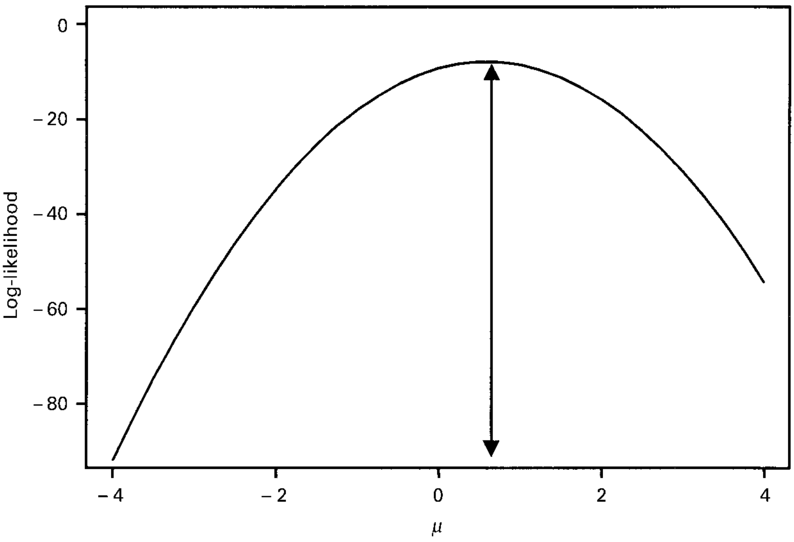


Figure 4.2. The log-likelihood function for the population mean (μ), given eight values (1.20, 0.08, 0.34, 0.57, 0.46, 0.48, 0.56, 1.01) from a normal distribution whose population variance (σ^2) is 1 over the range −4 to +4.

We see that this function is maximal at around 0.58. This is the sample mean of our eight values, showing that the standard formula for the sample mean – an unbiased estimator – is also a maximum likelihood estimator. Maximum likelihood estimates are not always unbiased in small samples (for instance the maximum likelihood estimate of the variance is not) but they are consistent, meaning that such estimates converge on the true value as the sample size increases. In other words maximum likelihood estimates are asymptotic estimates.

In general, the maximum (or minimum) of the likelihood function occurs when its first derivative is zero. To see whether one has found a maximum, one then checks to see whether the second derivative is negative.

In SEM, the usual assumption is that the data are multivariate normal. In other words, we assume that the probability density function is multivariate normal. The likelihood function for this distribution is:

$$\mathrm{L}(\boldsymbol{\mu},\boldsymbol{\Sigma};X)=\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\mathrm{e}^{-[X-\mu]\boldsymbol{\Sigma}^{-1}[X-\mu]}.$$

In SEM we usually centre the variables about their means, so the only parameters whose likelihood estimates we need to estimate are those in the population covariance matrix $\boldsymbol{\Sigma}$. These parameters are the free parameters that we have already derived. With this much more complicated function we are not able to derive the maximum likelihood estimates directly, and so we have to use numerical methods that involve iteration.

For instance, in Table 4.1 we had free parameters for the variances of the two independent observed variables, the three coefficients for the error variables and the one free covariance between X_4 and X_5. Let's group all these free parameters together in a vector called $\boldsymbol{\theta}$. Now, if we take a first guess at the values of these free parameters then we can calculate the predicted covariance matrix based on these initial values; let's call the predicted covariance matrix that results from this guess $\boldsymbol{\Sigma}_{(1)}(\boldsymbol{\theta})$ to emphasise that this matrix will change if we change our values in $\boldsymbol{\theta}$. We now calculate the value of the log-likelihood function. Next, we change our initial estimates of the free parameters and recalculate the predicted covariance matrix, $\boldsymbol{\Sigma}_{(2)}(\boldsymbol{\theta})$, in such a way as to increase the log-likelihood. We continue in this way until we can't increase the value of the log-likelihood anymore.

Problems can occur if the log-likelihood function contains 'potholes' or local maxima. If this happens then the iterative procedure can become 'trapped' without finding the global maximum. The only way to determine whether one has found a global maximum is to try different starting values and see whether they all converge on the same values. Problems can also occur if the iterative procedure wanders into areas of parameter space that are illegitimate, for instance, negative variances. Most computer programs will warn you when this happens, and this is usually a sign of a poorly fitting model.

Let $\boldsymbol{S}$ be the observed covariance matrix, involving p dependent (endogenous) and q independent (exogenous) variables. Let $\boldsymbol{\Sigma}$ be the maximum likelihood estimate of the model covariance matrix. Since these maximum likelihood estimates depend on the values of the free parameters, which we group together in a vector $\boldsymbol{\theta}$, we will write the model covariance matrix as $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The maximum likelihood fitting function, F_{ML}, that compares the difference between the observed and predicted covariance matrices is:

$$F_{\text{ML}} = \ln|\boldsymbol{\Sigma}(\theta)| + \text{trace}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}(\theta)) - \ln|\boldsymbol{S}| - (p+q)$$

This function has three important properties. First, the values of the free parameters, $\boldsymbol{\theta}$, that minimise it are also the values that make the predicted covariance matrix as similar as possible to the observed covariance matrix while respecting the constraints implied by the causal model. Second, the values of $\boldsymbol{\theta}$ that *minimise* this function are the same values that *maximise* the multivariate normal likelihood function and so such values of the free parameters that define the population covariance matrix ($\boldsymbol{\theta}$) are called *maximum likelihood estimates*. Third, and most importantly, if the observed data (and therefore $\boldsymbol{S}$) really were generated by the causal process that the structural equations are modelling, then the only remaining differences between $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{S}$ at the minimum of F_{ML} will be due to normally distributed random sampling variation. Given these assumptions then $(N-1)F_{\text{ML}}$ is asymptotically distributed as a chi-squared distribution.

I said that one probable reason why biologists did not accept Wright's method of path analysis was that his original method could derive the logical consequences of a causal model but could not test it. The method described above, developed by Jöreskog (1970), was the first to solve this important shortcoming of path analysis.

Step 5: Calculate the probability of having observed the measured minimum difference, assuming that the observed and predicted covariances are identical except for random sampling variation

The central chi-squared distribution has only one parameter: the degrees of freedom. In testing a structural equations model we are comparing the fit between the observed and predicted elements of the covariance matrix. If we have v variables then there will be v^2 elements in the covariance matrix. Since this matrix is symmetrical about its diagonal, some of these elements are redundant. The number of unique elements is $v(v+1)/2$. If we were to compare the observed and predicted values of these unique elements using $(N-1)F_{\text{ML}}$ and all of the predicted values were obtained independently of the observed values then this would define the degrees of freedom for the chi-squared test. However, we have had to use our data to estimate the free parameters that partly determine the predicted covariance matrix. Each free parameter that we have to estimate 'uses up' one degree of freedom. The degrees of freedom available to test the model are:

$$\frac{v(v+1)}{2} - (p+q)$$

As before, q is the number of free variances of exogenous variables (including the error variables) in the model and p is the number of free path coefficients in the model. I say 'free' because it may sometimes be possible to specify the value of a variance or path coefficient based on theory or prior experience and therefore constrain the model to have the specified value no matter what the data say.

So we specify, as the null hypothesis, that there is no difference between the observed and predicted covariance matrices except what would be expected given the random sampling variation of N independent observations all taken from the same multivariate normal distribution. Given this hypothesis, then the following statistic (the *maximum likelihood chi-squared statistic*) will asymptotically follow a central chi-squared distribution with the degrees of freedom given above:

$$(N-1)F_{\mathrm{ML}} \xrightarrow{N\to\infty} \chi^2_{[v(v+1)/2]-(p+q)}$$

In practice one uses a computer program to do all of these calculations.

Step 6: If the calculated probability is sufficiently small (say below 0.05) then one concludes that the model was wrong. If the probability is sufficiently large (say above 0.05) then one concludes that the data are consistent with such a causal process

At first blush this step appears much easier to understand than the previous ones. In fact, it is the step than causes the greatest confusion. The previous steps are more mathematically involved but they are largely automated and so the user does not need more than an intuitive grasp of what is happening. This last step requires that the user interpret the meaning of the resulting probability for the biological model. This interpretation can often lead to confusion.

In most of the statistical tests used by biologists, the biologically interesting hypothesis is the alternative hypothesis; the null hypothesis functions as a strawman that is erected only to see whether we have sufficiently strong evidence to knock it down. This is useful because it forces us to have strong evidence (evidence beyond reasonable doubt) before we can accept the biologically interesting alternative hypothesis. In SEM on the other hand, models are constructed based on biological arguments in such a way as to reflect what we hypothesise to be correct. In other words, our model and the resulting predicted covariance matrix embodies what we view to be biologically interesting. The null hypothesis, not the alternative, is therefore the biologically interesting hypothesis. A probability below the chosen significance level means that the predicted model is wrong and should be rejected (i.e. the null hypothesis should be rejected). Although the flipping of the null and alternative hypotheses might seem strange, it is exactly the same logic as testing the null hypothesis that the slope of a simple linear regression equals, say, 0.75. Notice that we are reversing the burden of proof: we are requiring strong evidence, evidence beyond reasonable doubt, before we are willing to reject our preferred hypothesis. This leads naturally to the temptation to conclude that the predicted model is correct simply because we have not obtained strong evidence to the contrary! In fact, all that we can conclude is that we have no good evidence to reject our model and that the data are consistent with it. The degree to which we have good evidence in favour of our model will depend on how well we can exclude other models that are also consistent with the data. This leads naturally to the subject of equivalent models (Chapter 8).

At this stage, some numerical examples will help. I will generate 100 independent 'observations' following the causal graph shown in Figure 4.1. Here are the generating equations; these are the same as those shown previously except that the free parameters have been replaced by actual values:

$$X_1 = \mathrm{N}(0,1)$$

$$X_2 = \mathrm{N}(0,1)$$

$$X_3 = 0.5X_1 + 0.5X_2 + 0.5\varepsilon_3$$

$$X_4 = 0.5X_3 + 0.707\varepsilon_4$$

$$X_5 = 0.5X_3 + 0.707\varepsilon_5$$

$$\mathrm{Cov}(X_1,X_2) = \mathrm{Cov}(X_1,\varepsilon_3) = \mathrm{Cov}(X_1,\varepsilon_4) = \mathrm{Cov}(X_1,\varepsilon_5) = \mathrm{Cov}(X_2,\varepsilon_3) = \mathrm{Cov}(X_2,\varepsilon_4) = \mathrm{Cov}(X_2,\varepsilon_5) = \mathrm{Cov}(\varepsilon_3,\varepsilon_4) = \mathrm{Cov}(\varepsilon_3,\varepsilon_5) = 0$$

$$\mathrm{Cov}(\varepsilon_4,\varepsilon_5) = 0.5$$

First, we look at the observed covariance matrix obtained from these 100 observations. This matrix is the observational 'shadow' that was cast by the causal process shown in Figure 4.1 and quantified by the above equations. Table 4.2 shows this covariance matrix.

The first step is to specify the hypothesised causal model. Imagine that we actually had two different competing models and wished to test between them. The first model is the model shown in Figure 4.1; this is the correct model that generated these data, although the model contains free parameters that have not been specified by our theory. The second causal

Table 4.2. *The observed unique variances and covariances between variables X_1 to X_5 from 100 simulated observations based on the causal process shown in Figure 4.1*

	X_1	X_2	X_3	X_4	X_5
X_1	0.931				
X_2	0.171	1.094			
X_3	0.630	0.762	1.350		
X_4	0.384	0.368	0.743	1.265	
X_5	0.324	0.385	0.611	0.624	0.949

model that we want to test is shown in Figure 4.3. The next step is to translate our two hypothesised causal graphs into structural equations. The translation of our first model has already been given. The translation of this second (incorrect) model is the following:

$$X_1 = \mathrm{N}(0,\sigma_1)$$

$$X_2 = \mathrm{N}(0,\sigma_2)$$

$$X_3 = a_{13}X_1 + a_{23}X_2 + b_3\varepsilon_3$$

$$X_4 = a_{34}X_3 + b_4\varepsilon_4$$

$$X_5 = a_{25}X_2 + b_5\varepsilon_5$$

$$\mathrm{Cov}(X_1,X_2) = \mathrm{Cov}(X_1,\varepsilon_3) = \mathrm{Cov}(X_1,\varepsilon_4) = \mathrm{Cov}(X_1,\varepsilon_5) = \mathrm{Cov}(X_2,\varepsilon_3) = \mathrm{Cov}(X_2,\varepsilon_4) = \mathrm{Cov}(X_2,\varepsilon_5) = \mathrm{Cov}(\varepsilon_3,\varepsilon_4) = \mathrm{Cov}(\varepsilon_3,\varepsilon_5) = 0$$

$$\mathrm{Cov}(\varepsilon_4,\varepsilon_5) = 0$$

Note the differences between these structural equations and the ones derived from the correct model. First, the population covariance between the residual errors of X_4 and X_5 (i.e. $\mathrm{Cov}(\varepsilon_4,\varepsilon_5)$) is zero in this incorrect model. Second, X_5 is hypothesised to be directly caused by X_2 rather than being indirectly caused by both X_1 and X_2 through their effects on X_3.

We next have to obtain the maximum likelihood estimates of the free parameters of each model. To do this we have to provide starting values for the iterative process. In this book I use the EQS program for structural equation models, although there are many other commercial programs on

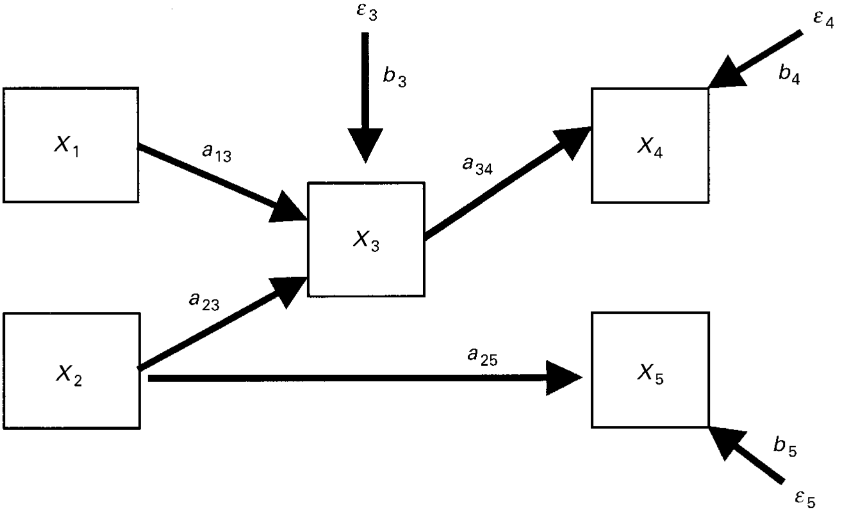


Figure 4.3. An alternative path model involving five variables.

the market. In the path models discussed in this chapter the choice of starting values for the free parameters is usually not critical and so I will use the default values of 1.0 for all of them. Remember that the fitting of these free parameters is an iterative process in which the estimates at each iteration are changed in such a way as to reduce the maximum likelihood fitting function, as described in Box 4.3. At the very first iteration, when all free parameters are equal to 1.0, both the correct model and the incorrect model produce a predicted covariance matrix that poorly fits the observed values; the maximum likelihood fitting function is 0.45707 for the correct model and 0.74821 for the incorrect model. The correct model took five iterations to converge on the maximum likelihood estimates, giving a final value of 0.04890 for the maximum likelihood fitting function. Since this value, multiplied by 99 (i.e. $N-1$) is the maximum likelihood chi-squared statistic, the final value of the chi-squared statistic was 4.8411. The incorrect model took four iterations to converge on the maximum likelihood estimates, giving a final value of 0.39369 for the maximum likelihood fitting function. Therefore, the final value of the chi-squared statistic for this incorrect model was 38.98.

To see whether these chi-squared statistics are significantly different from what one would expect given a correct model, we next need to determine the degrees of freedom. In both models we had five measured variables, giving a total of 15 unique variances and covariances, i.e. $v(v+1)/2$, or $5(6)/2$. In the correct model we had to estimate the variances of X_1 and

X_2, and the three error variances as well as four path coefficients and one free covariance (between X_4 and X_5); this makes 10 free parameters to estimate. The correct model therefore had $15-10=5$ degrees of freedom. In the incorrect model we had to estimate the variances of X_1 and X_2, and the three error variances as well as four path coefficients but no free covariance; this makes nine free parameters to estimate. The incorrect model therefore had $15-9=6$ degrees of freedom. SEM programs do all of these calculations for you.

Remember what we are testing. *If* the hypothesised model is correct, *then* the maximum likelihood chi-squared statistic will follow a chi-squared distribution. If the predicted and observed covariance matrices were identical then the maximum likelihood chi-squared statistic would be zero. The further the predicted covariance matrix deviates from the observed covariance matrix, the larger the maximum likelihood chi-squared statistic will be. Of course, even if our causal model were correct we would not expect the two matrices to be identical because of sampling variation; the predicted covariance matrix contains the predicted population values but the observed matrix is from a random sample of 100 observations. However, if the only differences were due to random sampling fluctuations then the maximum likelihood chi-squared statistic would closely follow (for large samples) a theoretical chi-squared distribution with the appropriate degrees of freedom. To evaluate our two models, we have only to hypothesise that each is the true model and then calculate the probability, based on this null hypothesis, of observing at least as large a difference between the observed and predicted covariance matrices as measured by our statistic.

First, let's look at the results for the correct model. The probability of observing a chi-squared value of at least 4.8411 with 5 degrees of freedom is 0.44. In other words, there is a probability of 0.44 of seeing such a result even if our null hypothesis were correct. In fact, our null hypothesis is correct, since we generated our data to agree with it. The result is telling us what we know to be true: the data are perfectly consistent with the model given normally distributed sampling variation. On the other hand, the maximum likelihood chi-squared statistic for the incorrect model was 38.98 with 6 degrees of freedom. The probability of observing such a large difference between the observed and predicted covariance matrices, assuming that the data were actually generated according to the incorrect model, is 7.2×10^{-7}. We either have to accept that an extremely rare event has occurred (one chance in about 1.5 million times) or reject the hypothesis that our data were generated according to the incorrect model. Again, the result is telling us what we know to be true: the data are not consistent with the model.

Compare the two predicted covariance matrices with the observed covariance matrix to see where the differences lie (Table 4.3). The biggest differences involve X_5. First, the predicted covariance between X_1 and X_5 is zero in the incorrect model while the observed value is 0.324. This is because X_1 is d-separated from X_5 in the incorrect model. The fitting procedure had to respect this constraint when fitting the incorrect model and so constrained this predicted covariance to be zero. The correct model allows X_1 to be an indirect cause of X_5 through its effect on X_3. The fitting procedure had to respect the constraint that the partial covariance between X_1 and X_5 be zero when controlling for X_3 but, since this constraint actually existed in the generating process, such a constraint did not distort the estimates. In the same way, the incorrect model required that the partial covariance between X_3 and X_5 as well as the partial covariance between X_4 and X_5 be zero when controlling for X_2. Since neither of these constraints actually existed in the correct causal process, the fitting procedure was forced to distort the estimates in order to meet these incorrect constraints.

Let's look next at the maximum likelihood estimates for the free parameters in the two different models. Here again are the true population values used to generate the data:

$$X_1 = \mathrm{N}(0,1)$$

$$X_2 = \mathrm{N}(0,1)$$

$$X_3 = 0.5X_1 + 0.5X_2 + 0.5\varepsilon_3$$

$$X_4 = 0.5X_3 + 0.707\varepsilon_4$$

$$X_5 = 0.5X_3 + 0.707\varepsilon_5$$

$$\mathrm{Cov}(X_1,X_2) = \mathrm{Cov}(X_1,\varepsilon_3) = \mathrm{Cov}(X_1,\varepsilon_4) = \mathrm{Cov}(X_1,\varepsilon_5) = \mathrm{Cov}(X_2,\varepsilon_3) = \mathrm{Cov}(X_2,\varepsilon_4) = \mathrm{Cov}(X_2,\varepsilon_5) = \mathrm{Cov}(\varepsilon_3,\varepsilon_4) = \mathrm{Cov}(\varepsilon_3,\varepsilon_5) = 0$$

$$\mathrm{Cov}(\varepsilon_4,\varepsilon_5) = 0.5$$

Here are the maximum likelihood estimates with their asymptotic standard errors in parentheses, based on the true model:

$$X_1 = \underset{(0.132)}{0.931}\ \mathrm{N}(0,1)$$

$$X_2 = \underset{(0.156)}{1.094}\ \mathrm{N}(0,1)$$

$$X_3 = \underset{(0.076)}{0.565}\ X_1 + \underset{(0.070)}{0.608}\ X_2 + \underset{(0.075)}{0.531}\ E_3$$

Table 4.3. *The observed covariance matrix for the 100 independent observations, along with the predicted maximum likelihood covariance matrices based on the correct model (Figure 4.1) and the incorrect model (Figure 4.3)*

	X_1	X_2	X_3	X_4	X_5
Observed covariance matrix					
X_1	0.931				
X_2	0.171	1.094			
X_3	0.630	0.762	1.350		
X_4	0.384	0.368	0.743	1.265	
X_5	0.324	0.385	0.611	0.624	0.949
Predicted values using the correct model					
X_1	0.931				
X_2	0.000	1.094			
X_3	0.526	0.665	1.233		
X_4	0.289	0.366	0.678	1.229	
X_5	0.238	0.301	0.557	0.594	0.925
Predicted values using the incorrect model					
X_1	0.931				
X_2	0.000	1.094			
X_3	0.526	0.666	1.234		
X_4	0.290	0.366	0.679	1.230	
X_5	0.000	0.385	0.234	0.129	0.949

$$\begin{array}{l} X_4 = 0.550\ X_3 + 0.856\ E_4 \\ \qquad (0.084) \qquad (0.122) \end{array}$$

$$\begin{array}{l} X_5 = 0.452\ X_3 + 0.673\ E_5 \\ \qquad (0.074) \qquad (0.096) \end{array}$$

$$\begin{array}{l} \text{Covariance}(X_4, X_5) = 0.287 \\ \qquad\qquad\qquad\qquad (0.082) \end{array}$$

Notice that each estimate is close to the population value. The standard errors are asymptotic, not exact, but with 100 observations these are quite close to the actual sample standard errors and so two times each value defines an approximate 95% confidence interval. For instance, the path coefficient from X_1 to X_3 is 0.565 with a standard error of 0.076 so an

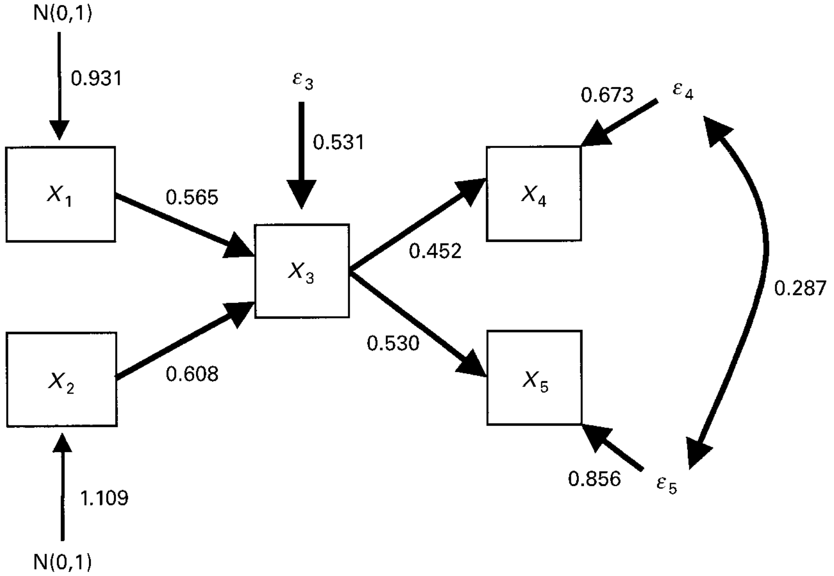


Figure 4.4. The fully parameterised path model of Figure 4.1. The numerical values are the maximum likelihood values of the free parameters based on centred, but not standardised, variables.

approximate 95% confidence interval would be 0.565 ± 2(0.076) or between 0.413 and 0.717; the true population value was 0.5. We could obtain the maximum likelihood estimates for the incorrect model as well, but since we already know that the data are very unlikely to have been generated by this incorrect model at least some of the estimates will be incorrect.

We can place the estimates of the free parameters of the correct model directly on the path diagram (Figure 4.4). I prefer this because the path diagram makes explicit that these estimates are based on a causal model with asymmetric relationships. The estimates shown in Figure 4.4 are not the ones Sewall Wright would have used. First, his estimates were not based on maximum likelihood methods but rather on least squares methods. Second, he used standardised variables so that the decomposition of direct and indirect effects were based on correlations rather than covariances. If the causal model is correct then the maximum likelihood and least squares estimates will be the same[12], since least squares (partial) regression coefficients are also maximum likelihood estimates, but if the causal model is wrong then the two types of estimate will differ. The standardised estimates are easily obtained by first standardising the variables to zero mean and unit var-

[12] This assumes, of course, that the data are multivariate normal.

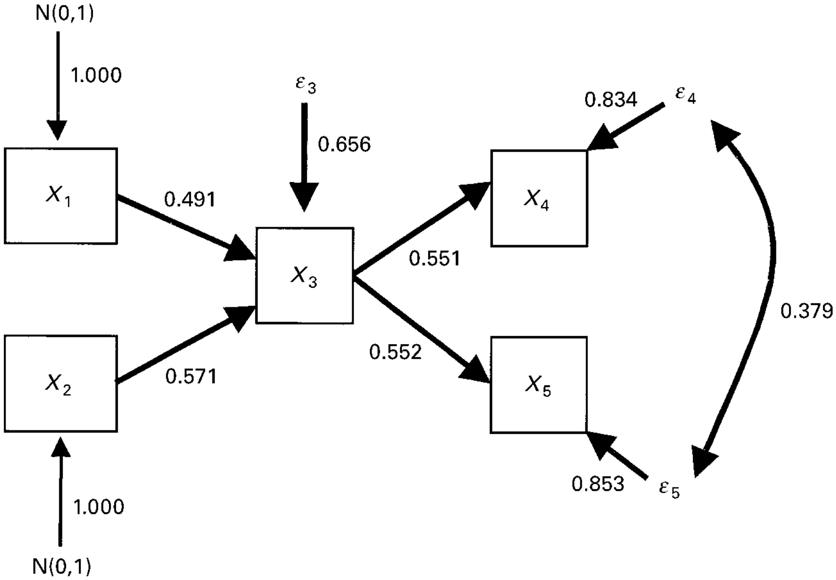


Figure 4.5. The fully parameterised path model of Figure 4.1. The numerical values are the maximum likelihood values of the free parameters based on centred and standardised variables. Units are therefore standard deviations from the mean.

iance. In fact, most SEM programs print these standardised estimates out. Figure 4.5 shows the path diagram for the correct model based on standardised variables.

4.2 Decomposing effects in path diagrams

One important use of path diagrams is to 'decompose' an association between variables into different types of causal relationship. In fact, this was the main goal of Wright's original method of path coefficients. Remembering the notions of causal graphs that were introduced in Chapter 2, we can differentiate between types of effect: direct causal effects, indirect causal effects, effects due to shared causal ancestors and unknown causal relationships[13]. One way of visualising this classification of associations is shown in Figure 4.6.

This decomposition of a statistical association into different types of causal relationship is based on the fundamental association linking causality with probability distributions, as described in Chapters 1 and 2. The overall

[13] Of course, there can always be associations due to random sampling fluctuations.

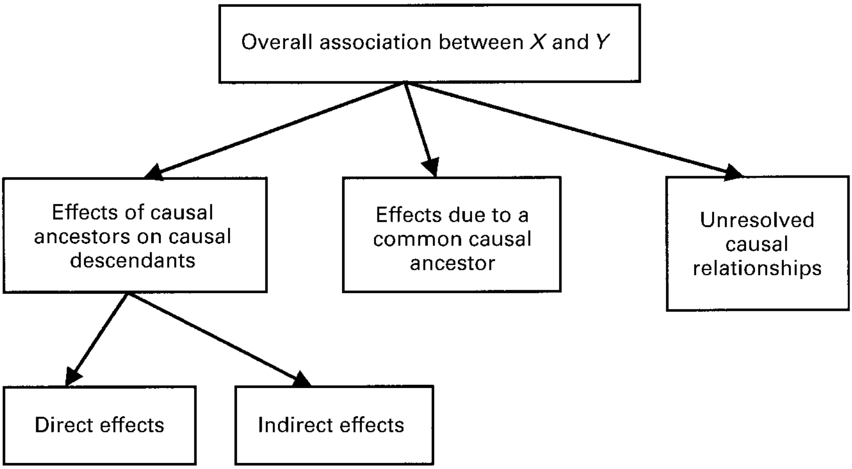

Overall association between X and Y
Effects of causal ancestors on causal descendants
Effects due to a common causal ancestor
Unresolved causal relationships
Direct effects
Indirect effects

Figure 4.6. A classification of associations (for example, correlations or covariances).

association between two variables is simply the overall correlation or covariance between them. This overall association can be generated by a number of different causal relationships at the same time. Since the consequences of interventions or manipulations will depend critically on these different types of relationship it is important to be able to distinguish and quantify them.

Ancestor–descendant relationships

If you can trace a path on the causal graph from a causal ancestor to a descendant by following the direction of the arrows then this path defines an effect from the ancestor to its descendant. These effects can be of two different types. A direct effect is an effect of the ancestor on its descendant that is not transmitted through any other variable in the model; of necessity this means that the relationship is one of parent and child. In other words, it is the effect that would occur if all other variables in the model did not change[14]. The magnitude of this direct effect is measured by the path coefficient on the arrow going from the parent to the child. The units of this effect are the same as those used to measure the variables. If the variables are not standardised then the path coefficient measures the number of unit changes in the child per unit change in the parent. For instance, if X is measured in grams, Y is measured in millimetres, there is an arrow from X to Y ($X \rightarrow Y$) and the path coefficient for this arrow equals 0.6 then this means that a 1 g change

[14] This may also be true even if other variables do change, as described below.

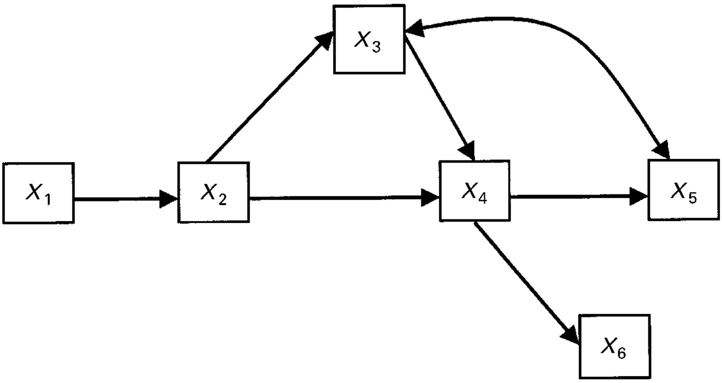


Figure 4.7. A path diagram used to illustrate the decomposition of associations.

in X will provoke a 0.6mm change in Y once X is d-separated from all other causes in the model. It is also the quantitative effect of X on Y when all other variables are held constant. If the variables are standardised, then the units are standard deviations from the mean. As usual, these points are much easier to grasp when looking at a causal graph. Consider Figure 4.7.

In Figure 4.7 there are six different direct effects. There are always as many direct effects as there are one-headed arrows in the path diagram. If we were to fit this model to data then the path coefficient from X_1 to X_2 would measure the direct effect of X_1 on X_2. If the three other variables were held constant then this direct effect would quantify by how much X_2 would change given a one unit change in X_1. However, since X_2 has no other causal ancestors then this direct effect would also quantify by how much X_2 would change given a one unit change in X_1 even if the other variables were not held constant.

Indirect effects are the effects of a causal ancestor on its descendant that are completely transmitted through some other variable. This intervening variable is sometimes called a *mediator* of the causal effect. For instance the effect of X_1 on X_3 along the path $X_1 \rightarrow X_2 \rightarrow X_3$ is an indirect effect of X_1 that is mediated by X_2. To quantify this effect one *multiplies* the path coefficients along this path. This indirect effect measures by how much X_3 would change following a change in X_1 if all causal parents of X_3 except for X_2 were held constant. In general an indirect effect measures by how much the effect variable would change following a change in the indirect cause when this effect is transmitted only along the path in question. It is possible for the same causal variable to exert both a direct and an indirect effect on the same descendent. An example of this is the effect that X_2 has on X_4 in

Figure 4.7; X_2 had a direct effect on X_4, since X_2 is the causal parent of X_4, but X_2 also exerts an indirect effect on X_4 through its effect on X_3.

Both direct and indirect effects involve variables in which one is a causal ancestor of the other. In these cases there is a directed path from one variable to the other. The third way in which an association can be decomposed in a path diagram is when the association between the two variables is due to another variable that is a causal ancestor of both. In Figure 4.7 the association between X_5 and X_6 is due to the effect of X_4 (their common ancestor) on both. To quantify this effect one *multiplies* the path coefficients along the path[15] from X_4 to X_5 and along the path from X_4 to X_6. Such effects do not measure any causal effect of one variable on the other and represent what Pearson might have called a 'spurious' association.

Finally, path diagrams can include unresolved causal relationships between variables; these are shown by double-headed arrows. Including such an effect in the model is an admission of ignorance; we do not know which is the cause, which is the effect, or whether the association is due to a common cause that is not included in the model. Such unresolved effects are quantified simply by the covariance between the two variables[16]. In tracing indirect effects along paths that include such double-headed arrows one can go in either direction but can traverse the double-headed arrow only once. Table 4.4 summarises the rules for decomposing the overall covariance or correlation between two variables in the path model and Table 4.5 lists the decomposition of Figure 4.7.

4.3 Multiple regression expressed as a path model

Since path analysis looks rather similar to multiple regression, let's look at how to represent a multiple regression as a path model. A multiple regression equation uses a series of predictor variables (say, X_1, X_2 and X_3) to predict, or account for, the observed variation in the dependent variable Y. The predictor variables are often called the 'independent' variables but this term can be misleading, since they do not have to be independent of one another at all. Except when these predictor variables are measured in controlled experiments, they are often not independent of one another. Figure 4.8 shows such a multiple regression in the form of a path model.

It is clear from the path diagram that the partial regression coefficients that are estimated with multiple regression are the direct effects of each predictor on the dependent variable. The indirect effects, which are

[15] The two paths together ($X_5 \leftarrow X_4 \rightarrow X_6$) are sometimes called a *trek*.

[16] Or a correlation coefficient if the variables are standardised, since a correlation is simply a standardised covariance.

Table 4.4. *Given two variables (X and Y) in a path model, the overall covariance or correlation (if using standardised variables) between them can be decomposed into three different causal sources. Shown are the rules of the estimation of each source*

Effect involving variables *X* and *Y*	Rule for its estimation
Direct effect	Value of the path coefficient on the arrow from *X* to *Y*
Indirect effect along a single path	Product of the path coefficients on the sequence of arrows along the path leading from *X*, through at least one intermediate variable, and into *Y*
Overall indirect effect along all paths	Sum of the indirect effects along all paths from *X* to *Y*.
Effect due to common causal ancestor (*Z*) of both *X* and *Y*	Multiply the path coefficients along a single path from *Z* to *Y* and the path coefficients along a single path from *Z* on *X* (called a *trek*). If there is more than one such trek linking *X* and *Y* due to common causes, sum these together
Effect due to unresolved causal relationship	Path coefficient on the double-headed arrow between *X* and *Y*
Effect due to all common causal ancestors of both *X* and *Y*	Sum together the effects due to each common causal ancestor of both *X* and *Y*
Overall effect	Sum together, the direct effect, the total indirect effects, the total effects due to common causal ancestors and any remaining unresolved causal relationship between *X* and *Y*. This will equal the covariance or correlation (if using standardised variables) between *X* and *Y*.

simply the unresolved causal relationships between the predictors, are ignored. If the free covariances, (the s_{ij} in Figure 4.8) really are zero, then the direct effects will also be the overall effects, but the regression equation will not tell you this[17]. Furthermore, the model in Figure 4.8 can't be tested as a causal claim. There are only four observed variables in this model and therefore there are $4(5)/2 = 10$ unique elements in the covariance matrix. There are also 10 free parameters that have to be estimated (the three path coefficients, the three free covariances, the error variance and the variances

[17] If these covariances are not zero then one can run into problems of collinearity.

Table 4.5. *Decomposition of the total association between each pair of variables in Figure 4.7 into direct effects, indirect effects, effects due to common causal ancestors and pure unresolved causal effects*

Variable pair	Direct	Indirect	Common causal ancestor	Unresolved causal relationship
X_1, X_2	$X_1 \rightarrow X_2$	None	None	None
X_1, X_3	None	(1) $X_1 \rightarrow X_2 \rightarrow X_3$ (2) $X_1 \rightarrow X_2 \rightarrow X_4 \rightarrow X_5 \leftrightarrow X_3$	None	None
X_1, X_4	None	(1) $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ (2) $X_1 \rightarrow X_2 \rightarrow X_4$	None	None
X_1, X_5	None	(1) $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5$ (2) $X_1 \rightarrow X_2 \rightarrow X_3 \leftrightarrow X_5$ (3) $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ $X_1 \rightarrow X_2 \rightarrow X_4 \rightarrow X_5$	None	None
X_1, X_6	None	(1) $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_6$ (2) $X_1 \rightarrow X_2 \rightarrow X_4 \rightarrow X_6$	None	None
X_2, X_3	$X_2 \rightarrow X_3$	(1) $X_2 \rightarrow X_4 \rightarrow X_5 \leftrightarrow X_3$	None	None
X_2, X_4	$X_2 \rightarrow X_4$	(1) $X_2 \rightarrow X_3 \rightarrow X_4$	None	None
X_2, X_5	None	(1) $X_2 \rightarrow X_3 \rightarrow X_3 \leftrightarrow X_5$ (2) $X_2 \rightarrow X_4 \rightarrow X_5$	None	none
X_2, X_6	None	(1) $X_2 \rightarrow X_3 \rightarrow X_6$ (2) $X_2 \rightarrow X_4 \rightarrow X_6$	None	None
X_3, X_4	$X_3 \rightarrow X_4$	None	(1) $X_4 \leftarrow X_2 \rightarrow X_4$	None

Table 4.5. (*cont.*)

Variable pair	Direct	Indirect	Common causal ancestor	Unresolved causal relationship
X_3, X_5	None	(1) $X_3 \rightarrow X_4 \rightarrow X_5$	None	$X_3 \leftrightarrow X_5$
X_3, X_6	None	(1) $X_3 \rightarrow X_4 \rightarrow X_6$	(1) $X_3 \leftarrow X_2 \rightarrow X_4 \rightarrow X_6$	None
X_4, X_5	$X_4 \rightarrow X_5$	None	(1) $X_4 \rightarrow X_3 \leftrightarrow X_5$	None
X_4, X_6	$X_4 \rightarrow X_6$	None	None	None
X_5, X_6	None	None	$X_5 \leftarrow X_4 \rightarrow X_6$	None

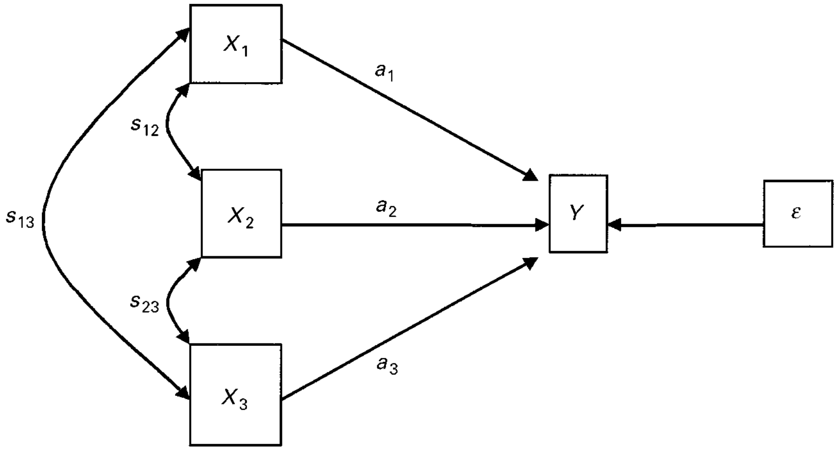


Figure 4.8. A multiple regression of X_1, X_2 and X_3 on Y, expressed as a path diagram.

of the three predictor variables). In other words, we have used up all of our degrees of freedom in estimating our free parameters and have none left over to test the causal implications of the model[18]. If you use the inferential test described in Chapter 3 you will find that no variable is d-separated from any other variable, either unconditionally or after conditioning on any set of other observed variables. The regression model places no statistical constraints with which to test the causal implications of the model. Multiple

[18] When we get to the topic of identification, we will see that multiple regression is an example of a just-identified model.

regression can certainly be used to decide whether the path coefficients (i.e. the partial regression coefficients) are different from zero. Multiple regression can help us to decide whether the error (or residual) variance is less than the total variance of Y (this is the F ratio). Multiple regression can't help us to decide whether the causal assumptions of the model are correct; it can't tell us whether the predictor variables are causes of Y. Multiple regression can allow us to predict but not to explain. Statistics texts are quite correct when they say that one can't draw causal conclusions from regression. The causal conclusions must come from somewhere else. The best way would be to conduct a controlled randomised experiment in which the values of the X variables are randomly assigned to the experimental units, since we would then have good reason to assume that the free covariances between them really are zero. If this is not possible then we have to construct our models, and collect our observations, in such a way that we can constrain the patterns of covariation based on our causal hypothesis and then test these constraints.

4.4 Maximum likelihood estimation of the gas-exchange model

In Chapter 3 we looked at the model of Shipley and Lechowicz involving specific leaf mass (SLM), leaf nitrogen concentration, stomatal conductance, net photosynthetic rate and the internal concentration of CO_2. Let's fit and test these same data (ln-transformed) to the proposed path model using maximum likelihood methods. Remember that 5 of the 40 species were actually C_4 species and that these were clear outliers in the data set. Because we require approximate multivariate normality, we can't include these 5 species in the data set. The analysis will be restricted to the remaining 35 species. Since the resulting chi-squared statistic and the standard errors of the free parameters are only asymptotically correct, we can expect that the estimated standard errors are somewhat narrower than they should be and the probability value of the chi-squared statistic will not be exact[19]. This model is reproduced in Figure 4.9.

The first step is to specify the structural equations and to indicate which parameters are free. There are five free path coefficients (a_1 to a_5) and five free variances (the variance of specific leaf mass and the four error variances ε_2 to ε_5). Since there are five measured variables there will be five

[19] Chapter 6 describes the effects of sample size on the maximum likelihood chi-squared statistic.

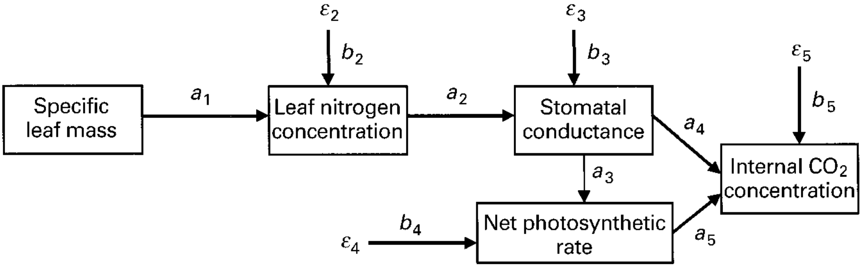


Figure 4.9. The proposed path model relating leaf morphology and leaf gas exchange. The letters with subscripts show the free parameters whose maximum likelihood estimates must be obtained.

degrees of freedom[20]. The free parameters are shown in Figure 4.9. Next, I have to specify initial values for these free parameters. In my experience one rarely has problems with convergence of the maximum likelihood estimates when there are no latent variables in the model, unless parts of the model are underidentified, and so I will make all free parameters equal to 1 except a_5, which I set at -1. I do this because I expect increasing photosynthetic rates to reduce the internal CO_2 concentration. One can sometimes have problems with convergence if some variances are much larger (i.e. orders of magnitude) than others, but I know that this is not the case with these variables.

The value of the chi-squared statistic, based on the initial values of the free parameters, was 150.96 – obviously a very poor fit. The numerical algorithm searched for changes in these initial values that would improve the fit while respecting the constraints and came up with a second set of values. The value of the chi-squared statistic, based on this second set of values of the free parameters, was 65.96 – obviously still a very poor fit but at least much better. Again the estimates of the free parameters were adjusted and after the third try the chi-squared statistic was 19.72. This process was repeated a forth, and then a fifth time, giving a chi-squared statistic of 4.72. The sixth attempt made such a small improvement (from 4.71954 to 4.71648) that the algorithm stopped; it had reached the valley floor. The final maximum likelihood chi-squared value was therefore 4.72 and, with 5 degrees of freedom, the asymptotic probability under the null hypothesis was 0.45. At this point we can obtain the estimates of the free parameters and their asymptotic standard errors. These estimates, divided by their asymptotic standard errors, can be used to test whether they are significantly

[20] $\frac{5(6)}{2} - 10 = 5$

different from zero, using a *z*-test. Remember, because we only have 35 observations, such a *z*-test will be somewhat liberal, since the real standard errors will be a bit larger that their asymptotic estimates. Here are the maximum likelihood estimates. The asymptotic standard errors are given in round brackets and the *z*-value (whose absolute value will be less that 1.96 95% of the time) is given in square brackets.

$$\ln(\mathrm{SLM}) = \mathrm{N}\begin{pmatrix} 0.183 \\ 0,(0.044) \\ [4.123] \end{pmatrix}$$

$$\ln(\mathrm{Nitrogen}) = \begin{array}{c} 0.898\ln(\mathrm{SLM}) \\ (0.096) \\ [9.338] \end{array} + \mathrm{N}\begin{pmatrix} 0.057 \\ 0,(0.014) \\ [4.123] \end{pmatrix}$$

$$\ln(\mathrm{Conductance}) = \begin{array}{c} 1.146\ln(\mathrm{Nit}) \\ (0.206) \\ [5.525] \end{array} + \mathrm{N}\begin{pmatrix} 0.300 \\ 0,(0.070) \\ [4.123] \end{pmatrix}$$

$$\ln(\mathrm{Photo}) = \begin{array}{c} 0.548\ln(\mathrm{Cond}) \\ (0.069) \\ [7.977] \end{array} + \mathrm{N}\begin{pmatrix} 0.091 \\ 0,(0.022) \\ [4.123] \end{pmatrix}$$

$$\ln(\mathrm{CO_2}) = \begin{array}{c} -0.162\ln(\mathrm{Photo}) \\ (0.020) \\ [-8.133] \end{array} + \begin{array}{c} 0.142\ln(\mathrm{Cond}) \\ (0.013) \\ [10.526] \end{array} + \mathrm{N}\begin{pmatrix} 0.001 \\ 0,(0.0002) \\ [4.123] \end{pmatrix}$$

The probability associated with the maximum likelihood chi-squared statistic (0.45) tells us that the data are consistent with the constraints that our model has placed on them. Given the small sample size we know that this probability estimate is not exact but is far from being significant. We already knew this based on the d-sep test (Chapter 3), whose probability estimates are exact. Although I discuss tests for multivariate normality in Chapter 6, one such test is due to Mardia (1970, 1974). The normalised version of Mardia's coefficient is asymptotically distributed as a standard normal variate, although this requires very large sample sizes. At least it gives a rough guide to deviations from multivariate normality and the value of this coefficient in the present data set, 0.174, is much smaller than the 1.96 needed for significant non-normality at the 0.05 level. These points tell us that the causal structure that was proposed is consistent with the data. We conclude that there is no good reason to reject the model and so we can go on to look at the maximum likelihood estimates for the free parameters.

The first thing to notice is that the *z*-scores, associated with each of the 10 free parameters, are all much larger than the 1.96 value that indicates significance at the 0.05 level. In other words, each of the free parameters is significantly different from zero. Since d-separation predicts that each of the five free path coefficients must be different from zero, this is good news. A path coefficient that is not significantly different from zero doesn't necessarily mean that it really is zero. It is always possible that the true value is close enough to zero that we can't detect the difference; this is the well-known problem of statistical power. None the less, a path coefficient that is very close to zero requires that we be able to provide reasonable doubt that it is not really zero as our model requires. The five variances (for specific leaf mass and the four error variables) are also significantly different from zero. The fact that the four error variances are not zero simply means that there are other unknown causes contributing to the variance of each of the dependent variables. We have no strong reason to suppose, however, that there are unknown variables that are common causes of two or more measured variables, since, if there were, the covariances between the error variables would not all be zero as specified by the model and the model X^2 value would be large. On the other hand, since we have a small data set, we have little statistical power to detect common unknown causes that are numerically weak.

We can obtain the sort of path model that Sewall Wright would have produced simply by first standardising each variable by subtracting its mean and dividing by its standard deviation. In this way, the variance of each variable is always 1.0. The residual variance of each variable is, as always, the variance of the error variable (ε, or the path coefficient from the error variable to the measured variable if the variance of the error variable is fixed at unity). This means that the explained variance is simply $1-\mathrm{Var}(\varepsilon)$ and the square root of this gives the Pearson correlation coefficient for the (multiple) regression, i.e. R. Here are the standardised equations:

$$\ln(\text{Nitrogen}) = 0.848\ln(\text{SLM}) + 0.530\varepsilon_1 \quad R = 0.845$$

$$\ln(\text{Conductance}) = 0.688\ln(\text{Nitrogen}) + 0.726\varepsilon_2 \quad R = 0.688$$

$$\ln(\text{Photo}) = 0.807\ln(\text{Conductance}) + 0.590\varepsilon_3 \quad R = 0.807$$

$$\ln(\text{CO}_2) = -1.144\ln(\text{Photo}) + 1.480\ln(\text{Conductance}) + 0.484\varepsilon_4 \\ R = 0.875$$

Now that we have the maximum likelihood estimates for the free parameters, we can estimate the effect of the variables on each other along different paths of influence. These effects are the amount by which the variable at the end of the path with change (in natural logarithmic units since they are transformed) after a one unit change in the variable at the beginning of the path, when all variables not involved in the path are held

Table 4.6. *Predicted changes, in units of natural logarithms, of the variable at the end of the path after a one unit change in the variable at the beginning of the path, when holding constant all variables not involved in the path*

Path	Effect along path
SLM→Nitrogen	0.898
SLM→Nitrogen→Conductance	1.029
SLM→Nitrogen→Conductance→Photosynthesis	0.564
SLM→Nitrogen→Conductance→Internal CO_2	0.146
SLM→Nitrogen→Conductance→Photosynthesis→Internal CO_2	−0.091
Nitrogen→Conductance	1.146
Nitrogen→Conductance→Photosynthesis	0.628
Nitrogen→Conductance→Internal CO_2	0.167
Nitrogen→Conductance→Photosynthesis→Internal CO_2	−0.102
Conductance→Internal CO_2	0.142
Conductance→Photosynthesis	0.548
Conductance→Photosynthesis→Internal CO_2	−0.089
Photosynthesis→Internal CO_2	−0.162

Note:
SLM, specific leaf mass.

constant. Table 4.6 summarises these effects based on the non-standardised variables.

Using the values in Table 4.6 we can see how the overall effects are decomposed into direct effects, indirect effects and the effects of common causal ancestors. For instance, the overall effect of a one ln-unit increase in stomatal conductance on the internal CO_2 concentration was to increase it by only 0.053 ln-units. However, the direct effect was 0.142 units. The small overall effect was due to the fact that stomatal conductance also had an indirect negative effect on the internal CO_2 concentration. Increasing stomatal conductance increased the net photosynthetic rate by 0.548 units and increasing photosynthetic rate decreased the internal CO_2 concentration by -0.162. The indirect effect of stomatal conductance on internal CO_2 concentration through the mediating effect of net photosynthesis was $0.548 \times (-0.162) \approx -0.089$. The overall effect was therefore $0.142 - 0.089 = 0.053$ ln-units.

The importance of decomposing effects can be seen more clearly when considering the standardised path coefficients. The overall effect (thus the overall predicted correlation) between stomatal conductance and internal CO_2 concentration was 0.557. This was due to a direct effect of 1.48 plus an indirect effect of $0.807 \times -1.144 \approx -0.923$. The overall effect (thus the overall predicted correlation) between net photosynthetic rate and internal CO_2 was only $(1.48 \times 0.81) - 1.14 \approx 0.06$, which is not even significantly different from zero at the 0.78 level. Clearly, it would be biologically absurd to suggest that the photosynthetic rate, which is removing CO_2 from the internal air spaces of the leaf, does not affect the internal concentration of CO_2. The apparent contradiction is resolved by decomposing this overall effect. The direct standardised effect of net photosynthetic rate on internal CO_2 was -1.144. However, both net photosynthetic rate and the internal CO_2 concentration have a common causal ancestor: the stomatal conductance. The standardised effect of the path: net photosynthesis←stomatal conductance→internal CO_2 was $0.807 \times 1.480 \approx 1.194$. The overall correlation between net photosynthesis and internal CO_2 was therefore $-1.144 + 1.194 = 0.050$. In other words, the direct effect and the effect of the common causal ancestor almost cancelled each other out. This is just what the physiological model of stomatal regulation of Cowan and Farquhar (1977) would predict (see Chapter 3).

5 Measurement error and latent variables

Ambient temperature affects the metabolic rate of animals. When it is cold a homeothermic animal has to burn stored energy reserves, first glycogen and fat and then (when these are exhausted) protein, in order to generate heat and maintain its body temperature. The scaling of surface area (the site of heat loss to the atmosphere) to body volume (where the heat is generated) means that small homeothermic animals such as song birds can lose up to 15% of their body fat in one cold night. To burn this fat the bird must increase its metabolic rate, which increases its O_2 consumption. Imagine that we conduct an experiment in which we place small birds inside metabolic chambers overnight and vary the air temperature. The hypothesised causal process is shown in Figure 5.1.

Unfortunately, we can't directly measure any of these three variables; they are unmeasured, or *latent*, and so I have enclosed them in circles following the conventions of path diagrams. If we measure the air temperature using a thermometer then we aren't directly measuring *temperature* – the average kinetic energy of the molecules in the air. Instead we are measuring the height of a column of mercury enclosed in a hollow glass tube. In fact, we can't even measure the actual height of the mercury exactly, since our observed height will include some measurement error. Nor can we directly measure metabolic rate. Typically, one measures the rate of gas exchange (O_2 decrease or CO_2 increase) between the air entering and leaving the metabolic chamber. If we measure oxygen consumption using an infrared gas analyser then we aren't even directly measuring *oxygen* consumption. Instead we are measuring differences in the amount of light of particular wavelengths that is absorbed as the light passes through the air. Again, even this variable is not perfectly measured, since the observed values will also contain measurement error. When we measure the fat reserves that are burned by the birds we might actually be measuring the difference in body weight over the course of the experiment and this too will include measurement error. One simplified representation of the actual causal process[1] is depicted in Figure 5.2.

[1] Some readers will recognise this as a particular example of the claim in Chapter 2 that

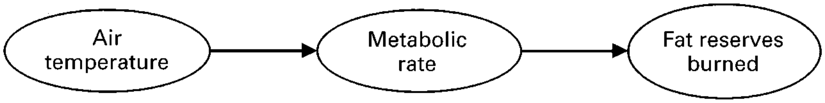


Figure 5.1. The causal structure relating air temperature, the metabolic rate of the bird and the amount of fat reserves that are used up.

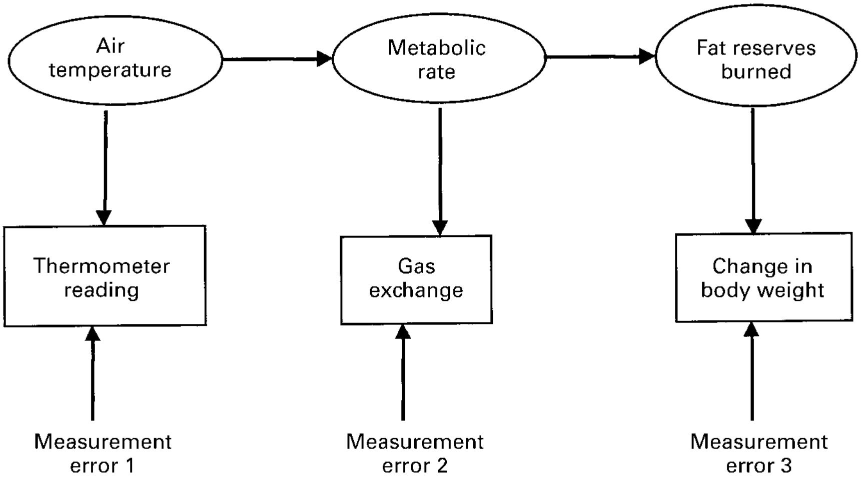


Figure 5.2. The causal structure relating air temperature, the metabolic rate of the bird and the amount of fat reserves that are used up when we conduct an experiment in the metabolic chamber and take measurements.

In this causal scenario the variables that we can observe and measure are not the variables that we hypothesise to form the causal chain of interest. The causal process involves variables that we cannot directly observe. We can obtain information about these latent variables by observing other variables but the observed variables are also affected by other independent causes that generate measurement errors. What are the consequences of this for path analysis? What are the consequences of this for causal analysis in general?

We know, based on d-separation, that if we could hold constant the actual metabolic rate of our birds then changes in the temperature of the ambient air would be independent of the changes in their fat reserves if the causal hypothesis were correct. Therefore the partial correlation of the unmeasured variables 'air temperature' and 'change in fat reserves' would be

even the simplest scientific hypotheses have, hidden behind them, a whole constellation of auxiliary hypotheses.

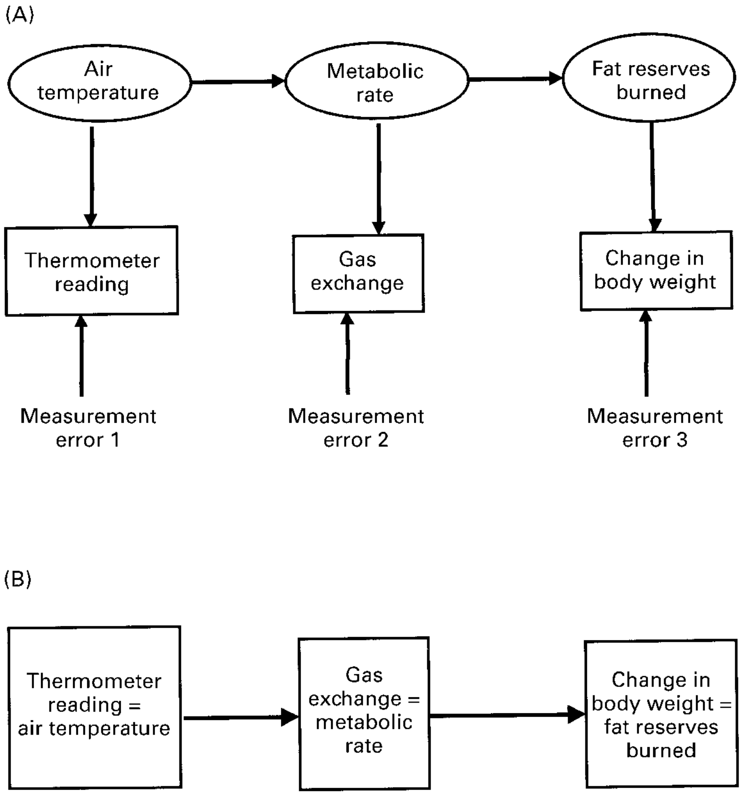


Figure 5.3. The first directed graph (A) shows the actual causal structure of the experiment and the second directed graph (B) shows the causal structure that is assumed when ignoring measurement error.

zero when conditioned on the unmeasured ‘metabolic rate’. This would be true whether we used experimental controls or statistical controls. However, we can’t observe the actual metabolic rate; we can only infer its constancy based on the observed rate of gas exchange. If gas exchange is perfectly correlated with metabolic rate then holding constant the rate of gas exchange would ensure that metabolic rate was also constant. What happens if the correlations between our measured variables and the variables of theoretical interest are not perfect?

5.1 Measurement error and the inferential tests

Figure 5.3A shows the causal scenario as we have conceived it, and Figure 5.3B shows it as it would look if we were willing to ignore the fact that our measured variables are imperfect measures of the causally important variables. Looking at Figure 5.3A, we see that there is only one conditional independence relationship in the basis set involving the three latent variables of interest; namely, 'air temperature' is independent of 'change in fat reserves' after conditioning on 'metabolic rate'. There are other conditional independence relationships that involve the observed variables but these observed variables don't interest us except in how they can test the underlying causal hypothesis involving the latents. Notice that 'thermometer reading' (one of our imperfect measures) is also independent of 'change in body weight' (another imperfect measure) upon conditioning on 'metabolic rate' as well. In other words, these two measured variables have the same causal implications as those involving their underlying latent variables and so the fact that these two measured variables are not perfect indicators of their latents makes no difference to our ability to test the causal hypothesis. However these two observed variables are not d-separated upon conditioning on 'gas exchange' in Figure 5.3A (the observed variable indicating metabolic rate) even though they are conditionally independent in Figure 5.3B which ignores measurement error. Therefore measurement error in the conditioning variable will introduce errors into our test of the underlying causal hypothesis. The error is not in the logic of the inferential test but in our incorrect assumption that our measure of gas exchange is a perfect indicator of metabolic rate.

We can conduct a numerical simulation based on the structural equations derived from the causal graph in order to see what happens when we include measurement error. I will represent the three latent variables (air temperature, metabolic rate and change in fat reserves) by X, Y and Z and the three corresponding observed variables by X', Y' and Z'. Here is one set of structural equations corresponding to the causal process:

$$X = \mathrm{N}(0,1)$$

$$X' = 1X + \mathrm{N}(0,\sigma_1)$$

$$Y = 2X + \mathrm{N}(0,2)$$

$$Y' = 1Y + \mathrm{N}(0,\sigma_2)$$

$$Z = 0.5Y + \mathrm{N}(0,\sqrt{0.75})$$

$$Z' = 1Z + \mathrm{N}(0,\sigma_3)$$

Since we have generated our data in accordance with the causal graph in Figure 5.3A, we know that about 5% of our simulated data sets would produce a probability of less that 0.05 when examined using the

Table 5.1. *The empirical rejection rates of 500 independent data sets, consisting of 1000 independent observations (X, X′,Y,Y′, Z, Z′) each, are shown. The true model is shown in Figure 5.3A with different amounts of measurement error and the inferential tests are based on the incorrect model in Figure 5.3B. The population variances of the latent X,Y and Z are 1, 8, and 4.75. The population variances of the observed X′,Y′, and Z′ are* $1+\sigma_1^2$, $8+\sigma_2^2$ *and* $4.75+\sigma_3^2$

Measurement error			Ratio of variances			d-sep rejection rate	ML rejection rate
σ_1^2	σ_2^2	σ_3^2	X/X'	Y/Y'	Z/Z'		
0.01	0.01	0.01	0.990	0.999	0.998	0.052	0.064
10.0	0.01	0.01	0.009	0.999	0.998	0.060	0.052
0.01	0.10	0.01	0.990	0.988	0.998	0.060	0.050
0.01	0.30	0.01	0.990	0.964	0.998	0.088	0.102
0.01	0.40	0.01	0.990	0.952	0.998	0.152	0.144
0.01	0.01	10.00	0.009	0.998	0.322	0.052	0.060

Note:
ML, maximum likelihood.

d-sep test. If we test 500 independent data sets, each with 1000 independent observations, then the 95% confidence interval for our empirical rejection rate would be between approximately 2% and 8% (Manly 1997). Table 5.1 summarises the results of these simulations as we increase the measurement errors. We see that even when the measurement error variance of Y' is 0.3, i.e. slightly less than 4% of the true variance of Y, the rejection rate is outside the 95% confidence limits. As the measurement error variance increases further, the rejection rate increases rapidly. In other words, even if the hypothesised causal process involving the theoretical variables were correct, we would tend to reject it too often if we were to incorrectly assume that our conditioning variable is measured without error. Here, ignoring measurement error increases the likelihood that one will incorrectly reject a model that is correct. The effect of measurement error on the accuracy of the probabilities associated with the maximum likelihood chi-squared statistic is the same in this example (Table 5.1).

5.2 Measurement error and the estimation of path coefficients

The effect of measurement error on the accuracy of the estimation of the path coefficients, using either the least squares regression methods of Chapter 3 or the maximum likelihood regression methods of Chapter 4, are perhaps somewhat better known to biologists. Let's first look at a simple example involving only two variables (X and Y) that are measured with error. When I say 'measured with error' I don't mean only the obvious case in which the measuring device (say an analytical balance) has a certain degree of error.

Imagine that you wish to measure the nitrate availability of the soil in the rooting zone of a plant, but only measure the total nitrogen content of samples of this soil at a single time. In such a case the error of measurement will include, not only the error involved in the analytical method for nitrate concentration, but also the error involved in using the sample measures of total nitrogen at one point in time as a proxy variable for the total nitrogen availability in the rooting zone. Let us imagine a causal process in which the nitrate absorption rate of a plant (Y) is caused by the amount of nitrate available in the rhizosphere of its roots (X). X is estimated as the average nitrate availability of a sample of soil cores taken directly beneath the plant at one point in time. Y is estimated as the change in the net total nitrogen concentration of the plant from the time the soil is sampled until the next day. Figure 5.4 shows the causal graph assuming measurement error (Figure 5.4A) and without measurement error (Figure 5.4B).

The path coefficient shown as '*a*' in Figure 5.4 is the regression slope of Y on X. By definition, this is:

$$a=\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}=\frac{a\mathrm{Var}(X)}{\mathrm{Var}(X)}$$

Now, the true value of a' in Figure 5.4B can be derived from the rules of path analysis:

$$a'=\frac{\mathrm{Cov}(X',Y')}{\mathrm{Var}(X')}=\frac{a(b_1)(b_2)\mathrm{Var}(X)}{b_1^2\mathrm{Var}(X)+\mathrm{Var}(e_2)}$$

Often the measured variables (X' and Y') will scale 1 : 1 with the underlying latent variable; that is, a unit increase in the underlying variable will result in a unit increase or decrease of the measured variable. In such a case (or if $b_1=b_2$) the formula can be simplified to:

$$a'=\frac{a\mathrm{Var}(X)}{\mathrm{Var}(X)+\mathrm{Var}(e_2)}$$

From this we see that the effect of measurement error (e_2) is to decrease a' relative to a. If we ignore the measurement error then a' will be a biased estimate[2] of a. The formula also shows why it is important to sample

[2] This is true in this simple bivariate case. When there is more than one cause of Y, each with measurement error, then the relationship between a and a' will also depend on the covariances between the measurement errors. See Bollen (1989) for the exact formulae.

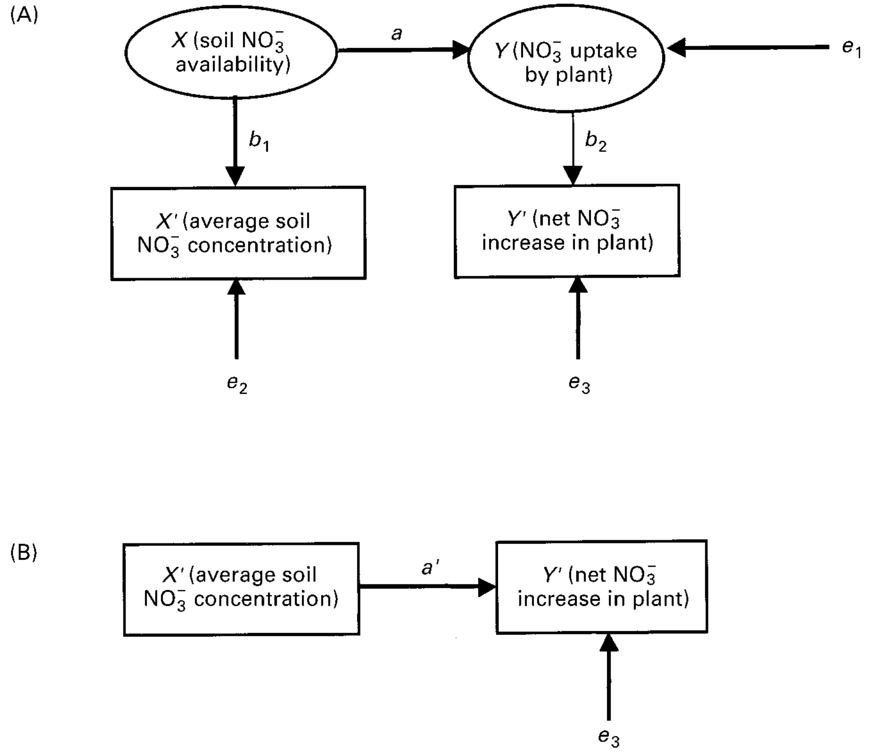

(A)
X (soil NO$_3^-$ availability)
a
Y (NO$_3^-$ uptake by plant)
e_1
b_1
b_2
X′ (average soil NO$_3^-$ concentration)
Y′ (net NO$_3^-$ increase in plant)
e_2
e_3
(B)
X′ (average soil NO$_3^-$ concentration)
a′
Y′ (net NO$_3^-$ increase in plant)
e_3

Figure 5.4. The first directed graph (A) shows the true causal structure in this scenario. The second directed graph (B) shows the causal structure that is assumed when ignoring measurement error.

in such a way as to allow the widest variation possible in the causal variable *X*. Presumably the measurement error will not change with the range of *X* and so as the variance of *X* increases the difference between *a*′ and *a* will decrease. Furthermore, measurement error in the effect variable (*Y*) has no effect on the bias of the path coefficient.

These measurement errors have different effects on the estimation of the path coefficients and on the probabilities of the overall inferential test of the causal model. For instance, in the causal model shown in Figure 5.3, measurement error in *X* (the air temperature) had no effect on the probability levels estimated in the inferential test when we ignored measurement error but would bias the path coefficient from air temperature to metabolic rate. Measurement error in *Y* (metabolic rate) did have an effect on the estimated probabilities but would not bias the path coefficient[3] from *X* to *Y*.

[3] Measurement error in *Y* would bias the estimation of the path coefficient from *Y* to *Z*, if we ignore it.

5.3 A measurement model

When measurement error can't be safely ignored, it must be explicitly included in the model and estimated. Most biologists deal with measurement error by ignoring it. Sometimes this is reasonable. After all, we are able to measure temperature, mass or CO_2 concentration with great accuracy. Sometimes ignoring measurement error is not at all reasonable. Trying to estimate the fat reserves of a large free-ranging mammal by palpating its ribs and giving it a score of 1 to 4 can hardly inspire confidence in its accuracy, yet this is a common measure of 'body condition'. Similarly, we do not possess the equivalent of a thermometer that we can put into the mouth of our animals to measure their evolutionary fitness. If we try to measure fitness using indirect measures then these measures will probably possess important measurement errors.

Although not generally known to most biologists, methods for dealing with measurement error have been developed in the social sciences. Almost all important variables are latent in these sciences and they can be measured only with substantial error. For instance, one might reasonably hypothesise that the degree to which a person can empathise with suffering might determine their career choice. It seems reasonable to suppose that a person with more empathy might choose to become a nurse rather than a mercenary soldier. Yet how can one measure 'empathy'? A common approach would be to devise a series of survey questions and develop an index of empathy based on the answers to these questions. It is obvious that choosing the answer 'A' in a multiple choice test does not cause one to become a nurse rather than a mercenary nor does it cause one to become more empathic. Rather, a psychologist might say that one's empathic tendency is a common latent cause both of the answers on the survey and of one's choice of career. The survey answers are imperfect measures, or indicators, of the underlying latent variable and the measurement model must separate those parts of the covariance between the answers that are due to the underlying latent cause from those parts of the covariance due to other causes. One simple type of measurement model is a factor model.

There is a huge literature devoted to measurement theory and to its many pitfalls. Many of these pitfalls are conceptual rather than statistical, so let's begin with an example (Dunn, Everitt and Pickles 1993) that doesn't pose any conceptual problems. You cut a number of pieces of string into different lengths and lay them on a table. Each string now has an attribute – length – that you ask four different people to measure. One person uses a ruler graduated in centimetres. One person uses a hand and measures in hand lengths. The third person uses a ruler graduated in inches and the

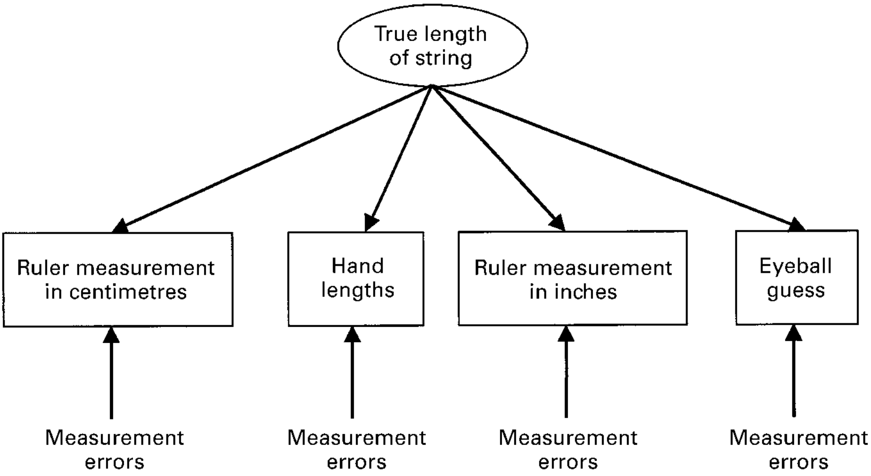


Figure 5.5. The causal structure generating the different measured lengths of the pieces of string.

fourth person simply looks at each string and tries to estimate it to the nearest centimetre. The length measurements have different units and each estimate has two different causes. The first cause is the true length of the string, since each person is trying to accurately measure this latent attribute. The other type consists of all those other causes that give rise to the measurement error (incorrectly calibrated rulers, tiredness, myopia . . .). Figure 5.5 shows the causal graph.

In this example everything is in plain sight. The true length of the string, although latent, is not hypothetical. We can see the strings on the table and know that each has a fixed value of the attribute 'length'. The only uncertainty is in knowing the actual length of each string. Let the j strings ($j=1,n$) measured by the four people be X_{j1} to X_{j4} and let the true length of each string be L_j. The structural equations representing this causal process are:

$$X_{j1}=\alpha_1 L_j+\mathrm{N}(0,\sigma_1)$$
$$X_{j2}=\alpha_2 L_j+\mathrm{N}(0,\sigma_2)$$
$$X_{j3}=\alpha_3 L_j+\mathrm{N}(0,\sigma_3)$$
$$X_{j4}=\alpha_4 L_j+\mathrm{N}(0,\sigma_4)$$

These structural equations, coupled with the path diagram in Figure 5.5, state that each person's measurement (X_{jk}) is a linear function of the true length of each string (L_j) plus a certain amount of other unknown causes whose variance is σ_k. The total variance in the kth person's measure of a given string is now separated into two parts: a part that is common to everyone (the *common variance*) that is due to the true lengths of the strings, and a part that is unique to each person (the *unique variance*, σ_k) that is due to those other causes of a particular measurement. Since these unique variances are d-separated in Figure 5.5, we know that they must be uncorrelated. Since there are four observed variables there are $4(5)/2=10$ unique covariances in the covariance matrix[4]. There are only nine free parameters that we have to estimate: the four path coefficients (the four α_i), the variances of the four measurement errors (the four σ_i^2) and the variance of the latent variable. We can therefore fit this model by maximising its likelihood and then test it using the maximum likelihood chi-squared test because we have 1 degree of freedom left. Before we do this, however, we have to overcome a problem of identification. Identification is a problem that is discussed in more detail in Chapter 6, after we have seen how to combine the measurement model with the structural model involving the latent variables, but it can be intuitively understood with the following example.

If you are given an equation, say $y=2x+z$, and are told only that x equals 1, then there is more that one combination of values for y and z that will solve this equation; in fact, there are an infinite number of such values. The equation is said to be *underidentified*. If you are told both that x equals 1 and that z equals 3, then there is only one value of y that is admissible: $y=5$. The equation is just *identified*[5].

In our current example with the string lengths the underidentification arises because we have to estimate both the path coefficients and the variance of the latent variable. We see in Figure 5.5 that d-separation predicts that the partial correlations (thus, the partial covariances) of each pair of observed variables must be zero when conditioned on the latent variable. Maximum likelihood estimation fits the data to the structural equations while respecting this constraint. The predicted covariance between the latent L and each observed X_j is given by $\mathrm{Cov}(L, X_j)=\alpha_i\mathrm{Var}(L)$. Since there is an infinite combination of α values and $\mathrm{Var}(L)$ that can solve the equations, we must choose one by imposing an additional constraint. In reality, the imposition of this constraint consists of choosing the units that you want for your latent variable.

[4] See Chapter 4.

[5] If we are told that x equals 1 and that two different estimates of z are 2.5 and 3.5, then the equation is *overidentified*. There is no unique solution to overidentified equations and in any empirical problem the objective is to find the combination of estimates that gives the 'best' solution; least squares regression or maximum likelihood estimation are both examples of this.

At this point you have a choice. If you want the latent variable to have the same units as one of your measures, then you can fix the path coefficient from the latent to this measure to 1. By doing this you are stating that a one unit change in the latent changes the measured variable by one unit of the chosen scale. For instance, if we wish to scale our latent string lengths in centimetres then we could fix α_1 to 1. Think carefully before doing this. Your measure might systematically underestimate small values of the latent variable and overestimate large values; in this case the slope (i.e. the path coefficient) would be greater than 1. If this is the case, or if the scales of none of the measured variables are inherently more reasonable or useful than any of the others, then you can express the scale of the latent variable in units of standard deviations. This is done by fixing the variance of the latent variable to unity and allowing all the path coefficients to be freely estimated. Remember that standardisation (dividing a variable by its standard deviation so that its variance is unity) removes the original unit of the variable and replaces it with a scale of standard deviations from the mean. This has the effect of defining the scale of the latent variable by the measured variable and the path coefficient.

Here is the full set of structural equations that are used in the likelihood maximisation using a standard deviation scale for the latent:

$$X_{j1} = \alpha_1 L + \mathrm{N}(0,\sigma_1)$$

$$X_{j2} = \alpha_2 L + \mathrm{N}(0,\sigma_2)$$

$$X_{j3} = \alpha_3 L + \mathrm{N}(0,\sigma_3)$$

$$X_{j4} = \alpha_4 L + \mathrm{N}(0,\sigma_4)$$

$$\mathrm{Var}(X_j) = \alpha_j^2 \mathrm{Var}(L) + \sigma_j^2 \quad j = 1,4$$

$$\mathrm{Var}(L) = 1$$

$$\mathrm{Cov}(X_i,X_j) = \alpha_i \alpha_j \mathrm{Var}(L) \quad i \neq j$$

These structural equations decompose the observed variances of the four measured variables into one part that is the same for all of them, due to the common cause of L, and one part that is different for each of them, due to the uncorrelated measurement errors. Let's simulate this causal process using the following generating equations:

$$X_{j1} = 1L + \mathrm{N}(0,0.5)$$

$$X_{j2} = 0.07L + \mathrm{N}(0,7)$$

$$X_{j3} = 0.39L + \mathrm{N}(0,3.3)$$

$$X_{j4} = 1L + \mathrm{N}(0{,}10)$$

$$L = \mathrm{N}(\mu = 50, \sigma = 10)$$

Notice that the real scale for the true latent length is in centimetres in this simulation, since the path coefficients leading from the latent to X_1 and to X_3 are unity. So, I generate 100 independent 'strings', whose true length (L) is measured in centimetres. Since the first person used a ruler graduated in centimetres, the path coefficient is 1. Since she rounded her estimates to the nearest half centimetre, I have given her measurement error a standard deviation of 0.5 cm. The second person used her hand, which was 14 centimetres long. Her measurement scale was hand-lengths, rounded to the nearest half-hand and so the path coefficient is 0.07, with standard deviation of the measurement error being 7 centimetres. The third person used a ruler calibrated in inches, resulting in a path coefficient of 0.39, which is the conversion from inches to centimetres. He took little care in his readings and so the standard deviation of the measurement error was 3.3 centimetres. The last person simply visually estimated the true length in centimetres and so the path coefficient is 1. He was accurate only to within 10 centimetres, resulting in the standard deviation of his measurement error being 10 centimetres. Figure 5.6 shows the scatterplot matrix[6] of the 100 strings.

The measured lengths taken by the four people are all correlated, since they are all trying to measure the same thing. The residual scatter in the graphs between each measured variable is due the measurement errors of both variables in the pair and the magnitudes of these measurement errors differ from one variable to the other. Here are the maximum likelihood estimates of the free parameters of the structural equations after fixing the variance of the latent variable to unity:

$$X_1 = 10.076L + \mathrm{N}(0{,}3.416)$$

$$X_2 = 1.829L + \mathrm{N}(0{,}7.167)$$

$$X_3 = 3.631L + \mathrm{N}(0{,}3.480)$$

$$X_4 = 12.120L + \mathrm{N}(0{,}9.075)$$

The chi-squared statistic is 3.283 with 2 degrees of freedom,[7] producing a probability of 0.19 under the null hypothesis, telling us that the

[6] A scatterplot matrix is like a correlation matrix except that the actual scatterplot of each pair of variables in the matrix is shown and a histogram of each variable is shown on the diagonal panels.

[7] I said above that there was only 1 degree of freedom. However, fixing the variance of the latent variable to 1 means that we did not have to estimate this parameter, adding one extra degree of freedom.

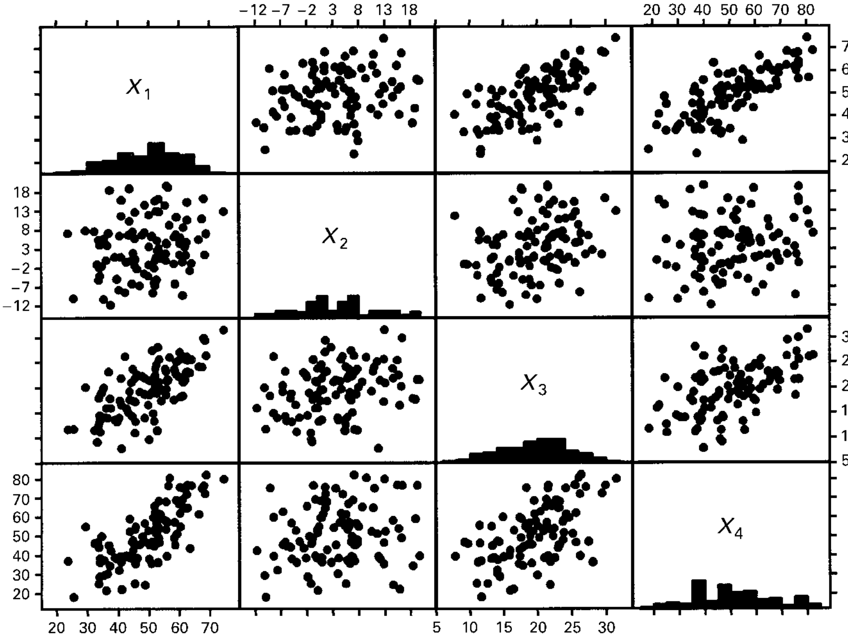


Figure 5.6. Scatterplot matrix of the simulated observations on the lengths of the four strings (X_1 to X_4); the histograms in the diagonal cells show the empirical distributions of the observations.

hypothesised causal structure is consistent with the data. The estimated variances of the measurement error of each variable agree well with the true values; the confidence intervals of these estimates, which are also given in most commercial SEM computer programs, include the true values. The estimated path coefficients are quite different from the true values. This is due to the fact that I fixed the variance of the latent variable to 1 even though we know that it is 100 (thus with a standard deviation of 10) in the simulations. This means that the path coefficients are proportional to the true values with the constant of proportionality being the inverse of the true standard deviation of the latent variable. Since we know the true variance of the latent variable in this simulation, we can convert the structural equations, obtaining:

$$X_1 = 1.0076L + \mathrm{N}(0,3.416)$$

$$X_2 = 0.1829L + \mathrm{N}(0,7.167)$$

$$X_3 = 0.3631L + \mathrm{N}(0,3.480)$$

$$X_4 = 1.2120L + \mathrm{N}(0,9.075)$$

Both these equations and the ones before are identical up to a constant (1/10) and so the conversion simply changed the mean of the latent variable. Since SEM is generally concerned with the relationships between the variables, not their means, this conversion will have no consequence on the model. However, I said above that we could have fixed the scale of our latent variable to centimetres by fixing the path from the latent to X_1 to unity and allowing the variance of the latent to be freely estimated. In an empirical study we would not know the true variance of the underlying latent variable and so this second strategy would be the one to use if the scale of the latent is important to you. We can calculate the correlation between the measured variables and the underlying latent variable in order to judge how well the measurement model has done. These correlation coefficients are routinely printed out in commercial SEM programs; Box 5.1 summarises the calculations.

Box 5.1. Correlating latents and indicators

By definition, the correlation coefficient between the latent variable, *L*, and its observed indicator variable, X_i, is:

$$\rho_{L,X_i} = \frac{\mathrm{Cov}(L,X_i)}{\sqrt{\mathrm{Var}(L)\mathrm{Var}(X_i)}} = \frac{\alpha_i \mathrm{Var}(L)}{\sqrt{\mathrm{Var}(L)\mathrm{Var}(X_i)}}$$

where α_i is the path coefficient from the latent (*L*) to its indicator (X_i).

The coefficient of determination, ρ^2_{L,X_i}, between the latent and its indicator is:

$$\rho^2_{L,X_i} = \frac{\alpha^2 \mathrm{Var}(L)}{\mathrm{Var}(X_i)} = \alpha_i^2 \theta_i$$

where θ_i is called the *reliability* of X_i.

If you want to obtain estimates of the latent variable, up to a scaling constant, then you can form a weighting function of the observed variables; see Bollen (1989) page 305, for the formula. However, no weighting function can estimate the latent without error and, in practice, the various weighting functions that have been proposed do not improve the accuracy of the estimation of the latent much beyond that obtained by choosing the measured variable whose correlation with the latent is highest.

The accuracy with which one can estimate the underlying latent variable (up to a scaling constant) will depend both on the reliability of the measured variables and on the number of such variables used to measure the latent. However, this predictive ability is really quite secondary in the context of testing causal models with measurement error. The most important point is that a measurement model allows one to explicitly account for measurement error, therefore providing unbiased estimates of the path coefficients. Because the predicted covariances between the observed variables are functions of the path coefficients linking them, then we can obtain unbiased estimates of the predicted covariance matrix and, therefore, of the asymptotic probability of the model under the null hypothesis. For those who like to see the algebraic details, Box 5.2 gives the generic factor model.

Box 5.2. Standard factor model

Consider the following model (Figure 5.7) with six observed (manifest) variables and two latent variables with an unresolved covariance between them.

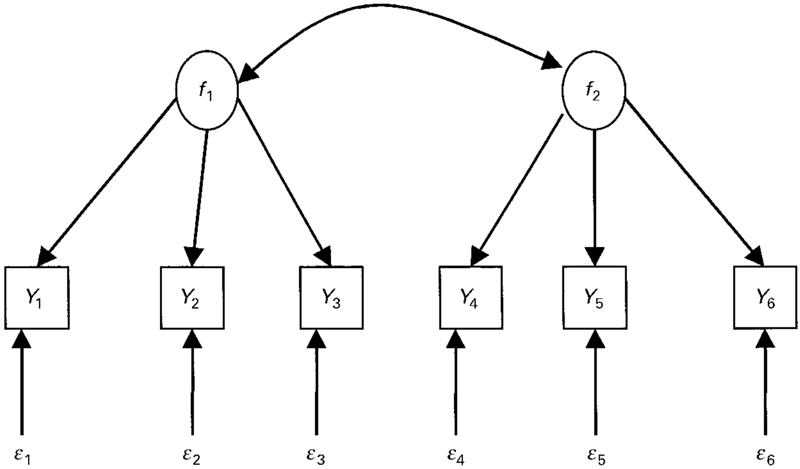


Figure 5.7. A measurement model involving two latent variables (f_1 and f_2), each measured by three observed variables (y_1 to y_3 and y_4 to y_6).

Here are the structural equations:

$$y_1 = a_{11}f_1 + 0f_2 + e_1$$

$$y_1 = a_{12}f_1 + 0f_2 + e_2$$

$$y_3 = a_{13}f_1 + 0f_2 + e_3$$

$$y_4 = 0f_1 + a_{24}f_2 + e_4$$

$$y_5 = 0f_1 + a_{25}f_2 + e_5$$

$$y_6 = 0f_1 + a_{26}f_2 + e_6$$

In matrix form, the equation is $Y = AF + E$. Multiplying both sides by the transpose of Y gives: $YY' = (AF + E)Y' = (AF + E)(AF + E)'$. Expanding this gives: $(AF)(AF)' + (AF)E' + E(AF)' + EE'$. Since the errors are indepents of the latent factors (f_1 and f_2) and of each other, when we take expectations the following two terms are zero: $(AF)E'$ and $E(AF)'$. This gives $E[YY'] = E[AFF'A'] + E[EE']$. From this comes the standard factor model equation:

$$\boldsymbol{\Sigma}_{YY} = A\boldsymbol{\Phi}A' + \boldsymbol{\Psi}$$

where $\boldsymbol{\Sigma}_{YY}$ is the model covariance matrix between the observed (Y) variables, $\boldsymbol{\Phi}$ is the model covariance matrix between the latent (F) variables, and $\boldsymbol{\Psi}$ is the model covariance matrix between the errors (which don't have to be mutually independent, as in the current example). In order to avoid indeterminacies, we can set the factor variances to unity. Putting this all together for our model we get the following matrix equation:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \vdots & \sigma_{16} \\ \sigma_{21} & \sigma_{22} & \vdots & \sigma_{26} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{61} & \sigma_{62} & \vdots & \sigma_{66} \end{bmatrix} = \begin{bmatrix} \alpha_{11} & 0 \\ \alpha_{21} & 0 \\ \alpha_{31} & 0 \\ 0 & \alpha_{42} \\ 0 & \alpha_{52} \\ 0 & \alpha_{62} \end{bmatrix} \begin{bmatrix} 1 & \phi_{42} \\ \phi_{21} & 1 \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{21} & \alpha_{31} & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha_{42} & \alpha_{52} & \alpha_{62} \end{bmatrix}$$

The minimum number of observed variables needed to fit, and test, a measurement model will depend on the number of hypothesised latent variables and the ways in which these latents are related to one another. For instance, if you have only one measured variable, the structural equation is $X_j = \alpha L + \varepsilon$. You have only one element in the covariance matrix (i.e. the variance of X) but you have three parameters to estimate: α, Var(ε) and Var(L). You can fix α to 1 to fix the scale of the latent, but this still leaves Var(L) and Var(ε). The equation is underidentified. If you can obtain an independent estimate of the error variance, then you can fix Var(ε) to this value. This can sometimes be done. For instance, one could physically extract the body lipids of a sample of animals to obtain a precise estimate of body fat and then regress an indirect measure of this body fat to obtain the residual error variance. This will allow you to separate measurement error in subsequent data (assuming that the measurement error doesn't change), but you still can't test this measurement model, since there would be no degrees of freedom.

What about two measures of a latent? With two observed variables we have three non-redundant elements of the covariance matrix (two variances and one covariance), but we also now have five free parameters (α_1, α_2, Var(ε_1), Var(ε_2) and Var(L)). We have not solved our problem of underidentification. With three measures of a latent we have six non-redundant elements of the covariance matrix and seven free parameters. Since we have to set the scale of the latent, for instance by fixing Var(L) at 1, we can now fit the structural equations, but we will have no degrees of freedom with which to test the measurement model. This is fine if we don't need to independently justify the measurement model (for instance, the relationship between a thermometer and the air temperature) but if there is any question about the causal relationships between the measured variables then we have defeated the whole purpose of modelling measurement error. Four measured variables per latent is the minimum number needed to both fit and test such a measurement model.

5.4 The nature of latent variables

So far, I have described latent variables simply as variables that we have not directly measured but that we can directly observe. In the previous examples there was no question but that the animals really did have lipid reserves or that the strings really did have a length. Our only concern was in accurately measuring these variables. In such situations the development of the measurement model involves choosing measurable indicator variables that are all linearly related to the same latent variable. Ideally, the *only* causal relationships between these indicators will be through the common effect of the latent variable. If there exist other causal relationships between the measured variables, through other latents or not, then these must also be included in the model.

Often nature is not that accommodating. What happens if we want to model latent variables that we *cannot* directly observe? In such cases, even the existence of the latent variable is hypothetical. The invocation of such theoretical entities presents much more difficult choices, since we can't rely on direct observation to know whether such things even exist, although the actual modelling is no different. None the less, the history of science is littered (or enriched, depending on your philosophical view) with such things. When Gregor Mendel invoked recessive and dominant alleles of genes to explain his patterns of inheritance in pea seeds, he did not measure or observe such things. Rather, he inferred them because the ratios of the resulting phenotypes agreed with the binomial proportions that would result if such things existed[8]. Genes were latent variables and still are; no one has ever directly observed a gene. Atoms, too, are latent; the Periodic Table was developed by inferring atomic structures from the numerical regularities that resulted from experiments. Ernst Mach, who was mentioned in Chapter 3 as one of the phenomenalists who influenced Karl Pearson's views, initially refused to accept the reality of shock waves caused by bullets going faster than the speed of sound. He accepted such waves only when he was able to devise an experiment in which a camera was rigged to take a picture just as a bullet cut a fine wire covered in soot, revealing a V-shaped pattern[9]. As these 'successful' latent variables attest, scientists have regularly invoked things that can only be indirectly observed through the use of proxy measures. The problem is that scientists have also invoked 'unsuccessful' latent variables. A classic example is the 'aether', through which light waves were supposed to cross outer space. The use of latent variables in measurement models or SEM is not so much a statistical controversy as a scientific and philosophical one. Think carefully before including latent variables in your models and be prepared to justify their existence.

Much of my personal discomfort with latent variable models comes from the causal claims that many (by no means all) latent variables make. It is one thing to invoke a theoretical unmeasured variable and quite another to demonstrate that such an entity has both a reality in nature and has causal efficacy. Choosing, developing and justifying such latent variables is, perhaps, the most difficult aspect of structural equations modelling. I don't know of any set of rules that can unfailingly guide us in this task either. The exploratory methods, described in Chapter 8, can help to alert us to the existence of latent variables. The statistical tests based on maximum likelihood allow us to compare our data with such hypothesised latent variable models and therefore potentially to reject them. However, scientists generally demand stronger evidence than an acceptable statistical fit before accepting the physical reality of unmeasured variables. Before continuing further, it is again useful to look briefly at the history of latent variable modelling in statistics. The hornets' nest of confusion involving latent variable models is due in part, I believe, to the historical link between latent variable models and factor models in the social sciences.

In 1904 the English psychologist Charles Spearman combined the new psychometric work of Alfred Binet on human intelligence with

[8] There has been a long-running debate as to whether Mendel 'cooked' his data by ignoring outliers, since the observed and predicted ratios are so remarkably close as to be highly unlikely to occur by chance.

[9] Needless to say, he did not really observe the shock waves, only the indirect effect on the soot particles.

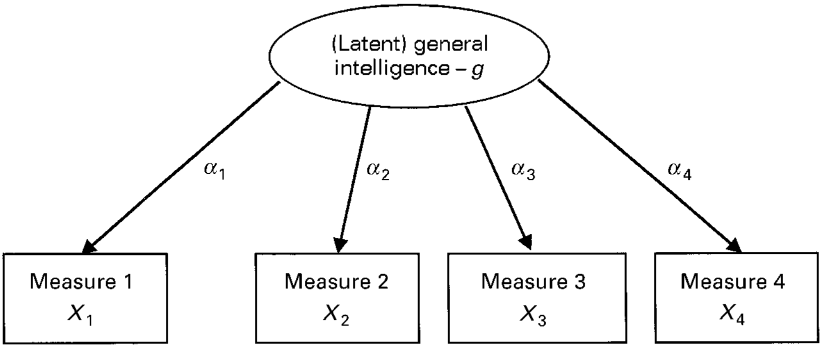


Figure 5.8. A measurement model for 'general intelligence' – a latent variable. Note that the errors for each of the observed measures (X_1 to X_4) are not shown but will exist unless the measures are all perfectly correlated with this latent variable.

correlation coefficients. In his provocatively titled paper (Spearman 1904), 'General intelligence objectively determined and measured', he hypothesised that the observed measures of intelligence of people, obtained from test questions, were all correlated because they were all due to a general latent intellectual capacity (*g*) that varied from person to person. If we have four different measures of intelligence (from an IQ test, say) then the causal graph would look like Figure 5.8.

Now, given this structure, the population correlation between any two observed variables, say X_1 and X_2, would be $\rho_{12} = \alpha_1\alpha_2$. It follows that the following three equations must be true if the model in Figure 5.8 is true:

$$\rho_{12}\rho_{34} - \rho_{13}\rho_{24} = 0$$

$$\rho_{13}\rho_{24} - \rho_{14}\rho_{23} = 0$$

$$\rho_{14}\rho_{23} - \rho_{13}\rho_{24} = 0$$

Spearman called these 'vanishing tetrads' because each involves four correlation coefficients and they are, in modern terms, constraints on the correlations due to the causal structure of the model. He argued (incorrectly, as explained in Chapter 8) that data obeying such vanishing tetrads was evidence for this unmeasured latent cause ('generalised intelligence'). Spearman apparently viewed this latent variable as a real, causally efficacious attribute of people. As more measured variables were added, and more complicated latent structures were hypothesised, one could derive the vanishing tetrads that were implied by the model, but this quickly became difficult to do, both conceptually and computationally.

In the 1930s Harold Hotelling invented principal components analysis and Louis Thurstone invented factor analysis. These methods were not based on any explicit causal model. Quite the contrary. Thurstone viewed science in much the same way as Pearson did: science consisted of erecting 'constructs' that could describe the data as simply as possible. Whether or not such 'constructs' actually existed in nature was irrelevant; they could economically summarise the patterns of correlation and replace a large number of observed variables by a single 'construct'. These constructs, or factors, had the property that they could partition the variances of each measured variable into one part (the construct) that was common to all and one part that was unique to each measured variable. Described in this way, we appear to be right back to our measurement model, as described above. However, such factor models (like principal component models) could perform this trick quite independently of whether the 'common variance' was really due to some unmeasured common cause. Moreover, the method, if drawn as a graph, always has the arrows going from the construct (factor, common variance) to the measured variables. Such a structure was a requirement of the method. Interpreted as Thurstone had intended, this was not a problem, since the constructs were simply mathematical functions designed to summarise data. Interpreted as causal models (as Thurstone most emphatically did not intend) factor models had the bizarre property of requiring that the direction of causality always went from the latent construct to the observed variables. The obvious advantage of factor analysis or principal components analysis[10] over Spearman's method of vanishing tetrads was that these former methods were easier to use and based on standard formulae. One had only to plug the data into the equations and out popped the construct or the principal component axis.

Thus vanishing tetrads became an historical footnote and factor analysis (with its requirement that the arrows go from the construct to the measured variables) took their place in psychometrics. Jöreskog (1967, 1969) applied maximum likelihood methods to factor analysis to develop an inferential test for such models, and then extended this to allow for cause-and-effect relationships between the latent variables based on econometric simultaneous equations models, giving rise to structural equations models (Jöreskog 1970, 1973). Although there is no longer any mathematical requirement that the arrows always go from the latents to the measured variables – as was the case with factor analysis – the formalism of factor analysis still persists in SEM along with some of its philosophical origins.

[10] Principal components analysis, another multivariate data-summary method, requires that the path coefficients always go in the opposite way to factor analysis.

As an example, consider the description given in Bollen's (1989) influential book on SEM. He stated that the measurement process begins with a concept and defines a concept as an idea that unites phenomena under a single term. He gave the example of 'anger', which provides the common element tying together attributes such as screaming, throwing objects, having a flushed face and so on. The concept of anger, he said, 'acts as a summarising device to replace a list of specific traits that an individual may exhibit'. This is a close paraphrase of Thurstone's original description of a 'factor' but remember that Thurstone's factor analysis was explicitly acausal. To Bollen's rhetorical question 'Do concepts really exist?', he answered '[c]oncepts have the same reality or lack of reality as other ideas . . . The concept identifies that thing or things held in common. Latent variables are the representations of concepts in measurement models.' If we are dealing with purely statistical models devoid of causal implications, then such a view might be fine. If our models are statistical translations of causal processes then the latent variables in our models must be something more than a mathematical summary; latent variables must represent variables with physical reality having causal relationships to the measured variables.

One of the early controversies amongst geneticists at the turn of the century concerned the inheritance of size differences in different body parts. W. E. Castle, an influential geneticist at the time and Sewall Wright's thesis supervisor, argued that there was a single 'size factor' that was inherited and that determined the allometric scaling of different body parts. Part of this argument was based on correlation coefficients, calculated by Wright while still a graduate student, relating five different bone measurements of rabbits. Davenport (1917), studying human stature, argued that the patterns of correlation between different lengths of different body parts suggested that these attributes of size were inherited independently. In 1918 Wright published 'On the nature of size factors' based on the rabbit measures, in which he calculated a series of partial correlations. Based on these calculations he concluded that his own supervisor was wrong and that 'These three correlations[11] suggest the existence of growth factors which affect the size of the skull independently of the body, others which affect similarly the length of homologous long bones apart from all else, and others which affect similarly bones of the same limb.' Since no one knew what these 'size factors' were, the entire argument concerned the number of latent variables controlling the inheritance of size in different body parts. The following example shows how one can test such claims.

[11] Actually, they were partial correlations.

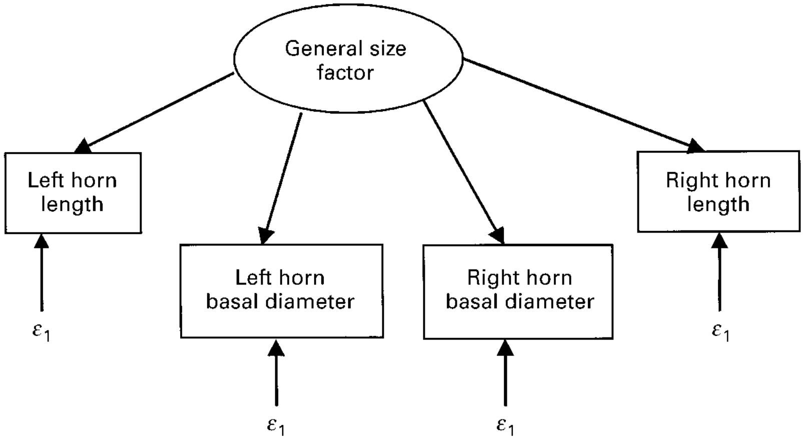


Figure 5.9. The hypothesised measurement model relating four observed attributes of the horns of male Bighorn Sheep.

5.5 Horn dimensions in Bighorn Sheep

Marco Festa-Bianchet and his students have been following a population of Bighorn Sheep from the Rocky Mountains of Alberta for many years. Horn size is very important in this species because, among other things, it affects the ability of males in combat during the rut and therefore their evolutionary fitness. Every year from 1981 until 1998 the researchers measured the total length and the circumference at the base of the two horns of the captured sheep. As one might expect, these four variables are highly correlated and display the sorts of allometric scaling pattern that are so ubiquitous in biology. Are these four variables simply responding to a single latent 'size factor'. In other words, are the patterns of correlation between the four variables simply due to a single common unmeasured cause that determines increases in linear dimensions, as Chase might have supposed?

Figure 5.9 shows this hypothesis, translated into a causal graph. To fix the scale of the latent, I fixed the variance of the latent variable to 1. The data do not follow a multivariate normal distribution, even after a ln-transformation, as shown by Mardia's normalised coefficient of kurtosis. Because of this, I use a robust estimation method for the chi-squared statistic (the Satorra–Bentler chi-squared); these statistics are explained in detail in Chapter 6. The data are clearly not consistent with the single common latent model in Figure 5.9, since the Satorra–Bentler chi-squared statistic is 759.106 with 2 degrees of freedom. The probability of observing this by chance is far lower than one in a million.

A look at the residuals shows why the fit is so bad. The residuals of the two length measures are highly correlated, indicating that there is something else that is affecting length independently of the basal circumference. Perhaps the 'General size factor' causes two 'Specific size factors' – one for length and one for circumference? In Chapter 6 I explain how one can statistically test these ideas, however I can't point to any specific biological mechanism to justify this proliferation of hypothetical unmeasured variables.

At this point I apply the SQUIRM test. When the hypothesised latent variable begins to resemble 'a summarising device to replace a list of specific traits that an individual may exhibit' rather than a physical 'thing' with causal efficacy then I begin to squirm. The statistical model has wandered too far away from any causal model to which it was supposed to be a translation. I can't justify one of the auxiliary assumptions (that the latent variable is not simply a statistical construct) beyond reasonable doubt. Each person will have their own tolerance for this, but in my experience most biologists (and reviewers!) have a very low SQUIRM tolerance indeed. My own (highly personal) opinion is that this is a good thing.

5.6 Body size in Bighorn Sheep

Body size is another important attribute of Bighorn Sheep. Large animals are less likely to fall prey to predators. Animals that have been able to amass sufficient fat reserves in the autumn are more likely to survive the severe winter conditions at the top of a mountain in Canada. Larger males are more successful in the rut and therefore are able to copulate with more of the females. The reproductive success of a female is affected by her fat reserves. Now imagine that you are a field biologist, perched at the top of a steep rocky slope with a temporarily subdued animal, and you need to estimate its body size. You do not have a balance (and have to keep your own balance!) but you can quickly take measurements of attributes associated with body size. Can you construct a measurement model that will be able to estimate the unmeasured 'body size'?

The following analysis, based on data provided by Festa-Bianchet, are from four indirect measures of body size based on 248 observations of Bighorn Sheep. The observed variables are the total length of the animal (snout to tail), the circumference of the neck, circumference of the chest just behind the front legs, and a visual estimate (which sounds better than 'guess') of body weight. These data, transformed to their natural logarithms, are consistent with multivariate normality, based on Mardia's coefficient. The measurement model consists of a single latent variable, which I have labelled body size, that is the single common cause of the four indirect meas-

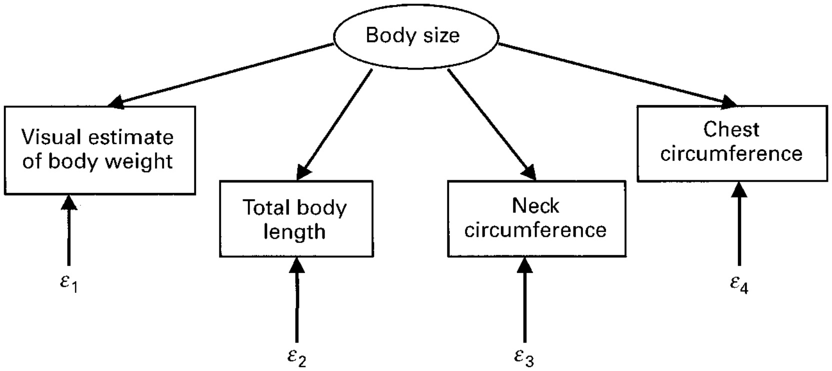


Figure 5.10. The hypothesised measurement model relating four observed size attributes of Bighorn Sheep.

ures listed above. As always, one can set the scale of the exogenous latent variable either by fixing its variance to unity or by fixing the path coefficient from it to one of the observed variables. Since 'body size' is usually interpreted as meaning mass, I have therefore chosen to fix the path coefficient from the latent to the estimated body weight at unity. This means that the latent 'body size' is measured in ln(kilograms) (the units of the estimated body weight). Figure 5.10 shows the directed graph for this measurement model.

The chi-squared statistic for this model is 0.971 with 2 degrees of freedom, giving a probability under the null hypothesis of 0.615. The data are perfectly consistent with the model. Here are the structural equations and the proportion of the variation of each measured variable that is accounted for by the latent 'body size':

$$\ln(\text{Estimated weight}) = 1\ln(\text{Body size}) + \text{N}(0{,}0.023) \quad R^2 = 0.893$$

$$\ln(\text{Total length}) = 0.370\ln(\text{Body size}) + \text{N}(0{,}0.003) \quad R^2 = 0.911$$

$$\ln(\text{Neck circumference}) = 0.424\ln(\text{Body size}) + \text{N}(0{,}0.005) \quad R^2 = 0.883$$

$$\ln(\text{Chest circumference}) = 0.387\ln(\text{Body size}) + \text{N}(0{,}0.001) \quad R^2 = 0.982$$

$$\ln(\text{Body size}) = \text{N}(0{,}0.191)$$

We see that the estimated error variance of the ln(estimated body weight) is 0.023 and the latent 'body size' accounts for 89.3% of the variance of this estimated weight. The guesses were not so bad after all. In fact, these guesses of the true body weight appear to be just as tightly correlated with the latent body size as the measures of neck circumference. The observed variable that was most highly correlated ($R^2 = 0.982$) with the latent body size was the chest circumference. If only one measurement is to be taken, the chest circumference should be the one to use.

If we now wish to test a causal model involving body size in a new data set we can either use one of the measured variables and explicitly include our estimates of the error variance or else use all four measured variables. The advantage of using all four variables is that we do not have to assume that the error variances will remain the same from one study to the next; we can instead estimate them. This might be a wise decision if, for instance, different people take these measurements from one study to the next and some people make more measurement errors than others.

The problem of different people making systematically biased estimates is the likely explanation for the lack of fit that occurs when another measured variable is included in the measurement model described above. This variable is the length of the hind foot. The chi-squared statistic for a new model that includes this new variable is 25.808 with 5 degrees of freedom, for a probability of about 1×10^{-4}. This five-variable measurement model can be made to fit the data only by letting the length of the hind foot covary freely with the weight estimate and the neck diameter[12]. In other words, there are other causes, independent of the latent 'body size' that are generating associations between these three measured variables. Although it is possible that these other causes are related to the biology of the animals, it seems more likely that the other causes are due to the way the data were collected.

The measurements were taken over many years by many different people – mostly graduate students with different levels of ability in field work. Unlike the other length measurements, the length of the hind foot requires that the foot and hoof be consistently extended to the same degree. These measurements must be taken quickly while the animal is still subdued. Imagine that you are sitting at the edge of a steep precipice on the top of a mountain, with an adult Bighorn Sheep about to wake up. It is safe to assume that the care with which the foot is extended will vary from person to person. So, if there are any systematic biases between people in how they measure the three variables in question (the length of the hind foot, the weight estimate and the neck diameter) then this would be a cause of correlations between them independent of differences in 'body size'.

[12] Actually, one permits covariation between the error variables of these measured variables.

5.7 Name calling

A certain Henry P. Crowell of Ravenna, Ohio, bought a bankrupt mill in 1881 and 'went into the business of convincing people to consume what previously only poverty-stricken Scotsmen, Germans and horses had eaten' (Burke 1996). How did he accomplish this feat of marketing[13]? Since people associated the word 'Quaker' with honesty and a healthy life-style, he simply called his new product 'Quaker Oats'. Thus began one of the longest-running breakfast cereals in the USA. A certain mouthwash company proudly proclaimed that its product, besides making one's breath taste fresh, also cured 'halitosis'. The dictionary definition of halitosis is 'bad-smelling breath'. So what is the latent variable, shown in Figure 5.10, that is the single independent 'cause' of the estimated body weight, the total length, the neck and the chest circumference of the sheep? I have labelled it 'body size' but is this misleading advertising? Just as saying that something 'cures halitosis' evokes connotations beyond simply 'curing bad breath', does calling my latent variable 'body size' evoke connotations beyond simply 'that theoretical variable that has the property of making the partial covariances between each unique set of measured variables equal to zero, when conditioned on it?'

Remember Bollen's (1989) claim that '[l]atent variables are the representations of concepts in measurement models', and that '[t]he concept identifies that thing or things held in common'. It is certainly reasonable to state that 'body size' is that which is common to weight, length and circumference of the body (the measured variables). However, if Figure 5.10 is to be interpreted as a description of a causal process, then the latent variable also represents a single common *cause* of body weights, lengths and circumferences. The causal claim must be that there is a single biological process that determines all of these body dimensions. It would obviously be better if we knew enough about the genetic and developmental processes determining body size that we could label our latent variable as 'hormone X' or 'gene Y'. If we can't, then we could at least label it as the 'unknown cause of body size'. So long as both you and I understand the name 'body size' as being a short form of saying this, then we can properly translate between the causal claim and the statistical model. It is particularly important that we choose our words carefully when dealing with latent variables and the burden of clarity is on the person proposing the model, not on the reader. If you see a latent variable in a structural equation and its meaning and causal justification are not clearly explained, think of bad breath.

[13] To show what Crowell was up against, consider that the first edition (1755) of Samuel Johnson's classic *Dictionary of the English language* defined 'oats' as a grain that sustained horses in England and people in Scotland.

6 The structural equations model

The structural equations model is commonly described as the combination of a measurement model and a structural model. These terms derive from the history of SEM as being a union of the factor analytical, or measurement, models of psychology and sociology and the simultaneous structural equations of the econometricians. In its pure form it therefore explicitly assumes that every variable that we can observe is an imperfect measure of some underlying latent causal variable and that the causal relationships of interest are always between these latent variables. As in many other things, purity is more a goal than a requirement. Using the example in Chapter 5 of the effect of air temperature on metabolic rate (Figure 6.1), the things that we can measure (the height of the mercury in the thermometer or the change in CO_2 in the metabolic chamber) always contain measurement error (ε_i). The measurement model, shown by the dotted squares in Figure 6.1, describes the relationship between the observed measures and the underlying latent variables (average kinetic energy of the molecules in the air and the metabolic rate of the animal). The structural model, shown by the dotted circle in Figure 6.1, describes the relationship between the 'true' underlying causal variables. If we have only one measured variable per latent variable, and we assume that the measured variable contains no measurement error (i.e. the correlation between the measured variable and the underlying latent variable is perfect) then we end up with a path model. If we have a set of measured variables for each latent variable and we do not assume any causal relationships between the latent variables, then we have a series of measurement models. If we have more complicated combinations in which we assume causal relationships between the latent variables, then we have a full structural equations model. Therefore, if you have understood Chapters 1 to 5, then you already know how to construct and test a structural equations model; you simply have to put the pieces together.

The goal of this chapter is therefore to deal with some technical details that I have ignored up to now. The first detail is the problem of identification. In models involving more complicated combinations of latent and observed variables, how can we make sure that no model parameters are

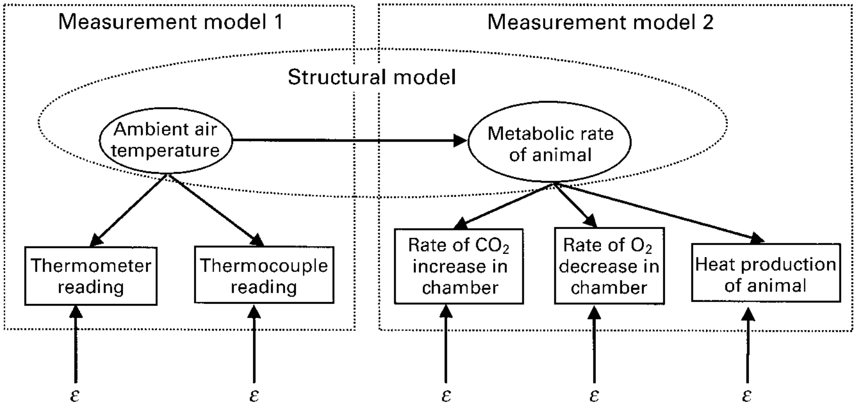


Figure 6.1. The relationship between the measurement model and the structural model in the causal structure relating the ambient air temperature and the metabolic rate of an animal.

underidentified? The second detail involves the robustness of SEM to violations of two important assumptions: large sample sizes and multivariate normal distributions. What happens when our data do not agree with these assumptions, and what can be done about it?

6.1 Parameter identification

We have already met the problem of underidentification in Chapter 5. Intuitively, a model is underidentified when more than one combination of parameter values can account for the same pattern of covariance. If a model is underidentified then you can't trust the parameter estimates, their standard errors, the chi-squared value or its probability level. If a model is underidentified then most commercial SEM programs will print a warning. For instance, if you are told that a parameter estimate 'is a linear combination' of some other set of parameters or that an estimated variance estimate is negative or 'set at zero', then this probably means that the model is underidentified.

A model can be *structurally* underidentified or *empirically* underidentified. Structural underidentification means that the model will be underidentified for any combination of parameter estimates – the problem is in the way the model itself is constructed. You will want to ensure that your model is not structurally underidentified before collecting data, in order to avoid wasting your time. Empirical underidentification means that the model is under identified only for some particular sets of parameter estimates – the problem is not in the general construction of the model but rather with the particular values found in the data. These points will be illustrated with examples later. Let's start with some useful rules for avoiding this problem.

6.2 Structural underidentification with measurement models

Following the notions introduced in Chapter 5, let's call a measurement model any factor analytical model consisting of a set of latent variables and a set of observed indicator (measurement) variables that are each caused by at least one of the latent variables. The latent variables can be allowed to freely covary (i.e. there can be curved double-headed arrows between them) but there are no cause–effect relationships between the latent variables (i.e. there can be no arrows from one latent to another). Bollen (1989) has summarised three rules to help in judging whether a measurement model is structurally identified. These rules cover many types of measurement model, but not all. All of these rules assume that the scale of each latent variable has been fixed, as described in Chapter 5, either by fixing one of the path coefficients to 1 or by fixing the variance of the latent variable to 1.

Rule 1: $t \leq n(n+1)/2$ where n is the number of observed variables and t is the number of free parameters (i.e. free path coefficients, free error variables and free covariances either between the latents or between the error variables). This rule is necessary for identification; if the rule doesn't hold in your model then you can be sure that the model is not identified. Unfortunately, this rule doesn't ensure that your model will be identified; even if the rule holds, the model might still be underidentified. The following two rules are sufficient (i.e. if they hold then the model is identified) but not necessary (i.e. there are still identified models that violate these rules).

Rule 2: A measurement model is identified if, along with rule 1:

1. There are at least three indicator variables per latent variable.
2. Each indicator variable is caused by only one latent variable.
3. There are no correlations between the error variables.

Rule 3: A measurement model is identified if, along with rule 1:

1. There is more than one latent variable.
2. There are at least two indicator variables per latent variable.
3. Each indicator variable is caused by only one latent variable.
4. Each latent variable is correlated with at least one other latent variable.
5. There are no correlations between the error variables.

To understand how these rules work, let's look at Figure 6.2, which shows six different measurement models. The scale of the latent variables

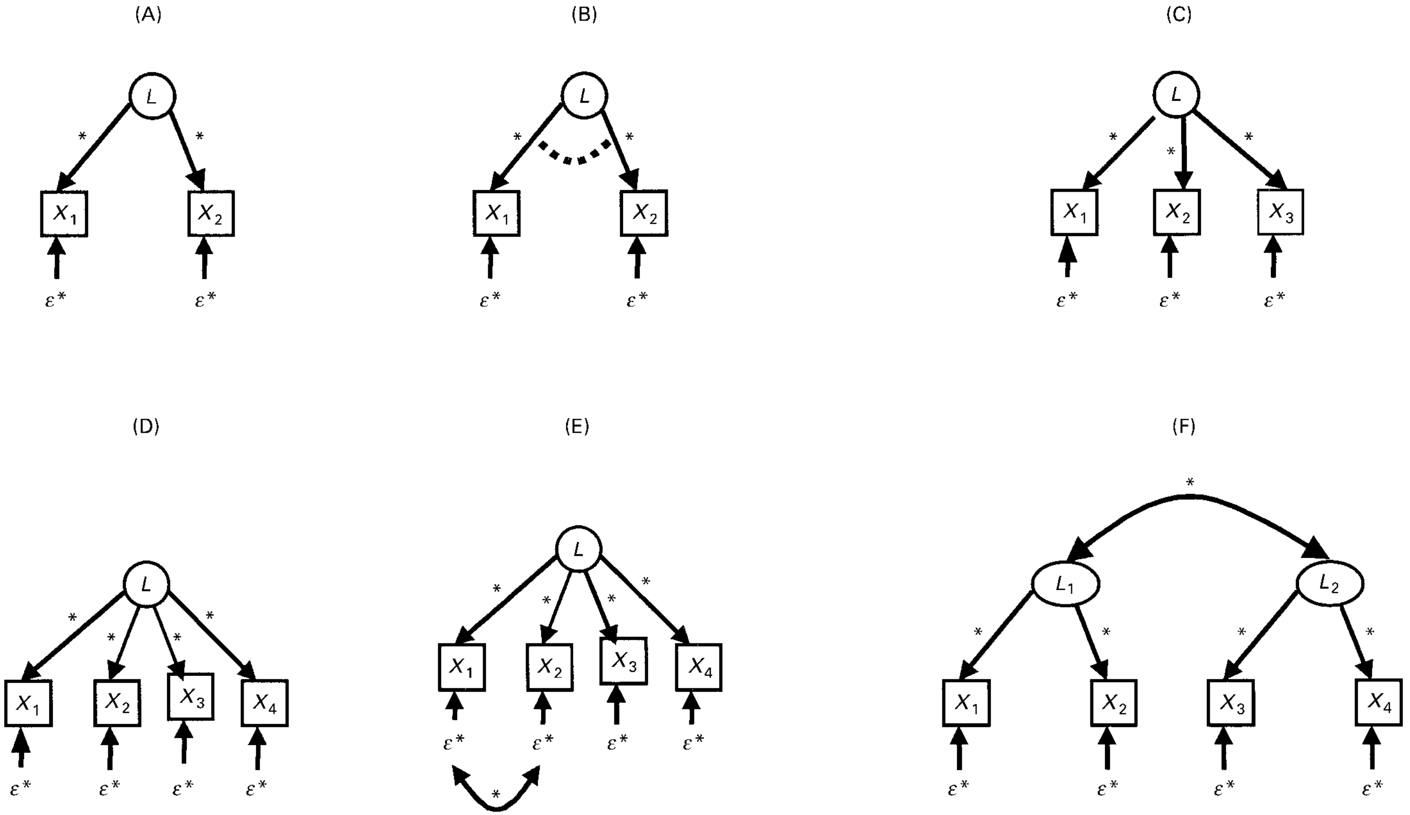


Figure 6.2. Six measurement models used to illustrate Bollen's rules for identification.

(*L*) have all been set by fixing the variances of the latent variables to unity. Free parameters (the path coefficients, the variances of the error variables, or the covariances indicated by curved double-headed arrows) are shown by an asterisk (⋆). The model in Figure 6.2A is underidentified; there are four free parameters ($t=4$) and two observed indicator variables ($n=2$) but rule 1 states that $t \leq n(n+1)/2$. If we constrain the values of the two path coefficients to be the same during the iterative procedure that minimised the maximum likelihood chi-squared statistic (shown in Figure 6.2B as the dashed line between the two free path coefficients) then we have only three free parameters, and the model is just identified[1]. This trick, while allowing us to identify the model, doesn't really allow us to get unbiased estimates of the measurement error or of the two path coefficients.

The model in Figure 6.2C is just identified. There are $t=6$ free parameters and $n=3$ observed variables, therefore $6=3(4)/2$ and rule 1 is fulfilled. Since there are no degrees of freedom left, we cannot test such a model using the maximum likelihood chi-squared. Such a model can always be fit even if the causal assumption of a single common latent cause is wrong, and we can't know whether or not the causal assumption is reasonable on the basis of statistical criteria.

The model in Figure 6.2D is overidentified. There are $t=8$ free parameters and $n=4$ observed variables, therefore $8<4(5)/2$ and rule 1 is fulfilled. Since there are $4(5)/2-8=2$ degrees of freedom left, we can also test such a model using the maximum likelihood chi-squared. Such a model can always be fit but if the causal assumption of a single common latent cause is wrong then we would obtain a significant probability estimate of the measured maximum likelihood chi-squared statistic and could therefore reject the model. Both the measurement models for the Bighorn Sheep horns and for the body dimensions, studied in Chapter 5, were of this form and we saw that the model for the horn dimensions was clearly rejected ($p<10^{-6}$) while the model for the body dimensions was not rejected ($p=0.615$).

The model in Figure 6.2E is like the model in Figure 6.2D except that it has a free covariance between the error variables of X_1 and X_2. Rule 1 is still satisfied since $t=9$, $n=4$ and $9<4(5)/2$. Rule 2 is not satisfied; although there are at least three indicator variables per latent (there are four) and each indicator variable is caused by only one latent, there is also a correlation between two of the error variables. Rule 3 can't be applied either, since there is only one latent variable. However, rules 2 and 3 are sufficient

[1] In fact, this model is equivalent to the so-called *major axis* (errors-in-variables) regression of Sokal and Rohlf (1981). *Reduced major axis* regression is obtained by simply standardising X_1 and X_2 to unit variance and a zero mean before fitting the model.

conditions, not necessary conditions. We can't state that the model is definitely not identified, only that we can't tell one way or the other. In fact, model E in Figure 6.2 is structurally identified.

The model in Figure 6.2F looks a bit like two models from Figure 6.2A combined. Remember that model A wasn't identified because rule 1 was violated. What about model F? There are $t = 9$ free parameters and $n = 4$ observed variables. Since $9 < 4(5)/2$, rule 1 is satisfied. Rule 2 is not satisfied (there are only two indicator variables per latent) but rule 3 is satisfied. Therefore, model F is structurally identified.

This model also provides a good example of how a model can be structurally identified but empirically underidentified. If, in reality, the covariance between the two latent variables (the curved, double-headed arrow) is close to zero then the estimated covariance in the data might be zero due to sampling fluctuations. If this occurs then the maximum likelihood procedure will be trying to fit two independent measurement models, each with only two indicators per latent. Since each separate measurement model has four free parameters (two error variances and two path coefficients) but only two indicator variables, rule 1 would be violated in this particular case.

Recall the measurement model for the length and basal diameter of the left and right horns of the Bighorn Sheep. We were quite confident that the correlations between these four measures were not due to a single common unknown cause because the measurement model with a single latent variable was strongly rejected. Perhaps the observed correlations are due to two correlated latent causes, as shown in Figure 6.2F? In bilaterally symmetrical organisms the left and right halves of the body should be mirror images in terms of size and shape. You have only to look into a mirror (when no one else is watching) to see that no one is really perfectly bilaterally symmetrical. Various environmental perturbations during embryonic development can cause random deviations from perfect bilateral symmetry, and the degree of this 'fluctuating asymmetry' is sometimes used as an index of pollution load or other forms of environmental stresses. Perhaps the model with a single latent cause of horn dimensions was rejected because there were additional causes of the left and right horns besides a single 'size' factor that generate deviations from bilateral symmetry? This hypothesis produces the model in Figure 6.2F and we know that this model is both structurally identified and has 1 degree of freedom left to test the model. The single 'size' factor is reflected in the free covariance between the two latent variables. The two latent variables, according to our present hypothesis, should represent the different causes of the left and right horns generating deviations from bilateral symmetry. The two latent variables in Figure

6.2F would then represent the causes specific to the left and right horns. When I fit this model to the data (using the Satorra–Bentler chi-squared, since there is significant multivariate kurtosis in the data), I get a chi-squared value of 146.1205 with 1 degree of freedom. Clearly, this model, too, is wrong[2].

If we think of how the measurements were taken, we are led to another measurement model with two latent variables. The horns are strongly curved and their length is measured with a measuring tape that has to properly follow the curve of the horn. It is possible that longer horns, with a more pronounced curve, would be systematically underestimated as the exasperated researcher tries to make the measuring tape follow the curve of the horn before the sheep regains control. The longer the horn, and the more it is curved, the greater the degree of underestimation, since the measuring tape will have more chance to slip down a bit along the horn. A similar systematic bias might occur for the two measures of basal diameter, since this measurement too requires a subjective decision as to where the base of the horn begins. If these speculations are correct, then each of the length measures and each of the diameter measures might have a separate cause (i.e. the way in which they are measured) besides a common 'size' factor as horn volume increases during development. When I fit this model to the data (again using the Sattora–Bentler chi-squared, since there is significant multivariate kurtosis in the data), I get a chi-squared value of 3.948 with 1 degree of freedom, giving a probability level of 0.05. This model has an ambiguous probability level and its true value is probably higher, given the large positive multivariate kurtosis of the data (Mardia's coefficient of multivariate kurtosis is 27.1); this point will be further discussed in the section on non-normality of the data. Therefore, I conclude that there is not sufficient evidence to reject it. The path coefficients from the latent 'diameter' to the two diameter measures were 0.522 and 0.519 for the left and right horns. Since the square of the diameter of a circle is proportional to its area (a dimension of 2), one would expect these path coefficients to be 0.5. The path coefficients from the latent 'length' to the two length measures were 0.875 and 0.869 with standard errors of about 0.03 for the left and right horns. Since length has a dimension of 1, one would expect these path coefficients to be 1. An approximate 95% confidence interval of the path coefficients is therefore about $0.87 \pm 2(0.03)$ and 1.0 is clearly outside

[2] I used this example simply for pedagogical reasons. In reality, the notion of fluctuating asymmetry is that the deviations are random from individual to individual. In some individuals the asymmetry will be on the left side and on others the deviation will be on the right side. In this case there would be no systematic difference in the population and this would not generate a latent cause that is systematic to either left or right horns. The causes of the asymmetry would simply be subsumed into the unique error variances.

this interval. Therefore the measurement model suggests that the lengths of the longer horns were systematically underestimated. The covariance (and correlation, since their variances were fixed at unity) between the two latents for 'length' and 'diameter' was 0.9995. If we accept this two-factor measurement model then all these points suggest that the horn dimensions are caused by a single latent 'size' factor, but that there was a systematic bias in measuring the longer horns that introduced a second latent cause of the lengths independent of the diameters.

Now apply your personal SQUIRM test. Does my interpretation of these latent variables seem reasonable to you? The acceptable fit of the measurement model with two latents says nothing about *what* these two latent variables represent. The explanation that I have outlined, that the two latents represent the systematic errors made in measuring lengths and diameters and that the covariance between the latents is due to the common 'size' factor, is an *interpretation* of the latent variables. This interpretation of the latent variables in the model is not supported by any statistical evidence; rather, my evidence comes from how the variables were measured and the sorts of error of measurement that might occur. The next step would be to search for an independent confirmation of this explanation. For instance, if the explanation is correct, then the two latent variables should disappear and be replaced by a single latent variable once we measure horn length in a way that does not systematically underestimate longer horns. One way would be to photograph the horns and then measure the lengths using image analysis.

Davis (1993) described a way of testing for identification in much more complicated measurement models, applicable to any measurement model in which each indicator is caused by only one latent variable. This method (the FC1 rule) requires that you be able to do matrix multiplication, but many statistical programs can do this[3]. A further requirement is that the scale of each latent be fixed by fixing one path coefficient to 1 rather than fixing the scale by fixing the variance of the latent variable to 1. Box 6.1 summarises the FC1 rule.

Box 6.1. FC1 rule for identification of a measurement model

The FC1 ('Factor Complexity 1') rule for the structural identification of a measurement model assumes that each observed indicator variable is caused by only 1 latent variable (hence its name).

For each latent variable, L_i, in the measurement model, construct a binary matrix $\boldsymbol{P}_i$ with q_i rows and t columns; q_i is the number of observed

[3] My Toolbox (Appendix) includes a program to carry out this test.

indicator variables of L_i and t is the total number of observed indicator variables in the model. Each element (p_{ij}) of $\boldsymbol{P}_i$ has a 1 if the error variables of indicators i and j are d-separated or if the covariance between them has been fixed and if the covariance of the latents associated with indicator variables i and j are free.

Form the matrix $\boldsymbol{D}_i = \boldsymbol{P}_i\boldsymbol{P}'_i$. Iteratively multiply $\boldsymbol{D}_i^{j+1} = \boldsymbol{D}_i\boldsymbol{D}_i^{j'}$ until you get the matrix $\boldsymbol{D}_i^{q_i-1}$.

The first requirement for structural identification is that every element of $\boldsymbol{D}_i^{q_i-1}$ be non-zero in the row corresponding to the indicator of L_i that defines its scale. This must be true for all latent variables in the model.

The second requirement for structural identification of the full measurement model is that, for every pair of latent variables whose covariance is to be estimated (i.e. that are not d-separated or whose covariance is not fixed) there must be at least one pair of indicator variables (one for each latent) whose error variables are independent (i.e. d-separated) or whose covariance is fixed.

The third requirement for structural identification of the full measurement model is that, for every latent variable whose variance is to estimated (i.e. is not fixed), there must be at least one pair of indicator variables (one for each latent) whose error variables are independent (i.e. d-separated) or whose covariance is fixed.

I will now show how this rule works with reference to Figure 6.3.

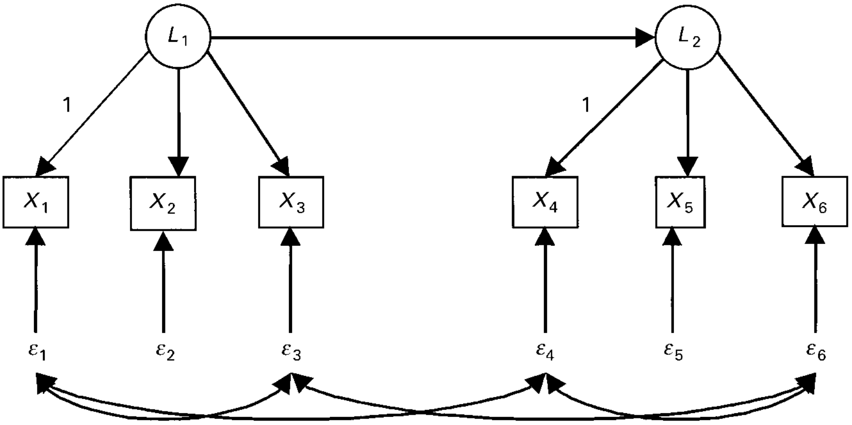


Figure 6.3. A structural equations model used to illustrate the FC1 rule for identification.

In this model there are two latent variables, and so we need two $\boldsymbol{P}$ matrices:

$$\boldsymbol{P}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix} \quad \boldsymbol{P}_2 = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Note that $p_{12} = p_{21} = 1$ in $\boldsymbol{P}_1$ because ε_1 has a fixed (zero) covariance with ε_2 and similarly for ε_4 and ε_5 in $\boldsymbol{P}_2$. There are three indicator variables for each latent and so $q_1 = q_2 = 3$. We must form $\boldsymbol{D}_1^{3-1}$ and $\boldsymbol{D}_2^{3-1}$. These are the same in this example, although this is not true in general:

$$\boldsymbol{D}_1^2 = \boldsymbol{D}_2^2 = \begin{bmatrix} 17 & 20 & 16 \\ 20 & 33 & 20 \\ 16 & 20 & 17 \end{bmatrix}$$

Now, the first requirement is that every element in the row of each matrix representing the scaling variable must be non-zero. The scaling variable of the first latent variable is X_1 and so every element in row 1 of the first matrix must be non-zero. The first part of this first requirement is fulfilled. The scaling variable of the second latent variable is X_4 and so every element in row 1 of the second matrix must be non-zero. The second part of this first requirement is also fulfilled.

The next requirement is that there be at least one pair of error variables (one associated with an indicator of each unique pair of latents) whose covariance is zero or fixed to some other value. Error variables ε_2 and ε_5 fulfil this second requirement.

The final requirement is that there must be at least one pair of error variables (associated with an indicator of each latent) whose covariance is zero or fixed to some other value. Error variables ε_1 and ε_2 fulfil this requirement for the first latent variable and error variables ε_4 and ε_5 fulfil this requirement for the second latent variable. Therefore this measurement model is structurally identified. ┘

6.3 Structural underidentification with structural models

Obtaining identification of the measurement model is necessary to fit a structural equations model. However, SEM also includes the causal relationships between the latent variables. In fact, you can think of the structural model as the 'path' model that is imbedded in the full model. A path model is therefore also a structural model. The rules for ensuring structural identification that I will describe come from Rigdon (1995). Rigdon's rules do not apply to models in which there are cyclic relationships involving more than two variables (for example, if X causes Y causes Z causes X). On the other hand, these rules are both necessary and sufficient for acyclic or block-acyclic structural models; a 'block-acyclic' model is defined below. This means that any acyclic or block-acyclic structural model that satisfies these rules is guaranteed to be structurally identified and any such structural model that does not satisfy these rules is guaranteed to be non-identified.

The first step is to conceptually divide the structural model into segmented blocks. The model is fully segmented when: (i) there are no cyclic relationships between the blocks and (ii) each block contains the minimum number of variables needed to satisfy (i). In other words, if the members of a set of variables in the model do not have a cyclic relationship then each variable defines a separate block. If, on the other hand, a set of variables does define a cyclic relationship (for example A causes B causes C causes A) then they must be included in the same block. If, once this has been done, there are more than two variables in any block then the identification status of the model can't be determined. If this is not the case, then the identification status of the whole structural model can be determined by verifying the identification status of each block. If each block is identified then the whole structural model is also identified[4]. In evaluating these blocks, we don't need to consider the exogenous variables. Figure 6.4 illustrates these points.

In Figure 6.4A there are seven variables. Since variables X_3 and X_4 have a reciprocal (cyclic) relationship they must be included in the same block. Since variables X_5 and X_6 have correlated errors they too must be included in the same block. Variables X_1 and X_2 also have correlated errors and form a block but they are exogenous in this model and so we don't have to worry about them. Finally, X_7 is in a block all by itself. Therefore Figure 6.4A can be decomposed into four blocks, the causal relationships between the blocks have no cyclic patterns, and the model fulfils the requirements for Rigdon's test.

In Figure 6.4B there are three variables (X_1, X_2 and X_3) that possess a feedback relationship. Therefore all three variables must be included in a single block. The last variable, X_4, forms a second block. Because there are more than two variables in one of the blocks, we can't determine the identification status of this model using Rigdon's rules.

Once the structural model has been reduced to these blocks, then you simply have to determine the identification status of each block. To do this, refer to Figure 6.5, which shows eight different patterns. To interpret these diagrams, you will need some notational conventions. The two variables indicated as '1' and '2' are the two variables in the block (if there is only one variable in the block then it is automatically identified). The variables indicated as 'P' are causal parents. If an arrow and a circle as shown with solid lines then the two 'P' variables must be present. If an arrow and circle are shown with broken lines then the two 'P' variables can be present but their existence is irrelevant to determining the identification status of the

[4] Such a model is called a *block-recursive* model.

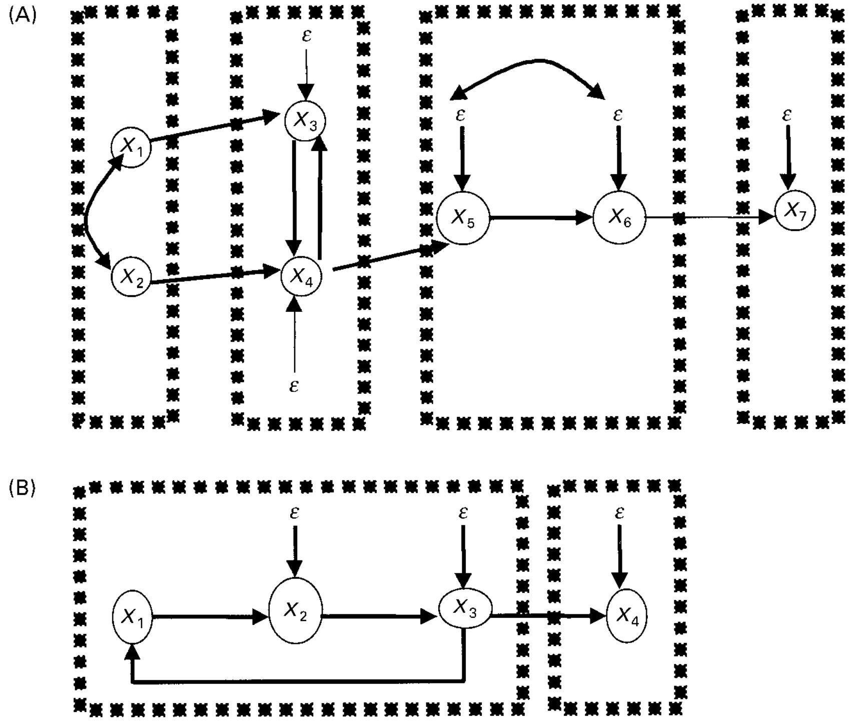


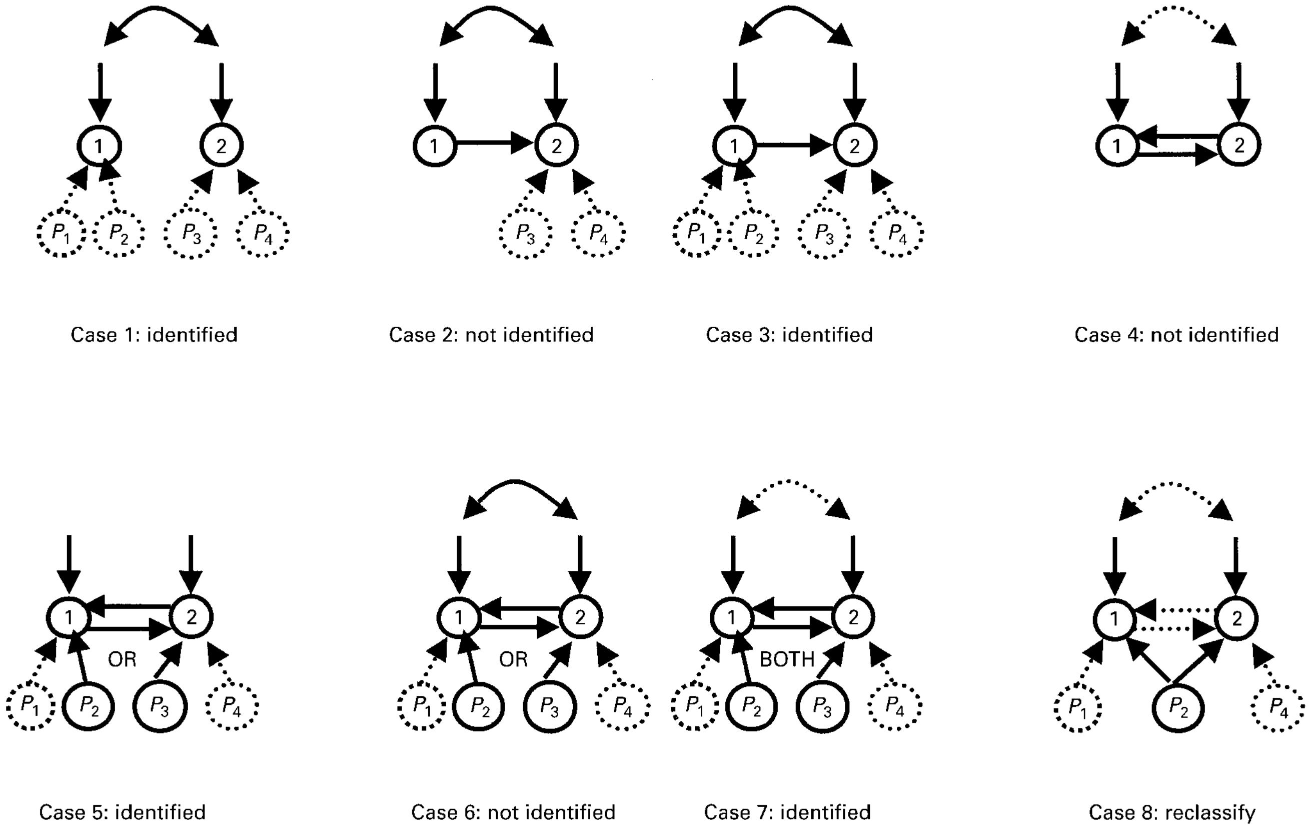
Figure 6.4. An illustration of Rigdon's rules for structural identification with cyclic models. (After Rigdon 1995.)

block. Finally, 'OR' means that at least one of the two '*P*' variables must be present and 'BOTH' means that both '*P*' variables must be present.

Now, look at each block of the structural equation that contains two variables in a cyclic relationship and classify it as belonging to one of the eight cases in Figure 6.5. If any of these blocks are not identified then the model is also not identified. The only complication is with case 8 in Figure 6.5. To determine the identification status of a block belonging to this case, ignore the common causal parent of 1 and 2 and then see which of the other seven cases corresponds to the block while ignoring the common causal parent.

6.4 Behaviour of the maximum likelihood chi-squared statistic with small sample sizes

Many of my in-laws like to make home-made wine. A superficial glance at bottles of these 'wines' might convince you that they are the real thing. When you taste them you realise that they vary along a gradient from 'gut-rot'



Case 1: identified

Case 2: not identified

Case 3: identified

Case 4: not identified

Case 5: identified

Case 6: not identified

Case 7: identified

Case 8: reclassify

Figure 6.5. Eight different cases used to evaluate Rigdon's rules for structural identification. (After Rigdon 1995.)

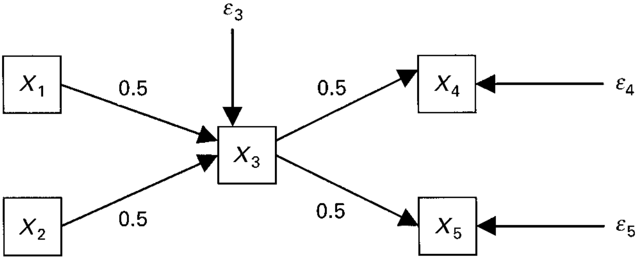


Figure 6.6. The path model used to simulate the data that are summarised in Figure 6.7.

through 'drinkable' to 'divine'. As with latent variables, giving something a name doesn't make it so. The so-called 'maximum likelihood chi-squared statistic' (MLX^2) is the statistical equivalent of home-made wine. It is not really distributed as a chi-squared variate at all and, unfortunately, its true sampling distribution is unknown. However, as the size of the sample of independent observations increases, the sampling distribution of this statistic becomes closer and closer to the theoretical chi-squared distribution. At very small sample sizes the MLX^2 statistic is like gut-rot wine; it bears an approximate resemblance to the true χ^2 distribution but there is no confusing the two. At moderate sample sizes the MLX^2 is like 'drinkable' home-made wine; it is a reasonable approximation of the real thing unless it is to be used for a special occasion. It is only when sample sizes are very large that one cannot distinguish between the two. So how big is 'big enough' and what can be done if one's sample is not big enough? In this section, I discuss the effects of sample size on the MLX^2 assuming that the data follow a multivariate normal distribution. To explore these questions I use simulations drawn from the path model shown in Figure 6.6, with all exogenous variables being drawn from a standard normal distribution (i.e. zero mean and unit standard deviation).

Figure 6.7 shows the empirical sampling distribution of the MLX^2 statistic, based on 1000 independent data sets. I fixed all path coefficients to their theoretical values (0.5) and all the error variances to their theoretical values (1). This way, the only free parameters were the variances of X_1 and X_2, and the model covariance matrix could be determined without iteratively minimising the MLX^2. There were therefore 13 degrees of freedom and the curve shown in Figure 6.7 is the theoretical χ^2 distribution with 13 degrees of freedom. The first histogram shows the distribution of the MLX^2 statistic in the 1000 data sets with 10 observations each. It is clear that this empirical distribution is not well approximated by the theoretical χ^2 distribution; the 95% quantile, corresponding to a 5% significance level,

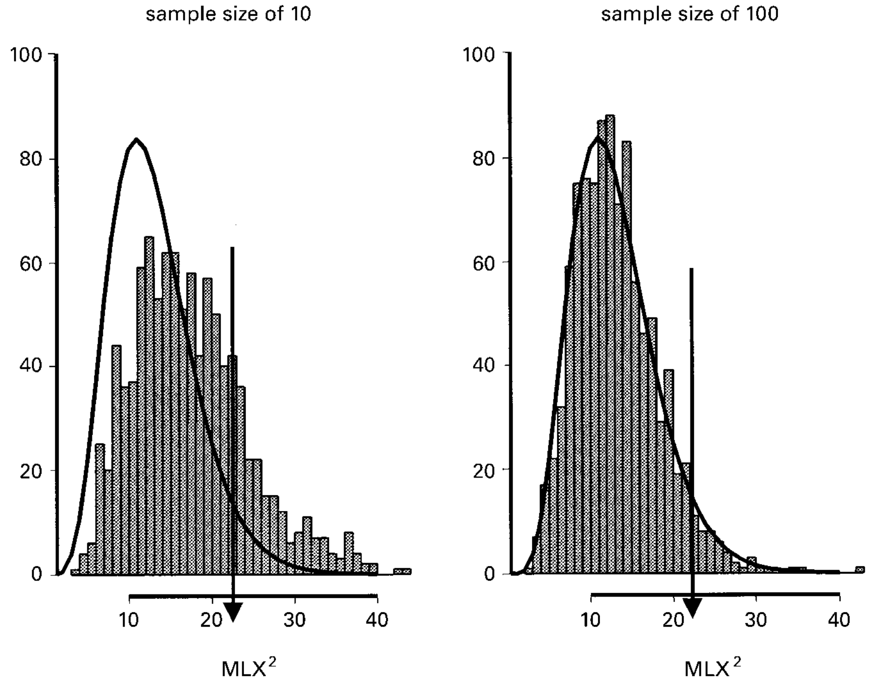


Figure 6.7. Empirical distributions of simulated data based on sample sizes of 10 (left) or 100 (right) observations per data set. The solid curve shows the theoretical chi-squared distribution and the arrow shows the 95% quantile corresponding to the 5% probability level.

is shown by the arrow. The second histogram shows the distribution of the MLX^2 statistic in the 1000 data sets with 100 observations each. The empirical 90%, 95%, 97.5% and 99% quantiles in the second simulation were 19.68, 21.94, 24.33 and 27.13, corresponding to theoretical probabilities of 0.103, 0.056, 0.028 and 0.012. Now, the empirical and theoretical distributions are quite close and, assuming that the MLX^2 is truly distributed as a χ^2 distribution, will introduce little error.

In general, small sample sizes result in conservative probability estimates. In other words, the true probability level will be larger than the value obtained when assuming a χ^2 distribution. If your model produces a MLX^2 value that is judged to be significant using the χ^2 distribution, then you have an ambiguous result and you will have to use a different method of estimating the true probability level. For instance, Table 6.1 shows the empirical quantiles and theoretical probability levels using the model shown in Figure 6.6. For this particular model a sample size of only 30 provides a passable estimate of the tail probabilities but with somewhat conservative probability estimates and a sample size of 50 is quite acceptable. In general, the more

Table 6.1. *Empirical quantiles from 1000 independent data sets, with different numbers of observations ('sample size'), are shown along with the theoretical probability levels assuming a χ^2 distribution with 13 degrees of freedom (see Figure 6.6)*

Sample size	50% quantile	90% quantile	95% quantile	97.5% quantile	99% quantile
10	16.18 (p=0.24)	26.80 (p=0.01)	30.58 (p=0.004)	33.70 (p=0.001)	38.65 (p=0.0002)
30	13.55 (p=0.41)	21.38 (p=0.07)	24.72 (p=0.025)	27.00 (p=0.012)	29.81 (p=0.005)
50	12.79 (p=0.46)	20.48 (p=0.08)	22.47 (p=0.05)	24.43 (p=0.027)	27.52 (p=0.011)

free parameters in the model that need to be estimated, the larger the sample size required. More complicated models may require sample sizes of 200 or more. One rule of thumb is that there should be at least five times more observations than free parameters (Bentler 1995).

What can be done if your sample size is too small to confidently assume that the sampling distribution of the MLX^2 statistic is close to the theoretical χ^2 distribution? If there are no latent variables in your model then you can use the method described in Chapter 3. If there are latent variables then you will need another way. One way is to use bootstrap methods; since this method is also useful in cases where the variables have other distributional problems, the bootstrap will be described later. Another way around the problem of small sample sizes (but not non-normal distributions in general) is to use Monte Carlo methods, as used in the simulations reported in Table 6.1.

The first step is to fit your model using any SEM program, obtain the MLX^2 statistic (call it X) and the degrees of freedom (df). Next, construct a model covariance matrix that has the same number of degrees of freedom as does your model. If your SEM program permits numerical simulations then just specify your original model with model parameters the same as those estimated by the program and whose random values are drawn from normal variates with the specified means and variances. Simulate a large number N (say 1000) data sets, each with the sample size (n) of your original data, following this model. Next fit each simulated data set to your original model with the same pattern of free and fixed parameters and save the calculated MLX^2 values of each run. Finally count the number (x) of these simulated MLX^2 values that are greater than the value (X) obtained in your original data. The proportion x/N will estimate the probability value

(p) that you are looking for. Since these simulated data sets are mutually independent and large in number you can obtain a 95% confidence interval (Manly 1997) around p by referring to a normal distribution whose mean is x and whose variance is $Np(1-p)$. Thus the 95% confidence interval is

$$p \pm 1.96\sqrt{\frac{p(1-p)}{N}}.$$

If your SEM program does not do Monte Carlo simulations, then you can still get an empirical probability estimate so long as you have access to a computer program that can generate standard normal random variates and do simple matrix operations (invert a matrix and calculate a determinant)[5]. Since the MLX^2 statistic requires that we calculate the determinant, and the inverse, of the model covariance matrix, it is useful to choose a matrix for which this can be easily done. The determinant of a square matrix whose non-zero values are all on the diagonal is simply the product of these diagonal values. Similarly, the inverse of such a diagonal matrix is simply a diagonal matrix whose diagonal values are the inverse of the original matrix. We therefore simulate data from a model consisting of v mutually independent variables, each of which is drawn from a standard normal distribution. The predicted covariance matrix, $\boldsymbol{\Sigma}$, of such a model has non-zero values only on its diagonal. There are $v(v+1)/2$ non-redundant elements. If we estimate the variance of q of the v variables, which will be on the diagonal of $\boldsymbol{\Sigma}$, then there will be $v(v+1)/2-q$ degrees of freedom. So, here are the steps needed to estimate an empirical probability level[6] for a MLX^2 statistic of X:

1. Given the desired degrees of freedom (df), find the smallest integer value of v such that $\text{df} \leq v(v+1)/2$. This will be the smallest integer value of v such that $v \geq \frac{-1\sqrt{1+8\text{df}}}{2}$. For example, if $\text{df}=9$ then we need the smallest integer value of v such that $v \geq \frac{-1+\sqrt{1+8(9)}}{2} = 3.8$. Thus $v=4$.
2. Find the integer value of c such that $c = v(v+1)/2 - \text{df}$. So, if $\text{df}=9$ and $v=4$ then $c=1$.
3. Construct a model covariance matrix $\boldsymbol{\Sigma}$ with $v^\star$ rows and columns. Estimate the variances of the first c variables and put these in the first c diagonal elements. Define all other diagonal elements (the remaining variances) to be 1 and all non-diagonal elements to be 0.

[5] Most commercial statistical programs can do this.

[6] My Toolbox (Appendix) contains a program to do this.

This is the population covariance matrix of v mutually independent standard normal variates, of which the variances of the first c of these variables have been estimated from the data. This model covariance matrix will have df degrees of freedom.

4. Now, generate a large number N (say 1000) independent data sets consisting of $v^\star$ mutually independent standard normal random variables with n observations in each data set.
5. For each of the i simulated data sets, calculate the sample covariance matrix $\boldsymbol{S}_i$ and also $\text{MLX}_i^2 = (n-1)(\ln|\boldsymbol{\Sigma}| + \text{trace}(S_i\boldsymbol{\Sigma}^{-1}) - \ln|\boldsymbol{S}_i| - c)$ ('trace' is the trace of the resulting matrix).
6. Count the number (x) of the N MLX^2 values that are greater than the value of the MLX^2 value obtained in your real data (X).
7. The estimated empirical probability of your data will be x/N and the 95% confidence interval of this estimate can be calculated as described before.

6.5 Behaviour of the maximum likelihood chi-squared statistic with data that do not follow a multivariate normal distribution

A biologist, a physicist and a statistician are shipwrecked on a deserted island. Besides themselves, only a crate of canned food has been washed ashore. After staring hungrily at the cans for a number of hours, the biologist suggests that they break open some cans with a large rock. The physicist suggests instead that they climb to just the right height in a palm tree. She explains that the kinetic energy, as the can hits the ground, should be just enough to crack it open without losing any food. Glancing over to the statistician, who has just finished writing some equations in the sand, they see him shaking his head in disapproval at their crude methods. He announces that he has just found a more elegant method of opening the cans, and points proudly at his equations. 'Now', he begins, pointing to the first equation, 'assume that we have a can opener . . .'.

Sometimes we don't have the statistical equivalent of a can opener. Knowing the assumptions of a statistical test is important but knowing what might happen if the assumptions are wrong can be just as important. Another assumption of the maximum likelihood chi-squared statistic is that the data follow a multivariate normal distribution[7]. We require methods of both testing and relaxing this assumption. First, let's look at how to test for a departures from multivariate normality.

[7] Actually, the assumption is that the endogenous variables follow a multivariate normal distribution. Exogenous variables (i.e. ones that are not caused by any others in the model) don't have this restriction (Bollen 1989).

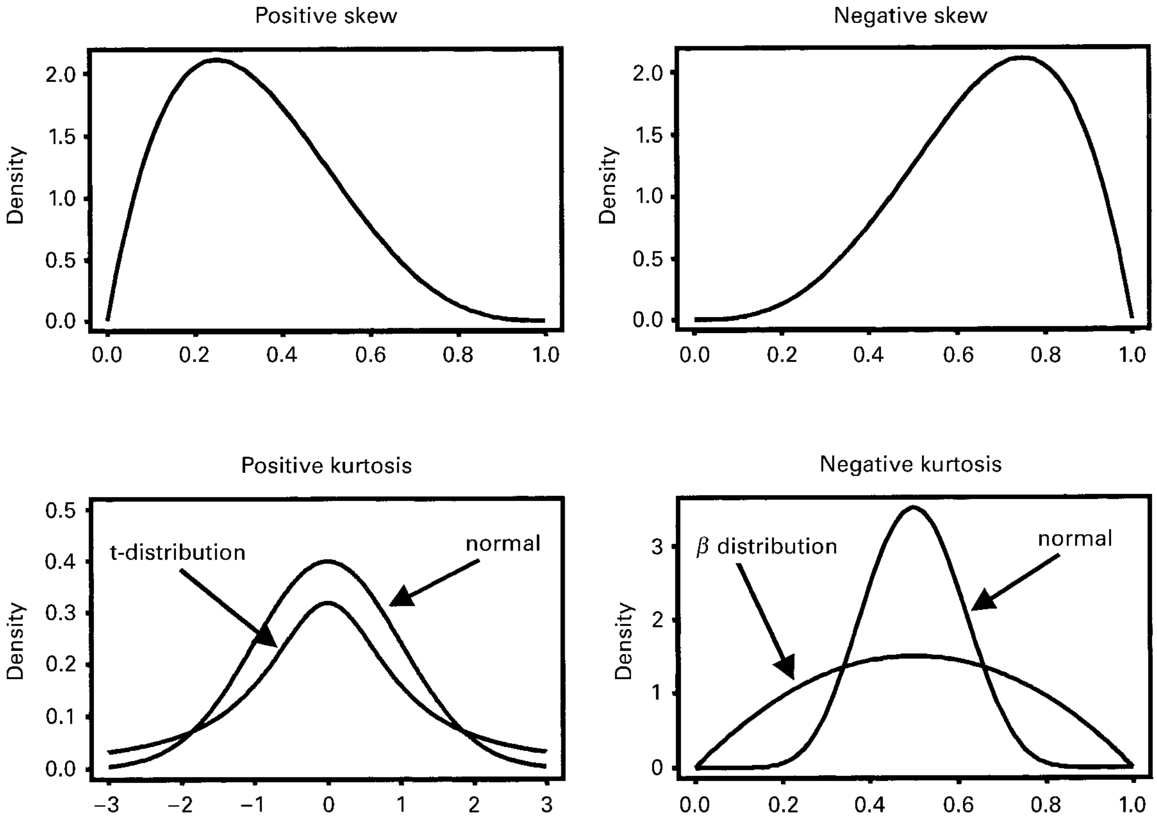


Figure 6.8. Examples of curves showing positive or negative skew and kurtosis.

The normal distribution is fully characterised by its mean and variance. Departures from normality can be characterised by non-zero skew and kurtosis. The skew measures the degree of asymmetry of the distribution. A negative skew occurs when a univariate distribution has a longer tail to the left and whose mode is to the right of centre. A positive skew occurs when a univariate distribution has a longer tail to the right and whose mode is to the left of centre (Figure 6.8). An index of skew for a series of N observations of a random vector X is:

$$g_{11} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^3}{Ns^3}$$

where s is the standard deviation of X. Box 6.2 summarises the calculations for using this statistic to test for skew in large (>150) sample sizes. D'Agostino, Belanger and D'Agostino (1990) provided more exact formulae that can be used for sample sizes as low as 8.

Kurtosis measures the concentration of values near the mean and at the extremes, relative to intermediate values. For symmetrical unimodal distributions, positive kurtosis indicates heavy tails and peakedness relative to the normal distribution and negative kurtosis indicates light tails and flatness (DeCarlo 1997). A familiar distribution with positive kurtosis is Student's t-distribution. With this distribution the kurtosis increases as the degrees of freedom decrease; the graph on the bottom left of Figure 6.8 shows a t-distribution with 3 degrees of freedom as well as a standard normal distribution. An index of kurtosis is:

$$g_{21} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^4}{Ns^4}$$

Box 6.2 summarises the steps for identifying significant deviations from normality with respect to kurtosis.

Box 6.2. Measures of skew and kurtosis

Univariate skew

The expected value of g_{11} for a normal distribution is 0. The following statistic is approximately distributed as a standard normal variate for large ($N > 149$) sample sizes, and values greater than 1.96 in absolute value would indicate skew:

$$z = g_{11}\sqrt{\frac{(N+1)(N+3)}{6(N-2)}}$$

For tests applicable to small samples, the reader is directed to D'Agostino, Belanger and D'Agostino (1990) or Bollen (1989).

Univariate kurtosis

The expected value of g_{21} is 3 for normally distributed variables. For this reason, many computer programs often report the centred version of the g_{21} statistic ($g_{21} - 3$) even though this is not always well documented. The following statistic follows a standard normal distribution only in very large samples ($N > 1500$) but at least provides a rough guide. A more complicated statistic, applicable to small sample sizes ($N > 19$) is given by D'Agostino, Belanger and D'Agostino (1990) or Bollen (1989).

$$\mathrm{E}(g_{21}) = \frac{3(N-1)}{(N+1)},\ \mathrm{Var}(g_{21}) = \frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)} \text{ and } z = \frac{(g_{21} - \mathrm{E}(g_{21}))}{\sqrt{\mathrm{Var}(g_{21})}}$$

Many SEM programs report either the standardised (z) or the asymptotic centred value of kurtosis ($g_{21} - 3$) as a benchmark for normality. In using these tests on all variables in your model, you should use a Bonferonni correction to the significance levels. If you want to test at an overall level of α, then test each of the V variables at a level of α/V. For instance, if you want to test at a 95% level ($\alpha = 0.05$), then test each variable at a level of $0.05/V$.

Multivariate measures of skew and kurtosis

The above measures of skew and kurtosis are applied separately to each variable. Since it is possible for the joint distribution to have skew or kurtosis even though each individual variable shows no evidence of this, Mardia (1970, 1974) developed multivariate analogs of these statistics. They are based on a matrix of squared Mahalanobis distances. For a single variable (X_i) with N observations, the squared Mahalanobis distance is simply

$$\frac{1}{\sigma_i^2}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2$$

If each of the j observations consists of a series of V variables then the resulting data set, $\boldsymbol{X}$, has N rows and V columns. The squared Mahalanobis distance matrix for the entire data set is: $\boldsymbol{X'SX}$, where $\boldsymbol{S}$ is the covariance matrix of $\boldsymbol{X}$. Looking at the Mahalanobis distance for each observation helps to identify outliers in the multivariate space. Based on the squared Mahalanobis distance, Mardia's multivariate measure of skew with V variables is:

$$g_{1v} = \left(\frac{1}{N^2}\right) \sum_{i=1}^{N} \sum_{j=1}^{N} (\boldsymbol{X'SX})^3$$

If the data follow a multivariate normal distribution then the expected value of this statistic is 0. The statistic $N \times g_{1v}/6$ asymptotically follows a chi-squared distribution with $V(V+1)(V+2)/6$ degrees of freedom if the data are multivariate normal. Mardia's multivariate measure of kurtosis is:

$$g_{2v} = \left(\frac{1}{N}\right) \sum \mathrm{trace}((\boldsymbol{X'SX})^2)$$

where 'trace' means the diagonal elements. If the data follow a multivariate normal distribution then the expected value of $g_{2\mathrm{v}}$ is $V(V+2)$ and the variance is $8V(V+2)/N$. The statistic

$$\frac{(g_{2v} - v(v+2))}{\sqrt{8v(v+2)\,/N}}$$

asymptotically follows a standard normal distribution. Bollen (1989) provides more complicated test statistics that are applicable to small data sets.

If any of the variables in your model have significant skew or kurtosis then the joint multivariate distribution also has significant skew or kurtosis. However, it is possible for the multivariate distribution to have either skew or kurtosis even though each variable, taken singly, is normally distributed. This means that we also require a multivariate version of our measures of skew and kurtosis. Such measures, along with their tests, are given by Mardia (1970, 1974). The calculations are explained in Box 6.2.

Most biologically oriented statistics texts describe the Box–Cox method of choosing a transformation to make data more closely follow a normal distribution. This is because statistical tests involving means (t-tests, ANOVA, etc.) are more sensitive to skew and the Box–Cox method helps to reduce skewness in data. However, tests involving variances and covariances, such as those used in SEM, are more sensitive to kurtosis than to skew (Mardia, Kent and Bibby 1979; Jobson 1992). Jobson (1992) described a modified power transformation that is designed to reduce kurtosis:

$$Y = \mathrm{SIGN}\frac{(|X - X_{\mathrm{M}}| + 1)^{\lambda} - 1}{\lambda} \qquad \lambda \neq 0$$

$$Y = \mathrm{SIGN}\ \ln(|X - X_{\mathrm{M}}| + 1) \qquad \lambda = 0$$

SIGN is the sign of the original value of $(X - X_{\mathrm{M}})$ and X_{M} is the median value of X.

To find the value of λ that reduces the kurtosis best, I calculate the sum of squared differences between a series of quantiles of Y (say, 5%, 10%,

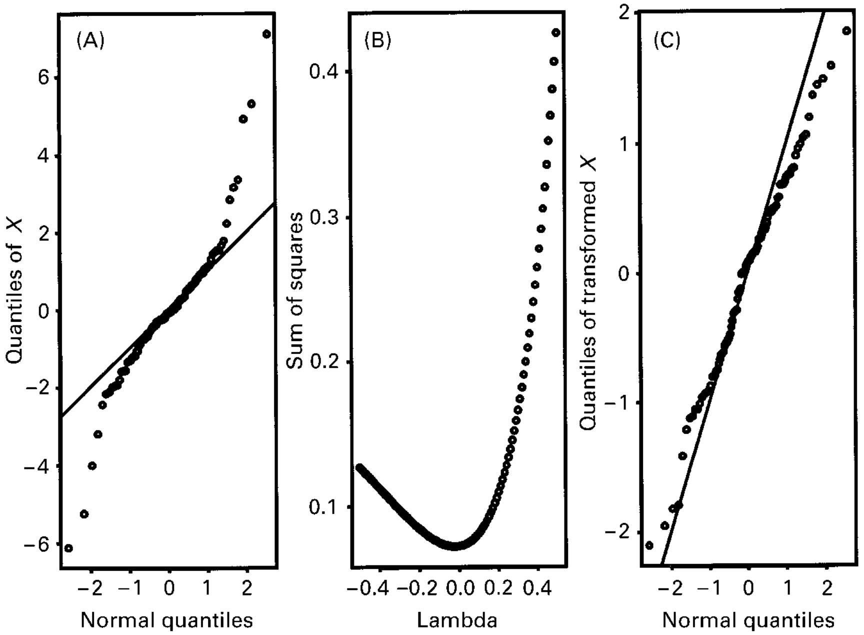


Figure 6.9. (A) A normal quantile plot for 100 values drawn from a t-distribution with 3 degrees of freedom. (B) The value of the sum of squared distances between the quantiles of these values and those of a standard normal distribution for various values of λ using Jobson's transformation. (C) The normal quantile plot after transforming the data using the best value of λ.

50%, 90% and 95%) for different values of λ and the quantiles of a normal distribution with the mean and standard deviation[8] of *Y*. The value of λ that minimises this sum of squared differences will best reduce the kurtosis of the original values. Most statistics packages allow one to plot the empirical quantiles (cumulative percentage) against these theoretical quantiles. Using these graphs, you can try different values of λ and choose the one in which the resulting graph looks most like a straight line. Figure 6.9A shows a quantile plot for 100 values drawn from a t-distribution with 3 degrees of freedom. These values have a centred kurtosis of 4.12, the standardised value is 9.06 and the probability of this occurring in a normally distributed variable is 6.6×10^{-5}. Notice the large deviations in the tails (the extreme values). Figure 6.9B shows the value of the sum of squared distances

[8] Alternatively, you can standardise your variable first and then refer to the quantiles of a standard normal distribution. These quantiles can be found in most tables of the standard normal distribution.

between the quantiles of these values and those of a standard normal distribution, as described above, for various values of λ between 0 and 1. The best value of λ is around 0, thus demanding a ln-transformation. Figure 6.9C shows the quantile plot for the values transformed using $\lambda = 0$. The centred kurtosis of this transformed variable is -0.10, the standardised value is -0.23 and the probability of observing such a kurtosis in a normally distributed variable is 0.95.

Non-normality can affect the accuracy of the maximum likelihood chi-squared statistic, the standard errors of the free parameters and the estimation of the free parameters themselves. As you might expect, researchers have spend a good deal of effort in exploring how different types and degrees of non-normality affect these statistics. To get an idea of robustness of the maximum likelihood chi-squared statistic, I again generated data from the model shown in Figure 6.6. The path coefficients were fixed at 0.5. The exogenous variances (i.e. the variances of X_1, X_2, ε_3, ε_4 and ε_5) were generated from different probability distributions with different degrees of skew and kurtosis: the standard normal distribution, a t-distribution with 3 degrees of freedom, a beta distribution with shape parameters of 2 and 5, a chi-squared distribution with 2 degrees of freedom and a uniform distribution between 0 and 1. The t-distribution has no skew but strong positive kurtosis. The uniform distribution has no skew but strong negative kurtosis. The chi-squared distribution with 2 degrees of freedom has both strong positive skew and kurtosis. The beta distribution with shape parameters of 2,5 has positive skew and negative kurtosis. For each distributional type, I simulated 1000 independent data sets of 50, 100 or 200 observations. Average values of Mardia's multivariate estimates of skew and centred kurtosis for these data were also estimated. The results are shown in Table 6.2.

The first thing to notice is that distributions with strong kurtosis produce conservative probability levels; there is a tendency for models to be rejected more often than they should be when one is assuming a theoretical χ^2 distribution. This is the same result as we saw for small sample sizes. The second thing to notice is that even models generated from very non-normal distributions produce quite acceptable probability levels as sample sizes increase. This shows the asymptotic robustness of the maximum likelihood chi-squared test statistic. If the errors are distributed *independently* of their non-descendants in the model, the test statistic should asymptotically follow a chi-squared distribution. The robustness conditions hold under independence and not necessarily under 'uncorrelatedness'[9]; for example, if the variance of the error variable changes systematically with respect to any

[9] I am sorry for using such an ugly word.

Table 6.2. *Simulation results (1000 data sets each) based on sample sizes of N with exogenous variables drawn from a standard normal, a t-distribution with 3 degrees of freedom, a beta distribution with shape parameters of 2 and 5, a uniform distribution between 0 and 1, and a χ^2 distribution with 2 degrees of freedom. Shown are the 50, 90, 95 and 97.5% quantiles of the 1000 maximum likelihood statistics for each simulation as well as the theoretical probabilities assuming a χ^2 distribution with 10 degrees of freedom. The average of Mardia's multivariate centred estimate of kurtosis is also shown*

		Quantiles (theoretical probability)				
N	Type	50	90	95	97.5	kurtosis
50	Normal	10.00 (0.440)	16.58 (0.084)	18.73 (0.044)	21.17 (0.020)	−1.26
50	t(3df)	9.72 (0.465)	17.87 (0.057)	20.89 (0.022)	24.81 (0.006)	17.83
50	Beta(2,5)	9.92 (0.448)	17.00 (0.074)	19.20 (0.038)	21.23 (0.020)	−1.51
50	χ^2(2 df)	9.91 (0.448)	17.24 (0.069)	20.11 (0.028)	23.41 (0.009)	11.98
50	Uniform	9.48 (0.487)	16.59 (0.084)	19.92 (0.030)	22.04 (0.015)	−5.38
100	Normal	9.45 (0.490)	15.87 (0.103)	18.34 (0.050)	20.76 (0.023)	−0.70
100	t(3df)	9.25 (0.509)	16.98 (0.075)	19.48 (0.035)	22.16 (0.014)	34.37
100	Beta(2,5)	9.68 (0.469)	16.43 (0.089)	19.10 (0.039)	20.79 (0.023)	−1.13
100	χ^2(2 df)	9.30 (0.504)	17.09 (0.072)	19.48 (0.035)	22.09 (0.015)	17.86
100	Uniform	9.24 (0.509)	16.35 (0.090)	18.53 (0.047)	20.09 (0.028)	−5.68
200	Normal	9.38 (0.497)	16.19 (0.094)	18.50 (0.047)	20.67 (0.024)	−0.17
200	t(3df)	9.25 (0.508)	16.45 (0.087)	19.61 (0.033)	22.74 (0.012)	64.24
200	Beta(2,5)	9.46 (0.489)	16.42 (0.088)	18.44 (0.048)	20.77 (0.023)	−0.76
200	χ^2(2 df)	9.64 (0.472)	16.67 (0.082)	18.37 (0.049)	22.04 (0.015)	23.51
95% confidence intervals for probabilities		0.531 to 0.479	0.119 to 0.081	0.064 to 0.036	0.035 to 0.015	

of its causal parents then this would undermine independence[10]. The robustness of the maximum likelihood chi-squared statistic depends on many different attributes of the model and the data: the number of free parameters, the distributional properties of each variable and (especially) non-independence of the error variables with respect to their non-descendants in the model[11].

6.6 Solutions for modelling non-normally distributed variables

Since non-normality can cause problems with the maximum likelihood chi-squared statistic, a number of alternative ways of fitting the model have been devised. Most commercial SEM programs will include statistics based on generalised least squares, elliptical estimators and distribution-free estimators, as well as a method of correcting for non-normality that produces 'robust' chi-squared statistics and confidence intervals[12]. The most popular, and best studied, correction method comes from Satorra and Bentler (1988).

There now exists an extensive literature that uses Monte Carlo methods to explore the relative merits of these different solutions for non-normality. Different studies have explored the effects of sample size, the number of free parameters, model type (measurement models, path models, full structural models) and distributional violations (kurtosis, skew and non-independence of errors and their causal non-descendants). Hoogland and Boomstra (1998) have done a meta-analysis of these studies. Their main recommendations are the following:

1. With respect to sample size, they recommend that there be at least five times as many observations as there are degrees of freedom in the model.
2. When the observed variables have an average positive kurtosis of 5 or more, the sample size may have to be increased by up to 10 times the degrees of freedom.
3. The generalised least squares chi-squared statistic has an acceptable performance for a sample size that is two times smaller than the sample size needed for an acceptable performance of the maximum likelihood chi-squared statistic.

[10] This is similar to heteroscedastic error variances in ordinary regression.

[11] Except, of course, when the non-independence is explicitly modelled.

[12] Another correction method is found in Browne (1984). This consists of dividing the maximum likelihood chi-squared statistic by the ratio of Mardia's multivariate measure of kurtosis to its expected value given normality. Little simulation work seems to have been done on this correction.

4. With small samples the standard errors of the estimates of the free parameters are biased. Positive kurtosis results in estimates of the standard errors that are smaller than they should be. Negative kurtosis results in estimates of the standard errors that are larger than they should be.
5. The degree of skew has little effect on the bias of the estimators.
6. The asymptotic distribution-free estimator should not be used except for very large sample sizes (>1000).
7. The Satorra–Bentler robust estimator, upon which is based their robust (S–B) chi-squared statistic and standard errors, largely corrects for excessive kurtosis and for problems in which the errors are not independent of their causal non-descendants. This is particularly important for models that include latent variables and measurement models, since the S-B chi-squared statistic can correct for cases in which the latent variables and the measurement errors are not independent.

Basically, unless your data are very strongly kurtotic and your sample sizes are very low, you can still perform a reasonable test of your causal model. As a last resort, you can use bootstrap methods (Bollen and Stine 1993). The bootstrap has been included in some commercial SEM programs, and most will soon have this option. Note, however, that the original data must be transformed using a technique called a Cholesky factorisation[13] and not all commercial SEM programs implement this step. The bootstrap is related to Monte Carlo methods except that, rather than sampling from some theoretical distribution (multivariate normal or otherwise), you sample from your own data to build up an empirical sampling distribution. See Manly (1997) for a discussion of bootstrap methods in biology. Box 6.3 summarises the steps required to generate a bootstrap distribution of the maximum likelihood chi-squared statistic. Note, however, that this method is very computer-intensive.

Box 6.3. Bootstrapping the sampling distribution

Here are the steps to take in order to generate a bootstrap sampling distribution and perform an inferential test.

1. Given your original data set (Y) with N rows and p variables centred about their means, calculate the sample covariance matrix ($\boldsymbol{S}$), obtain

[13] See Press *et al.* (1986) for a description of the Cholesky factorisation of a square positive definite matrix and numerical algorithms to calculate this.

the predicted model covariance matrix ($\boldsymbol{\Sigma}$) and the maximum likelihood chi-squared statistic, MLX^2.

2. Calculate the Cholesky factorisation of $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ to give $\boldsymbol{S}^{-1/2}$ and $\boldsymbol{\Sigma}^{1/2}$.
3. Form a new data set: $Z = Y\boldsymbol{S}^{-1/2}\boldsymbol{\Sigma}^{1/2}$.
4. Randomly choose N observations from Z with replacement to form a bootstrap sample $Z^{\star}$. Form the covariance matrix from this bootstrap sample, fit the model to these data, and save the bootstrap value of the maximum likelihood chi-squared statistic (MLX^{2*}).
5. Repeat step 4 a large number of times (at least 1000).
6. Count the proportion of times that MLX^{2*} is greater than MLX^2. This proportion is the empirical estimate of the probability of observing the data given the model. Note that this probability does not assume any particular sampling distribution.

6.7 Alternative measures of 'approximate' fit

This section deals with various methods of assessing the degree of 'approximate' fit between data and a theoretical model. I don't like these methods and don't advise you to use them either, for reasons that I will explain below. However, they are popular with many users of SEM and are always printed out in commercial SEM programs. These measures of approximate fit are generally used once the model has already been rejected and the purpose of these approximate fit measures are to determine the degree to which the rejected model is 'approximately' correct.

The origin and rationale behind the use of these approximate fit indices comes from a consideration of statistical power. The power of a statistical test can be defined as the probability that the test will reject the null hypothesis when it is indeed false. To illustrate this notion, imagine that we wish to test the null hypothesis that two random variables, X and Y, are uncorrelated (H_0:$\rho = 0$). I generated 100 independent data sets each with 10, 50, 100 or 500 observations in which the true population correlation coefficient was either 0, 0.1, 0.2, 0.3, 0.4 or 0.5. We know that, if H_0 is true then we should reject about 1 out of 20 tests at the $\alpha = 0.05$ level. If we had perfect statistical power then we should reject all data sets in which ρ is different from 0. In other words, we should reject a proportion α when the null hypothesis is correct and proportion 1 whenever ρ deviates, however slightly, from 0. Figure 6.10 shows the actual proportion of the 100 data sets for which the null hypothesis ($\rho = 0$) was rejected at $\alpha = 0.05$.

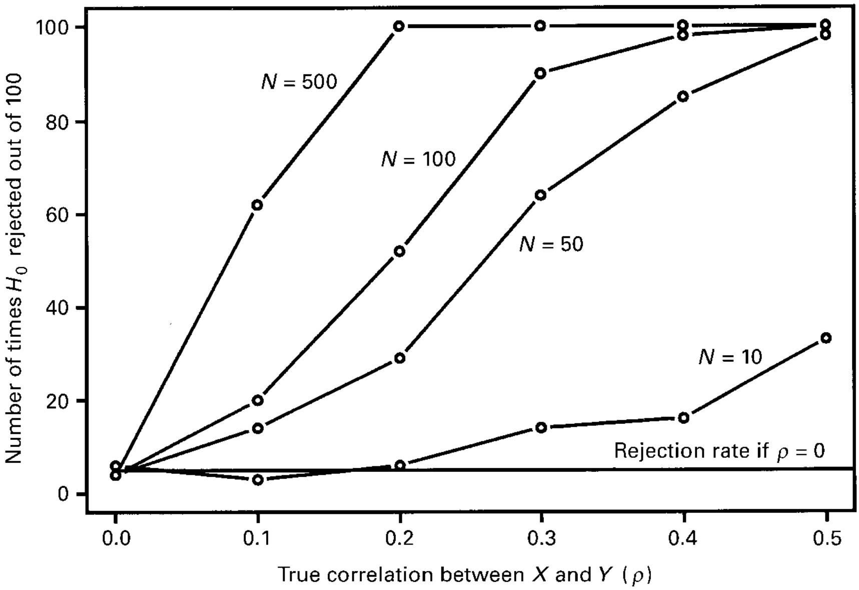


Figure 6.10. The proportion of the 100 data sets of various sample sizes (N) for which the null hypothesis H_0 ($\rho = 0$) was rejected at $\alpha =$ 0.05 when the true population value of the correlation coefficient took various values between $\rho = 0$ and $\rho = 0.5$.

In Figure 6.10 we see that when the sample size is very small ($N =$ 10) then even when the null hypothesis is false (i.e. when the true correlation between X and Y is not zero) the null hypothesis won't be rejected a large proportion of the time; even when $\rho = 0.5$ only 33 out of 100 tests[14] rejected the null hypothesis that $\rho = 0$. As the number of observations per data set increases then the number of times that the test correctly rejects the hypothesis that $\rho = 0$ also increases. The curves in Figure 6.10 are called power functions and the proportion of times that a test will reject H_0:$\rho = 0$ when, in fact $\rho = \theta$, is called the *power* of the test. From Figure 6.10 we can see that if we have 50 observations then the test has at least a 90% chance of rejecting our null hypothesis (thus, a power of 0.9) when ρ is greater than about 0.5. If we have 100 observations then we have a power of 0.9 as soon as ρ is greater than about 0.3 and if we have 500 observations then we have a power of 0.9 as soon as ρ is greater than about 0.16. In other words, as the sample size increases we have a greater and greater chance of detecting a

[14] This is a simple example of why failing to reject a null hypothesis is not the same as showing that it is true.

smaller and smaller difference between the hypothesised value and the true value. At very large sample sizes even minuscule differences (say $\rho = 0.01$) will almost surely be detected and the null hypothesis would be rejected almost always.

Usually therefore, more power is a good thing. Tests of structural equations models, based on the chi-squared distribution, also have power properties. The justification for using alternative tests of fit is based on the premise that statistical power is not always such a good thing. If you remember the section in Chapter 2 dealing with the logic of inference in science then you will recall that no hypothesis is ever really tested in isolation. Every hypothesis contains within it many other auxiliary hypotheses. In the context of testing structural equations models with reference to a chi-squared distribution we are really interested in knowing whether the causal structure of the model is wrong. Unfortunately, when we conduct our statistical test we are testing all aspects of the model: the causal implications, the distributional properties of the variables, the linearity of the relationships and so on. Now, when we add the notion of statistical power to our argument we realise that, as sample size increases, we run a greater and greater risk of rejecting our models because of very minor deviations that might not even interest us. This point was raised early in the history of modern SEM by Jöreskog (1969).

What might these uninteresting and minor deviations be? They can't be minor deviations from multivariate normality, since the maximum likelihood chi-squared statistic is asymptotically robust against non-normality. In any case, we have already seen ways of dealing with this. Small amounts of non-linearity could be one such minor deviation that would not interest us. If some parameter values (for instance, path coefficients or error variances) are fixed to non-zero values in our model then small deviations from these fixed values might be another minor difference that would not interest us. For instance, we might have only a single indicator of some latent variable whose error variance we fix at 1.1, perhaps based on previous experience. If the true error variance of this indicator was 1.15 and we use a large enough sample size, then our model would be rejected. However, the principal 'minor deviation' that is evoked in the justification for measures of approximate fit is a minor deviation in the causal structure of the model. The theoretical objective of the various indices of approximate fit is therefore somehow to quantify the degree of these deviations. The various alternative fit indices attempt to quantify the degree of such deviations by measuring the difference between the observed covariance matrix and the predicted (model) covariance matrix. The most popular fit indices do this in a way that standardise for differences in sample size.

At first blush then, these indices of approximate fit have a seductive quality. Wouldn't it be nice, after having found that one's preferred causal explanation (as translated by the structural equations model) has been rejected, to be able to say: 'but it is almost right! The remaining lack-of-fit is only due to minor errors that are not really very important anyway.' This, I suspect, is the real (psychological) objective of these fit indices. Even this weakness of the flesh could be tolerated if there were any justification for the implicit assumption that minor errors in specifying the causal structure will translate into only minor differences between the observed and predicted covariance matrices. Unfortunately, no such one-to-one relationship has ever been demonstrated for these indices of approximate fit. To me, evoking such an argument of approximate fit to justify accepting a causal model is like the old joke about the drunk in the parking lot[15]. The alternative fit indices measure different aspects of the ability of the *observational* model (the structural equations) to *predict* the data, not the *explanatory* ability of the *causal* model. As such, the indices of approximate fit commit the sort of subtle error of causal translation that I discussed in Chapter 3: small (but real) differences between the observed and predicted covariances of the observational model do not necessarily mean only small (but real) differences between the actual causal structure and the predicted causal structure.

Now that I have given my reasons why you should not use these alternative fit indices, you can read the justifications of those who promote them and decide for yourself (Bentler and Bonnett 1980; Browne and Cudeck 1993; Tanaka 1993). Below, I describe two of the more popular alternative fit indices. The book by Bollen and Long (1993) contains a number of chapters that deal with these alternative indices of approximate fit.

6.8 Bentler's comparative fit index

Let's go back to the maximum likelihood chi-squared statistic for a moment. This statistic, and its inferential test, measure exact fit between the observed and predicted covariance matrices. The logic is that if the data are generated by the process specified by the structural equations (and therefore the causal structure of which these equations are a translation) then the observed and predicted covariance matrices will be identical except for random sampling variation. If this assumption is true then the maximum likelihood chi-squared

[15] You enter a parking lot late at night and see a drunk causal modeller on his knees underneath the only street light. He explains that he is looking for his car keys. 'Are you sure that you lost your keys here?' you ask. 'No', he answers. 'In fact, I have no idea where they are, but at least here I have enough light to see'.

statistic will asymptotically follow a chi-squared distribution with the appropriate degrees of freedom (υ). Actually, it is more precise to say that this statistic will asymptotically follow a *central* chi-squared distribution (χ^2_v) with the appropriate degrees of freedom. The central chi-squared distribution is a special case of a more general chi-squared distribution called the *non-central* chi-squared distribution. The non-central chi-squared distribution ($\chi^2_{v,\lambda}$) has two parameters: the degrees of freedom (υ) and the non-centrality parameter (λ). A central chi-squared distribution is simply a non-central chi-squared distribution whose non-centrality parameter (λ) is zero.

Now, if the degree of mis-specification of the model covariance matrix is not zero (as assumed in the test for exact fit) but is small relative to the sampling variation in the observed covariance matrix, then the maximum likelihood chi-squared statistic actually asymptotically follows a non-central chi-squared ($\chi^2_{v,\lambda}$) distribution and the non-centrality parameter (λ) measures the degree of mis-specification. The expected value of the non-central chi-squared distribution is simply the expected value of the central chi-squared distribution plus the non-centrality parameter: $E[\chi^2_{v,\lambda}] = E[\chi^2_v] + \lambda = \upsilon + \lambda$. In practice, the non-centrality parameter is estimated as the value of the maximum likelihood chi-squared statistic (MLX^2) minus the degrees of freedom of the model (i.e. the expected value that the maximum likelihood chi-squared statistic would have if there were no errors of mis-specification). Because the non-centrality parameter can't be less than zero, negative values are replaced with zero. Therefore $\lambda = \max\{(\mathrm{MLX}^2 - \nu),0\}$.

The Bentler comparative fit index uses this fact to measure by how much the proposed model has reduced the non-centrality parameter (thus, the degree of mis-specification) relative to a baseline model. The most common baseline model is one that assumes that the variables are mutually independent. If λ_{i} is the estimate of the non-centrality parameter for the model of interest and λ_0 is the estimate of the non-centrality parameter for the baseline model, then the comparative fit index is defined as:

$$\mathrm{CFI} = \frac{\lambda_0 - \lambda_i}{\lambda_0}$$

If the model of interest fit exactly then the expected value of its non-centrality parameter (λ_i) would be zero and the CFI value would be 1.0. Therefore, the CFI index varies from 0 (the proposed model fits no better than the baseline model) to 1.0. The sampling distribution of this index is unknown and users of this index consider a value of at least 0.95 as being an acceptable 'approximate' fit. There is no theoretical justification for this value; it is simply a rule of thumb.

Actually, the description above is for the sample-based CFI. Although this is the index usually reported in most commercial SEM programs, it is known that the sample-based CFI is a biased estimator of the population-based CFI. The result of this bias is to exaggerate the degree of misfit. Steiger (1989) explained how to calculate the unbiased estimator of the population CFI from the information provided by most commercial programs. If p is the number of observed variables in the model, df is the degrees of freedom and n is the sample size then first get the model fit index, or calculate it as:

$$\hat{F} = \frac{(X^2 - \text{df})}{n - 1}$$

The unbiased estimator of the population CFI is

$$\frac{p}{p + 2\hat{F}}$$

6.9 Approximate fit measured by the root mean square error of approximation

A second measure of approximate fit was developed by Steiger (1990) and expanded by Browne and Cudeck (1993). This measure also relies on the non-centrality parameter. The root mean square error of approximation (RMSEA, ε) is defined as:

$$\varepsilon = \sqrt{\frac{\lambda}{n\nu}} = \sqrt{\frac{\max\{\text{MLX}^2 - \nu),0\}}{n\nu}}$$

where λ is the non-centrality parameter and ν is the degrees of freedom of the model. If we propose a null hypothesis for the RMSEA (H_0: $\varepsilon_a \leq a$) then we can test this hypothesis using the non-central chi-squared distribution and produce confidence intervals around it. If your favourite statistics program doesn't have this probability distribution then you can use the algorithm given by Farebrother (1987). Of course, if your null hypothesis is that $\varepsilon_a = 0$ then you are doing a test of exact fit with reference to the central chi-squared distribution. Here are the steps:

1. Specify the null hypothesis H_0: $\varepsilon_a \leq a$
2. Calculate the maximum likelihood chi-squared statistic (MLX^2) and the non-centrality parameter $\lambda^\star = n \times \nu \times \varepsilon_a^2$.
3. Find the probability of having observed MLX^2 given a non-central chi-squared distribution with parameters $\nu, \lambda^\star$.

4. If the probability is less than your chosen significance level, reject the null hypothesis and conclude that ε_a is greater than that specified in the null hypothesis.

An obvious problem with this test is in choosing the null hypothesis. Remember that these indices of approximate fit are used when one has already rejected the null hypothesis of exact fit (i.e. $\varepsilon_a = 0$). We already know that there is something wrong with the model. Browne and Cudeck (1993) recommend the null hypothesis of $\varepsilon_a \leq 0.05$, but this is only their rule of thumb. Again, there is no compelling reason for choosing this value as a reasonable level of 'approximate' fit.

Quite apart from using the RMSEA to measure 'approximate' fit, there is a very useful property of the inferential test for this fit statistic. If we have not been able to reject our model then it is still important to be able to estimate a confidence interval for the RMSEA. In such a case the confidence interval will have a lower bound of zero. The upper bound will reflect the statistical power of our test. A large upper bound indicates that the test had little statistical power to reject alternative models. A 90% confidence interval for RMSEA would not reject the null hypothesis of exact fit at the 5% level. This interval can be calculated as the values of λ for a non-central chi-squared distribution whose 5% and 95% quantiles equals the calculated MLX^2 statistic (Browne and Cudeck 1993). These confidence intervals, and a whole plethora of approximate fit statistics, are provided by most SEM programs.

6.10 An SEM analysis of the Bumpus House Sparrow data

Natural selection was in the air during the last decade of the nineteenth century. According to Bumpus (1899) natural selection was literally *in the air* during a New England snow and ice storm one cold night. Many House Sparrows (*Passer domesticus*) were immobilised during that storm and 136 of the unfortunate birds were collected and transported to the Brown University Anatomical Laboratory. Seventy two birds (51 males and 21 females) subsequently recovered but 64 birds (36 males and 28 females) died. Bumpus determined the sex of all 136 birds and also measured nine phenotypic attributes of each bird, alive or dead. He used these data to show the selective elimination of individuals in a population based on their characteristics.

These data have been subsequently analysed by many different people[16]. In particular, Lande and Arnold's (1983) influential paper on the statistical estimation of selection gradients used this particular data set as an

[16] Pugesek and Tomer (1996) provide a brief history of these analyses.

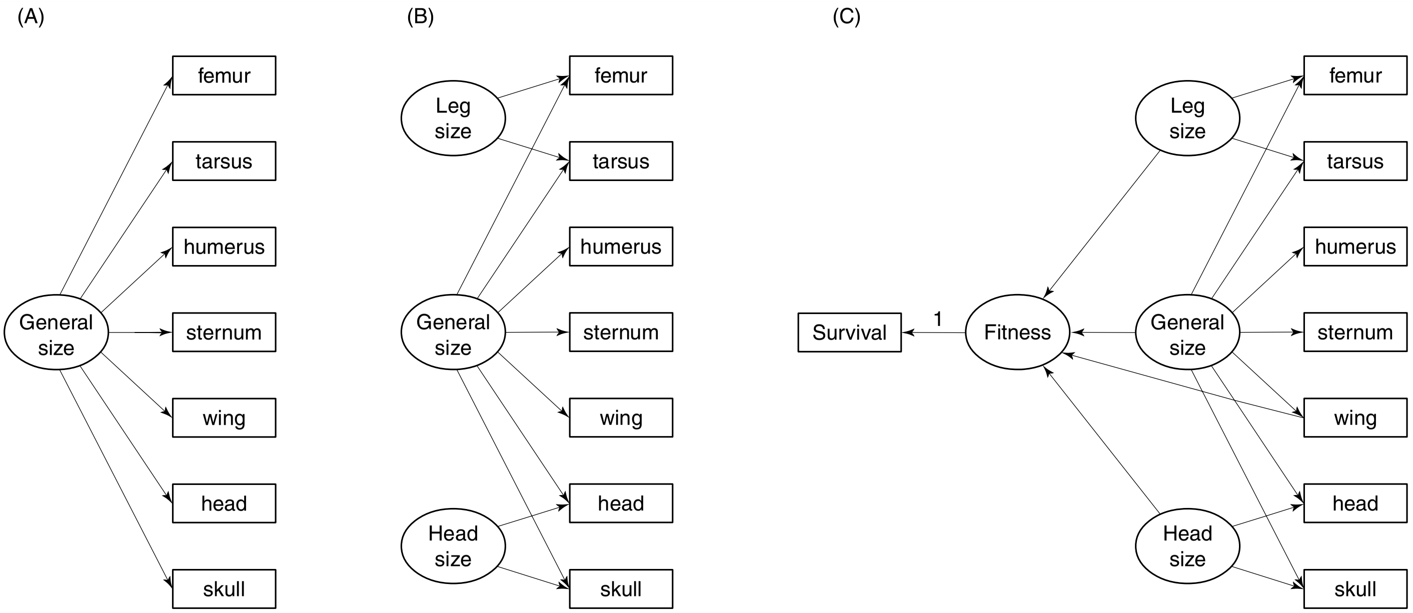


Figure 6.11. Pugesek and Tomer's model of the classic Bumpus House Sparrow data.

example. The method of Lande and Arnold was essentially an application of multiple regression of a suite of correlated characters on a measure of evolutionary fitness. The regression coefficients were interpreted as causal measures of the selection gradient. In Chapter 2 we have seen the problems that can occur when we use multiple regression in such a context.

Pugesek and Tomer (1996) reanalysed the Bumpus data using SEM. Besides two binary variables representing sex (male/female) and survival (alive/dead) they used seven other observed variables of various body measurements, transformed to natural logarithms: length of femur, tarsus, humerus, sternum, wing and head, and width of skull. They began with the measurement model involving all birds, living or dead. The first model that they considered (Figure 6.11A) was that all seven length measures were due to a single latent variable. Since there were two sexes they actually used a two-group model with across-group constraints (see Chapter 7). This first measurement model did not fit well ($MLX^2 = 43.59$, 28 df, $p = 0.03$). Their second measurement model was a three-factor model (Figure 6.11B). Pugesek and Tomer interpreted these latents as a latent general 'size' factor that is a common cause of all seven body measurements, a latent 'leg size' factor that is an additional common cause only of the femur and tarsus lengths, and a latent 'head size' factor[17] that is an additional common cause only of the head length and skull width. As we have seen before, giving the latents these names doesn't necessarily mean that the names are accurate; it is also possible that some more mundane causes, systematic measurement errors for instance, are the source of these latent variables. Whatever the source of the latent variables, the model provided a good fit ($MLX^2 = 28.82$, 24 df, $p = 0.227$). A series of nested models (see Chapter 7) showed that there were not significant differences between the males and females in any of the free parameters. Fixing all these free parameters to be equal in the two sexes provided a final measurement model with an acceptable fit ($MLX^2 = 49.29$, 40 df, $p = 0.149$).

The next step was to relate the measured and latent variables to survival. Pugesek and Tomer allowed the three latent 'size' variables to be direct causes of a fourth latent variable that they call 'fitness' and which then determines the death or survival of the individual bird. Since they fixed the path from the latent 'fitness' to the observed 'survival' at 1, and the residual error of 'survival' at zero, the latent 'fitness' variable is redundant since it will be

[17] Since the two specific latent variables only had two observed variables each, identification of the model was obtained by constraining the two path coefficients of each latent to be equal. This is simply a mathematical trick and means that the actual values of the path coefficients cannot be interpreted.

perfectly correlated with 'survival'[18]. This two-group model provided a marginal fit to the data ($MLX^2 = 62.72$, 48 df, $p = 0.075$), but the authors added an edge from 'wing length' directly to the latent 'fitness' that significantly improved the fit of the model ($MLX^2 = 52.37$, 46 df, $p = 0.241$). Finally, a nested sequence of models showed no significant differences in the path coefficients or error variances leading into, and out of, the latent 'fitness' variable and so these were constrained to be equal in the males and females.

The final model is shown in Figure 6.11C. The parameter values of this final model can be found in Figure 8 of Pugesek and Tomer (1996). The path coefficients (based on standardised variables) allow one to determine by how much a change in one morphological variable will change the probability of survival of the bird. For instance from their model one can calculate that an individual whose general size was one standard deviation larger than average increased its chances of survival by 0.564 standard deviations more than the average. On the basis of their model, it seems that larger birds were less likely to die during the storm than the smaller birds, birds whose legs and head were even larger than average given their general size were even less likely to die (although the path coefficients from these latent variables were not significant at the 5% level), but birds whose wings were shorter than average given their general size were also less likely to die.

[18] Unless there was some detail that was omitted from the paper, it was not necessary to fix the error variance of 'survival' to zero. It is not clear either what the latent 'fitness' means. Presumably it represents the propensity (or probability) of a bird to pass on a greater or lesser number of offspring to the next generation. If this is the case, then the causal interpretation of the model makes no sense. One's survival is not caused by such a propensity; rather, the propensity is partially caused by one's survival. Given the constraints that were put on the model, the latent 'fitness' should be removed and replaced by the observed 'survival'. If one also had data on reproductive output of these individuals, then 'survival' and 'reproductive output' could be modelled as causes of a latent 'fitness'.

7 Nested models and multilevel models

Like successful politicians, good statistical models must be able to lie without getting caught. For instance, no series of observations from nature are *really* normally distributed. The normal distribution is just a useful abstraction – a myth – that makes life bearable. In constructing statistical models we pretend that the normal distribution is real and then check to ensure that our data do not deviate from it so much that the myth becomes a fairy tale. In Chapter 6 we saw how far we could stretch the truth about the distributional properties of our data before our data called us a liar. The goal of this chapter is to describe how SEM can deal with two other statistical myths that people often relate with respect to their data.

Two important assumptions made by all of the models that we have studied up to now is that the observations in our data sets are (i) independent draws generated by (ii) the same causal process. Consider first the assumption of causal homogeneity. It is easy to imagine cases in which different groups of observations might be generated by partially different causal processes. For instance, a behavioural ecologist studying a series of variables related to aggression and social dominance in primates would not necessarily want to combine together the observations from males and females, since it is possible that the behavioural responses of males and females are generated by different causal stimuli. When we sample from populations with different causal processes, either in terms of the causal structure or of the quantitative strengths between the variables, and we wish to compare the causal relationships across the different groups, we require a model that can explicitly take into account these differences between groups. Such a model is called *multigroup* SEM and this, in turn, requires the notion of *nested* models.

The assumption of independence of observations can often be violated as well. Natural selection itself suggests a way in which we can get non-independence of observations (Felsenstein 1985; Harvey and Pagel 1991). The attributes of organisms, if they have a genetic component, will tend to be more similar to those of close relatives than to genetic strangers. The process of speciation therefore generates a hierarchical structure to data when we combine observations from different families, populations or species. If we ignore this hierarchical structure, and therefore ignore the non-independence of the observations, then we will obtain incorrect probability estimates. The application of *multilevel* SEM can deal with this complication.

7.1 Nested models

Given two SEM models with the same set of variables, then one model is *nested* within a second one if (i) all of the fixed parameters in the first model are also fixed to the same values in the second, but (ii) some of the free parameters in the first are still fixed in the second. In other words, the fixed parameters in the first model are a subset of the fixed parameters in the second model. The notion of nesting can be grasped most easily by comparing some path diagrams. In Figure 7.1 model B is nested within A and model D is nested within C.

Model A has two fixed parameters. The path coefficients for the edges between X_1 and X_2 and between X_1 and X_3 have each been fixed to zero, therefore there is no edge between X_1 and X_2 or between X_1 and X_3. There are two fixed parameters in model A, all others being freely estimated[1]. There is only one fixed parameter in model B – the path coefficient for the edge between X_1 and X_3 is still fixed to zero – and all others, including the path from X_1 to X_2, are freely estimated. So the fixed parameters of model B are a subset of those in model A and model B is nested within model A.

Model C also has two fixed parameters. The path coefficient for the edge between X_1 and X_3 is still fixed to zero and the path coefficient for the edge from X_1 to X_2 has been fixed to 0.5. Note that model C is not nested within model A; it is true that the path coefficients between X_1 and X_2 are both fixed but they are not fixed to the same *value*. Model D is, however, nested within model C. This is because every fixed parameter in model D – the path coefficient for the edge between X_1 and X_3 – is also fixed in model C.

Nested models are useful because the difference in the maximum likelihood chi-squared values between nested models is, itself, asymptotically distributed as a chi-squared variate if the freely estimated parameters are equal to their associated fixed parameters. The degrees of freedom of this change in chi-squared are the number of parameters that have been freed in the nested model, which is the same as the change in the degrees of freedom between the nested models.

[1] The free parameters representing the variances are not shown in Figure 7.1.

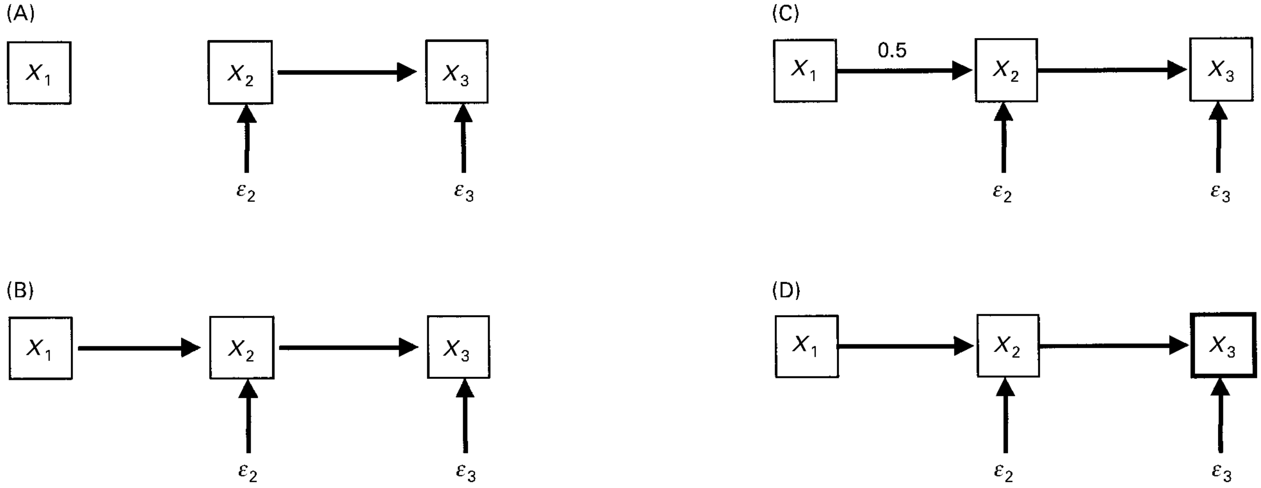


Figure 7.1. Four path models used to illustrate the concept of nesting.

Intuitively, the testing of a nested model uses the following logic. One starts with a model (call it model 1) in which a set of parameters are fixed to particular values (zero or otherwise). Now, we define a new nested model (call it model 2) by freeing some previously fixed parameters but without changing anything else relative to model 1. If we allow some of these previously fixed parameters to be freely estimated, but these newly freed parameters really do have the values to which they had previously been fixed, then the only difference in the estimated covariance matrices between models 1 and 2 will be due to random sampling variation. If this is true then the difference between the maximum likelihood chi-squared statistics will also follow a chi-squared distribution with degrees of freedom equal to the number of previously fixed parameters that have been freed in the nested model 2. Here are the steps:

1. Fit the model at the top of the nested sequence, obtain its chi-squared value (MLX_1^2) and its degrees of freedom (df_1).
2. Fit the model at the bottom of the nested sequence, obtain its chi-squared value (MLX_2^2) and its degrees of freedom (df_2).
3. Calculate the change in the chi-squared value and the change in the degrees of freedom: $\Delta\text{MLX}^2 = \text{MLX}_1^2 - \text{MLX}_2^2$ and $\Delta\text{df} = \text{df}_1 - \text{df}_2$.
4. Determine the probability of having observed this change in the chi-squared value ($\Delta\chi^2$)assuming that the freed parameters in the second (nested) model are equal to those in the first model, except for random sampling variation.
5. If this probability is less than the chosen significance level, conclude that the freed parameters were not the same as those fixed in the first model.

Tests of nested models are used in a number of different research contexts. One reason might be if you want to test for the equality of a set of parameters to some theoretical values but don't care whether the model as a whole is acceptable. Two exploratory methods in SEM (the Wald and Lagrangian multiplier tests) are based on this logic. Perhaps the most useful application of nested models is in the context of multigroup models and multilevel models.

7.2 Multigroup models

Another assumption of the tests for structural equations models that have been described so far is that all of the observations come from the same statistical population. In other words, we are assuming that the same causal process has generated all of our observations even if we don't know what this causal process might be. Often we know (or suspect) that this is not the case. For instance, if we are studying attributes related to reproductive success then we might suspect that different causal processes are at work for males and females. Even if the causal structure is the same in males and females, it is possible that the two sexes differ in the numerical strength of the causal relationships. If we were to combine males and females into one data set then we would obtain incorrect parameter estimates and might incorrectly reject the model even though the qualitative structure of the model is correct. Perhaps our data come from three different geographical regions and we are not willing to assume that the same causal forces (with the same numerical strengths) apply to the observations in these different regions. Perhaps our data come from groups that we have subjected to different experimental treatments. All of these examples require that we explicitly include the group structure into our analysis. Such analyses are called 'multigroup SEM'.

The first impulse (which is not always wrong) is to analyse the data in each group separately. The real strength of multigroup SEM is the ability to compare statistically between groups and determine which parts of the models in each group (i.e. which parameters) are the same and which parts differ. In this sense multigroup SEM is analogous to ANOVA except that, rather than testing for differences in the means between groups, we are testing for differences in the covariance structure between the groups. To do this we construct a series of nested multigroup models.

A multigroup model can be fit with a minor modification of the method that you already know. Since the standard structural equation model is simply a multigroup model with only one 'group', let's start there. With only one group we have only one observed covariance matrix ($\boldsymbol{S}_1$). We then set up the model covariance matrix ($\boldsymbol{\Sigma}_1$) using covariance algebra and iteratively find values of the free parameters of $\boldsymbol{\Sigma}_1$ that minimise the maximum likelihood chi-squared statistic: $(N_1 - 1)(\ln|\boldsymbol{\Sigma}_1(\theta_1) + \mathrm{trace}(\boldsymbol{S}_1\boldsymbol{\Sigma}_1^{-1}(\theta_1)) - \ln|\boldsymbol{S}_1| - p_1)$. This is the same formula that you saw in Chapter 4 (p is the number of variables in the model) except that I have added subscripts to emphasise that we are referring to group 1. When our data are divided into g groups with $N_1, N_2, \ldots, N_g$ observations in the different groups then we have g sample covariance matrices ($\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_g$) and also g population covariance matrices ($\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_g$). Each population covariance matrix can potentially have different sets of free and fixed parameters or even different sets of variables. We iteratively choose values of all of these free parameters simultaneously to minimise:

$$[(N_1 - 1)(\ln|\boldsymbol{\Sigma}_1(\theta_1) + \mathrm{trace}(\boldsymbol{S}_1\boldsymbol{\Sigma}_1^{-1}(\theta_1)) - \ln|\boldsymbol{S}_1| - p_1)] + [(N_2 - 1)$$
$$(\ln|\boldsymbol{\Sigma}_2(\theta_2) + \mathrm{trace}(\boldsymbol{S}_2\boldsymbol{\Sigma}_2^{-1}(\theta_2)) - \ln|\boldsymbol{S}_2| - p_2) + \ldots$$
$$[(N_g - 1)(\ln|\boldsymbol{\Sigma}_g(\theta_g) + \mathrm{trace}(\boldsymbol{S}_g\boldsymbol{\Sigma}_g^{-1})(\theta_g)) - \ln|\boldsymbol{S}_g| - p_g)$$

Although this equation looks intimidating, it is simply the sum of the maximum likelihood chi-squared statistics for each group.

The value of this multigroup maximum likelihood chi-squared statistic at the minimum also asymptotically follows a chi-squared distribution with degrees of freedom equal to the sum of the degrees of freedom of the model in each group. Even if a particular parameter is free in each group, we can constrain the fitting procedure to choose the *same* value for all groups (this generates $g-1$ extra degrees of freedom). In this way we are stating that, although we don't know what the numerical value of the free parameter is, it must be the same numerical value in all groups. Viewed in this way, we see that multigroup models define a continuum. If we propose the same causal structure and the same numerical values for all free parameters across the groups, then we get the same result as if we had centred each variable around its group mean and then put all of our data into one big group. If we allow all of the free parameters to differ between groups then we get the same result as if we had tested each group separately and summed the maximum likelihood chi-squared statistics and degrees of freedom. By constraining the estimation of different sets of free parameters across groups then we can define a series of nested models. In this way, we can test for the equivalence of various free parameters in the different groups. If we do this more than once then we should adjust our significance level using a Bonferonni correction[2].

The following example comes from Meziane (1998). Although it is a path model without latent variables, the logic and approach are identical with full SEM models. The study consisted of 22 species of herbaceous plants grown under controlled conditions in four different environmental conditions: a high (N) and low (n) nutrient concentration in hydroponic culture crossed with a high (L) and low (l) light intensity. This gave four different groups of data corresponding to the four different environments: NL, Nl, nL, nl. Two leaves on each plant were harvested and a series of four morphological attributes were measured: the water content of the leaf, the thickness of the lamella, the thickness of the midvein and the specific leaf area (the ratio between the projected leaf area and its dry weight). The values of the two leaves per plant were averaged. Owing to a few missing values, there were a total of 80 independent observations in the final data set. A

[2] If we do this t times with a significance level of α then we should test the change in the chi-squared of each test at a significance level of α/t.

previously published study (Shipley 1995) had described a path model relating these variables, and one objective of Meziane (1998) was to see whether the previous path model could be applied under different environmental conditions. If Meziane had simply combined the data from all four environments and tested his path model then he would have implicitly assumed that the different environments had no effect on the relationships between the four variables. By 'no effect' I mean both that the structure of the relationships (their presence or absence) and their numerical strengths do not change. If you remember that each variable is centred around its mean in the data set, then combining the data from all four environments would also implicitly require that the treatments did not affect the mean values of the variables either. By separating the data into the four groups, the variables are centred around their respective group means. In this way, the treatment effects on the means are removed and only the relationships between the variables are analysed.

In his multigroup analysis he specified four models, each with the same structure but each potentially differing in the numerical strengths of the free parameters. I have shown this in Figure 7.2, in which I have included the free parameters. In this model there are five free path coefficients and four free error variances in each of the four models. There are therefore from 9 (if all free parameters are constrained to be equal across groups) to potentially 36 different free parameters to estimate (if no free parameters are constrained to be equal across groups). Using the rule of thumb requiring five times more observations than free parameters, we see that any multigroup model with more than 16 free parameters will not be well approximated by a chi-squared distribution and will have true probability values that are somewhat higher than those obtained using a chi-squared distribution. The data, after transforming to natural logarithms, had reasonably low values of Mardia's multivariate index of multivariate kurtosis (-4.37, -2.70, -3.34 and 1.40 for the NL, Nl, nL and nl groups, respectively).

The first step was to fit the data to the most constrained model; namely, in which all nine free parameters are forced to be equal across the four groups[3]. This fully constrained model gave a maximum likelihood chi-squared statistic of 48.271 with 31 degrees of freedom ($p=0.02$). Why 31 degrees of freedom? Each covariance matrix was composed of four variables, so there were $4(5)/2=10$ non-redundant elements in each matrix. There were four independent matrices for a total of 40 non-redundant

[3] Note that, although the free parameters are constrained to be equal, the variables are still centred about their group means, not the overall mean of all data taken together. Therefore, differences in the means between the groups are still removed.

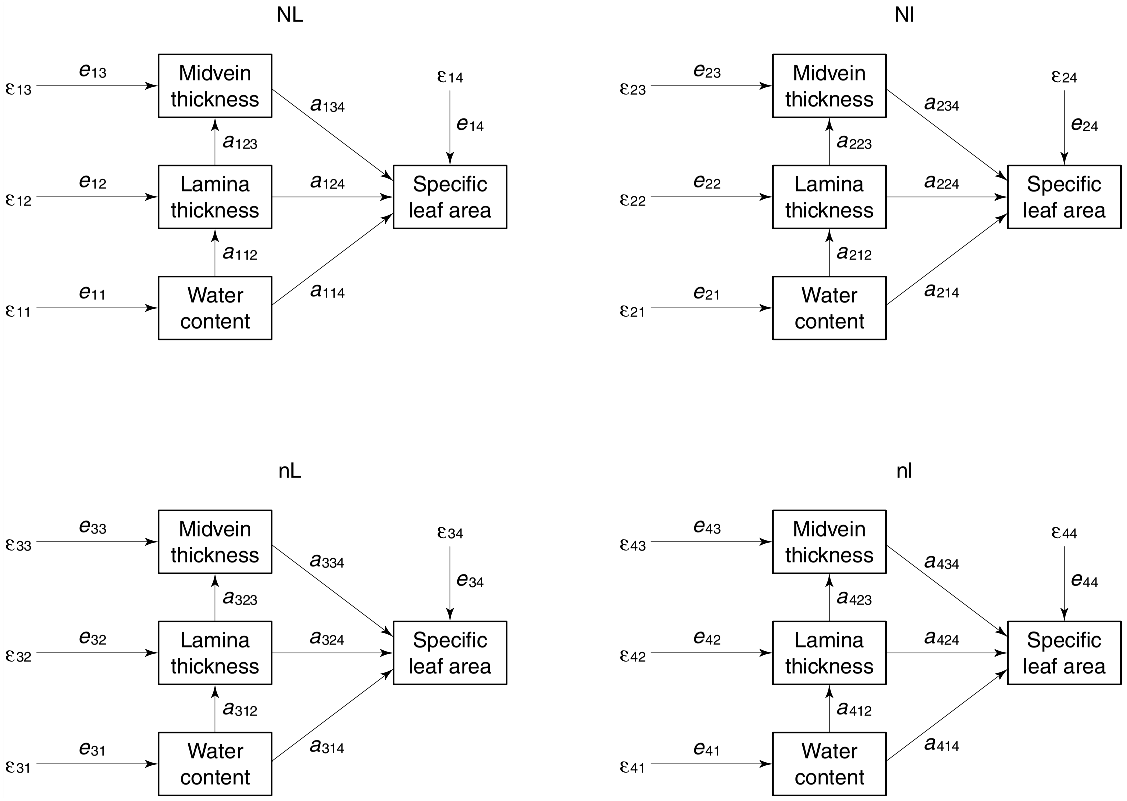


Figure 7.2. Meziane's four-group path model relating four attributes of the leaves of herbaceous plants. Each group refers to plants grown in different environments of hydroponic nutrient solution and light intensity (see the text for symbols).

elements in all. Since we fixed all free parameters to be equal across groups, we estimated only nine different free parameters. This gives a total of 40 − 9 = 31 degrees of freedom. Since we had 80 observations and 9 free parameters, and the data did not show strong kurtosis, we can be fairly confident that this fully constrained multigroup model has some errors in it. Since there were no obvious nonlinearities in these data and the distributional assumptions do not seem to cause any problems, the remaining problems reside either in the causal structure of the model or in the equality constraints that we have imposed on the data. If we remove all equality constraints across groups, then we will be testing only that the same qualitative structure applies to all four groups. The maximum likelihood chi-squared statistic, when all equality constraints are removed, is 3.224 with 4 degrees of freedom ($p = 0.52$). Even though we now have few observations per estimated free parameter (we have 36 free parameters now, so the ratio is only 2.2 : 1) the probability level gives us no good reason to reject this multigroup model with no between-group equality constraints for the parameter estimates[4].

Since the lack of fit that was detected in the fully constrained multigroup model appears to be in the equality constraints between the four groups, we can use a series of nested models to detect which of the equality constraints is unreasonable. If we remove only one between-group equality constraint then this new model will be nested within the fully constrained model. There will be 3 degrees of freedom less in this new model, since we now have to independently estimate the value of this free parameter in all four groups, rather than simply estimating one value for all four groups. The difference in the two maximum likelihood chi-squared statistics, compared to a chi-squared distribution with 3 degrees of freedom (the difference in the degrees of freedom between the fully constrained model and the new model), will test for a difference in the value of this parameter between the four groups. We do this nine times, each time removing the equality constraint for a different free parameter. Since we have done this test nine times at a significance level of 5%, we adjust the overall significance level to 0.05/9 = 0.0056.

Table 7.1 summarises the results. The first column lists the free parameter whose between-group equality constraint has been removed. The second column lists the maximum likelihood chi-squared statistic for this new model (always with 28 degrees of freedom). The third column lists the change in the maximum likelihood chi-squared statistic relative to the

[4] Remember that the effect of small sample sizes is to produce conservative probability estimates.

Table 7.1. *The results of comparisons of a series of nested models based on the four-group model shown in Figure 7.2. The first row gives the results of the fully constrained model assuming all free parameters are the same in the four groups. The remaining rows show the result of relaxing one constraint at a time*

Free parameter whose between-group equality constraint was released	MLX^2	Change in MLX^2 (ΔMLX^2)	Probability of ΔMLX^2
None	48.271		
Variance of leaf water content	42.141	6.130	0.105
Error variance of specific leaf area	45.995	2.276	0.517
Error variance of lamina thickness	46.387	1.884	0.597
Error variance of midvein thickness	47.195	1.076	0.783
Path coefficient from leaf water content to specific leaf area	39.411	8.860	0.031
Path coefficient from lamina thickness content to specific leaf area	27.710	20.561	0.0001
Path coefficient from midvein thickness content to specific leaf area	35.122	13.149	0.004
Path coefficient from leaf water content to leaf lamina thickness	44.34	3.931	0.269
Path coefficient from leaf lamina thickness to midvein thickness	47.646	0.625	0.891

Note:
MLX^2, maximum likelihood chi-squared statistic.

model with all between-group equality constraints applied. The fourth column lists the (asymptotic) probability for the change in the maximum likelihood chi-squared statistic.

From Table 7.1 we see that only two of the path coefficients (those between the thickness of the lamina and midvein and the specific leaf area) differ between the four groups given our chosen significance level. Note that, although the path coefficient from leaf water content to specific leaf area had a probability level of 0.031, we had to adjust our individual significance levels to $0.05/9 \approx 0.0056$ in order to maintain an overall significance level of 0.05. Our final multigroup model fixes all free parameters except for these two path coefficients to be equal across groups. This model has a maximum likelihood chi-squared statistic of 21.463 with 25 degrees of freedom, giving a probability level of 0.667. The confidence interval of the

RMSEA for this model is (0,0.074). Since the original purpose of the analysis of Meziane (1998) was to see whether the original path model that I had proposed (Shipley 1995) could be applied to plants growing in different resource environments, the conclusion is that the model appears to apply in its general structure, but that the numerical strengths of the two thickness measures on the specific leaf area change in the different environments. Of course, given the rather small number of observations and therefore low power, we must temper this conclusion, since small differences between groups might not have been detected.

In interpreting the results of a multigroup SEM it is important to remember that we are testing only for differences in the relationships between the variables within each group. Differences in the mean values of the variables between the groups are never detected because the variables are centred about their mean values within each group. In the example from Meziane (1998) the mean values of every one of the variables differed between the groups, on the basis of analyses of variance. In other words, the different levels of nutrients and light intensities did cause changes in the average values of the leaf attributes (from the ANOVA) and did change the numerical strengths of the effects of the two thickness measures on specific leaf area, but did not change the numerical strengths of the other relationships, the error variances or the causal structure (i.e. the topology) between the variables.

7.3 The dangers of hierarchically structured data

Let's now turn to the problem of analysing data when the observations are not independent. To illustrate the problems caused by partially dependent data, consider first the following naïve analysis[5]: I wish to test the hypothesis that people with blue eyes (such as myself) have shorter hair than do people with beautiful green eyes (such as my wife). To test this hypothesis I randomly choose 20 hairs from my head and 20 hairs from my wife's head, measure them, and then conduct a t-test on this set of 40 observations using (40 – 2) degrees of freedom. Of course, I find a highly 'significant' difference in hair length between the two groups and, of course, the probability level associated with this test would be profoundly wrong. The problem is that both variables, eye colour and hair length, are nested within individuals, of which there are only two. A large proportion of the total variation in hair length and all of the total variation in eye colour resides at the level of

[5] I actually present this problem to students in a first-year undergraduate course in biometry. They instinctively recognise the nonsensical nature of the result, but are not able to explain why the result is flawed.

individual people, not at the level of individual observations. Clearly, I do not have 40 independent observations of the two variables (eye colour, hair length).

If two values of some variable X (say X_1 and X_2) are independent then knowing the value of X_1 tells us nothing about the likely value of X_2. Two independent values give us two 'pieces' of information and n independent values give us n 'pieces' of information. If, in some group, every individual had exactly the same values of X, then as soon as you knew X_1 then you would also know the values of all other X in the group. No matter how many observations of X that you took from such a group, you would have only one 'piece' of non-redundant information.

Now, imagine that we create two groups. We randomly choose 20 values of X to form the first group and these values are independently and normally distributed with a mean of 1 and a standard deviation of 0.5. We randomly choose 20 values of X to form the second group and these values are also independently and normally distributed but with a mean of 2 and a standard deviation of 0.5. If the values within each group were exactly the same then we would still only have two 'pieces' of information, but this is not the case. If knowing that an observation came from a particular group told us nothing about what values it might have, then we would have 40 'pieces' of information, but this is not the case either. So, we have more than 2 and fewer than 40 'pieces' of information[6]. This is the nature of hierarchically structured data, and we require a way of determining how many 'pieces' of information different variables possess at different levels of the hierarchy. This is the goal of multilevel models, also called 'random coefficient models', 'variance component models' or 'hierarchical linear models'. Such models, in the generalised linear context, have a large literature[7]. Detailed discussions and statistical derivations can be found in a number of books (Bryk and Raudenbush 1992; Longford 1993; Goldstein 1995).

Hierarchies and partial dependence are the rule, rather than the

[6] The ratio of the variation of a variable at a given level in a hierarchy to its total variation is given the unfortunate name of 'intraclass correlation' (Muthén and Sattora 1995). If we let the estimate of the variance of a mean, when ignoring the hierarchical nature of data, be $\mathrm{Var}_{\mathrm{SRS}}$ ('simple random sampling'), the correct variance estimate be $\mathrm{Var}_{\mathrm{C}}$, the number of observations per group be c and the intraclass correlation be ρ, then the relationship between them is $\mathrm{Var}_{\mathrm{C}} = \mathrm{Var}_{\mathrm{SRS}}(1 + (c - 1)\rho)$. A similar formula exists for the variance of a linear regression slope with hierarchical structure. If ρ_ε is the intraclass correlation for the residuals and ρ_X is the intraclass correlation for the predictor, then Scott and Holt (1982) showed that $\mathrm{Var}_{\mathrm{C}} = \mathrm{Var}_{\mathrm{SRS}}(1 + (c - 1)\rho_\varepsilon \rho_X)$.

[7] These models were mostly developed for the field of educational research and commercial statistical programs are available. One such program is MLwiN (Goldstein *et al.* 1998).

exception, in nature. If this is so, then we need to incorporate this structure into our models of nature. One way to do this is through the use of multi-level models. Before going on to the mechanics of fitting such models, or even of interpreting them, it is useful to have a simple concrete example of such a structure. A good example is the relationship between seed size and seedling relative growth rate. Relative growth rate (RGR) is the amount of new biomass produced by a plant over a unit time period, relative the amount of biomass that the plant had as an initial 'capital' at the beginning of the time period. If we plot the weight of different seeds of individuals within a single species against the RGR of the seedling that emerges, we often find a positive relationship between the two. For individuals within a given species, having larger seeds translates into more rapidly growing seedlings, with all of the attendant benefits. When we compare across species, however, we see that both average seed size and average potential RGR varies much more between species than within species. For instance, the seeds of some orchids are almost microscopic while it might require two hands to hold the seeds of a Coconut Palm. Under constant environmental conditions and resource levels, the variation of seedling RGR within a single species usually varies by less than 10% of the variation in the mean RGR values between different species. Curiously, the relationship between the *average* values of seed size and seedling RGR across species usually shows a negative relationship. Figure 7.3 plots some simulated data showing this pattern. There are 10 simulated species, each having a different plotting symbol. Figure 7.3A shows the relationship between the two variables for the first species, along with its regression line. Figure 7.3B shows the relationship between the two variables when the data for all 10 species are combined.

This example, although very simple, demonstrates many of the challenges of analysing data that are hierarchically ordered. It is clear that part of the variation in each variable, and the covariation between the two, is generated by differences between individuals within each species, as shown by Figure 7.3A. It is also clear that part of the variation and covariation involving these variables is generated by differences between the species means. We would like to model this covariation both within and between species, taking into account the fact that individuals within a given species tend to resemble each other more than the do individuals of different species. Finally, we would like to model how the different levels in the hierarchy interact and constrain each other.

Since there are two levels to this data hierarchy, let's call 'level 2' the level of species and call 'level 1' the individual level. If we had only one species, then we could write the regression equation as:

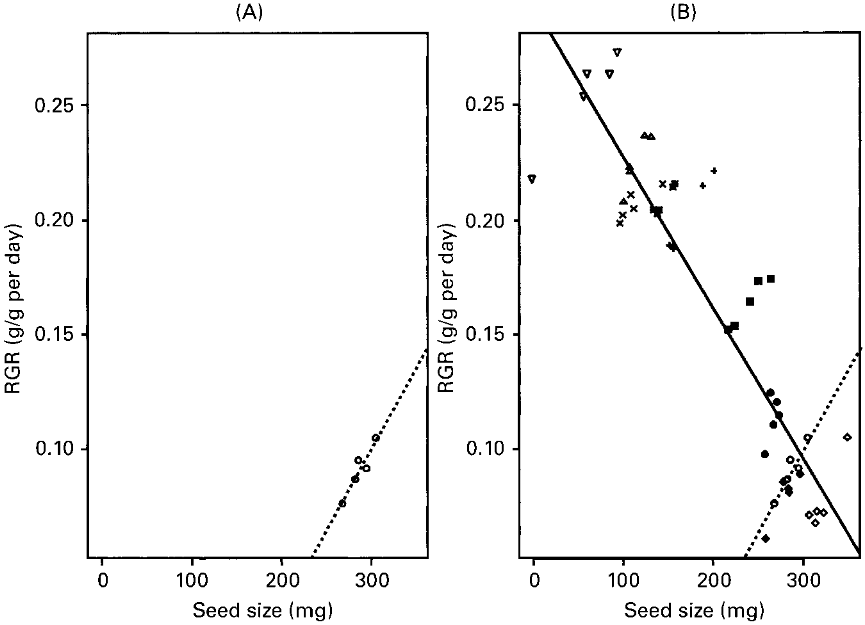


Figure 7.3. Hypothetical relationships between the relative growth rate (RGR) of 50 plants and their seed size. (A) This relationship for one species. (B) The relationship when all 10 species are combined. Notice that the relationship is positive within every species but the overall trend shows a negative relationship because the relationship between the species' means is negative.

$$y_i = a + bx_i + e_i$$

In this equation the subscript *i* refers to each individual, *a* is the value of the average individual when x = 0 (i.e. the intercept) and e_i is the amount by which the value of *y* for the *i*th individual deviates from its expected value given x_i. As usual, we assume that these deviations are random. Since we have 10 species in our data set, we could write:

$$y_{i1} = a_{01} + bx_{i1} + e_{i1}$$

$$y_{i2} = a_{02} + bx_{i2} + e_{i2}$$

$$y_{i10} = a_{010} + bx_{i10} + e_{i10}$$

In this simple example I have assumed a common slope for all species (thus *b* has no subscript) because that is how I generated these data. In general, multilevel models allow for random slopes as well as random intercepts. The assumption of a common slope across species in a structural equation model could be tested with a multigroup model, as described in the previous section, remembering that the slope is the path coefficient between x and y.

Here, we have a different regression equation for each species. We could, at this point, simply introduce a dummy variable and conduct an analysis of covariance. In the context of SEM, this would be the equivalent of doing a multigroup model. However, if we have chosen our 10 species at random, and want to extrapolate to a larger population of species, then our regression intercepts are, themselves, random variables and we might want to model how these species-level random variables change as well. In this case, the 10 intercept terms (a_1 to a_{10}) are random variables which we can model as: $a_{0j}=a_{00}+u_{0j}$. Here, a_{00} is the overall intercept term for the entire population of species and u_{0j} is the random deviation of the intercept for the jth species from this overall intercept term. Putting this all together we obtain:

$$y_{ij}=(a_{00}+u_{0j})+bx_{ij}+e_{ij}$$

Here the i-subscript refers to the level-1 units (the individual plants) and the j-subscript refers to the level-2 units (the different species). Rearranging, we obtain:

$$y_{ij}=a_{00}+bx_{ij}+u_{0j}+e_{ij}$$

This equation expresses each y_{ij} as a function of a systematic component ($a_{00}+bx_{ij}$), a random component due to the differences in the mean values of y_{ij} between the species (u_{0j}) and a random component due to the differences of individuals within each different species (e_{ij}).

Up to now the model that we have developed is perhaps familiar to some readers, since it is simply a variance components model[8] with a between-species variance component (the variance of the u_{0j}) and a within-species component (the variance of e_{ij}). However, since the intercepts (u_{0j}) are random variables, the relationships between these intercepts may be determined by other, species-level variables. For instance, in Figure 7.3 it is clear that as the mean value of seed size for a given species increases the mean value of RGR for that species decreases.

Secondary succession in plant communities starts with some major

[8] Actually, the classical variance component model for RGR would not include the seed size variable. Such a model would be of the form: $\mathrm{RGR}_{ij}=a_{00}+u_{0j}+e_{ij}$, the variance of u_{0j} and of e_{ij} being the two variance components and the total variance of RGR_{ij} being the sum of these two variances. Using the simulated data for RGR in Figure 7.2 these values are: $a_{00}=0.162(\pm 0.020)$, $\mathrm{Var}(u_{0j})=0.0039(\pm 0.0018)$ and $\mathrm{Var}(e_{ij})=0.0002(\pm 0.00004)$. One usually expresses these variance components as percentages of the total variation. In this example, the species-level variance is 95% (i.e. $0.0039/(0.0039+0.0002)$).

disturbance event, for instance a field that has been cultivated and then abandoned. As different species reinvade the site the relative abundance of each species changes. Because of this one often finds that particular species tend to be most abundant in abandoned fields of a specific age. Immediately following a major disturbance event one typically finds annual species that have rapid relative growth rates and that produce a large quantity of small seeds. As secondary succession proceeds dominance shifts to species with larger seeds and slower relative growth rates (Grime 1979). This is a selection process in which different frequencies and intensities of density-independent mortality (the result of disturbance events) select for different suites of plant attributes. As such, selection pressures represented by such variables as 'average time since the last major disturbance event' affect species-level properties by determining the mean and variation of individual-level attributes.

Actually, I generated the data shown in Figure 7.3 by simulating the scenario described above. I defined a species-level variable – the frequency of major disturbance events – that quantifies how often a habitat experiences a major event of density-independent mortality. The causal effect of this variable is to select for individuals with both larger seeds and lower RGRs as the frequency of disturbance events decreases and as successional age of the vegetation increases. In other words, although RGR increases with increasing seed size within each species, the *average* seed size and the *average* relative growth rates of a given species are both determined by the common cause 'disturbance frequency'. This is shown in Figure 7.4.

As Figure 7.4 makes clear, the relationship between individual seed size and individual relative growth rate consists of causes that operate at two different hierarchical levels. At the level of individual plants there is a *positive* direct effect (seed size→relative growth rate). At an interspecific level there is a *negative* indirect effect between the two variables that is generated by the common cause of selection for habitats experiencing different disturbance frequencies. Whether the overall relationship between seed size and relative growth rate is positive, negative or ambiguous depends on the relative strengths of these different paths. In the data that I have simulated, the species-level effect dominated, which is why the overall trend in Figure 7.3 is a downward sloping cloud of points. How can we incorporate these hierarchical effects into our models? We could ignore the individual level and simply work with the species means. If we did this then we would not only lose a great deal of information (by reducing our data set from 50 observations to 10) but we would also ignore the fact that the relationship at the individual level is quite different from the relationship at the species level. We could ignore the fact that 'disturbance frequency' is a variable that

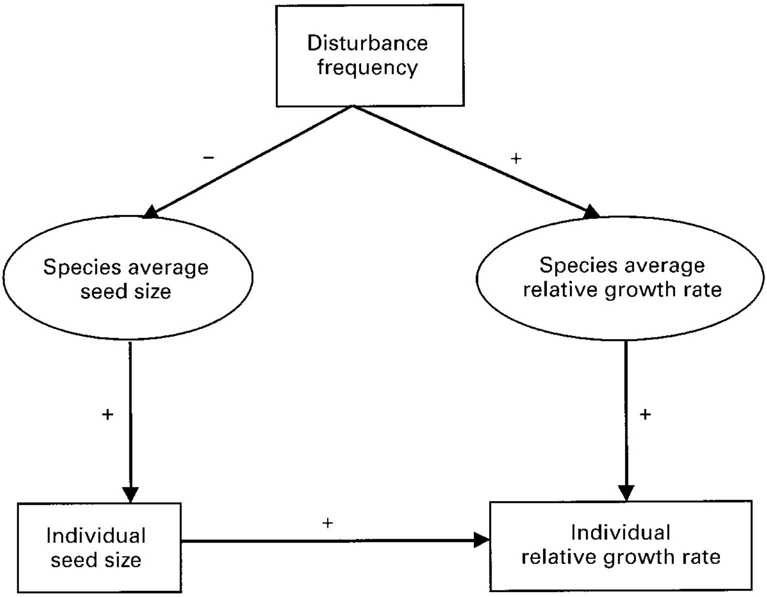

Disturbance frequency
–
+
Species average seed size
Species average relative growth rate
+
+
Individual seed size
+
Individual relative growth rate

Figure 7.4. The causal process determining the relationship between seed size and relative growth rate operated at two levels. At the level of individuals, seed size causes relative growth rate. However, the average seed size and relative growth rate of each species is caused by the typical disturbance frequency of the habitat occupied by each species.

is only relevant to the species-level process and simply conduct a standard multiple regression of RGR on both seed size and average disturbance frequency. This is like correlating eye colour and hair length – the average disturbance frequency is the same for all individuals within a given species and so we would be inflating the number of 'pieces' of information that we really possess. Instead, we should take an explicit multilevel approach.

First, we model the individual level relationship between RGR and seed size:

$$\mathrm{RGR}_{ij} = \beta_{0j} + \beta_1 \mathrm{seedsize}_{ij} + e_{ij}$$

Here, we are specifying that RGR varies at both the individual (*i*-level) and the species (*j*-level). The intercept (β_{0j}) varies only at the level of species. The slope of the relationship between RGR and seed size (β_1) is constant across all species. Finally, the residual variation in RGR within each species (e_{ij}) is assumed to be normally distributed with a constant variance[9]. Since

[9] Both the assumption of normality and the assumption of constant variance can be relaxed in multilevel modelling.

the intercepts of RGR are, themselves, random variables that change from species to species, we next model this species-level variation in these intercepts:

$$\beta_{0j} = \alpha_{00} + \alpha_{01}\text{disturbance}_j + \mu_{0j}$$

Now we are specifying that the species-level intercepts are functions of the average disturbance frequency, which is a species-level variable. There is a constant intercept (α_{00}) term which represents the average seed size across all species in the statistical population. The species-level slope between average disturbance frequency and the intercepts of each species is α_{01}. Finally, the deviations of each species' intercept from that predicted by average disturbance frequency is the random variable μ_{0j}. Putting this all together we get:

$$\text{RGR}_{ij} = \alpha_{00} + \beta_1\text{seedsize}_{ij} + \alpha_{01}\text{disturbance}_j + \mu_{0j} + e_{ij}$$

The standard error of the slope of seed size is based on the error variance at the level of individuals within species (e_{ij}), which has been corrected for the species-level variation. The slope of the average disturbance frequency is based on the error variance at the level of species (μ_{0j}), which has been corrected for the error variance at the individual level.

We next specify a multilevel model for seed size. According to Figure 7.4, seed size is caused only by average disturbance frequency and this effect occurs only at the species level. Our multilevel model is therefore:

$$\text{seedsize}_{ij} = \alpha_{00} + \alpha_{01}\text{disturbance}_j + \mu_{0j} + e_{ij}$$

If I fit these models using the MLwiN software (Goldstein *et al.* 1998), I obtain the following results:

$$\text{Seed size} = 443.76 - 16.32 \text{ disturbance frequency}$$

$$\text{RGR} = -0.284 + 0.0006 \text{ seed size} + 0.0217\text{disturbance frequency}$$

The residual variation of the mean seed sizes per species was 55.28 (13%) and the residual variation of individual seed sizes was 361.76 (77%). The residual variation of the mean RGR values per species was 4.96×10^{-5} (64%) and the residual variation of the individual RGR values was 2.79×10^{-5} (46%). The big difference between this multilevel model and an ordinary regression model is best seen in the standard errors of the parameters of the equation for RGR. If we were to do an ordinary regression then we get an estimate of the standard error of the slope for average disturbance frequency of 1×10^{-4} while the standard error estimated from the multilevel model was 8.3×10^{-4}. In other words, by ignoring the hierarchical nature of RGR the ordinary regression overestimated the precision of the slope by eight times. Since significance tests of the effects of a variable in a regression are based on these standard errors, the effect of ignoring the partially dependent nature of observations is to produce probability estimates that are much smaller than they should be.

Hierarchically ordered data not only cause problems for parameter estimation and inferential tests of significance. The patterns that can be generated with multilevel data can often be downright counterintuitive. To give you a feeling for such patterns, I have generated data from a very simple two-level model, shown in Figure 7.5. In this scenario there are two attributes (X and Y) of each individual i (1 to 11) from each species j (1 to 10). Each species has a mean value of each variable (μ_{X_j} and μ_{Y_j}) and each individual has a value of each variable that varies around its species mean ($X_{ij} = \mu_{X_j} + \varepsilon_{X_{ij}}$ and $Y_{ij} = \mu_{Y_j} + 1X_{ij} + \varepsilon_{Y_{ij}}$). Variable $\varepsilon_{X_{ij}}$ takes values from -0.25 to $+0.25$. Since we are interested in comparing the within-species and between-species patterns, I will ignore the random variation of Y at the individual level ($\varepsilon_{Y_{ij}}$) and concentrate on the expected value of y ($\mathrm{E}[Y|X] = \mu_{Y_j} + 1X_{ij}$). Substituting for X_{ij} we get:

$$\mathrm{E}[Y|X] = \mu_{Y_j} + 1\mu_{X_j} + 1\varepsilon_{X_{ij}}$$

This equation represents the intraspecific (i.e. within-group) level. The interspecific (i.e. between-group level) will be generated in different ways[10] and the combined data are then shown in Figure 7.6. We imagine that the mean value of variables X and Y (μ_X and μ_Y) differ randomly between the 10 species and that the interspecific relationship between these interspecific means follows the following equation: $\mu_Y = a_{\mu_X} + \varepsilon_{\mu_X}$.

First, let's see what happens when there is no species-level variation or covariation. To simulate this, I set a in Figure 7.5 equal to zero and make the values of μ_X and μ_Y the same (zero) for all 10 species (thus, the variance of these two variables is zero). If we put these values into our generating equation we obtain:

$$\mathrm{E}[Y|X] = (0) + 1(0) + 1\varepsilon_{X_{ij}}$$

Figure 7.6A shows the pattern that results. There are actually 10 lines in this figure but they are superimposed on each other, since we have exactly the same values for all 10 species (remember that I am plotting the expected values). The solid circles show the mean values of X and Y for each species.

[10] Ignore for now the fact that the path coefficient from each μ to the observed variable is fixed at the square root of the number of individuals per species. This will be explained later.

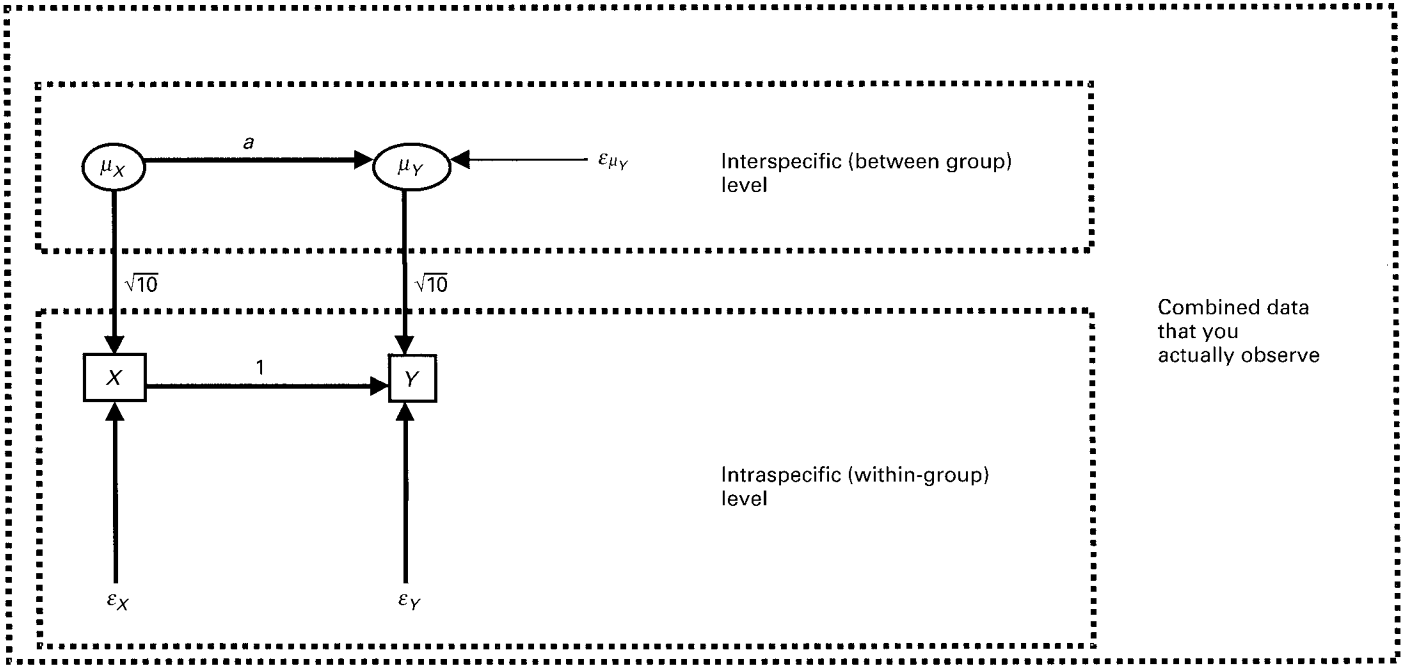

μ_X
a
μ_Y
ε_{μ_Y}
Interspecific (between group) level
$\sqrt{10}$
$\sqrt{10}$
Combined data that you actually observe
X
1
Y
Intraspecific (within-group) level
ε_X
ε_Y

Figure 7.5. A very simple two-level model. In this scenario there are two attributes (*X* and *Y*) of each individual from each species. Each species has a mean value of each variable (μ_{X_j} and μ_{Y_j}) and each individual has a value of each variable that varies around its species mean ($X_{ij} = \mu_{X_j} + \varepsilon_{X_{ij}}$ and $Y_{ij} = \mu_{Y_j} + 1X_{ij} + \varepsilon_{Y_{ij}}$). Variable $\varepsilon_{X_{ij}}$ takes values from −0.25 to +0.25.

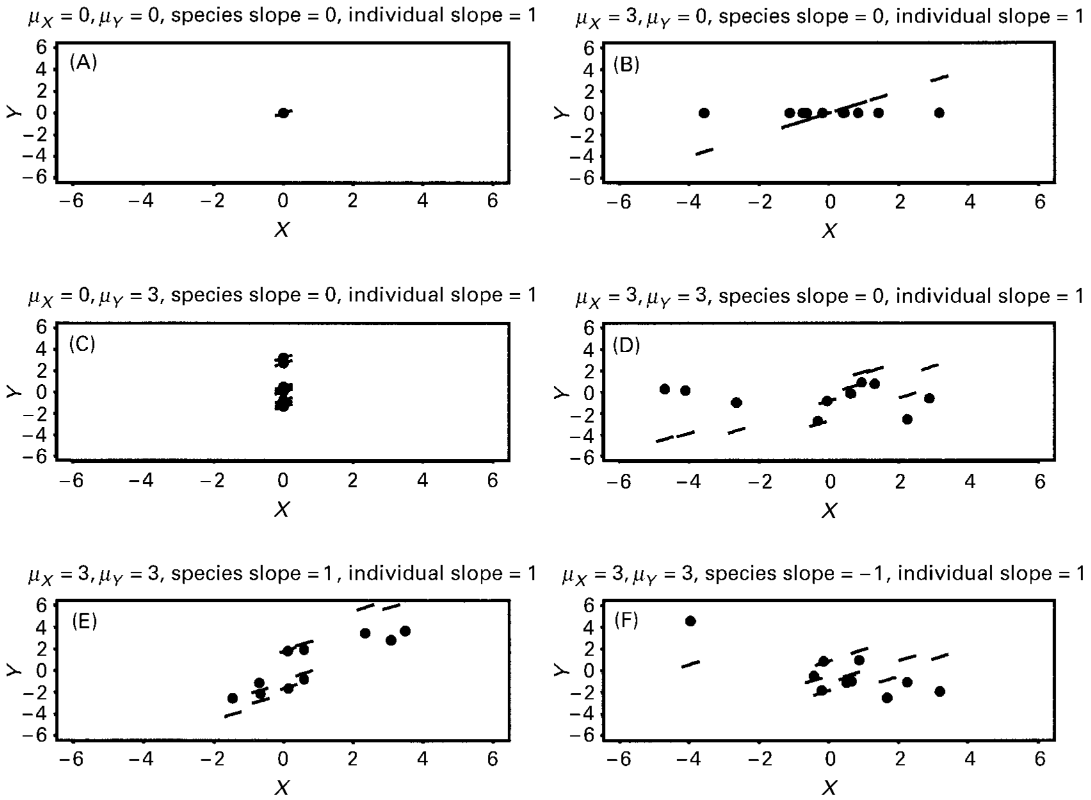


Figure 7.6. Simulated data based on Figure 7.5 based on six different scenarios involving the population mean values (μ_X, μ_Y) the species-level slope linking these population values, and the individual-level slope. The lines show the systematic relationships within each species and the solid circles show measured mean values of X and Y for each species.

Since all 10 species have the same means, these 10 circles are also superimposed.

In Figure 7.6B I simulate what happens if each species has a different mean value for *X* but has the same mean value for *Y*; that is, I allow μ_X to vary randomly but not μ_Y. All 10 lines – one for each species – appear to line up along the same trend as that observed at the intraspecific level. Notice that species whose mean for *X* (μ_{X_j}) is less than other species have their individual values of both *X* and *Y* (their line in the graph) in the lower left. Similarly, those species whose mean for *X* (μ_{X_j}) is greater than other species have their individual values of both *X* and *Y* (their line in the graph) in the upper right. In other words, there is a positive correlation between the mean values of *X* and *Y* of these 10 species even though there is no real relationship between the species means μ_{X_j} and μ_{Y_j}. To see why, we have only to write the generating equation for this simulation:

$$\mathrm{E}[Y|X]=(0)+1\mu_{X_j}+1\varepsilon_{X_{ij}}$$

Whenever the mean value of *X* for a given species (μ_{X_j}) happens, by chance, to be less than average, this decreases the values of *Y* that individuals of this species will possess. Similarly, whenever the mean value of *X* for a given species (μ_{X_j}) happens, by chance, to be greater than average, this increases the values of *Y* that individuals of this species will possess. The observed interspecific correlation that we observe is simply an artefact of mixing together the two levels of variation.

In Figure 7.6C I simulate what happens if each species has a different mean value of *Y* but the same mean value of *X*; that is, I allow μ_Y to randomly vary but not μ_X. Returning to our generating equation and substituting, we get:

$$\mathrm{E}[Y|X]=\mu_{Y_j}+1(0)+1\varepsilon_{X_{ij}}$$

The result is a series of lines stacked on top of each other. The overall correlation between *X* and *Y* is severely diluted.

In Figure 7.6D I simulate what happens if each species has a different mean value of both *X* and *Y* but there is still no true interspecific relationship between these mean values; that is, I allow both μ_X and μ_Y to vary randomly and independently. The result is intermediate between graphs B and C.

In Figure 7.5E I simulate what happens when each species both has a different mean value of *X* and *Y* and there is also a positive interspecific relationship between these mean values. To do this, I allow μ_X and μ_Y to randomly vary but link them: $\mu_{Y_j}=1\mu_{X_j}+\varepsilon_{\mu_{X_j}}$. Substituting this into the generating equation, I get:

$$E[Y|X]=\mu_{Y_j}+1\mu_{X_j}+1\varepsilon_{X_{ij}}=2\mu_{X_j}+\varepsilon_{\mu_{X_j}}+1\varepsilon_{X_{ij}}$$

The result is that the slope between the means appears twice as large as it really is.

Finally, in Figure 7.6F I simulate what happens when each species has a different mean value of *X* and *Y* and there is also a negative interspecific relationship between these mean values. In other words, the interspecific relationship is the opposite of the intraspecific relationship. To do this, I allow μ_X and μ_Y to vary randomly but link them: $\mu_{Y_j}=-1\mu_{X_j}+\varepsilon_{\mu_{X_j}}$. Substituting this into the generating equation, I get:

$$E[Y|X]=\mu_{Y_j}+1\mu_{X_j}+1\varepsilon_{X_{ij}}=0(\mu_{X_j})+\varepsilon_{\mu_{Xj}}+1\varepsilon_{\mathrm{Xij}}$$

The result is that the correlation between the means disappears even though there are really strong (but opposite) relationships between the variables at both hierarchical levels. The moral of this simple set of simulations is that combining data that have relationships at different levels and analysing it as if the hierarchical structure did not exist can lead to incorrect conclusions.

There is much more to be said about multilevel regression than has been said so far. What I have described is not so much an introduction as an appetiser and for more information the interested reader should consult the references given earlier in this chapter. Now that we recognise the problem that hierarchically organised data can cause, and have an intuitive understanding of how multilevel regression deals with it, let's see how these notions can be incorporated into SEM[11].

7.4 Multilevel SEM

Suppose that we have a variable that has been measured on *N* observations. These observations are organised into *G* groups with N_1, N_2, . . ., N_G observations in each group; for the moment we will assume that there are the same number (C) observations in each group. I write Y_{ij} to mean the *i*th observation in group *j* and I write $\overline{Y}_{\cdot j}$ to mean the mean of the value in the *j*th group. The deviation of this value from the overall mean is $(Y_{ij}-\overline{Y})$. We can decompose this deviance as follows: $(Y_{ij}-\overline{Y})=(\overline{Y}_{\cdot j}-\overline{Y})+(Y_{ij}-\overline{Y}_{\cdot j})$. This leads to the one-way ANOVA table that is fondly remembered by everyone who has taken an introductory course in statistics (Table 7.2).

[11] Although I do not discuss multilevel tests of path models based on the method presented in Chapter 3, the reader should be aware that multilevel regression methods can be used to obtain probability estimates of conditional independence by fitting a series of regressions in accordance with the hypothesised causal graph. These probability estimates can then be combined using Fisher's *C* statistic.

Table 7.2. *Decomposition of the variance in a one-way analysis of variance*

Source of variance	Sum of squares	Degrees of freedom	Variance
Total	$\sum_{i=1}^{G}\sum_{j=1}^{N_i}(Y_{ij}-\overline{Y})^2$	$N-1$	$\dfrac{\sum_{i=1}^{G}\sum_{j=1}^{N_i}(Y_{ij}-\overline{Y})^2}{N-1}$
Between groups	$C\sum_{j=1}^{G}(\overline{Y}_{.j}-\overline{Y})^2$	$G-1$	$\dfrac{C\sum_{j=1}^{G}(\overline{Y}_{.j}-\overline{Y})^2}{G-1}$
Within groups	$\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{ij}-\overline{Y}_{.j})^2$	$N-G$	$\dfrac{\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{ij}-\overline{Y}_{.j})^2}{N-G}$

The above decomposition has the useful property that the between-group deviations have zero correlation with the within-group deviations. Remembering that a variance is simply the covariance of a variable with itself, we can do the same trick with covariances. In this way, we can define both a pooled within-group covariance matrix ($\boldsymbol{S}_{\mathrm{PW}}$) and a between-group covariance matrix ($\boldsymbol{S}_{\mathrm{B}}$) for our data. The pooled within-group covariance matrix is constructed by first centring each variable by its group mean, calculating the sum of squares and cross-products of these group-centred variables, and then dividing by $N-G$. One easy way to obtain this matrix from any statistical program is simply to calculate the covariance matrix of the centred variables (which has a denominator of $N-1$), and multiply by $(N-1)/(N-\mathrm{G})$. The between-group covariance matrix is constructed by calculating the sum of squares and cross-products of the group means and dividing by $G-1$. An easy way to obtain this matrix from any statistical program is simply to calculate the covariance matrix of the group means, which has a denominator of $G-1$.

The sample pooled within-group covariance matrix ($\boldsymbol{S}_{\mathrm{PW}}$) is an unbiased estimate of the population within-group covariance ($\boldsymbol{\Sigma}_{\mathrm{PW}}$) matrix. Unfortunately, the sample between-group covariance matrix ($\boldsymbol{S}_{\mathrm{B}}$) is not a simple estimator of the population between-group covariance matrix ($\boldsymbol{\Sigma}_{\mathrm{B}}$); instead it estimates $\boldsymbol{S}_{\mathrm{B}}=\boldsymbol{\Sigma}_{\mathrm{PW}}+\mathrm{C}\boldsymbol{\Sigma}_{\mathrm{B}}$. If you look carefully at this equation then you will notice that it looks suspiciously like the multigroup structural equations formulation that we studied earlier in this chapter, with two 'groups'. We can therefore trick commercial SEM programs into fitting a multilevel SEM by treating it like a multigroup SEM with particular cross-group constraints.

To set up the analysis we tell the program that it is actually conducting a multigroup analysis with two groups. The first 'group' represents the group-centred data, for which there is the pooled sample covariance matrix obtained from the group-centred variables ($\boldsymbol{S}_{\mathrm{PW}}$) based on $N-G$ 'observations'. This is the level 1 covariance matrix. We specify our within-group causal structure for this 'group'. Next, we define a second 'group', for which we have the sample between-group matrix ($\boldsymbol{S}_{\mathrm{B}}$) based on G 'observations', where G is the number of groups in the multilevel model. For this second 'group' we specify both the within-group causal structure and the between-group causal structure. These two causal structures are linked by latent variables that represent the true values of the group means in the statistical population; remember that our calculated group means are only estimates of these underlying parameters. Since the variances and covariances of the group means are multiplied by the constant C (the number of individuals within each group), we fix the path coefficients leading from these latent variables to the individual variables by $\sqrt{C}$[12]. Finally, we must constrain all of the free parameters in the first 'group' (i.e. the model at the level of the individual) to be equal to the equivalent parameters in the second group (i.e. those parameters in the second group dealing with the model for the individual level). When we fit this model to our data the estimation procedure will correct for the partial non-independence of our data due to its hierarchical nature.

When we have different numbers of observations per group, we can calculate an approximate scaling factor due to Muthén:

$$C=\frac{N^2-\sum_{i=1}^{G}N_i^2}{N(G-1)}$$

Estimates based on this scaling factor have been shown to be fairly accurate so long as the group sizes are not extremely different (Hox 1993; McDonald 1994; Muthén 1994b). Of course, when group sizes are equal this reduces to the common group size. The parameter estimates, standard errors and maximum likelihood chi-squared statistics are still asymptotic values but now the requirement for sufficient sample sizes applies both at the level of

[12] If we have an equation: $Y=aX+e$ then the variances are $\mathrm{Var}(Y)=a^2\mathrm{Var}(x)+\mathrm{Var}(e)$. By setting the path coefficient from the latent variable representing the group mean to the individual-level variable at $\sqrt{C}$ (i.e. $L=\sqrt{C}X+e$) then we obtain $\mathrm{Var}(L)=C^\star\mathrm{Var}(X)+\mathrm{var}(e)$.

individuals and at the level of groups. Note that, at the level of groups, we are considering random samples of means. This means that the central limit theorem applies and the distribution of means will be closer to multivariate normal than is the distribution of the actual values.

In order to better understand how to interpret such multilevel structural equations models, I will analyse simulated data generated by different models. First, let's see what happens in the simplest case when all of the observations are really independent observations generated by the same causal process; in other words, there is really no group-level structure and all variances and covariances exist at the individual level. To do this, I generate 200 independent observations from the following equations:

$$X = \mathrm{N}(0,1)$$

$$Y = 0.5X + \mathrm{N}(0,0.75)$$

$$Z = 0.5Y + \mathrm{N}(0,0.75).$$

Now, I randomly divide these 200 independent observations into 40 groups of 5 observations each. Since this assignment is completely random, the only variation between the group means is due to sampling variation and the systematic variation in the group means is zero. Figure 7.7 shows the multilevel model. The variables M_1, M_2 and M_3 are latent variables representing the population means of each variable centred around the overall mean of each variable. Since there are 5 observations per group the scaling constant C is 5. Since there are no causal paths linking these latent variables, we are assuming that there is no covariance between them, although we could allow for such covariances; this will be shown in a later example.

I now fit two models, one nested within the other. First, I fit the model shown in Figure 7.7 (remembering to constrain all the free parameters at level 1 to be equal in the model associated with the within-group covariance matrix and the model associated with the between-group covariance matrix) while fixing the variance of the latent M_1, M_2 and M_3 to zero. The model in Figure 7.7 is the between-groups model that necessarily contains the within-groups model ($X \rightarrow Y \rightarrow Z$) nested within it. By fixing the variance of the latents M_1, M_2 and M_3 to zero I am assuming that the variance in the population means between groups are zero for all three variables and that any observed variance at this group level is due only to sampling variation. This assumption is, of course, correct for these data. The resulting model gives a maximum likelihood chi-squared statistic of 7.611 with 7 degrees of freedom, for a probability of 0.368.

There were two covariance matrices, the within-groups covariance matrix and the between-groups covariance matrix, and each had $(3 \times 4)/2 =$

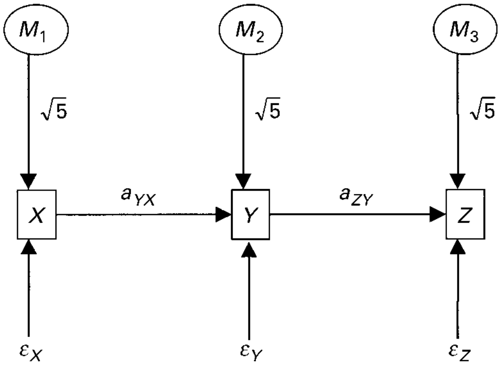


Figure 7.7. A two-level model involving three variables (*X*, *Y* and *Z*) and their population means.

6 non-redundant elements. Therefore, we had 12 non-redundant elements in total. The within-groups model had to estimate 5 free parameters (a_{YX}, a_{ZY}, ε_X, ε_Y and ε_Z). The between-groups model also had to estimate the same 5 free parameters since the variances of the latents M_1, M_2 and M_3 were fixed at zero. I also had to constraint the 5 free parameters associated with the within-groups model to be equal to these same free parameters in the between-groups model. Therefore, I had $12-5=7$ degrees of freedom.

Since this model, with the variances of the latents M_1, M_2 and M_3 fixed to zero, provides a non-significant probability level, we could stop there. However to test the hypothesis that there is really no group-level variance contributing to *X*, *Y* or *Z*, I now re-fit the model by allowing the variances of the latent M_1, M_2 and M_3 to be freely estimated. This new model gives a maximum likelihood chi-squared statistic of 7.475 with 4 degrees of freedom, for a probability level of 0.113. Note that we have reduced the degrees of freedom from 7 to 4 because we are now estimating three parameters that were previously fixed. Since this model is nested within the first, we can calculate the probability that the variances of M_1, M_2 and M_3 really were zero by calculating the difference in the chi-squared statistics (0.136) with 3 degrees of freedom. The resulting probability level (0.987) tells us that the observed variation in the group means was very likely to have been due only to sampling variation. If we go back to our first model and look at the estimated variances of M_1, M_2 and M_3 and the standard variation of these estimates, we find that each is very small and less than 1 standard error from zero. I have not explained the actual code needed to fit models in this book because each program does this differently, and most have user-friendly interfaces that hide much of the code anyway. However, since the multilevel model is more complicated, I show some ‘pseudo-code’ for the EQS program (Bentler 1995) in Box 7.1.

Box 7.1. EQS program code for a multilevel model

The following is the program code of the EQS program needed to fit the multilevel model shown in Figure 7.4. Note that the actual code generation is done automatically in EQS from a user-friendly interface. My comments are shown in italics.

```
/TITLE
   within-groups model
/SPECIFICATIONS
   DATA = ‘WITHIN.ESS’;
   VARIABLES = 3; CASES = 160;
   GROUP = 2
   METHODS = ML;
   MATRIX = COVARIANCE;
```

This section specifies that the first input data file, containing the pooled within-group covariance matrix, is called ‘WITHIN.ESS’, that there are 3 variables in this file, that it is a covariance matrix, and that it is based on 160 observations. Remember that there are really 200 observations, but the data are grouped into 40 groups. The within-group covariance matrix has 200 − 40 − 1 degrees of freedom. Finally, the code tells us that the overall model has 2 groups and the parameters are to be estimated using maximum likelihood techniques.

```
/LABELS

   V1 = X; V2 = Y; V3 = Z;
```

There are only four legal types of variable in EQS. Observed variables are called V, latent variables are called F, errors of observed variables are called E and the errors of latent variables are called D. The LABELS section just tells us how our variables' names (X, Y, Z) map onto the EQS variables.

```
/EQUATIONS
   V2 = + 1*V1 + E2;
   V3 = + 1*V2 + E3;
```

These are the equations for the within-groups section of the overall model. Note that there are no latent variables here. The asterisk indicates that there is a free parameter to be estimated (slope in this case), and that the starting guess for its value in the iterations is the number before the asterisk.

```
/VARIANCES
   V1 = *;
   E2 = ★;
   E3 = ★;
```

These are the variances whose free values are to be estimated. Again, the asterisk indicates that it is a free parameter whose value must be estimated. Since X (V1) is exogenous, its variance is estimated. Since Y (V2) and Z (V3) are endogenous, their error variances must be estimated.

```
/COVARIANCES
```

Since no entries have been given in the COVARIANCE section, this means that no free covariances are to be allowed.

```
/END
```

Now, we enter the model for the second model of the multigroup model. Remember that we are actually trying to trick the program into fitting a multilevel model using the syntax of a multigroup model.

```
/TITLE
   full between-group and within-group model
/SPECIFICATIONS
   DATA = 'between.ESS';
   VARIABLES = 4; CASES = 40;
   METHODS = ML;
   MATRIX = COVARIANCE;
```

This gives the same type of information as in the first section, except that we are using the file 'between.ESS' which holds the covariance matrix of the 40 group means.

```
/LABELS
   V1 = X; V2 = Y; V3 = Z;
/EQUATIONS
   V1 = +2.236F1 + E1;
   V2 = +2.236F2 + 1★V1 + E2;
   V3 = +2.236F3 + 1★V2 + E3;
```

Compare these equations with those in the first section; the variables F1, F2 and F3 are EQS code for three latent variables. These latents represent the population group means in this case. The difference is that both the within-group model and the between-group model are combined in this second 'group'. The between-group part links X(V1), Y(V2) and Z (V3) at level 1 to their group means, which are represented by the three latent variables. The path coefficients from these latents to their respective level-1 variables are fixed at $\sqrt{5} \approx 2.236$*. You can tell that these are fixed values because there is no asterisk after these values.*

```
/VARIANCES
   E1 = ★;
   E2 = ★;
   E3 = *;
   F1 = *;
   F2 = *;
   F3 = *;
```

Notice that we now specify the variances of the three latent variables representing the group means. Here, these latent variances are allowed to be freely estimated. If we want to fix them at zero, we replace the asterisks with 0. Note also that we don't specify a variance for V1 since, in this model, it is no longer exogenous (it is now caused by F1, its group mean). Instead, we have to specify its error variance (E1).

```
/COVARIANCES
```

*Again, we don't allow any free covariances. If we wanted to allow a free covariance between the population means of (say) X(V1) and Y(V2) then we would add the following line: 'F1, F2 = *;'.*

```
/CONSTRAINTS
   (1,V1,V1) = (2,E1,E1);
   (1,E2,E2) = (2,E2,E2);
   (1,E3,E3) = (2,E3,E3);
   (1,V1,V2) = (2,V1,V2);
   (1,V3,V2) = (2,V3,V2);
```

This is a critical part of the overall model. This section specifies which free parameters in the two models are constrained to be equal. We must make all free parameters in the first model (the within-group model) be equal to their equivalents in the second model. Thus, we state for instance that the error variance of V2 in model 1 – (1,E2,E2) – be equal to the error variance of V2 in model 2 – (2,E2,E2). Note also that V1 is exogenous in model 1 and so it has a variance (1,V1,V1) but it is endogenous in model 2 so it is represented by its 'error' variance – (2,E1,E1).

```
/END
```

The first simulation exercise was simply to show you that, if there really is no group-level variation that results in partial non-independence of the observations, then the multilevel model will detect this fact. Next, let's look at a case in which there really is group-level variation, but not group-level covariation. This time, we generate our 200 observations according to the following equations:

(A) For the 40 groups:

$$\mu_{X_j} = \mathrm{N}(0,1)$$

$$\mu_{Y_j} = \mathrm{N}(0,1)$$

$$\mu_{Z_j} = \mathrm{N}(0,1)$$

(B) For each of the five observations within each of the 40 groups:

$$X_{ij} = \mu_{X_j} + \mathrm{N}(0,1)$$
$$Y_{ij} = \mu_{Y_j} + 0.5X_{ij} + \mathrm{N}(0,1)$$
$$Z_{ij} = \mu_{Z_j} + 0.5Y_{ij} + \mathrm{N}(0,1)$$

Note that, in this simulation, each variable receives variation from two sources: the variation at level 1 (the group level) due to the random variation of each of the group means and the variation at level 2. I again fit two nested models, as in the first example. In the first model I fix all of the group-level variation associated with the latent M_1, M_2 and M_3 to zero (see Figure 7.7). This model does poorly (MLX^2=197.079, 7 df, $p<10^{-7}$), as it should. Since the variance of the group-level latent means was fixed at zero then all of the group-level error variance is incorrectly forced down to the within-group level. The resulting estimates of the three within-group error variances are 1.647, 1.904 and 1.857 instead of the correct value of 1. Now, I re-fit the model but allow the variances of the latent population means (M_1, M_2 and M_3) to be freely estimated. This time, the model provides an adequate fit (MLX^2=6.931, 4 df, p=0.140), as it should. Since this model is nested within the first, the difference in the maximum likelihood chi-squared statistic tests the hypothesis that there was group-level variance. This difference is highly significant (MLX^2=190.148, 3 df, $p<10^{-7}$). Both the group-level variances and the individual variances are correctly estimated. If we ignore the multilevel nature of these data and simply put all 200 observations into the same data set and fit the model $X \rightarrow Y \rightarrow Z$, then the model fails (MLX^2=6.206, 1 df, p=0.013). This is a general result (Muthén 1994a).

Up to now, we have seen how ignoring the multilevel nature of data can result in improper parameter estimates and probability levels. The real strength of multilevel SEM is that we can actually model how the group-level variables interact separately from the level 1 variables, and how these two levels interact together. To show this, I will simulate data from the process shown in Figure 7.8.

This model has four observed variables. Three of these variables, the number of seeds produced per plant, the seed size for an individual plant and the relative growth rate (RGR) of an individual plant, are properties of individual plants, and form the within-group model. The fourth variable, the average disturbance frequency of the habitat, is a property of each species; that is, individuals within a given species tend to be found in habitats with the same average frequency of disturbance. The model proposes that, at the level of individuals, increasing seed size causes an increase in

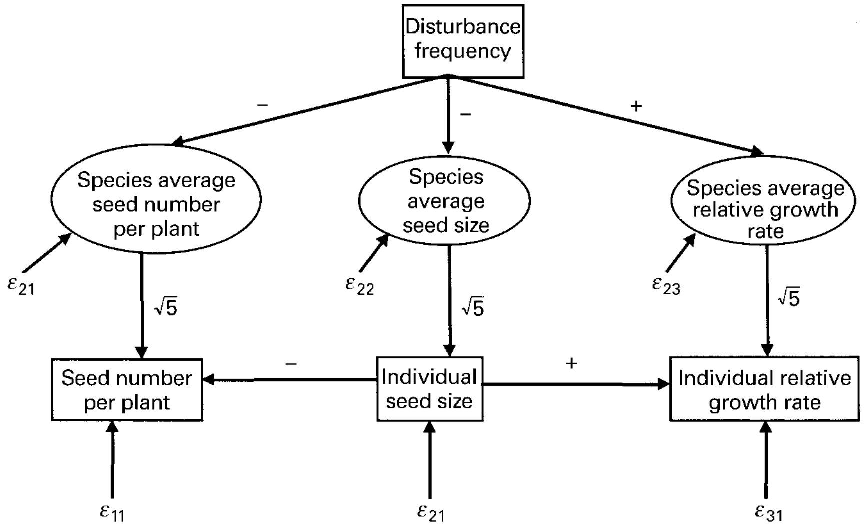

Disturbance frequency
−
−
+
Species average seed number per plant
Species average seed size
Species average relative growth rate
ε21
ε22
ε23
√5
√5
√5
Seed number per plant
Individual seed size
Individual relative growth rate
−
+
ε11
ε21
ε31

Figure 7.8. A hypothetical causal structure involving four observed variables and three latents. One of these observed variables (disturbance frequency) is a species-level variable that is a common cause of the three latent species' means. The other three observed variables are individual-level variables that are caused by the latent species' means.

RGR but a decrease in the number of seeds produced. At the level of species, selection for habitats of different disturbance frequencies results in species that are adapted to frequently disturbed habitats (early during secondary succession) producing fewer seeds (because they tend to be smaller plants), also smaller seeds on average, and seedlings having faster average RGRs. Note that the relationship between seed size and RGR is positive at the level of individuals but negative at an interspecific level. Figure 7.9 shows the simulated data set.

Before analysing these simulated data, I want you to notice some interesting trends. First, there is a negative relationship between seed size and RGR. This is because the species-level effect of selection, based on average disturbance frequency, dominates the within-species tendency for larger seeds to increase RGR. Second, notice that there is a positive relationship between seed number and disturbance frequency even though the direct effect of disturbance frequency on these two variables, at an interspecific level, is negative. This is because an increasingly disturbed habitat selects for species with smaller seeds on average and this, in turn, reduces the seed size within such species. However, smaller seeds increase seed number within a given species resulting in an overall effect along this path

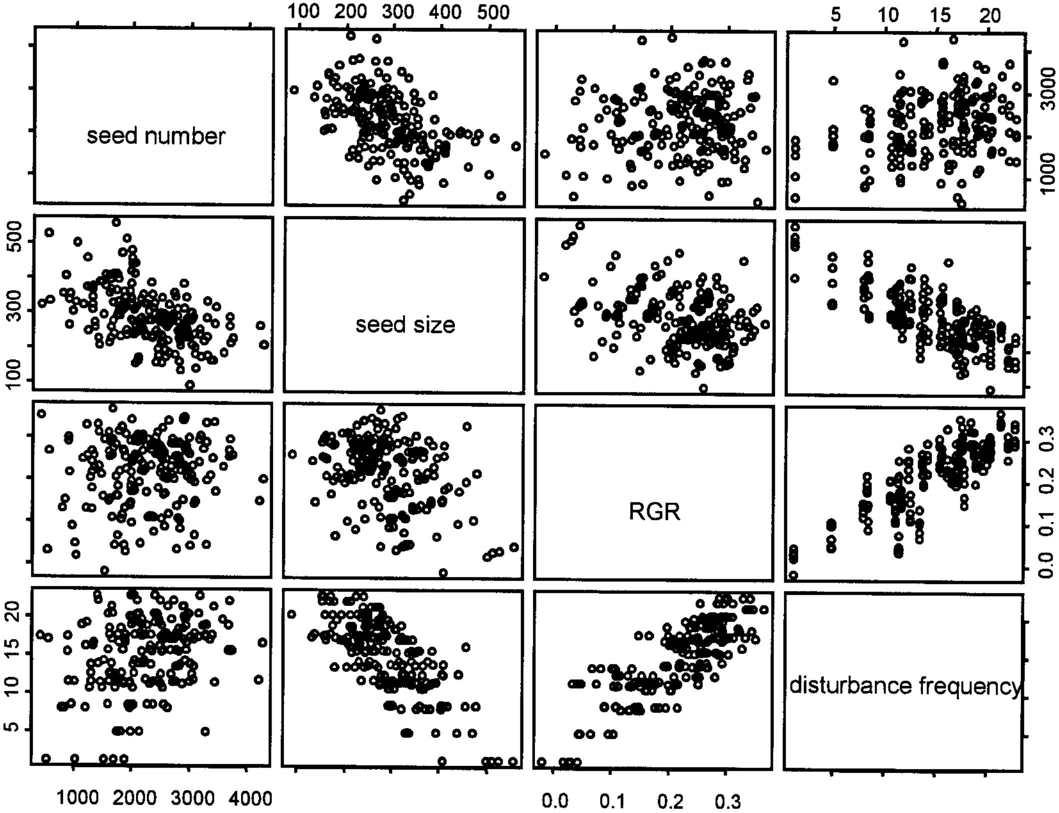


Figure 7.9. The scatterplot matrix of the simulated data generated by the causal structure shown in Figure 7.8.

being positive. Since this path dominates the direct effect at the species level, we see a positive overall relationship.

Now, we fit a series of nested models. First, we specify no variance or covariance at the species level. This model is rejected (MLX$^2 = 746.249$, 7 df, $p < 10^{-10}$). Next, we allow variance, but not covariance, at the species level. This nested model is also rejected (MLX$^2 = 51.385$, 4 df, $p < 10^{-10}$) but the change in the maximum likelihood chi-squared statistic is significant (MLX$^2 = 694.864$, 3 df, $p < 10^{-10}$) showing that there is significant species-level variation. Finally, we allow the three latent variables to freely covary amongst themselves. This model, nested within the second, provides an acceptable fit (MLX$^2 = 2.096$, 1 df, $p = 0.148$) and the change in the maximum likelihood chi-squared statistic is significant (MLX$^2 = 49.289$, 3 df, $p = 1.13 \times 10^{-10}$) showing that there is also species-level covariation.

Since there is significant covariation between the latent species-level means, we introduce the species-level variable 'average disturbance frequency' and specify that this variable is the sole common cause of these three species-level variables. This model (which is the true model that generated these data) also provides an acceptable fit (MLX$^2 = 5.123$, df$= 4$, $p = 0.275$). Figure 7.10 shows the final parameter estimates and their standard errors in parentheses.

The parameter estimates are all close to the true values that I had simulated, and all are within the approximate 95% confidence intervals (two times the standard errors). This model is quite remarkable. Not only have we been able to account for the partial non-independence of the data within each species due to their hierarchical nature, but we have also been able to separate the within-species structure from the between-species structure and link the two hierarchical processes together. Although the data were simulated, they are biologically realistic. Natural selection based on some average environmental property determines the average values of attributes shown by different species and the covariation between these average values. These average values then limit the range of values shown by particular individuals within each species but still allow variation between individuals and co-variation of the attributes at the level of individuals.

The multigroup model can be extended to more than two levels. For simplicity, let's imagine that our data are grouped into G different genera, S different species and I different individuals. The value for our attribute in individual i of species j of genus k is Y_{ijk}. Now, we can write:

$$Y_{ijk} = (\overline{Y}_{..k} - \overline{Y}) + (\overline{Y}_{.jk} - \overline{Y}_{..k}) + (Y_{ijk} - \overline{Y}_{.jk})$$

where $\overline{Y}$ is the grand mean over all I observations, $\overline{Y}_{..k}$ is the mean of Y for each genus, and $\overline{Y}_{.jk}$ is the mean of Y for each species. It follows that we can

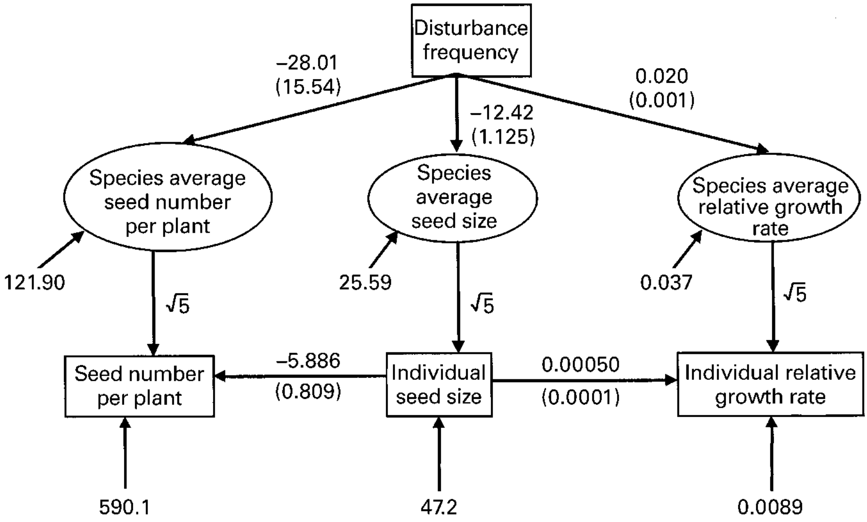


Figure 7.10. The maximum likelihood estimates of the free parameters of Figure 7.8, based on the simulated data. Values in parentheses are the standard errors of the parameter estimates.

decompose the sum of squares of Y_{ijk} into a term representing the deviations of each genus mean from the grand mean, $(\overline{Y}_{..k}-\overline{Y})$, a term representing the deviations of each species from its genus mean $(\overline{Y}_{.jk}-\overline{Y}_{..k})$ and a term representing each individual from its species mean $(Y_{ijk}-\overline{Y}_{.jk})$. Following exactly the same logic as we used to derive the two-group multilevel model, we can therefore obtain a genus-level sample covariance matrix ($\boldsymbol{S}_{\mathrm{G}}$), a species-level pooled sample covariance matrix ($\boldsymbol{S}_{\mathrm{PS}}$) and an individual-level pooled sample covariance matrix ($\boldsymbol{S}_{\mathrm{PI}}$)[13]:

$$\boldsymbol{S}_{\mathrm{PI}}=\frac{\sum_{k=1}^{G}\sum_{j=1}^{S}\sum_{i=1}^{I}(Y_{ijk}-\overline{Y}_{.jk})^2}{N-S}$$

$$\boldsymbol{S}_{\mathrm{PS}}=\frac{\sum_{k=1}^{G}\sum_{j=1}^{S}N_{.jk}(\overline{Y}_{.jk}-\overline{Y}_{..k})^2}{S-G}$$

$$\boldsymbol{S}_{\mathrm{G}}=\frac{\sum_{k=1}^{G}N_{..k}(\overline{Y}_{..k}-\overline{Y})^2}{G-1}$$

[13] My Toolbox (see Appendix) contains a program to calculate these multilevel covariance matrices.

If our data are completely balanced with N_G species per genus and N_S individuals per species, we can write: $\boldsymbol{\Sigma}_T = N_G\boldsymbol{\Sigma}_G + N_S\boldsymbol{\Sigma}_S + \boldsymbol{\Sigma}_I$. If this is not the case then we have to use Muthén's approximate scaling factor for each level of the hierarchy:

$$C_L = \frac{N^2 - \sum_{i=1}^{N_L} N_{iL}^2}{N(N_L - 1)}$$

Here, C_L is the scaling factor for level L of the hierarchy, N is the total number of observations in the data set, N_L is the total number of units at level L (for instance, the number of species or genera) and N_{iL} is the number of units within group i of level L. In this way, we can model structural relationships at various levels of organisation and account for partial non-independence due to common ancestry at various taxonomic levels. Of course, these levels do not have to represent traditional taxonomic classifications. For instance, you might have measures at different times for the same individual, defining a within-individual level. Some readers might have noticed that the above description looks much like a nested Type II ANOVA. Box 7.2 describes the relationship for those who are interested.

Box 7.2. Variance components in a nested ANOVA

Consider a typical balanced ANOVA table for a nested analysis with three levels. There are 160 observations. These observations are grouped into $N_1 = 40$ level 1 groups, $N_2 = 80$ level 2 groups (i.e. two level 2 groups per level 1 group) and $N_3 = 160$ level 3 observations (two level 3 observations per level 2 group).

Source	df	SS	MS	Expected MS
Level 1	$N_1 - 1 = 40 - 1$	SS_1	$SS_1/(N_1 - 1)$	$\sigma^2_{L_3 \subset L_2} + C_2\sigma^2_{L_2 \subset L_1} + C_1\sigma^2_{L_1}$
Level 2	$N_2 - N_3 = 80 - 40$	SS_2	$SS_2/(N_2 - N_3)$	$\sigma^2_{L_3 \subset L_2} + C_2\sigma^2_{L_2 \subset L_1}$
Level 3	$N_3 - N_2 = 160 - 80$	SS_3	$SS_3/(N_3 - N_2)$	$\sigma^2_{L_3 \subset L_2}$
Total	$N_3 - 1 = 160 - 1$	SS_T	$SS_T/(N_3 - 1)$	

Note:
SS, sum of squares; MS, mean square.

Here, the notation $\sigma^2_{L_3 \subset L_2}$ means the variation of the level 3 units nested within the level 2 units. C_2 is the number of level 3 units within each level 2 unit and C_1 is the number of level 2 units within each level 1 unit. We see that the higher level mean squares (which are sample variances if the units are randomly sampled) do not estimate variation unique to that level but a weighted sum of variation at that level and all levels below it.

If we subtract the variance (i.e. MS) at a given level with the variance directly below it in the hierarchy, and divide by the number of observations per unit (i.e. C_i), then we obtain an estimate of the *variance component* at that level. For instance, to obtain the variance component at level 1 we write:

$$\frac{[(\sigma^2_{\mathrm{L}_3\subset\mathrm{L}_2} + C_2\sigma^2_{\mathrm{L}_2\subset\mathrm{L}_1} + C_1\sigma^2_{\mathrm{L}_1}) - (\sigma^2_{\mathrm{L}_3\subset\mathrm{L}_2} + C_2\sigma^2_{\mathrm{L}_2\subset\mathrm{L}_1})]}{C_1} = \sigma^2_{\mathrm{L}_1}$$

We estimate this variance component by calculating:

$$\frac{[\mathrm{MS}_{\mathrm{L}_1} - \mathrm{MS}_{\mathrm{L}_2}]}{C_1}$$

The variance (i.e. MS) at a given level measures the total amount of variation that is found at that level. However, such variation is due to the combined effect of variation at lower levels and the added variation contributed at that level. The variance components measure the amount of added variation at each level. Usually, one expresses these variance components as percentages of the total variation.

A variance (or a sum of squares) is simply a special type of covariance (or sum of cross-products); namely the covariance of a variable with itself. We can therefore apply the same logic to each variance and covariance in a covariance matrix. If we measure a whole set of variables on each observational unit instead of only one then we can produce a table that summarises the decomposition of the entire covariance matrix. Rather than sums of squares (SS), we would calculate sums of squares and cross-products (SSCP). Rather than mean squares (variances) we would calculate mean squares and cross-products (MSCP), i.e. covariances.

Source	df	SSCP	MSCP	Expected MSCP
Level 1	$N_1 - 1 = 40 - 1$	SSCP_1	$\mathrm{SSCP}_1/(\mathrm{N}_1 - 1)$	$\boldsymbol{\Sigma}^2_{\mathrm{L}_3\subset\mathrm{L}_2} + C_2\boldsymbol{\Sigma}^2_{\mathrm{L}_2\subset\mathrm{L}_1} + C_1\boldsymbol{\Sigma}^2_{\mathrm{L}_1}$
Level 2	$N_2 - N_3 = 80 - 40$	SSCP_2	$\mathrm{SSCP}_2/(\mathrm{N}_2 - \mathrm{N}_3)$	$\boldsymbol{\Sigma}^2_{\mathrm{L}_3\subset\mathrm{L}_2} + C_2\boldsymbol{\Sigma}^2_{\mathrm{L}_2\subset\mathrm{L}_1}$
Level 3	$N_3 - N_2 = 160 - 80$	SSCP_3	$\mathrm{SSCP}_3/(\mathrm{N}_3 - \mathrm{N}_2)$	$\boldsymbol{\Sigma}^2_{\mathrm{L}_3\subset\mathrm{L}_2}$
Total	$N_3 - 1 = 160 - 1$	SSCP_T	$\mathrm{SSCP}_\mathrm{T}/(\mathrm{N}_3 - 1)$	

The variance components can be extracted from the diagonal elements of these covariance matrices.

One important type of multilevel model involves repeated measurements over time on the same set of individuals. In such a sampling design, the first level would be the intra-individual level (i.e. variation over time in the same individual); this is analogous to 'repeated-measures' analyses with which many biologists are familiar. Unfortunately, I am not aware of any multilevel models that have been used in a biological context and only very few in any other context (Muthén 1990, 1994a,b; Muthén and Satorra 1995). None the less, I suspect that multilevel models will become very important in biology, since hierarchies are so ubiquitous.

8 Exploration, discovery and equivalence

8.1 Hypothesis generation

If this were a textbook of statistics then this chapter would not exist. Modern statistics is almost entirely concerned with *testing* hypotheses, not *developing* them. This bureaucratic approach views science as a compartmentalised activity in which hypotheses are constructed by one group, data are collected by another group and then the statistician confronts the hypothesis with the data. Since this book is a user's guide to causal modelling such a compartmentalised approach will not do. One of the main challenges faced by the practising biologist is not in testing causal hypotheses but in developing causal hypotheses worth testing.

If this were a book about the philosophy of science then this chapter might not exist either. The philosophy of science mostly deals with questions such as: 'How can we know whether a scientific hypothesis is true or not?' or 'What demarcates a scientific hypothesis from a non-scientific hypothesis?'. For most philosophers of science the question of how one looks for a useful scientific hypothesis in the first place is someone else's problem. For instance, Popper's (1980) influential *Logic of scientific discovery* says that 'there is no such thing as a logical method of having new ideas, or a logical reconstruction of this process. My view may be expressed by saying that every discovery contains "an irrational element", or "a creative intuition". . .' Later, he says that '[scientific laws] can only be reached by intuition, based on something like an intellectual love of the objects of experience.' Again, one gets the impression that science consists to two hermetically sealed compartments. One compartment, labelled 'hypothesis generation', consists of an irrational fog of thoughts and ideas, devoid of method, out of which a few gifted people are able to extract brilliant insights. The other compartment, labelled 'hypothesis testing', is the public face of science. Here, one finds method and logic, in which established rules govern how observations are to be taken, statistically manipulated and interpreted.

At a purely analytical level there is much to be gained by taking this schizophrenic view of the scientific process. After all, how a scientific idea is developed is irrelevant to its truth. For instance, the history of science documents many important ideas whose genesis was bizarre[1]. Archimedes reportedly discovered the laws of hydrostatics after jumping into a bathtub full of water. Kukulé discovered the ring structure of benzene after falling asleep before a fire and dreaming of snakes biting their tails. These curious stories are entertaining but we remember them only because the laws of hydrostatics hold and benzene really does have a ring structure. As a public activity, science is interested in the result of the creation, not in the creative act itself.

The day-to-day world of biology does not exist at such a purely analytical level. Although it is possible to conceptually divide science into distinct hypothesis-generation and hypothesis-testing phases, the two are often intimately intertwined in practice. When the two are not intertwined the science can even suffer. Peters (1991), in his *A critique for ecology*, pointed out that because empirical and theoretical ecology are often done by different people, the result is that much ecological theory is crafted in such a way that it can't be tested in practice and much of field ecology can't be generalised because it is not placed into a proper theoretical perspective. In this context I like the citation, attributed to W. H. George, given at the beginning of Beveridge's (1957) *The art of scientific investigation*: 'Scientific research is not itself a science; it is still an art or craft.' Unlike the assembly-line worker who receives a partly finished object, adds to it, and then passes it along to someone else, the craftsman must construct the object from start to finish. In the same way the craft of causal modelling consists as much of the generation of useful hypotheses as of their testing. Certainly hypothesis generation is more art than method, and hypothesis testing is more method than art, but this does not mean that we must relegate hypothesis generation to a mystical world of creative intuition in which there are no rules. The purpose of this chapter is to describe reliable methods of generating causal hypotheses.

8.2 Exploring hypothesis space

How does one go about choosing promising hypotheses concerning causal processes? To place the problem in context, imagine that you have collected data on N variables and at least some of these variables are not amenable to controlled randomised experiments. Why you suspect that these N variables

[1] The appendix of *The art of scientific investigation* (Beveridge 1957) lists 19 cases in which the origin of important scientific ideas arose from bizarre or haphazard situations. In fact, Beveridge devotes an entire chapter to the importance of chance in scientific discovery.

Table 8.1. *The number of different cyclic causal graphs without latent variables that can be constructed given N variables*

N	Number of graphs
2	4
3	64
4	4096
5	1 048 576
6	1 073 741 824

possess interesting or important causal relationships may well be due to the irrational creative intuition to which Popper referred, but you are still left with the problem of forming a multivariate hypothesis specifying the causal connections linking these variables.

To simplify things, let's assume that all of the data are generated by the same unknown causal process (i.e. causal homogeneity), that there are no latent variables responsible for some observed associations (i.e. causal sufficiency) and that the data are faithful[2] to the causal process. How many different causal graphs could exist under these conditions? Each pair of variables (X and Y) can have one of four different causal relationships: X directly causes Y, or Y directly causes X, or X and Y directly cause each other, or the two have no direct causal links. We now have to count up the number of different pairs of variables, which is just the number of combinations of two objects out of N. The combinatorial formula is therefore

$$4^{\frac{N!}{2!(N-2)!}}$$

Table 8.1 gives the number of different potential causal graphs of this type that can exist given N variables.

If we think of the full set of potential causal graphs having N variables as forming an 'hypothesis space', and your research program as a search through this space to find the appropriate causal graph, then Table 8.1 is bad news. Even if we could test one potential graph per second it would take us

[2] See Chapter 2 for the definition of faithfulness. In fact, much of the present chapter makes use of notions introduced in Chapter 2, and the reader might want to re-read that chapter before continuing.

almost 32 years to test every potential graph containing only six variables! If we were to restrict our problem to acyclic graphs then the numbers would be smaller, but still astronomical (Glymour *et al.* 1987). If it is true that the process of hypothesis generation (in this case, proposing one casual graph out of all those in the hypothesis space) is pure intuition, devoid of method, then it is a wonder that science has made any progress at all. That science *has* made progress shows that efficient methods of hypothesis generation, although perhaps largely unstated, do exist.

So how should we go about efficiently exploring this hypothesis space? To go back to my previous question: how does one go about generating promising hypotheses concerning causal processes (Shipley 1999)? One way would be to choose a graph at random and then collect data to test it. With five variables there is a bit less than one chance in a million of hitting on the correct structure. There is nothing logically wrong with such a search strategy; we will have proposed a falsifiable hypothesis and tested it. However, no thinking person would ever attempt such a search strategy because it is incredibly inefficient. We need search strategies that have a good chance of quickly finding those regions of hypothesis space that are likely to contain the correct answer. What would be our chances of hitting on the correct structure if we were to appeal only to 'pre-existing theory', as recommended by many SEM books? Clearly that would depend on the quality of the pre-existing theory. If, however, the theory really was so compelling that the researcher did not feel a need to search for alternatives then the problem would be firmly within the 'hypothesis testing' compartment and no question of a search strategy would be posed.

Very often biologists find themselves in the awkward position of straddling the 'hypothesis-generation' and 'hypothesis-testing' compartments. Often, we have some background knowledge that excludes certain causal relationships and suggests others, but not enough firmly established background knowledge to specify the full causal structure without ambiguity. In such situations the goal is not to test a pre-existing theory – which might not be sufficiently compelling to justify allocating scarce resources and time to testing it – but rather in developing a more complete causal hypothesis that would be worth testing with independent data. The real problem is less in testing hypotheses than in finding hypotheses that are worth testing in the first place. We need search strategies that can be proved to be efficient at exploring hypothesis space, at least given explicitly stated assumptions. Until very recently such search strategies, which are described in this chapter, did not exist. You will see that these search strategies rely heavily on the notion of d-separation and on how this notion allows a translation from causal graphs to probability distributions.

8.3 The shadow's cause revisited

I have repeatedly compared the relationship between cause and correlation to the relationship of an object and its shadow. There is something missing in this analogy when applied to actual research projects. When we measure a correlation in a sample of data we are almost never interested in the value as such. Rather, we use the value to infer what the correlation might be in the population from which we randomly chose our sample data. It is as if, in nature's Shadow Play, not only do the causal processes cast potentially ambiguous correlational shadows, but these shadows are randomly blurred as well. We therefore have two problems. First, we have to find a way of provably deducing causal processes from correlational shadows and, second, we have to take into account the inaccuracies caused by using sample correlations to infer population correlations. It is important to keep these two problems distinct. The second problem, that of dealing with sampling variation, is a typical problem of mainstream statistics. For this reason, we will first see how to go from correlations to causes when there is no sampling variation. In other words, we will consider asymptotic methods.

The history of the development of these exploratory methods, or 'search' algorithms, is fascinating. The word 'history' has connotations of age but, in fact, all of these methods date to less than 10 years before the writing of this book. The mathematical relationships between graphs, d-separation and probability distributions were worked out in the mid 1980s by Judea Pearl and his students at the University of California at Los Angeles (UCLA) (Pearl 1988). This was the translation device between the language of causality and the language of probability distributions that had been missing for so long. As soon as it became possible to convert causal claims into probability distributions the dam was burst and the conceptual flood came pouring out. It became immediately obvious that one could also convert statements concerning probabilistic independencies into causal claims. Pearl and his team at UCLA developed a series of algorithms to extract causal information from observational data during the period 1988–1992[3]. Interestingly, a group of people at the Philosophy Department at Carnegie–Mellon University (Clark Glymour, Peter Spirtes, Richard Scheines and their students) had also been working on the same goal. In the late 1980s they had published a book (Glymour *et al.* 1987) in which zero partial correlations and vanishing tetrad differences were used to infer causal structure, but without the benefit of d-separation or the mathematical link between

[3] This brief history, and the algorithms of Pearl and his students, are given in Chapter 2 of Pearl (2000).

causal graphs and probability distributions. As soon as the Carnegie-Mellon group encountered Pearl's work on d-separation (they didn't know about the discovery algorithms of Pearl) they immediately began to independently derive and prove almost identical search algorithms. These algorithms (and much more) were proved and published in Spirtes, Glymour and Scheines (1993) and incorporated into their TETRAD II program. An algorithm called the Inductive Causation algorithm was proved and published by Verma and Pearl (1991) and is very similar to the Causal Inference algorithm of the Carnegie–Mellon group that is presented in this chapter. I will leave it to the people involved to sort out questions of priority. I think that it is fair to say that once the d-separation criterion was developed the various algorithms were 'in the air' and had only to be brought down to earth by those with the knowledge. The philosopher's dream of inferring (partial knowledge of) causation from observational data had been realised.

In Chapter 2 I explained how to translate from the language of causality, with its inherently asymmetric relationships, to the language of probability distributions with its inherently symmetric relationships. The Rosetta Stone allowing this translation was the notion of d-separation. Using d-separation we could reliably convert the causal statements expressed in a directed acyclic graph into probabilistic statements of dependence or independence that are expressed as (conditional) associations. This translation strategy was used in Chapters 3 to 7 to allow us to test hypothesised causal models using observational data.

Now that we are attempting to discover causal relationships, the problem has been turned on its head. We have to start with probabilistic statements of (conditional) dependence or independence and somehow back-translate into the language of causality. As you will see, this back-translation is almost always incomplete. There is almost always more than one acyclic causal graph that implies the same set of probabilistic statements of (conditional) dependence or independence. In other words, there are almost always different acyclic causal graphs that make different causal predictions but exactly the same predictions concerning probabilistic dependence or independence. This gives rise to the topic of equivalent models, a topic that has been recognised in SEM for a long time and generally ignored for just as long.

The methods that I describe in this chapter are based on the strategy of back-translation that I described above. The first step is to obtain a list of probabilistic statements of (conditional) dependence or independence involving the variables in question. From this list, we construct an *undirected dependency graph*. An undirected dependency graph looks like a causal graph in which all of the arrows have been converted into lines without arrowheads. However, the lines in the undirected dependency graph have a very different meaning. Two variables in this graph have a line between them if they are probabilistically dependent conditional on every subset of other variables in the graph. The lines in the undirected dependency graph express symmetrical associations, not asymmetrical causal relationships. Since we can't measure associations involving variables that we have not measured, the undirected dependency graph can't have latent variables. The next step is to convert as many of the symmetrical relationships in the undirected dependency graph as possible into asymmetrical causal relationships. This is called *orienting* the edges and uses the notion of d-separation[4]. Generally, not all of the undirected lines will be converted into directed arrows and so we do not end up with a directed graph. Rather, we end up with a partially oriented graph.

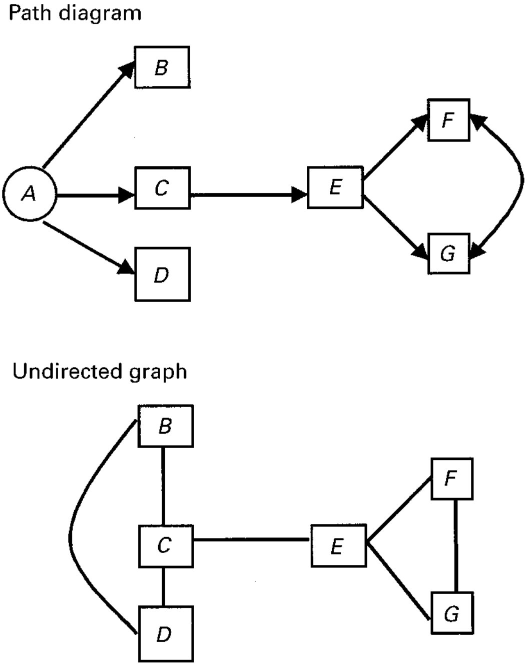
8.4 Obtaining the undirected dependency graph

Before I explain how to obtain an undirected dependency graph from observational data, it is useful to explore how to convert a directed acyclic graph into an undirected dependency graph involving only measured variables. Doing this will help to underscore the difference between the undirected dependency graph and the causal graphs with which you are now familiar. In acyclic graphs without latent variables, the undirected dependency graph is simply the directed acyclic graph in which all of the arrows are replaced with lines lacking arrowheads. However, for the method to be useful in discovery, we can work only with those variables that we have actually measured. If the directed graph contains latent variables then the resulting undirected dependency graph, involving only observed variables, will usually require modifications and these modifications help to illustrate the proper interpretation of such graphs. To get the undirected graph from a directed acyclic graph[5] (Figure 8.1), or from a typical acyclic path diagram if it contains correlated errors, do the following things.

1. If there is not already an arrow or curved double-headed arrow between any two observed variables, but d-separation of the pair requires conditioning on latent variables, then draw a line (not an arrow) between the pair; see *inducing paths* (pp. 250–253).
2. If there are curved double-headed arrows between any pairs of variables (i.e. correlated errors), then replace these with a line (not an

[4] Vanishing Tetrad differences (Chapter 5) are also used, but these zero tetrad equations can be reduced to statements concerning d-separation involving latent variables.

[5] The case of cyclic directed graphs will be dealt with later.



Path diagram

Undirected graph

Figure 8.1. A path diagram (top) involving six observed variables and one latent variable. Below is the undirected dependency graph corresponding to this path diagram.

arrow).

3. Remove the latent variables and also any arrows going into, or out of, these latent variables.
4. Change all remaining arrows to lines.

The top of Figure 8.1 shows a path diagram with both latent variables and correlated errors. The bottom of Figure 8.1 shows the undirected dependency graph that results when considering only the observed variables. In Figure 8.1 there is a line between $\{B,C\}$, $\{B,D\}$ and $\{C,D\}$ in the undirected dependency graph even though these pairs of variables were not adjacent in the original path diagram. This is because, following the first rule, d-separation of each of these pairs required conditioning on a latent variable (A). Since we have not measured variable A we can't condition on it, and so the three pairs of observed variables remain probabilistically associated even after conditioning on any set of other observed variables. Similarly, there is a line between $\{F,G\}$ in the undirected dependency graph even though the two variables were not adjacent in the original path diagram. This is because, following the second rule, this pair of variables has correlated errors. After changing all remaining arrows in the path diagram to lines, we end up with the undirected graph.

A *direct* cause between two variables in a causal graph is a causal relationship between them that can't be blocked by other variables or a set of variables involved in the causal explanation. Similarly, we can define a *direct association* between two variables in an undirected dependency graph as an association that can't be removed upon conditioning on any other observed variable or set of variables[6]. The undirected dependency graph is therefore a graph that shows these direct associations. Of course, if we are attempting to discover causal relationships then we will not already have the directed acyclic graph. Our first task is therefore to discover the undirected dependency graph from the data alone (remembering that, for the moment, we are assuming that our sample size is so large that we can ignore sampling variation) when we don't know what the true directed acyclic graph looks like.

Let's begin with the following assumptions:

1. Every unit in the population is governed by the same causal process (i.e. causal homogeneity).
2. The probability distribution of the observed variables measured on each unit is faithful to some (possibly unknown) cyclic[7] or acyclic causal graph.
3. For each possible association, or partial association, among the measured variables, we can definitely know whether the association or partial association exists (is different from zero) or does not exist (is equal to zero). This is simply the assumption that there is no sampling variation.

We don't have to assume that there are no unmeasured variables generating some associations (this assumption is called *causal sufficiency*) or that the variables follow any particular probability distribution, or that the causal relationships between the variables take any particular functional form. No assumptions of an acyclic structure are needed, although the algorithms for cyclic structures require linearity in the functional relationships between variables. The method uses d-separation and we know the d-separation implies zero (partial) associations (Spirtes 1995; Pearl and Dechter 1996) under such conditions. Unfortunately, we don't yet know whether the converse is true; that is, whether there can be independencies generated

[6] This will be more formally defined as an inducing path later on.

[7] The subsequent orientation phases will differ depending on whether or not we assume an acyclic structure.

by cyclic causal processes that are not implied by d-separation (Spirtes 1995). The assumption concerning causal homogeneity can be partly relaxed as well, as is described later.

Given these assumptions, Pearl (1988) has proved that there will be an edge (a line) in our undirected dependency graph between a pair of variables (X and Y) if X and Y are dependent conditional on every set of variables in the graph that does not include X or Y. We can therefore discover the undirected graph of the causal process that generated our data by applying the algorithm below. Following the definition of the order of a partial correlation, let's define the conditioning *order* of an association as the number of variables in the conditioning set. So, a zero-order association is an association between two variables without conditioning, a first-order association is an association between two variables conditioned on one other variable, and so on. How one measures these associations will depend on the nature of the data; the various methods described in Chapter 3 can be used for different types of data.

8.5 The undirected dependency graph algorithm[8]

The first step is to form the *complete* undirected graph involving the V observed variables. In other words, add a line between each variable and every other variable. Since latent variables are, by definition, unmeasured, we can't include them in our complete undirected graph. Now, for each unique pair of observed variables (X, Y) that have a line between them in the undirected dependency graph at any stage during the implementation of the algorithm, do the following:

1. Let the order of the association be zero.

2.1 Form every possible set of conditioning variables, containing the number of variables specified by the order, out of the remaining observed variables in the graph.

2.2 If the association between the pair of variables (X,Y) is zero when conditioned on any of these sets, then remove the line between X and Y from the undirected dependency graph, move on to a new pair of variables and then go to step 1.

2.3 If the association between the pair of variables (X,Y) is not zero when conditioned on all of these sets, then increase the order of the association by one, and go to step 2.1. If you cannot increase the conditioning order, then the line between your two variables is kept. Move on to a new pair of variables.

[8] This algorithm is included in the SGS algorithm of Spirtes, Glymour and Scheines (1990).

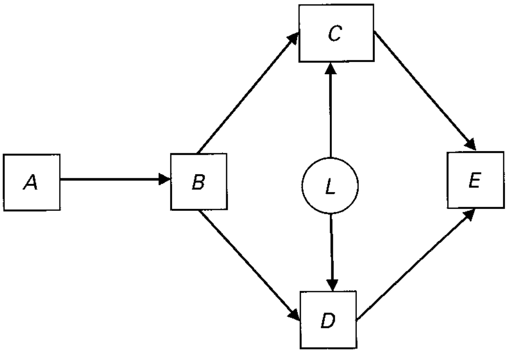


Figure 8.2. A directed graph, including one latent variable, used to illustrate the undirected dependency graph algorithm. This causal graph is unknown to the observer.

Once you have applied this algorithm to every set of observed variables, the result is the undirected dependency graph. Given the assumptions listed above, you are guaranteed to obtain the correct undirected dependency graph of the causal process that generated the data if the algorithm is properly implemented.

To illustrate this algorithm, let's imagine that we have been given data (lots of it so that we do not have to worry about sampling variation) that, unknown to us, was generated by the causal graph shown in Figure 8.2.

Now, we don't know about Figure 8.2; this causal structure is hidden behind the screen of nature's Shadow Play. In fact, we might not even know of the existence of the latent variable (L), since, had we known about it, we probably would have measured it. All that we have is a (very large) data set containing observations on the variables A to E and a series of measures of association and partial association between them; these are the shadows that we can observe on the screen. Our task is to infer as much about the structure of Figure 8.2 as we can. To begin, we create the complete undirected dependency graph of these five variables (Figure 8.3).

Notice that the latent variable (L) doesn't appear in Figure 8.3 because we are dealing only with observed variables at this point. Let's begin with the pair (A,B) and apply the algorithm. Since A and B are adjacent in the true causal structure (Figure 8.2) then these two variables are not unconditionally d-separated. We will therefore find that the pair is associated in our data when we test for a zero association (independence) without conditioning (zero-order conditioning). Therefore the line between A and B in Figure 8.3 remains after the zero-order step. We increase the conditioning

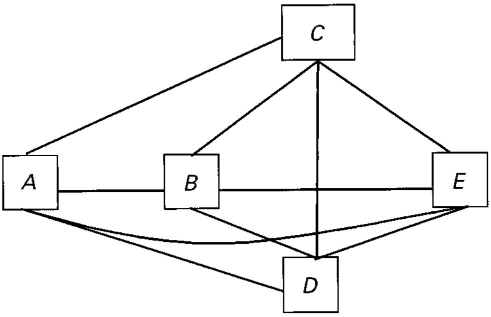


Figure 8.3. Step 1 in the construction of the undirected dependency graph. There is an undirected edge between each pair of observed variables.

order to 1 and see whether *A* and *B* become independent upon conditioning on the following first-order sets: $\{C\}$, $\{D\}$, $\{E\}$. These are the only first-order conditioning sets that we can form from five variables while excluding variables *A* and *B*. From Figure 8.2 we know that *A* and *B* are not d-separated given any of these sets. Therefore they will not be independent in our data upon first-order conditioning and the line between them in Figure 8.3 remains after this step. We continue by increasing the conditioning order to 2 and test for a zero association relative to the following sets: $\{C,D\}$, $\{C,E\}$, $\{D,E\}$. These are the only second-order conditioning sets that we can form. Given the true causal structure in Figure 8.2 we will find that the second-order association between *A* and *B* remains. We increase the conditioning set to 3 and test for a zero association relative to the following conditioning set: $\{C,D,E\}$ but still the association between *A* and *B* would remain. Since we cannot increase the conditioning order any more, we conclude that there is a line between *A* and *B* in the final undirected dependency graph.

We then go on to a new set of variables; in this case, *A* and *C*. When we apply the algorithm to the pair (*A*,*C*) we will find that *A* and *C* are still zero-order associated since they are d-connected in Figure 8.2. When we increase to order 1 and form the sets $\{B\}$, $\{D\}$ and $\{E\}$ we will find that *A* and *C* become independent upon conditioning on *B*. This is because, in Figure 8.2 (the true graph) *A* and *C* are d-separated given *B* and d-separation implies probabilistic independence. So, we remove the line between *A* and *C* in Figure 8.3, giving Figure 8.4. Since we have removed the line we don't have to go any further with this pair.

If we apply the algorithm to the pair (*A*,*D*) we would find that *A* and *D* also become independent upon conditioning on *B*, and so we would

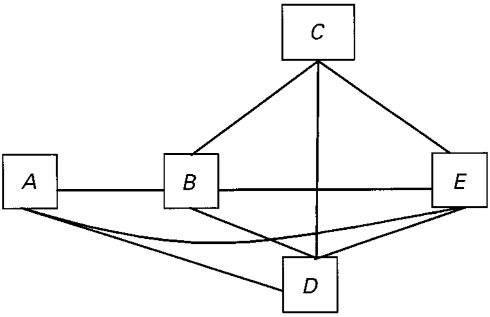


Figure 8.4. The undirected edge between *A* and *C* has been removed because we have found a subset of observed variables – {*B*} – that makes *A* and *C* independent upon conditioning.

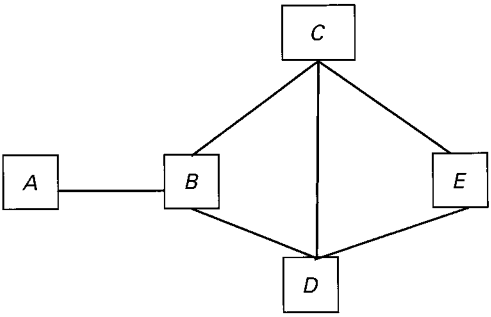


Figure 8.5. The completed undirected dependency graph. The undirected edges between *A* and *D*, between *A* and *E*, and between *B* and *E* have been removed because we have found a subset of observed variables that renders each pair independent upon conditioning.

remove the line from *A* to *D* in Figure 8.4. *A* and *E* would also become independent either upon conditioning on *B* or on the sets $\{B,C\}$, $\{B,D\}$ or $\{B,C,D\}$. This is because *A* and *E* are d-separated by any of these conditioning sets. The pair (*C*,*D*) would never become independent, since, in Figure 8.2, they are both caused by a latent variable that will never therefore appear in any of the conditioning sets. Similarly, the pair $\{C,E\}$ will always remain associated as will the pair $\{D,E\}$. The undirected dependency graph that results after applying the algorithm to every possible pair is shown in Figure 8.5; this is the correct undirected dependency graph given the causal process shown in Figure 8.2.

8.6 Interpreting the undirected dependency graph

The undirected dependency graph informs us of the pattern of direct associations in our data. It doesn't inform us of the pattern of direct *causes* in our data. For instance, there is a line between *C* and *D* in Figure 8.5 even though, peeking at the causal process that generated the data (Figure 8.2) we know that the association between *C* and *D* is due only to the effect of the latent variable (*L*). Just as the term 'direct' cause can only have meaning relative to the other variables in the causal explanation, a 'direct' association can only have meaning relative to the other variables that have been measured. However, we can infer from the undirected dependency graph that if two variables have a line between them then there is:

1. a direct causal relationship between the two and/or
2. a latent variable that is a common cause of the two and/or
3. a more complicated type of path between the two, called an inducing path; this will be explained in more detail later.

At the same time, we can exclude other types of latent variables. For instance, we know that there is no latent variable that is a common cause of *A*, *B* and *C* in Figure 8.5. If there were, then *A* and *C* would not be d-separated given any set of other observed variables and there would therefore be a line between *A* and *C* in the undirected dependency graph.

The first two explanations for a direct association in an undirected dependency graph should be understandable by now. The third possibility is less obvious but can be illustrated by an example given by Spirtes, Glymour and Scheines (1993). Consider Figure 8.6. On the left is the directed acyclic graph with a latent variable *F*. Neither variable *A* nor variable *B* has any direct causal link with variable *D*. On the right is the undirected dependency graph. Notice that there is a line between *A* and *D* and also between *B* and *D* in the undirected dependency graph even though there are neither direct causal links between the pairs nor latent variables that are causes common to both. This is because *A* and *D* would never be d-separated given any subset of variables *B* and *C* and thus would always be probabilistically associated. Similarly, *B* and *D* would never be d-separated given any subset of variables *A* and *C* and thus would always be probabilistically associated. For instance, if we look at the pair (*A*,*D*) in the true causal graph then *A* and *D* would be both unconditionally associated through the path $A \rightarrow C \rightarrow D$ and associated conditional on *B*, associated conditional on *C* through the path $A \rightarrow C \leftarrow F \rightarrow D$ and also associated conditional on both *B* and *C* for the same reason.

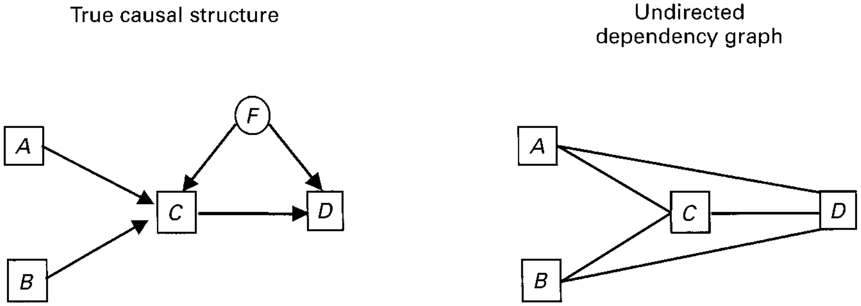


Figure 8.6. The true causal structure is shown on the left and the resulting undirected dependency graph is shown on the right. Notice that there are edges between *A* and *D* and between *B* and *D* in the undirected dependency graph even though no such directed edges exist in the directed graph. This is due to the presence of the latent variable *F* generating an inducing path between these pairs of variables.

To better understand how this third possibility can arise in general requires more definitions before the explanation can be understood.

Directed versus undirected paths

Look at the directed acyclic graph (DAG) in Figure 8.7. Imagine that this DAG is a road map consisting only of one-way streets whose direction is shown by the arrows. Normally, to go from one variable to another we have to respect the traffic rules and follow the arrows. If we can go from one variable to the other by following these rules then we will call our route a *directed path*. For instance, we can go from *A* to *D* by following the directed path $A \rightarrow B \rightarrow C \rightarrow D$. The following is *not* a directed path: $A \rightarrow B \leftarrow F \rightarrow D$ because we have gone the wrong way on a one-way road ($B \leftarrow F$) when going from *B* to *F*. However, if we ignore the rules of the road and drive in whatever direction we want, irrespective of the direction of the arrows, then we can go from *A* to *F* along the path $A \rightarrow B \leftarrow F \rightarrow D$. Such a path, in which the direction of the arrows is ignored, is called an *undirected path*. We haven't erased the arrows, we have simply decided to ignore them[9]. Of course, a directed path must also be an undirected path but an undirected path might not also be a directed path. For instance, the following are undirected paths in Figure 8.7 but they are not directed paths: $A \rightarrow B \leftarrow F \rightarrow D$, $A \rightarrow B \rightarrow C \leftarrow F \rightarrow D$.

[9] Pretend you are a diplomat who is working at the United Nations HQ in New York City. You can't change the laws about one-way streets but you can safely ignore them because of diplomatic immunity.

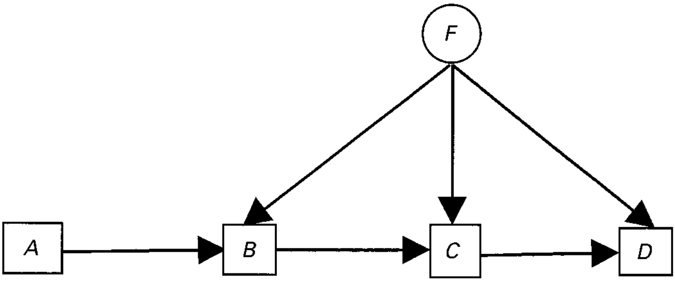


Figure 8.7. A directed graph with one latent variable (*F*).

Inducing paths[10]

List all of the variables in the DAG and call it the set ***V***. In Figure 8.7 the set ***V*** is $\{A,B,C,F,D\}$. We can call this complete DAG the graph G. Now choose some subset of variables in the DAG and call it **O**. For instance, you might choose the set $\mathbf{O}=\{A,B,C,D\}$ thus leaving out the variable *F*. By doing this you will have a new graph (call it G′) in which the variable *F* is latent; that is, the variable *F* still has the same causal relationships to the other (**O**) variables as before, but variable *F* doesn't appear in G′. Because G′ doesn't show the variable *F*, it is not a complete description of the full causal process. Now, choose two variables in your chosen set (***O***) of variables and find an undirected path between them in the complete graph (G). For instance, if we choose *A* and *D* in Figure 8.7 then we can find the undirected paths $A{\rightarrow}B{\leftarrow}F{\rightarrow}D$, $A{\rightarrow}B{\rightarrow}C{\leftarrow}F{\rightarrow}D$, $A{\rightarrow}B{\leftarrow}F{\rightarrow}C{\rightarrow}D$ and $A{\rightarrow}B{\rightarrow}C{\rightarrow}D$. Some of these undirected paths might be a special type of path called an *inducing path relative to* **O**. To determine whether a given undirected path is an inducing path relative to **O**, look at those variables in the undirected path in G that are also in your chosen set **O**. If (i) every variable in ***O*** along the undirected path except for the endpoints (here, *A* and *D*) is a collider along the path, and if (ii) every collider along this undirected path is an ancestor of either of the endpoints, then the path is an inducing path between the endpoints relative to **O**. Such an inducing path has the property that the endpoints will never be d-separated given any subset of other variables from the set **O**.

Let's look at the first undirected path between *A* and *D* ($A{\rightarrow}B{\leftarrow}\mathrm{F}{\rightarrow}D$) and choose $\mathbf{O}=\{A,B,C,D\}$. The only other variable along this undirected path, except for *A* and *D* (the endpoints), that is in ***O*** is *B*, since F has been left out (i.e. is latent). Since *B* is a collider along this path and is

[10] The properties of inducing paths were described by Verma and Pearl (1991), but the name for such a path was introduced by Spirtes, Glymour and Scheines (1993).

an ancestor (because of the path *B*→*C*→*D*) of *D* (one of the endpoints) the path *A*→*B*←F→*D* is an inducing path relative to **O** = {*A*,*B*,*C*,*D*}. To see that this inducing path results in *A* and *D* never being d-separated given any subset of variables from **O**, we have only to look at each possible conditioning set. The empty set (i.e. unconditional conditioning) allows d-connection though the path *A*→*B*→*C*→*D*. The set {*B*} allows d-connection through the undirected path *A*→*B*←F→*D*. The set {*C*} allows d-connection through the undirected path *A*→*B*→*C*←*F*→*D*. The set {*B*,*C*} allows d-connection through either of these last two undirected paths.

None of the other undirected paths between *A* and *D* are inducing paths relative to **O** = {*A*,*B*,*C*,*D*}. For instance, the undirected path *A*→*B*→*C*←*F*→*D* has the variable *B* that is in **O** but is not a collider along the path. Therefore, conditioning on *B* will d-separate *A* and *D*. Similar reasons exclude the paths *A*→*B*←*F*→*C*→*D* and *A*→*B*→*C*→*D*.

Notice that the variables at the ends of such an inducing path will never be d-separated given the variables in ***O*** because one will always be conditioning on a collider and thus opening a path through some variable not in **O**. Therefore the undirected dependency graph involving the variables in ***O*** will always have a line between two variables if there is an inducing path between them. Noting that **O** will usually consist of the set of 'observed' variables, you might start to see the usefulness of the notion of an inducing path. If you see a line between two variables in the undirected dependency graph then you will know that there is an inducing path between them.

One practical problem with the algorithm that I have presented for obtaining the undirected dependency graph is that, as the number of observed variables increases, the number of sets of conditioning variables increases geometrically. When faced with large numbers (say 50) observed variables, even fast personal computers might take a long time to construct the undirected graph if the topology of the true causal graph is uncooperative. A slightly modified version of the algorithm[11] is presented by Spirtes, Glymour and Scheines (1993); it is more efficient when one is dealing with many observed variables. The two algorithms are equivalent given population measures of association, but the more efficient algorithm can make more mistakes in small data sets.

We sometimes have independent information about some of the causal relationships governing our data. In such cases it is straightforward to modify the algorithm for the undirected dependency graph to incorporate

[11] This modified algorithm is incorporated in their PC algorithm. The algorithm that I have described forms part of their SGS algorithm.

such information. If we know that the association between two observed variables is due only to the fact that another measured variable, or set of measured variables, is a common cause of both then we simply remove that edge before applying the algorithm. Similarly, if we know that two observed variables either have a direct causal relationship, or share at least one common latent cause, then simply forbid the algorithm from considering that pair. Note that it is not enough to know (say from a randomised experiment) that one measured variable is a cause of another; we must know that it is (or is not) a *direct* cause. A randomised experiment will not be able to tell us this when some of the observed variables are attributes of the experimental units, as explained in Chapter 1.

8.7 Orienting edges in the undirected dependency graph using unshielded colliders assuming an acyclic causal structure

In Chapter 2 I discussed how d-separation predicts some counterintuitive results concerning statistical conditioning. Consider a simple causal graph of the form $X \rightarrow Z \leftarrow Y$. X and Y are causally independent and, since they are unconditionally d-separated, they are also probabilistically independent. However, if we condition on Z (the common causal descendant of both X and Y) then X and Y become conditionally dependent. This is because X and Y are not d-separated conditional on Z. In general[12], if we have two variables (X and Y) and condition on some set of variables $\mathbf{Q}$ that contains at least one common causal descendant of both X and Y, then X and Y will not be d-separated. Because of this X and Y will not be probabilistically independent upon conditioning on $\mathbf{Q}$ even if X and Y are causally independent.

This fact allows us to determine the causal direction of some lines in the undirected dependency graph. In Chapter 2 I defined an *unshielded collider* as a causal relationship between three variables (X, Y and Z) such that both X and Y are direct causes of Z ($X \rightarrow Z \leftarrow Y$) but there is no direct causal relationship between X and Y (i.e. there is no arrow going from one to the other[13]). Let's now define an *unshielded 'pattern'* in an undirected dependency graph as one in which we have three variables (X, Y and Z) such that there is a line between X and Y, a line between Y and Z (X—Z—Y), but no line between X and Y. Since there is no line between X and Y we know that X

[12] This is true for acyclic causal structures but not for cyclic causal structures. This is discussed in more detail later.

[13] If we were dealing with a path diagram rather than a directed acyclic graph then there must not be any edge at all, either an arrow or any double-headed arrows, between X and Y.

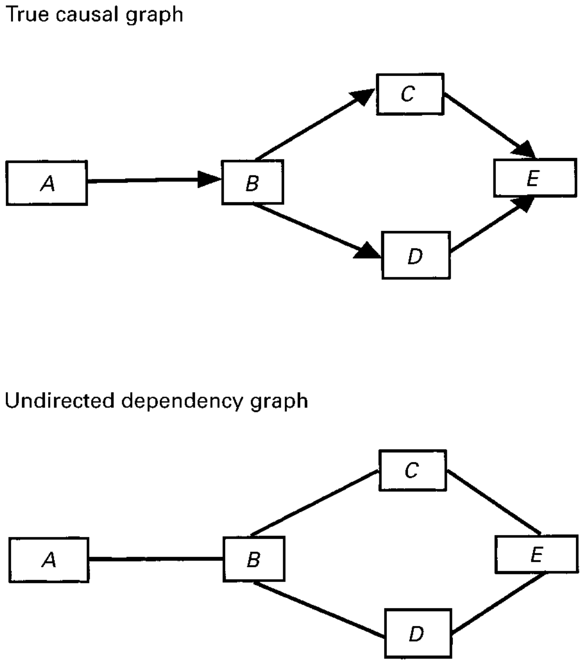


Figure 8.8. The true (unknown) causal graph and the resulting undirected dependency graph.

and Y are d-separated given some subset of other variables in the undirected dependency graph. Given such an unshielded pattern we can decide whether there are arrowheads pointing into Z from both directions or not in the causal graph that generated the data. If there were arrowheads pointing into Z from both directions in the actual causal process generating the data, then X and Y would never be probabilistically independent conditional on any set of other observed variables that includes Z.

To illustrate this method of orienting our undirected paths in the undirected dependency graph, imagine that the unknown causal process generating our observed data is as shown in Figure 8.8. Even though the causal process is hidden from us, we will obtain the undirected dependency graph shown in Figure 8.8 once we apply the algorithm to our data. Now, since we don't know what the actual causal process looks like, we don't know whether there are latent variables generating some of the direct associations.

Before going on, let's introduce some more conventions for modifying our undirected dependency graph. A graph in which only some of the edges are oriented is called a *partially directed* graph or a *partially oriented* graph. Since we don't yet know whether or not there are arrowheads at the ends of any of the lines in our undirected dependency graph (i.e. we don't yet know the directions of the causal relationships shown in the causal graph at the top of Figure 8.8), let's admit this fact by adding an open circle (Xo—oY) at the end of each line (Figure 8.9). By doing this we are no longer dealing with an undirected graph; rather, we are dealing with a partially oriented graph whose directions are not yet known. An open circle simply means that we don't know whether or not there should be an arrowhead. Therefore, given Xo—oY the oriented edge in the true causal graph might be '$X \rightarrow Y$', '$X \leftarrow Y$' or '$X \leftrightarrow Y$'. The final oriented edge ($X \leftrightarrow Y$) doesn't mean a feedback relationship between X and Y (remember our assumptions). Rather, it means that there is an unmeasured (latent) common cause generating the direct association between X and Y. It doesn't necessarily mean that there is a common latent cause of X and Y either, as Figure 8.6 makes clear.

The partially oriented graph in Figure 8.9 has six unshielded patterns, as given in Table 8.1. To orient some of the edges by detecting an unshielded collider, apply the following algorithm to each unshielded pattern (Xo—oZo—oY).

8.8 Orientation algorithm[14] using unshielded colliders

Let the conditioning number (i) be 1.

1. Form all possible conditioning sets of i observed variables consisting of the variable in the middle of the unshielded pattern (Z) plus any observed variables other than the variables at the ends of the unshielded pattern (X and Y). Call each such conditioning set **Q**.

2.1 If the partial association between X and Y, conditioned on any set **Q** of other variables, is zero then stop and conclude that the three variables forming the unshielded pattern do not form an unshielded collider in the true causal graph (i.e. not Xo$\rightarrow Z \leftarrow$oY). We can call such a pattern a *definite non-collider.*

2.2 If the partial association between X and Y, conditioned on every set **Q** of other variables, is not zero, then increase the conditioning number (i) by one and go to step 1.

[14] This algorithm is used in Pearl's IC (Inductive Causation) algorithm. The related algorithm in Spirtes, Glymour and Scheines (1993) uses a set called **Sepset**(X,Y) that reduces the computational burden. The output is identical in acyclic causal structures, but can be different in cyclic causal structures.

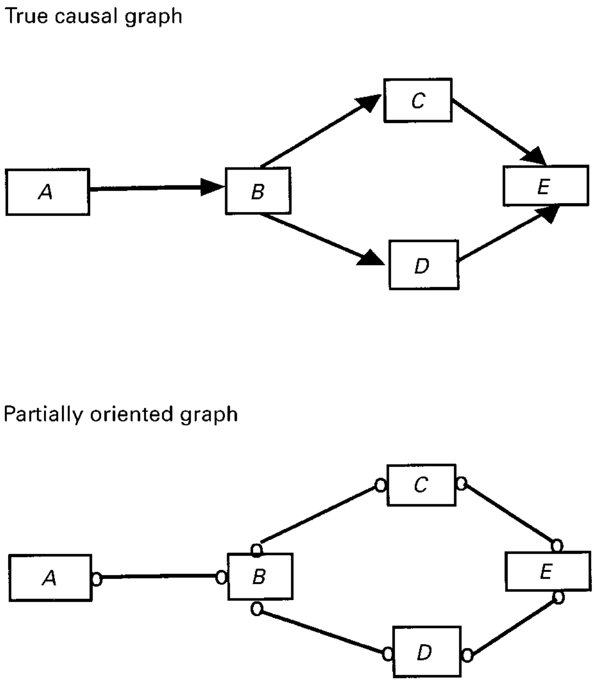


Figure 8.9. The true causal graph and a corresponding partially oriented graph with no orientations of the edges specified.

After cycling through all possible orders of *i*, if we have not declared the unshielded pattern to be a definite non-collider, then it is a collider. Orient the pattern as: *X*o→*Z*←o*Y*.

Since, in this example, we can peek at the true causal graph (top of Figure 8.9), we can use d-separation to predict what would happen if we applied the above algorithm to each of the six unshielded patterns that we found in our partially oriented graph. For instance, when we test the unshielded pattern *A*o—o*B*o—o*C* we would begin the algorithm by testing for a zero association (probabilistic independence) between *A* and *C* given *B*. The order of the conditioning set is initially 1 and we already have one variable (*B*). Since *A* and *C* are d-separated by *B* we will find *A* and *C* to be probabilistically independent given *B* and therefore stop right away, concluding that the causal effects do not collide at *B*. This unshielded pattern is a definite non-collider. The full results, and their explanations, are given in Table 8.2. You will notice one more new notation in Table 8.2. If we have concluded that an unshielded pattern is a definite non-collider (i.e. that there definitely are not arrowheads pointing into the middle variable), then underline the middle variable. Thus, the notation $X\text{o—}\underline{\text{o}Y\text{o}}\text{—o}Z$ means

Table.8.2. *Applying the orientation algorithm using unshielded colliders to the partially oriented graph in Figure 8.9*

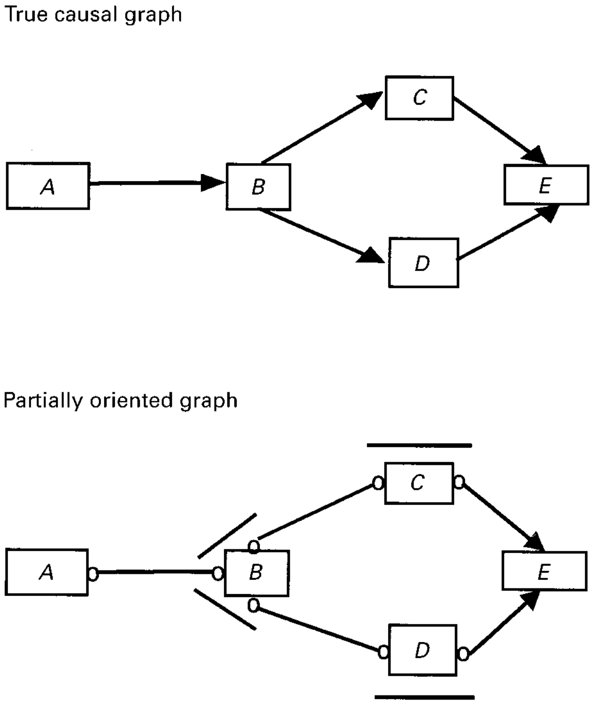
Unshielded pattern	Partially oriented pattern	Explanation
Ao—oBo—oC	Ao—$\underline{\text{o}B\text{o}}$—o$C$	B must be in **Q**. A and C are always d-separated given B and any other observed variable
Ao—oBo—oD	Ao—$\underline{\text{o}B\text{o}}$—o$D$	B must be in **Q**. A and D are d-separated given B and any other observed variable
Co—oBo—oD	Co—$\underline{\text{o}B\text{o}}$—o$D$	B must be in **Q**. C and D are d-separated given B and any other observed variable
Bo—oCo—oE	Bo—$\underline{\text{o}C\text{o}}$—o$E$	C must be in **Q**. B and E are d-separated given $\{C,D\}$ and any other observed variable
Bo—oDo—oE	Bo—$\underline{\text{o}D\text{o}}$—o$E$	D must be in **Q**. B and E are d-separated given $\{D,C\}$ and any other observed variable
Co—Eo—oD	Co→E←oD	E must be in **Q**. C and D are never d-separated given E plus any other observed variable

that we still don't know what the actual orientation is, but it is definitely not Xo→Z←oY.

The final partially oriented graph that results is shown in Figure 8.10. In fact, since the partially oriented edges indicate inducing paths, Spirtes, Glymour and Scheines (1993) called these *partially oriented inducing path graphs* or POIPGs.

At this point, other information about some of these partially oriented relationships might help us. If, for instance, we knew from previous work that the direct association between A and B was due at least in part to a common latent cause, then we could orient this as: A↔B. This would immediately restrict the orientations of the other lines, since we know that the three unshielded patterns of which B is the middle variable are all definite non-colliders. Therefore we can exclude A↔B←oC and A↔B←oD.

Because a randomised experiment, when it can be done, can give us information about causal direction, the combination of prior information from randomised experiments and these search algorithms can often be very useful. For instance, imagine that the five variables in Figure 8.10 represent five attributes of a plant and that we can't perform randomised


True causal graph

Partially oriented graph

Figure 8.10. The true, but unknown causal graph is shown at the top. The final partially oriented graph, with orientation using unshielded colliders, is shown below.

experiments to untangle the causal relationships between them. However, we can introduce a new variable that is a property of the external environment, for instance light intensity. It is possible to randomly allocate plants to the different treatment groups representing light intensity and so we can tell, for each of the five plant attributes, whether changes in light intensity cause changes in the attribute. For the reasons given in Chapter 1 we can't say that light intensity is a *direct* cause but we can say that, if the values of the attribute differ between treatments, then the causal signal (direct or indirect) goes from light intensity to the attribute and not the other way around. Now if, in an observational study, we measure the five attributes plus light intensity, and find that there is an edge in the partially oriented graph between light intensity and some of the plant attributes, then we can use this information to orient such edges. Once some edges are oriented this will usually help to orient others.

If we are willing to assume that there are no latent variables responsible for some of the lines in an undirected graph (i.e. causal sufficiency)

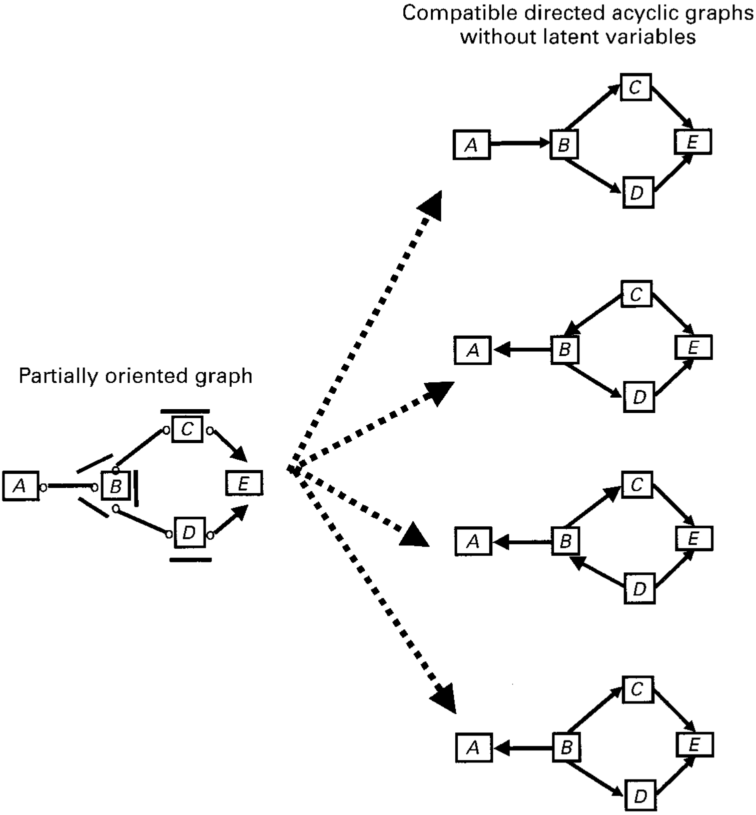


Figure 8.11. On the left is the partially oriented graph of Figure 8.10. On the right are the four possible completely directed acyclic graphs without latent variables that are consistent with the partially oriented graph.

then we can further restrict the number of possible graphs. For instance, if we take the partially oriented graph in Figure 8.10 and assume causal sufficiency then there are only four different directed acyclic graphs that are compatible with the partially oriented graph (Figure 8.11). There were a huge number of potential causal graphs involving five variables in our initial hypothesis space and we have reduced this number to four.

8.9 Orienting edges in the undirected dependency graph using definite discriminating paths

In order to orient edges using unshielded colliders we require that the pattern be unshielded. When we have a shielded pattern (three variables with lines between each of the three pairs, forming a 'triangle') we can sometimes still orient the pattern if it is embedded within a special type of partially oriented path called a *definite discriminating path*.

Let's start with some undirected path (call it U) between two variables (X and Y) in a partly oriented graph that contains some other variable B. Even though the graph is only partially oriented (thus we don't know all of the asymmetrical relationships between the variables) there is a special type of undirected path that contains important information about the variable B. Before giving the formal definition, I have to introduce yet another symbol. If we look at a single variable and the edge coming into it then we can have three different symbols at the end of the edge. For instance, we can have $X\leftarrow$, Xo— or X—. The three different symbols are the 'arrowhead', the 'o' and the 'empty mark'. Now, if I write '$X\star$—' then the star is simply a placeholder that can refer to any of the three different symbols. So, if I say 'replace $X\star$— oY by $X\star\!\!\rightarrow Y$' then I mean 'keep whatever symbol was next to the X but change the "o" symbol next to the Y to a ">" symbol'.

Here is the definition of a definite discriminating path. An undirected path U is a *definite discriminating path* for variable B if and only if:

1. U is an undirected path between variables X and Y containing B.
2. X and Y are not adjacent.
3. B is different from both X and Y (i.e. B cannot be an endpoint of the path).
4. Every variable on U except for B and the endpoints (X, Y) is either a collider or a definite non-collider on U.
5. If two other variables on U (V and V') are adjacent on U, and V' is between V and B on U, then the orientation must be: $V\star\!\!\rightarrow V'$ on U.
6. If V is between X and B on U and V is a collider on U then the orientation must be either $V\rightarrow$Y or else $V\leftarrow\!\!\star X$.
7. If V is between Y and B on U and V is a collider on U then the orientation must be either $V\rightarrow X$ or else $V\leftarrow\!\!\star Y$.

To see the usefulness of such definite discriminating paths, consider Figure 8.12. Each of the four unshielded patterns in the undirected dependency graph (X—V—V', V—V'—A, V'—A—B and V'—A—Y) derived from this partially oriented graph allowed us to apply the algorithm for unshielded colliders; in this case all were determined to be definite non-colliders. Unfortunately, the shielded pattern involving A, B and Y can't be oriented this way. However, the undirected path Xo—$\underline{\text{o}V\text{o}}$—$\underline{\text{o}V'}$o—$\underline{\text{o}A\text{o}}$—o$B$o—o$Y$ is also a definite discriminating path for the variable B. What happens if, in the underlying causal graph, X and Y are d-separated

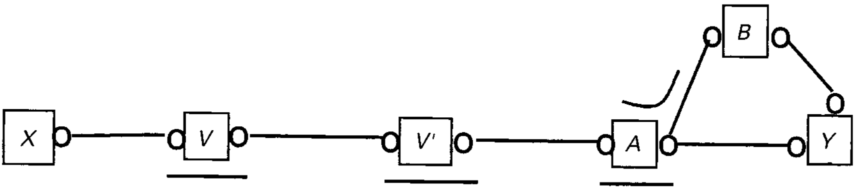


Figure 8.12. A partially oriented graph involving six observed variables. The undirected path Xo—$\underline{\text{o}V\text{o}}$—$\underline{\text{o}V'\text{o}}$—$\underline{\text{o}A\text{o}}$—o$B$o—o$Y$ is a definite discriminating path for the variable B.

given A and B, but not given A alone? The only way that this could occur is if B were a definite non-collider along the undirected path (Ao—$\underline{\text{o}B\text{o}}$—o$Y$), since, if the orientation was really Ao→B←oY then conditioning on A and B would not d-separate X and Y. So we can definitely state that the partial orientation is Ao—$\underline{\text{o}B\text{o}}$—o$Y$. Yet this is not all. Since we have assumed that the unknown causal graph is acyclic, there are only two different partially oriented acyclic causal graphs that accord with this information (Figure 8.13).

Now we can put all of the pieces together and state the Causal Inference algorithm of Spirtes, Glymour and Scheines (1993).

8.10 The Causal Inference algorithm[15]

1. Apply the algorithm to obtain the undirected dependency graph.
2. Orient each edge in the undirected dependency graph as o—o.
3. Apply the orientation algorithm using unshielded colliders. For each unshielded pattern (A—B—C) orient unshielded colliders as Ao→B←oC and orient each definite non-collider as Ao—$\underline{\text{o}B\text{o}}$—o$C$.
4. If there is a directed path from A to B, and an edge A★—★B, orient A★—★B as A★→B.
5. If B is a collider along a path A★→B←★C, B is also adjacent to another variable D (i.e. B★—★D), and A and C are conditionally independent[16] given D, then orient B★—★D as B←★D.
6. If there is an undirected path U that is a definite discriminating path between variables A and B for variable M, and variables P and R are adjacent to M along U, and P★—oMo—★R forms a triangle, then

[15] My description of the Causal Inference algorithm differs from the original formulation only in replacing **Sepset** sets with the actual d-separation claim.

[16] There was an error in Spirtes, Glymour and Scheines (1993) at this point, which was corrected in a subsequent Erratum.

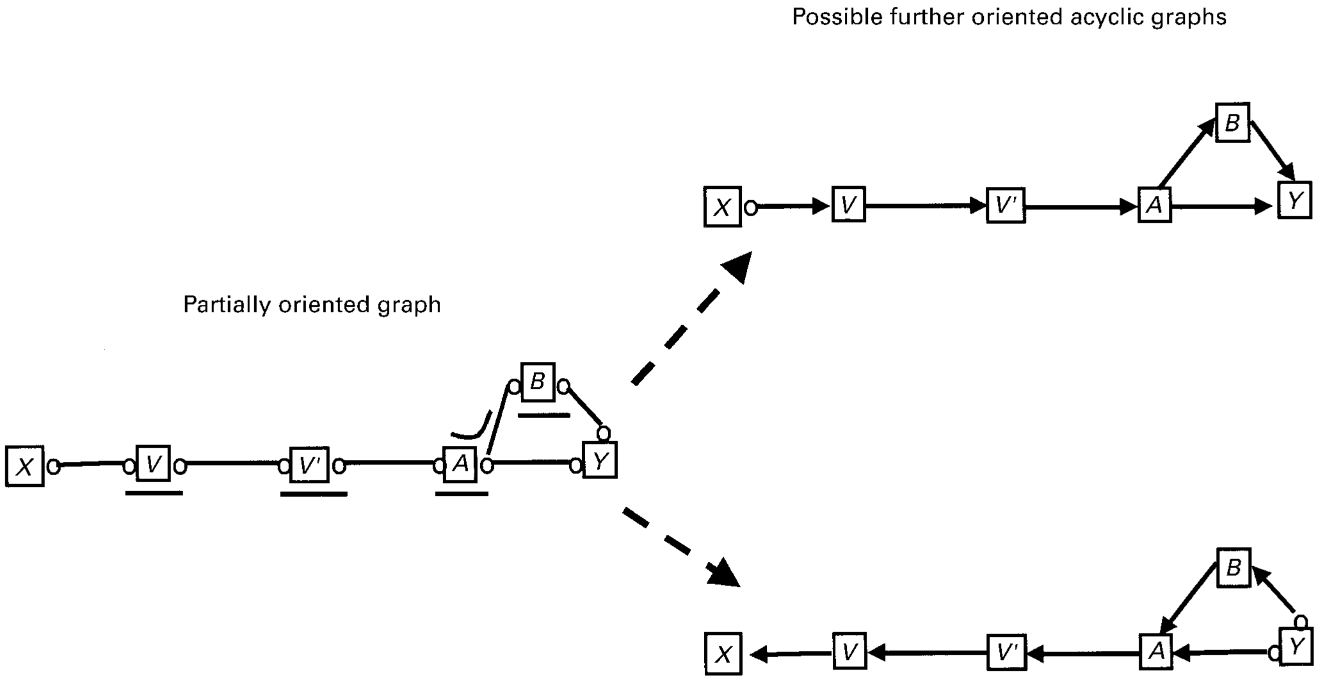


Figure 8.13. The partially oriented graph on the left implies only two alternative partially oriented acyclic graphs on the right.

i. If A and B are conditionally independent given M plus any other variable except A and B, then $P{\star}\text{—o}M\text{o—}{\star}R$ along U is oriented as a non-collider: $P{\star}\text{—}\underline{\text{o}M\text{o}}\text{—}{\star}R$.

ii. If A and B are never conditionally independent given M plus any other variable except A and B, then $P{\star}\text{—o}M\text{o—}{\star}R$ along U is oriented as a collider: $P{\star}{\rightarrow}M{\leftarrow}{\star}R$.

iii. If the triangle is already oriented as $P{\star}{\rightarrow}\underline{M}{\star}\text{—}{\star}R$ then orient it as $P{\star}{\rightarrow}M{\rightarrow}R$.

Repeat steps 4–6 until no further changes can be made.

The result is a partially oriented inducing path graph. You should be able to understand steps 1 to 3 by now. Step 4 is justified by the assumption that there are no cyclic relationships in the causal structure. If there is a directed path from variable A to variable B and we were to also orient the direct edge as $B{\rightarrow}A$ then this would create a cyclic path. Step 5 is simply a generalisation of the reason for orienting an unshielded collider. Remember that the two variables (X, Y) in this group will never be d-separated if conditioned on any of their causal descendants. Since we have already established that the orientation is $A{\star}{\rightarrow}B{\leftarrow}{\star}C$ and there is another edge oriented as $B\text{o—}{\star}D$, and that both A and B are (conditionally) causally independent of D (i.e. they are not d-connected), then d-separation predicts that A and D would become probabilistically dependent when conditioned on D if the orientation was $B\text{—}{\star}D$ and remain probabilistically independent if the orientation was $B{\leftarrow}{\star}D$. Steps 6i and 6ii derive from the notion of a definite discriminating path, as described before. In step 6iii we have already established that M is a non-collider along $P{\star}{\rightarrow}\underline{M}{\star}\text{—}{\star}R$. Therefore there cannot be arrowheads pointing into M from both directions and we can orient the triplet as $P{\star}{\rightarrow}\underline{M}\text{—}{\star}R$. There is now only one orientation possible, namely $P{\star}{\rightarrow}\ M{\rightarrow}R$.

8.11 Equivalent models

The inferential testing of structural equations models, described in Chapters 3 to 7, consisted of deriving the observational predictions of the hypothesised causal process (the correlational shadows) and then comparing the observed and predicted patterns of correlation or covariation. I have emphasised that failing to reject such an hypothesised model provides support for it, but does not allow us to accept it without other (non-statistical) evidence. One reason might be that the sample size was too small to permit us to detect a real (but small) deviation between the observed and predicted patterns. However, the search algorithms in this chapter should alert us to another reason: different causal processes can cast the same observational shadows.

This leads to the topic (rarely discussed in the SEM literature) of observationally equivalent models; that is, different causal models that can't be distinguished on the basis of observational data. Such *equivalent models* will produce exactly the same chi-squared values, and exactly the same probability levels, when tested against the same data set. This is true no matter how big your data set is. In fact, it will be true even if you have the population values rather than sample values. When we test a structural equations model we are really testing the entire set of observationally equivalent models against all non-equivalent models. In one sense this might be disappointing; we can't distinguish between some competing causal explanations. In another sense this is useful; when we reject a particular model we are also simultaneously rejecting all of the observationally equivalent models as well.

The search algorithms in this chapter can allow us to find all of the causal models that are observationally equivalent[17] to our hypothesised one. Given your path diagram, here are the steps:

1. Change all arrows (even double-headed ones[18]) to lines.
2. Draw the **o** symbol at either end of each line.
3. Redraw each unshielded collider that was in the original path diagram; that is, if there was an unshielded collider in the path diagram ($X \rightarrow Z \leftarrow Y$) then replace $X\text{o—o}Z\text{o—o}Y$ with $X\text{o}\rightarrow Z \leftarrow \text{o}Y$.
4. For each non-collider triplet that was in the original path diagram, add an underline; that is, if there was either $X \rightarrow Z \rightarrow Y$, $X \leftarrow Z \leftarrow Y$ or $Z \leftarrow Z \rightarrow Y$ in the path diagram then replace $X\text{o—o}Z\text{o—o}Y$ with $X\text{o—}\underline{\text{o}Z\text{o}}\text{—o}Y$.

Figure 8.14 summarises these steps.

At this point you can permute the different possible orientations so long as you never introduce an unshielded collider that was not in your original path diagram, and never remove an unshielded collider that was in your original path diagram.

[17] The algorithm for observational equivalence in acyclic models was first published by Verma and Pearl (1991).

[18] A model having two variables sharing correlated errors (i.e. a double-headed arrow between them) is equivalent in its d-separation consequences to a model having a latent variable that is a common cause of both (Spirtes *et al.* 1998).

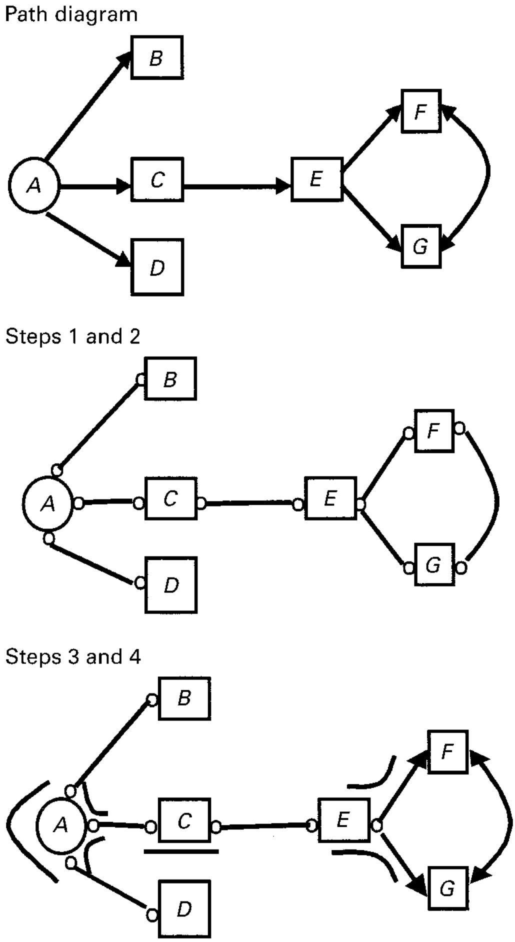
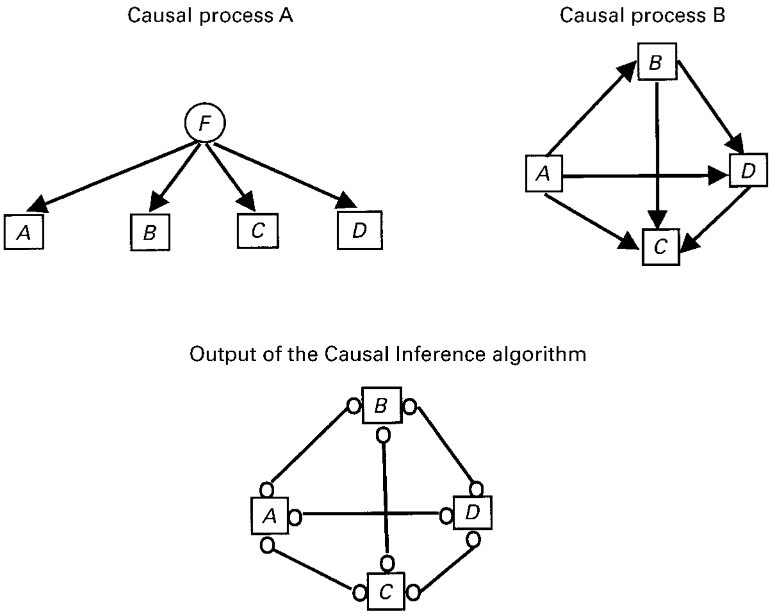


Figure 8.14. The top graph shows a path diagram. The following two graphs show the steps in obtaining all models that are equivalent to the path diagram.

8.12 Detecting latent variables

One practical problem with the Causal Inference algorithm is that it can be quite uninformative when many observed variables are all caused by a small number of latent variables. In such cases the application of the Causal Inference algorithm will not be very informative. Consider the simple measurement model (graph A) shown in Figure 8.15 and the resulting output from the Causal Inference algorithm.


Causal process A

Causal process B

Output of the Causal Inference algorithm

Figure 8.15. On the top are directed acyclic graphs of two different causal processes that both imply the same partially oriented graph on the bottom.

The output of the Causal Inference algorithm tells us that each of the observed variables (*A* to *D*) is probabilistically associated with each of the others. Since there are no unshielded patterns among the observed variables in either of the two output graphs, we can't orient any of the edges. Whenever you see a set of observed variables that form such a pattern (I will call this a *saturated* pattern) you should suspect latent variables. However, it is possible for such saturated patterns to arise even without latent variables, as causal process B shows. Is there any way of differentiating between the two? Yes. For this, we need to look again at vanishing tetrad equations, which we studied briefly in Chapter 5.

You will recall that Spearman (1904) derived a set of equations, called vanishing tetrads, which must be true given the type of structure shown in the causal graph A in Figure 8.15. He argued that if such vanishing tetrad equations held then this was evidence for the presence of a common latent cause of the observed variables. As will be explained below, this claim is not true, but a modification of it can indeed be used to detect such a common latent cause if the relationships between the latent variable and the observed variables are linear.

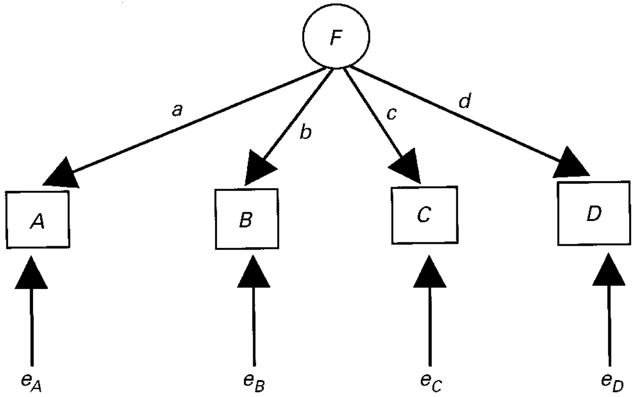


Figure 8.16. A directed graph involving four observed variables and one latent variable (*F*).

A vanishing tetrad equation is a function of four correlation (or covariance) coefficients. Because of the causal structure of models like those in Figure 8.16, and because of the rules of path analysis, such a vanishing tetrad equation must be zero in the population regardless of the (non-zero) values of the path coefficients. For instance, the population correlation coefficient (ρ_{AB}) between the observed variables *A* and *B* is $a\star b$. The population correlation coefficient (ρ_{CD}) between the observed variables *C* and *D* is $c\star d$. Therefore, $\rho_{AB}\rho_{CD}=a\star b\star c\star d$. However, we also know that $\rho_{AC}=a\star c$ and that $\rho_{BD}=b\star d$. Therefore $\rho_{AC}\rho_{BD}=a\star c\star b\star d$. It follows that $\rho_{AB}\rho_{CD}-\rho_{AC}\rho_{BD}=0$, since $a\star b\star c\star d-a\star c\star b\star d=0$. The tetrad equation $(\rho_{AB}\rho_{CD}-\rho_{AC}\rho_{BD})$ becomes zero, or *vanishes*, because of the way the observed variables relate to the latent variable. The causal process shown in Figure 8.16 implies three different vanishing tetrad equations (of which only two are independent). In fact, every set of four variables can have three possible tetrad equations regardless of the true causal process, although they don't have to be zero.

$$\rho_{AB}\times\rho_{CD}-\rho_{AD}\times\rho_{BC}=0$$

$$\rho_{AC}\times\rho_{BD}-\rho_{AD}\times\rho_{BC}=0$$

$$\rho_{AC}\times\rho_{BD}-\rho_{AB}\times\rho_{CD}=0$$

Unfortunately, a causal structure like the one in Figure 8.17 also implies vanishing tetrad equations. For instance, using the rules of path analysis we will find that $\rho_{AC}\times\rho_{BD}-\rho_{AD}\times\rho_{BC}=(ab)(bc)-(abc)(b)=0$. Clearly, simply showing that a vanishing tetrad equation holds is not evidence for the presence of a common latent cause of the observed variables. Although

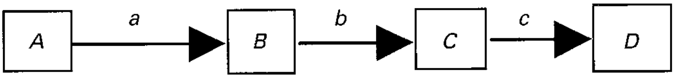


Figure 8.17. This directed acyclic graph also implies a vanishing tetrad ($\rho_{AC} \times \rho_{BD} - \rho_{AD} \times \rho_{BC} = (ab)(bc - (abc)(b) = 0)$ even though there are no latent variables.

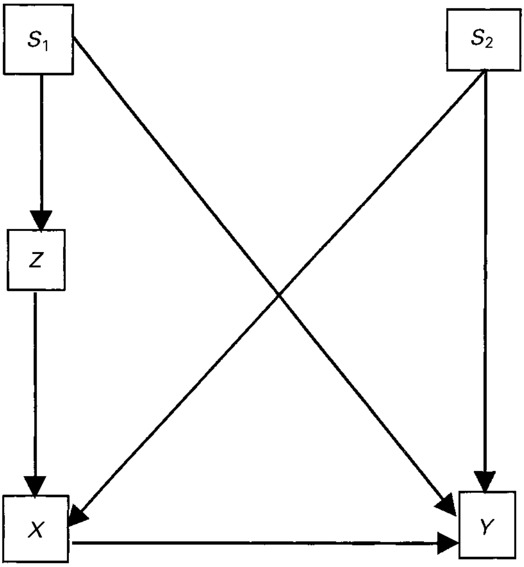


Figure 8.18. A directed acyclic graph used to illustrate the concept of a *trek*.

it may not seem immediately obvious, there is a close relationship between d-separation and vanishing tetrad equations[19].

A vanishing tetrad equation can be given a graphical interpretation. Let's define a *trek* between two variables (X,Y) as a pair of directed paths; one directed path goes from a *source* variable (S) to X and the other directed path goes from the same source variable to Y. One of the two directed paths can be of length 0 (i.e. $S=X$ or $S=Y$). For instance, in Figure 8.18 there are three treks between X and Y. One is from the source variable S_1 ($X \leftarrow Z \leftarrow S_1 \rightarrow Y$), one is from the source variable S_2 ($X \leftarrow S_2 \rightarrow Y$) and one is from the 'source' variable X in which one directed path is of length zero

[19] Theorem 6.11 of Spirtes, Glymour and Scheines (1993) states that a vanishing tetrad equation of the type $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is linearly implied by an directed acyclic graph only if either ρ_{IJ} or ρ_{KL} equals zero and if either ρ_{IL} or ρ_{JK} equals zero or there is a (possibly empty) set $\mathbf{Q}$ of variables in the directed acyclic graph such that $\rho_{IJ.\mathbf{Q}} = \rho_{KL.\mathbf{Q}} = \rho_{IL.\mathbf{Q}} = \rho_{JK.\mathbf{Q}} = 0$.

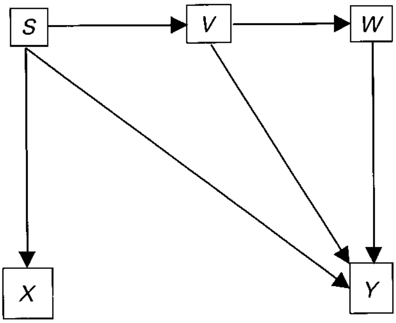


Figure 8.19. A directed acyclic graph used to illustrate the concept of a *choke point* for a set of treks.

($X \rightarrow Y$). I will write '$\boldsymbol{T}(X,Y)$' to mean a trek between X and Y, '$\boldsymbol{T}(X,Y)$' to mean the set of all treks between X and Y and I will write '$X(\boldsymbol{T}(X,Y))$' to mean the directed path in a trek between X and Y that goes into X.

In Figure 8.19 there are three different treks between X and Y: $X \leftarrow S \rightarrow Y$, $X \leftarrow S \rightarrow V \rightarrow Y$ and $X \leftarrow S \rightarrow V \rightarrow W \rightarrow Y$. Notice that all the directed paths in all these treks leading into X pass through S. When this occurs we say that S is a *choke point* for $X(\boldsymbol{T}(X,Y))$. There was no choke point for $X(\boldsymbol{T}(X,Y))$ in Figure 8.18.

To see what all this has to do with vanishing tetrads let's consider a set of four variables (I,J,K,L). If we have a set of treks $\boldsymbol{T}(I,J)$ between two variables (I,J) and a set of treks $\boldsymbol{T}(K,L)$ between two other variables (K,L) and all of the directed paths in $\boldsymbol{T}(I,J)$ that are into J (i.e. $J(\boldsymbol{T}(I,J))$) and all of the directed paths in $T(K,L)$ that are into L (i.e. $L(\boldsymbol{T}(K,L))$) intersect at the same variable Q, then Q is called a *JL choke point*. The Tetrad Representation Theorem (Spirtes, Glymour and Scheines 1993) states that if we see a vanishing tetrad in the statistical population ($\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$) then this means that there is either a JL choke point or an IK choke point.

How can vanishing tetrads help to detect the presence of latent variables? If you see a saturated pattern in your undirected dependency graph involving four variables[20], then test to see whether there are any vanishing tetrads between these variables. If vanishing tetrads exist then this is evidence for a latent variable. To see why, consider that if the choke point implied by this vanishing tetrad was an observed variable then the two variables (J,L) or (I,K) would be d-separated by this choke point and therefore could not form part of a saturated pattern[21].

[20] If there are more than four variables forming a saturated pattern, then take each unique set of four variables.

[21] In Figure 8.17 there was a vanishing tetrad but no latent variable. The treks between each

This fact provides a simple algorithm to test whether the observed correlations among a set of four observed variables is due to a common latent cause. These are the steps.

8.13 Vanishing Tetrad algorithm

Given a set ***O*** of observed variables and a set ***T*** of four observed variables from ***O*** that form a saturated pattern in the undirected dependency graph, assume that there is no reason to invoke a common latent cause for these four variables in ***T*** and then do the following:

1. Choose one of the three tetrad equations that are possible given the four chosen variables in ***T***. If you have tried all three then stop.
2. If the tetrad equation does not equal zero, go to step 1.
3. If the tetrad equation does equal zero then there is a latent variable that forms the *IK* choke point of $IK(\boldsymbol{T}(I,J),\boldsymbol{T}(K,L),\boldsymbol{T}(I,L),\boldsymbol{T}(JK))$ or the *JL* choke point of $JL(\boldsymbol{T}(I,J),\boldsymbol{T}(K,L),\boldsymbol{T}(I,L),\boldsymbol{T}(J,K))$.

To illustrate this algorithm, let's go back to the two graphs in Figure 8.15. The undirected dependency graph will contain a saturated pattern for variables *A*, *B*, *C* and *D*. Here again are the three tetrad equations:

$$\rho_{AB}\times\rho_{CD}-\rho_{AD}\times\rho_{BC}=0$$

$$\rho_{AC}\times\rho_{BD}-\rho_{AD}\times\rho_{BC}=0$$

$$\rho_{AC}\times\rho_{BD}-\rho_{AB}\times\rho_{CD}=0$$

All three tetrad equations vanish in graph A of Figure 8.15. Because the first equation vanishes we know that there is either an *AC* and/or a *BD* choke point. Because the second equation vanishes we know that there is either an *AB* and/or a *CD* choke point. Because the third equation vanishes we know that there is either an *AD* or a *BC* choke point. In fact all these choke points exist in this graph and all are the same variable (*F*). If we do the same thing to graph B of Figure 8.15 we will see that no tetrad equation vanishes[22].

Let's go on and apply the Vanishing Tetrad algorithm to a causal

of the four pairs of variables (*AC*, *BD*, *AD* and *BC*) all had directed paths of zero length (*A*→*B*→*C*, *B*→*C*→*D*, *A*→*B*→*C*→*D* and *B*→*C*). The choke point for these four treks was the variable *B*, which was an observed variable. This is part of the reason why these four variables do not form a saturated pattern.

[22] Unless the graph is unfaithful. It is always possible to choose path coefficients in such a way as to make a particular tetrad equation vanish, but the vanishing tetrad equation is not implied by the topology of the graph.

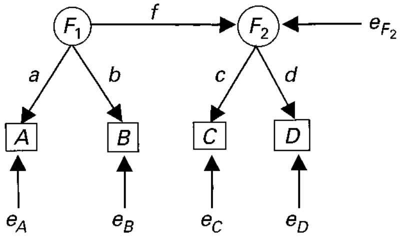


Figure 8.20. A path diagram involving four observed variables and two latent variables; *a*, *b*, *c*, *d* and *f* are path coefficients.

graph involving two latent variables (Figure 8.20). Here are the three tetrad equations:

$$\rho_{AB}\times\rho_{CD}-\rho_{AD}\times\rho_{BC}=(ab)(cd)-(afd)(bfc)=abcd(1-f^2)\neq 0$$

$$\rho_{AC}\times\rho_{BD}-\rho_{AD}\times\rho_{BC}=(afc)(bfd)-(afd)(bfc)=0$$

$$\rho_{AC}\times\rho_{BD}-\rho_{AB}\times\rho_{CD}=(afc)(bfd)-(ab)(cd)=f^2(1-abcd)\neq 0$$

Notice that the only tetrad equation that vanishes is the one with either an *AB* and/or a *CD* choke point. In fact, both choke points exist. All the directed paths leading into *A* and *B* of all treks between the four pairs of variables (*AC*, *BD*, *AD* and *BC*) pass through F_1. All the directed paths leading into *C* and *D* of all treks between these four pairs of variables also pass through F_2.

8.14 Separating the message from the noise

The ancients knew how to discover causal relationships. Things happened in the world because the gods willed them. One had only to ask and, if the gods were willing and the diviner gifted, the causes would be revealed. Unfortunately, the gods were capricious and their words couched in allegory. A good seer had to be able to separate the message from the noise, to know when a bump in a goat's intestine foretold war and when it was simply undigested grass[23]. If the methods presented in this chapter are the modern version of the diviner's art then we still need to separate the causal message from the sampling noise.

The various algorithms all require that we know whether or not sets of random variables are independent. We are constantly being asked: 'Is

[23] The ancient Greek philosophers were the first to conceive of a world governed by natural causes rather than by divine will. They then confronted the subject of this chapter. Democritus (460–370 BC) is reported to have said: 'I would rather discover one causal law than be King of Persia' (Pearl 2000).

the statistical association zero or different from zero'? So far I have assumed that we can always answer such a question unambiguously because I have assumed that we have access to the entire statistical population. If correlations are the shadows cast by causes then I have assumed that these shadows are always crisp and well defined. Given such an assumption we can extract an amazing amount of causal information from purely observational data; certainly much more than is intimated by the old mantra that 'correlation does not imply causation'.

Let's get back to reality. We almost never have access to the entire statistical population. Rather, we collect observations from random samples of the statistical population and these random samples are not perfect replicas of the entire population. If correlations are the shadows cast by causes then *sample correlations* are randomly blurred correlational shadows. We have to find a way of dealing with the imperfect information contained in these blurred correlational shadows. Inferring population values from sample values is the goal of inferential statistics and inferential statistics is the art of drawing conclusions based on imperfect information. In practice we can never unambiguously know whether the statistical association is zero or different from zero. How can we deal with this problem when applying the various discovery algorithms and what sort of errors might creep into our results? To see this, we first need to review some basic notions of hypothesis testing.

Consider the problem of determining whether the population value of a Pearson correlation coefficient (ρ_{XY}) between two random variables, X and Y, is zero based on the measured sample value (r_{XY}). There are only two possible choices: either it really is or it really isn't. Similarly, we can only give one of two answers based on our sample measure: either we think it is or we think it isn't. These define four different outcomes (Table 8.3).

Normally, the types of biological hypothesis that interest us are ones in which variables are associated, not independent. Because we want evidence beyond reasonable doubt before accepting this interesting hypothesis (see Chapter 2) we usually begin by assuming the contrary – that there is no association – and then look for strong evidence against this assumption before rejecting it and therefore accepting our biologically interesting one. In other words, we want to see a value of r_{XY} that is sufficiently large that there is very little probability that it would have come from a statistical population in which $\rho_{XY} = 0$ is true.

So what is a small enough probability that we would be willing to declare that ρ_{XY} really is different from zero? This is a somewhat subjective decision, as described in Chapter 2, but Table 8.3 shows that part of our decision will depend on how important it is for us to avoid either a Type I

Table 8.3. *Possible combinations of decisions to a null hypothesis and its alternative, giving rise to Type I and Type II errors*

		True value in the statistical population	
		$\rho_{XY} = 0$	$\rho_{XY} \neq 0$
Your answer after looking at r_{XY} and calculating its probability	$\rho_{XY} = 0$	Right choice	Type II error
	$\rho_{XY} \neq 0$	Type I error	Right choice

error (incorrectly declaring that $\rho_{XY} \neq 0$ when, in reality, $\rho_{XY} = 0$) or a Type II error (incorrectly declaring that $\rho_{XY} = 0$ when, in reality, $\rho_{XY} \neq 0$). Because the presence of a real association usually (but not always) gives us useful biological information, and because we know that our ever-present sampling variation can sometimes fool us into observing a large value of r_{XY} even when the variables are independent, we usually place more importance in reducing our Type I errors than in reducing our Type II errors. Therefore, we usually choose a small probability before we are willing to declare our value of r_{XY} as being 'significantly' different from zero. For instance, choosing a significance level of 0.05 means that we are only willing to accept a 5% chance of making a Type I error. Notice, however, that by decreasing our significance level to the low value of 0.05 we are simultaneously willing to accept a larger chance of making a Type II error. This is usually okay because we have already decided that it is more important to be quite sure that ρ_{XY} is *not* zero than to be quite sure that ρ_{XY} *is* zero.

In each of the algorithms described in this chapter you are repeatedly asked to decide whether the measure of association is zero or not. You have to make this choice based on a random sample of data and, therefore, you have to conduct statistical tests and choose how important it is to minimise either Type I or Type II errors. Here's the rub: by definition, these are discovery algorithms. You don't already have a preferred causal hypothesis that you wish to test. You can't have any *a priori* preference for either $\rho_{XY} = 0$ or $\rho_{XY} \neq 0$ and neither outcome provides more information than the other to the various algorithms[24].

This brings us to the notion of statistical power. The hypothesis that

[24] One exception might be in the orientation phase of the algorithms. It might be better to plead ignorance, and leave edges unoriented, than to make a definite choice about declaring an unshielded pattern to be a collider or a definite non-collider.

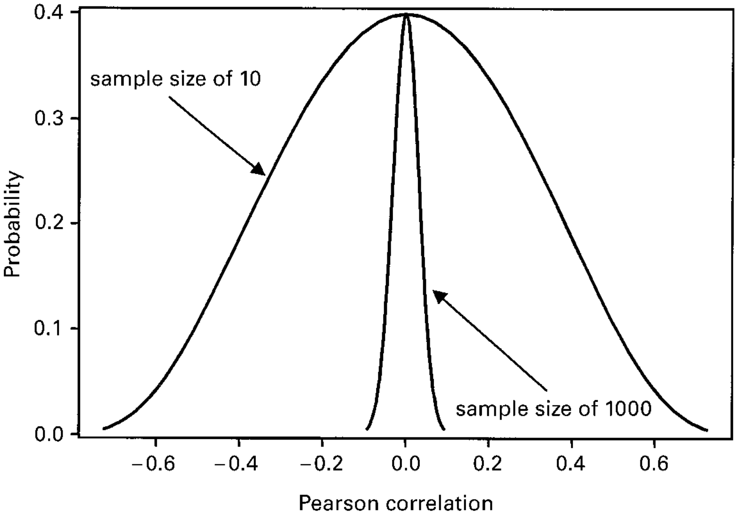


Figure 8.21. The probability of observing a Pearson correlation coefficient of various values when the population value is zero at two different sample sizes.

$\rho_{XY} \neq 0$ is not really a single hypothesis at all; rather, it is a composite hypothesis that includes $\rho_{XY} = 0.01$, $\rho_{XY} = 0.1$, $\rho_{XY} = 0.9$ and an infinite number of other individual hypotheses. Intuitively, it is obvious that it would be much more difficult to distinguish between 0.0 and 0.01 than between 0.0 and 0.9 in any sample of data. If we had a huge data set (say a thousand observations) and the population value was 0.0 then our sample correlation would almost always be extremely close to zero (Figure 8.21). Sampling variation would only very rarely result, by chance, in a value greater than even a low number such as 0.1. Therefore, if the population value was even slightly different from 0.0 then we would almost always find a very small probability value for our measured r_{XY} and would almost never conclude that $\rho_{XY} = 0$ when, in reality, $\rho_{XY} \neq 0$. Our test would be very powerful in detecting even very slight associations between *X* and *Y*. If, on the other hand, we had only a small data set (say 10 observations) and our population value was 0.0 then our sample correlation would fluctuate quite widely around zero simply due to sampling variation (Figure 8.21). Therefore, even if the population value was quite different from 0.0 (say $\rho_{XY} = 0.4$) we would often observe sample values close to zero simply due to sampling variation. Because of this, the probability of incorrectly concluding that $\rho_{XY} = 0$ would not be negligible. Our test would not be very powerful in detecting even moderate associations between *X* and *Y* and we would make Type II errors more often.

The *power* of a statistical test is the probability of rejecting the null hypothesis when, in fact, it is false. It is defined as $1-\beta$, where β is the probability of a sample statistic, taken from a statistical population in which the null hypothesis is false, falling within the acceptance region of the null hypothesis. The power is affected by sample size, by the significance level chosen for rejecting the null hypothesis and by the difference (the 'effect size') between the true value of the test statistic and the value assumed by the null hypothesis. Figure 8.22 plots the statistical power to reject the null hypothesis that $\rho=0$ when the true value varies from -0.9 to $+0.9$ at two sample sizes (30 and 300 observations) and three different significance levels ($\alpha=0.05$, 0.10 and 0.20).

Figure 8.22 clearly shows the compromise that must be made. If the null hypothesis ($\rho=0$) is true, then increasing the significance level from $\alpha=0.05$ to $\alpha=0.2$ increases our chances of incorrectly rejecting the null hypothesis (Type I error). We will be incorrectly declaring associations to exist more often when they really do. On the other hand, increasing the significance level from $\alpha=0.05$ to $\alpha=0.2$ increases our power to reject the null hypothesis when associations really do exist but are weak. As sample size increases then power increases irrespective of the chosen significance level. This is a good thing because now we can increase our power to detect real, but weak associations without increasing our significance level and therefore without increasing our chances of falsely accepting associations that don't really exist. At large sample sizes we are best to set a low significance level (say $\alpha=0.05$ or even $\alpha=0.01$), since at such large sample sizes we will keep both Type I and Type II error rates low. At small sample sizes we are best to increase our significance level if, in fact, we don't have any preference for the presence or absence of a real association. This is because, at a low significance level, only very large values of the correlation coefficient would have a reasonable chance of being detected. As the sample size increases, that power approaches 1.0 even as α approaches 0, meaning that the chances of committing both Type I and Type II errors approach zero.

You will see that significance levels of $\alpha=0.2$, 0.4 or even higher might be used with very small sample sizes. Clearly, applying these algorithms to small samples means accepting more and more errors due to sampling fluctuations. Remember that these are *exploratory* methods, not methods designed to test a preconceived hypothesis. If you set out to hike through an unfamiliar area then you would probably take a map. The search algorithms are like imperfect maps to a causal landscape. At large sample sizes these maps will give you all the detail that can be obtained even though no one map might be able to provide all the information that is wanted. At small sample sizes these maps will only give you the major hiking trails, may

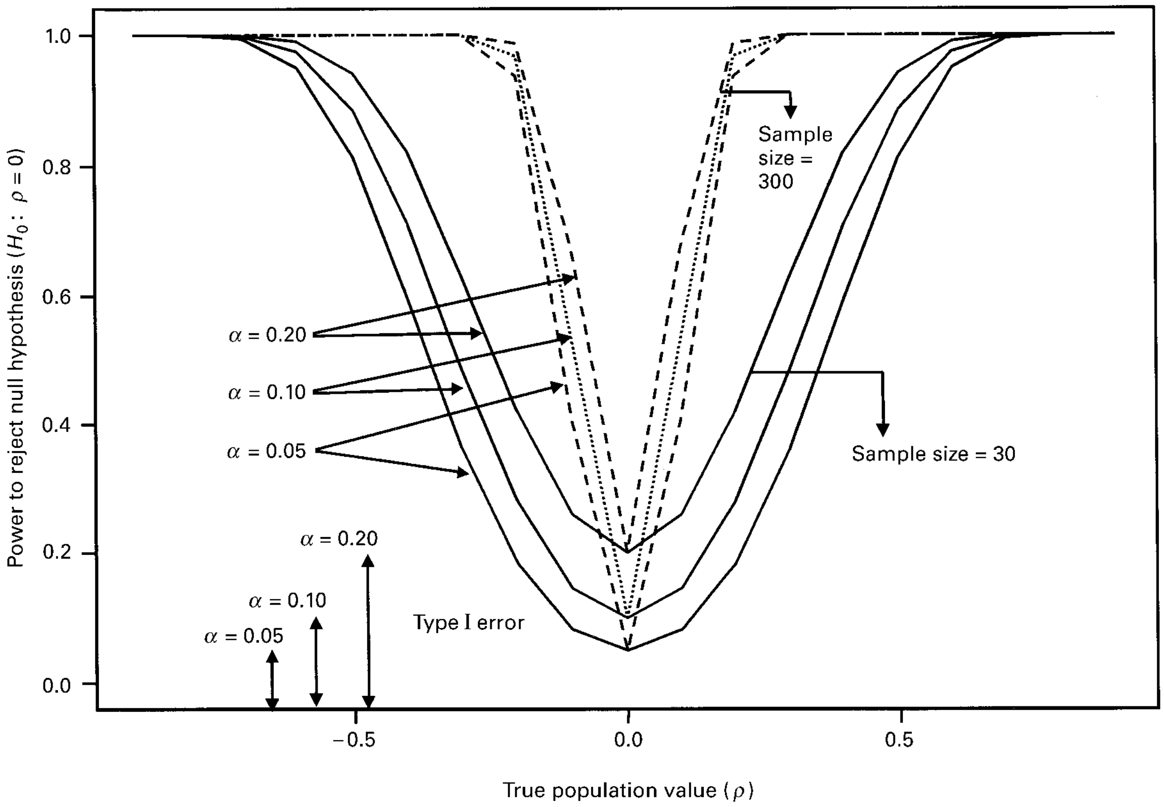


Figure 8.22. The power to reject the null hypothesis that the Pearson correlation coefficient is zero in the population, when the true value of the population Pearson correlation takes on different values. This power is affected by the sample size and the chosen significance level (α) used to reject the null hypothesis.

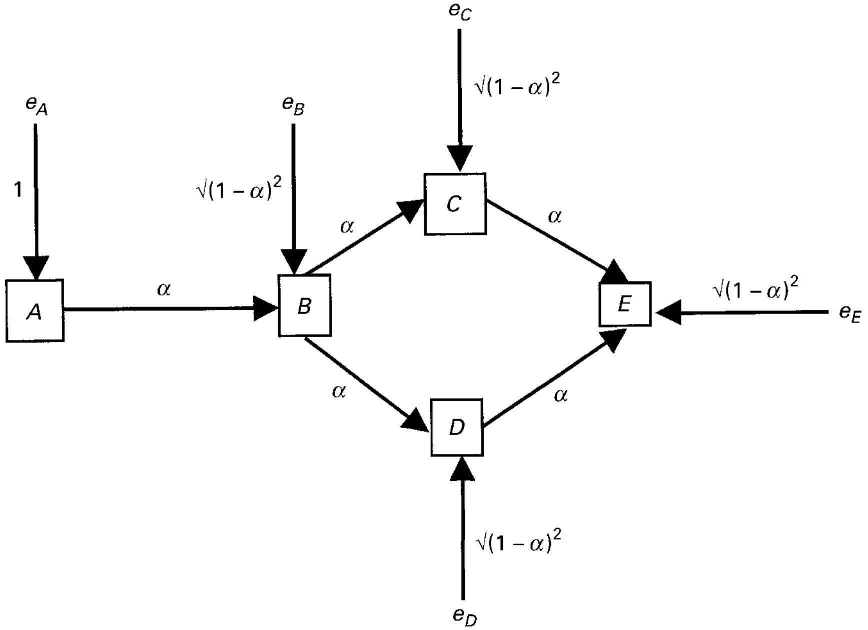


Figure 8.23. Path model used to generate sample data.

quite possibly miss some of the smaller trails and even include some incorrect paths. Independent tests using new data are always important after applying the search algorithms but this is especially important as sample size decreases. None the less, you will see that the error rates are not that bad, especially for constructing the undirected dependency graph, even at small sample sizes. With these points in mind, let's look at the Causal Inference algorithm in the presence of sampling error.

8.15 The Causal Inference algorithm and sampling error

At this stage it is useful to look at a numerical example. Figure 8.23 shows a path model from which I will generate sample data and apply the Causal Inference algorithm. I will generate two different data sets, one with 30 observations and one with 300 observations. The coefficient α equals 0.4. Let's begin with the larger sample size ($N=300$). Table 8.4 shows the variances on the diagonal, the covariances below the diagonal, and the correlations above the diagonal, for a simulated data set with $N=300$.

The first step is to obtain the undirected dependency graph. I will

Table 8.4. *Variances (diagonal), covariances (lower subdiagonal) and correlations (upper subdiagonal) of 300 simulated data from a multivariate normal distribution generated according to the causal structure shown in Figure 8.23*

	A	*B*	*C*	*D*	*E*
A	0.93	0.35	0.14	0.08	0.10
B	0.35	1.03	0.42	0.38	0.35
C	0.15	0.45	1.11	0.22	0.52
D	0.07	0.37	0.22	0.94	0.52
E	0.11	0.38	0.59	0.54	1.12

choose a significance level[25] of 0.05. After constructing the saturated undirected graph, I then remove any lines between variables whose zero-order correlations are not significantly different from zero at a rejection level of 0.05; that is, if $|r| < 0.113$. There is one correlation coefficient in Table 8.4, between *A* and *E*, that is judged to be zero. Note that this decision is actually a Type II error, since *A* and *E* are really associated with a weak population coefficient of 0.128. However, this does not introduce any errors in the undirected dependency graph, since this graph is concerned only with direct associations. In other words, the algorithm is robust to these types of error.

I next look at each of the 10 pairs of variables ($\{A,B\}$, $\{A,C\}$, . . ., $\{D,E\}$) and, for each, test for zero first-order partial correlations. That is, for each pair I calculate the partial correlation conditional on each of the remaining three variables in turn. With each test I see whether the absolute value of the partial correlation is less than 0.113 and, if it is, I remove the line joining the two variables in the pair. When I do this I find only four first-order partials that are judged to be zero: $r_{AC|B} = -0.008$ ($p = 0.89$), $r_{AD|B} = -0.062$ ($p = 0.28$), $r_{AE|B} = -0.022$ ($p = 0.70$) and $r_{CD|B} = 0.069$ ($p = 0.23$). Note that we can't remove the line between *A* and *E*, since it was already removed (by error) when looking at the zero-order correlations[26]. This is why I said that the algorithm is robust.

[25] This significance level refers to each individual statistical test, not the final partially oriented graph. From Figure 8.17 I know that I will have almost 100% power to detect correlations whose population values are greater than 0.2 in absolute value, although, in an empirical study, I would not know what the population values were.

[26] Actually the algorithm would not even calculate this partial correlation, since the two variables are not adjacent at this stage.

I next look at each of the remaining pairs of variables that are still adjacent and, for each, test for a zero second-order partial correlation. That is, for each pair I calculate the partial correlation conditional on each possible pair of the remaining three variables in turn and remove the line between any pair whose absolute value of the second-order partial is less than 0.113. There is only one such zero second-order partial: $r_{BE|\{CD\}} = 0.009$ ($p = 0.88$). I then go on to test for zero third-order partials but I do not find any. The result is the correct undirected dependency graph (Figure 8.24, top).

I then go on to the orientation phase. If I maintain the same significance level (0.05) then the algorithm makes a mistake. The path Co→E←oD collides at E but, in order to detect this, I must find that the partial correlation between C and D, conditioned on every other possible subset of the other variables that includes E, is never zero. However, even at a sample size of 300 we don't have much power to detect small non-zero associations. What's even worse, we have to conduct four different tests ($r_{CD|\{E\}}$,$r_{CD|\{EA\}}$ $r_{CD|\{EB\}}$, $r_{CD|\{EAB\}}$). If these were independent tests then our actual rejection level for all four tests together would be $0.05^4 = 6.25 \times 10^{-6}$. If these four tests were perfectly correlated (i.e. the probability level for each is the same) then we would have maintained an overall significance level of 0.05. In other words, by setting the significance level of each test at 0.05 we were really demanding very strong evidence – perhaps as low as six chances in a million and certainly less than five chances in a hundred – before we were willing to recognise a collider triplet. Of course, we don't know the degree to which these four tests are correlated but if they were independent then we should choose a significance level of about 0.47 for each test in order to maintain an overall level of 0.05. This is because $0.47^4 = 0.049$.

These considerations emphasise the exploratory nature of the method. The best approach is to try different significance levels and see how the undirected graph, and the partially directed graph, change. In any real study you could also include any prior information about the data, for instance if some variables occurred earlier in time than others, in order to choose between the different results. Figure 8.24 shows the undirected graph, and the partially directed graph that results from our data using different significance levels. The structure of this undirected dependency graph is very stable; varying the significance level from 0.001 to 0.6 always gives the same result. The structure of the partially oriented graph is less stable. Below a significance level of around 0.15 all unshielded patterns are declared unshielded colliders, including the one that is really a collider. Between 0.15 and 0.3 the correct partially oriented graph is obtained. Beyond 0.3 a second (incorrect) unshielded collider is detected.

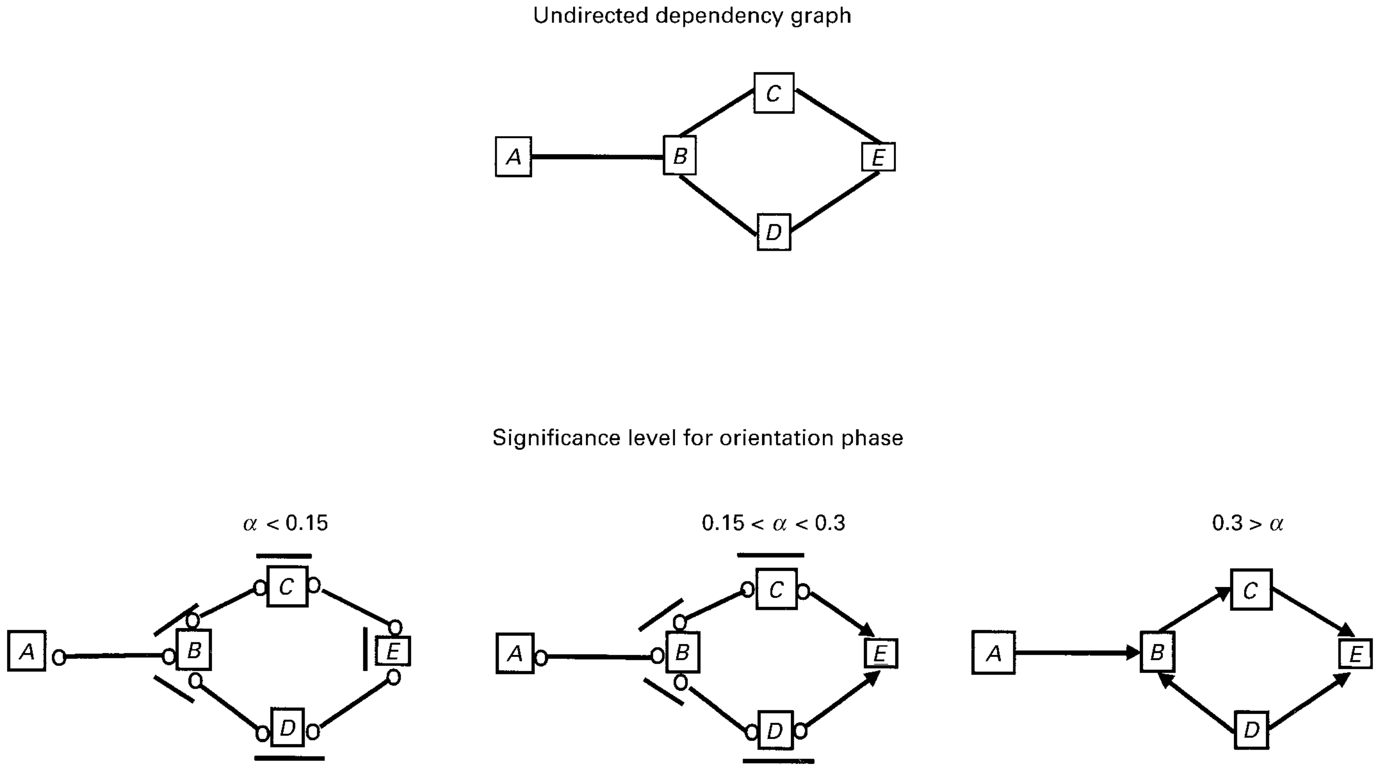

Undirected dependency graph
A
B
C
D
E
Significance level for orientation phase
α < 0.15
0.15 < α < 0.3
0.3 > α

Figure 8.24. The top graph is the undirected dependency graph obtained by applying the Causal Inference algorithm to the sample data. The three graphs on the bottom show the final output of the algorithm when using different significance levels for the orientation phase.

At this stage I can use my d-sep test to evaluate each of the three partially oriented graphs[27]. When I do this I find that the equivalent graphs that result when the significance level used in the orientation stage is 0.3 or larger are clearly rejected. This is because these graphs all predict that *A* and *C* are unconditionally independent. In fact, $r_{AC}=0.35$ and, with a sample size of 300, this would occur much less than once in a million times if the data were really generated according to this graph. Therefore, we have to choose between the first two partially oriented graphs. When we look at the first partially oriented graph we see that it is impossible for all of the unshielded patterns to be definite non-colliders and for there to also be no cycles in any equivalent DAG that is consistent with it. We are led to accept the middle partially oriented graph as the most consistent with our data.

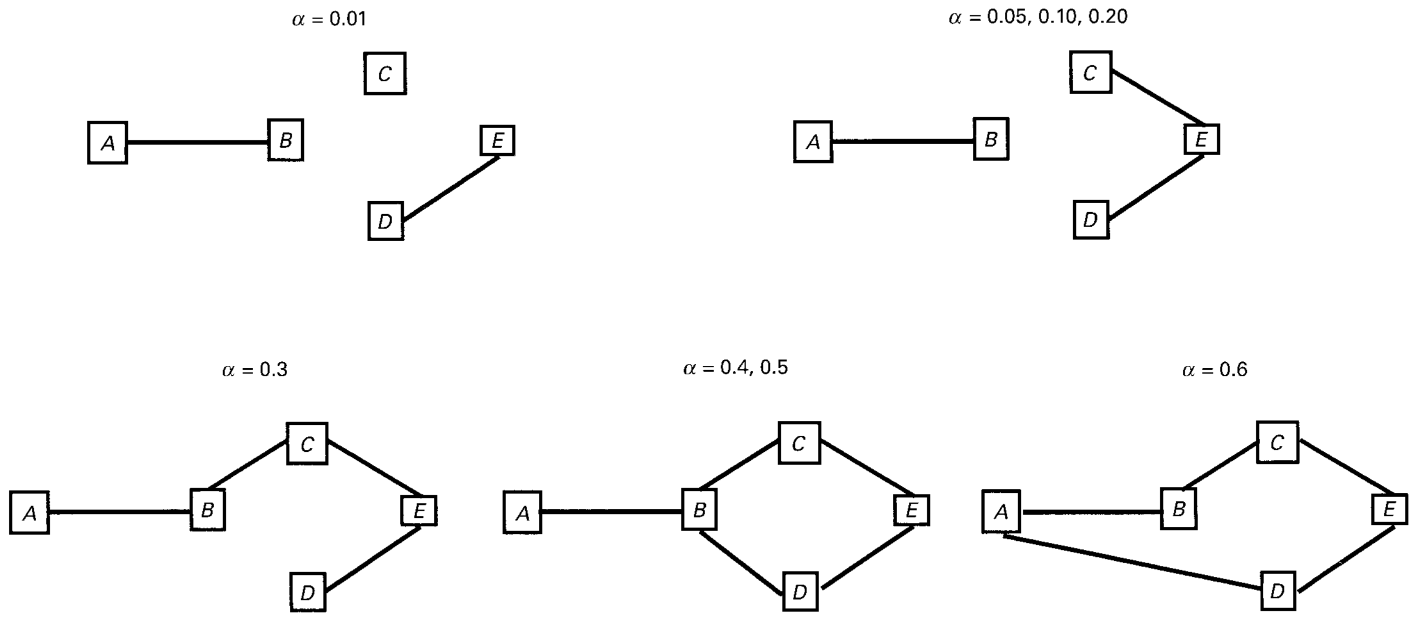
Figure 8.25 shows the results when the undirected dependency graph algorithm is applied to a sample data set, generated from Figure 8.23, but with a small sample size of 30 observations. Now we see many errors. No equivalent model from the undirected dependency graph, obtained using $\alpha=0.01$, provides an acceptable fit to the data. However, all of the remaining undirected graphs have equivalent models that do produce an acceptable fit, based on the d-sep test and a significance level of 0.05. To go any further requires information beyond that which exists in this little data set. For instance, the undirected graphs at $\alpha=0.05$ to 0.2 all predict that *A* and *B* are independent of *C*, *D* and *E*[28]. If, in other studies, either *A* or *B* was found to be correlated with *C*, *D* or *E*, then this undirected graph could be rejected[29]. Once you decide upon a particular undirected graph as being most consistent with all of the information, then you can begin to explore the orientation phase.

The best way to see what types of error these search algorithms will make at different sample sizes is to generate data with different characteristics and count the types of error that occur. Spirtes, Glymour and Scheines (1993) have conducted such simulation studies for the case of causal sufficiency and using both the algorithms described here and also others that are more efficient with large numbers of variables, but which commit more

[27] Every possible partially oriented graph can be tested in this way. Simply choose one of the equivalent DAGs consistent with the partially oriented graph and apply the d-sep test. Since every equivalent graph will give the same probability level under the null hypothesis it doesn't matter which one you choose.

[28] Since there are no undirected paths between these two sets of variables, there can be no directed paths either. Therefore they must be independent.

[29] In Shipley (1997) I described how imbedding the algorithm for the undirected dependency graph inside a bootstrap loop helps to reduce the effects of small sample sizes. This option is included in the EPA program of my Toolbox (see Appendix).



$\alpha = 0.01$

$\alpha = 0.05, 0.10, 0.20$

$\alpha = 0.3$

$\alpha = 0.4, 0.5$

$\alpha = 0.6$

Figure 8.25. Output of the undirected dependency graph algorithm when applied to a sample data set, generated from Figure 8.23, with a small sample size of 30 observations.

errors at small sample sizes. I have also explored the error rates of the Causal Inference algorithm with small sample sizes, and using bootstrap techniques (Shipley 1997). The general results that come from these simulation studies are:

1. Error rates are lower for constructing the undirected dependency graph than for orienting the edges.
2. The error rates for adding a line in the undirected dependency graph when there shouldn't be one are quite low. Even at very small sample sizes (say 30 observations), if a line appears then it probably exists unless the rejection level is very high (say 0.5 or more).
3. The error rates for missing a line in the undirected dependency graph when there should be one are higher. As the strength of the direct causal relationships decrease, this error rate increases. As the number of other variables to which a given variable is a direct cause increases, this error rate increases. As the sample size increases, this error rate decreases.
4. The error rates for orienting edges are higher than those related to the undirected dependency graph. This is to be expected, since the orientation phase depends on the number and types of unshielded pattern; therefore any errors in the undirected dependency graph will be propagated into the orientation phase.
5. The rejection level used in constructing the undirected dependency graph should increase as the sample size decreases. At very small sample sizes values of 0.2 or higher should be used. At sample sizes of around 100 to 300 a rejection level of 0.1 should be used. At higher sample sizes a value of 0.05 is fine, since statistical power does not have to be traded off against the ability to avoid Type I errors.

8.16 The Vanishing Tetrad algorithm and sampling variation

The Vanishing Tetrad algorithm has been much less studied than the other algorithms that I have described. In part, this is because the assumption of a linear relationship between the latents and the observed variables limits its application. Another reason is perhaps because it is less informative than the other algorithms; it can alert us to the presence of latent variables but can't tell us exactly how these latents connect to the observed variables. Another reason is that, unlike the tests for (conditional) independence that are used in the other algorithms, the test for a zero tetrad equation is only asymptotic. In other words, in order to get accurate probability estimates based on the null hypothesis that a tetrad equation is zero, you need a certain minimum sample size. No one (to my knowledge) has formally studied the asymptotic requirements for this test, and so I will present some Monte Carlo results to give you some rules of thumb when interpreting the asymptotic probability levels of the test statistic.

The test statistic is $\tau = \rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK}$, where I, J, K and L are the four variables involved in the tetrad; remember that there are always three tetrad equations for each set of four variables. Under the null hypothesis this value will be zero in the population. Wishart (1928) derived the asymptotic sampling variance of this statistic in the first part of the twentieth century but no one (to my knowledge) has ever derived the exact sampling variance. The asymptotic sampling variance[30] is:

$$\left(\frac{D_{IK}D_{JL}(N+1)}{(N-1)} - D\right)\left(\frac{1}{N-2}\right)$$

where D is the determinant of the population correlation matrix of the four variables, D_{IK} is the determinant of the 2×2 matrix consisting of the population correlation matrix of variables I and K, D_{JL} is the determinant of the 2×2 matrix consisting of the population correlation matrix of variables J and L, N is the sample size and the four variables follow a multivariate normal distribution. There are six possible pairs of four variables, four of these pairs define a tetrad equation and the other two pairs define the 2×2 submatrices whose determinants are used in calculating the asymptotic variance. If the null hypothesis is true then the test statistic is asymptotically distributed as a normal variate with a zero mean and the given variance. Therefore, the value $\tau/\sqrt{\text{var}(\tau)}$ asymptotically follows a standard normal distribution.

To conduct the statistical test you replace the population values by the sample values. In doing this you are only approximating the true probability level and so it is important to know how good (or bad) this approximation is. Table 8.5 shows some results of Monte Carlo simulations in which a four variable measurement model, of the sort shown in Figure 8.16, was used to generate 500 independent data sets. Each such simulation used a different sample size per data set and a different value for the path coefficients (α) between the latent variable and the observed variables. Remember that, according to the rules of path analysis, the population correlation coefficient (ρ) between each pair of observed variables in such a model, is α^2.

[30] The formula given by Spirtes, Glymour and Scheines (1993) is incorrect, but these authors give the correct formula in their earlier book (Glymour *et al.* 1987).

Table 8.5. *Each line summarises the results of 500 independent data sets generated from a model like that of Figure 8.16 with path coefficients α between the latent and each measured variable. ρ gives the population correlation between each measured variable, N is the sample size, $\sigma(\tau)$ is the asymptotic standard deviation of the tetrad equation and SD(τ) is the average standard deviation of the 500 data sets. Also shown are the 20%, 10% and 5% quantiles of the standardised tetrad equations* $\left(\tau/\sqrt{Var(\tau)}\right)$

					Quantiles		
α	ρ	N	$\sigma(\tau)$	SD(τ)	0.20	0.10	0.05
0.1	0.01	25	0.0603	0.0974	0.09	0.05	0.02
		50	0.0293	0.0483	0.11	0.06	0.02
		1000	0.0015	0.0024	0.12	0.05	0.02
0.3	0.09	25	0.0603	0.0974	0.09	0.05	0.02
		50	0.0374	0.0514	0.13	0.06	0.03
		100	0.0218	0.0285	0.16	0.07	0.03
		500	0.0079	0.0085	0.21	0.10	0.05
0.4	0.16	25	0.0811	0.1061	0.12	0.06	0.03
		50	0.0481	0.0582	0.17	0.09	0.05
0.5	0.25	25	0.0964	0.112	0.20	0.09	0.04
0.6	0.36	25	0.1094	0.1175	0.23	0.11	0.05

It is clear that when the correlations between the four observed variables are very low then the approximation is not good. When the population correlation between them is only 0.01 then even a sample size of 1000 produces conservative probability levels and you would reject the null hypothesis that the tetrad is zero too often. By the time that the population correlation between the four observed variables is about 0.25 then even small sample sizes cause no problems.

In using the Vanishing Tetrad algorithm, you would first construct the undirected dependency graph. If no set of four (or more) variables in the undirected dependency graph is saturated (i.e. in which each variable has a line to each other variable in the set) then there is no need to apply the Vanishing Tetrad algorithm. If you do see such a pattern then apply the Vanishing Tetrad algorithm to this set of variables, keeping in mind the approximate nature of the calculated probabilities. If the correlations between the variables are very weak then you should increase the significance level.

8.17 Empirical examples

In Chapters 3 and 4 I used data from Jordano (1995) consisting of 5 variables measured on 60 trees of St Lucie's Cherry, to test a path model. In fact, the path model was derived from the Causal Inference algorithm. The five variables were (1) the area of the tree canopy projection (a measure of the photosynthetic biomass), (2) the total number of ripe fruit produced per tree during the year, (3) average fruit diameter, (4) average seed weight and (5) the number of seeds dispersed from the tree by birds. The primary variable of interest was the number of seeds dispersed from the tree. This is because seeds that fall directly beneath the tree and germinate will generally die owing to shading from the parent and so the evolutionary fitness of the tree will be more closely related to the number of seeds that are dispersed away from the parent. However, it is reasonable to suppose that the other measured variables will interact to affect the number of seeds dispersed and so we wish to understand how these variables relate to one another.

There are approximately 59 000 potential acyclic graphs to consider given our five measured variables. I could propose a specific causal model, and even provide reasonable biological arguments to back it up. I am quite sure that most readers could also propose a specific causal model and provide reasonable biological arguments. I am also quite confident that the proposed causal models will not be the same[31]. In fact, I could propose alternative different causal models and come up with equally good reasons for each one! The problem is that my biological knowledge of the phenomenon is not sufficiently detailed to strongly favour one model over other reasonable models. This is a situation in which it does not make much sense to waste a great deal of effort in applying an inferential test to a specific model. When applying an inferential test (either the d-sep test or that based on SEM) the question is: 'Are the data consistent with *this* model?'. However, since we can easily come up with a number of different reasonable models, and have no idea whether there might be others, we are really at a stage in which we want to ask: '*Which* models are consistent with the data and *which* aren't?'. It is analogous to the difference between testing whether a regression coefficient is zero and obtaining a confidence limit for the regression coefficient.

After transforming the original variables to their natural logarithms, I gave these data to the Causal Inference algorithm. Figure 8.26 shows the resulting undirected dependency graph for various rejection levels. The undirected graph that results when the rejection level is low (0.05 or 0.10)

[31] In fact, the model that I proposed in Chapters 3 and 4 differed from the model proposed in Jordano (1995); see Chapter 3.

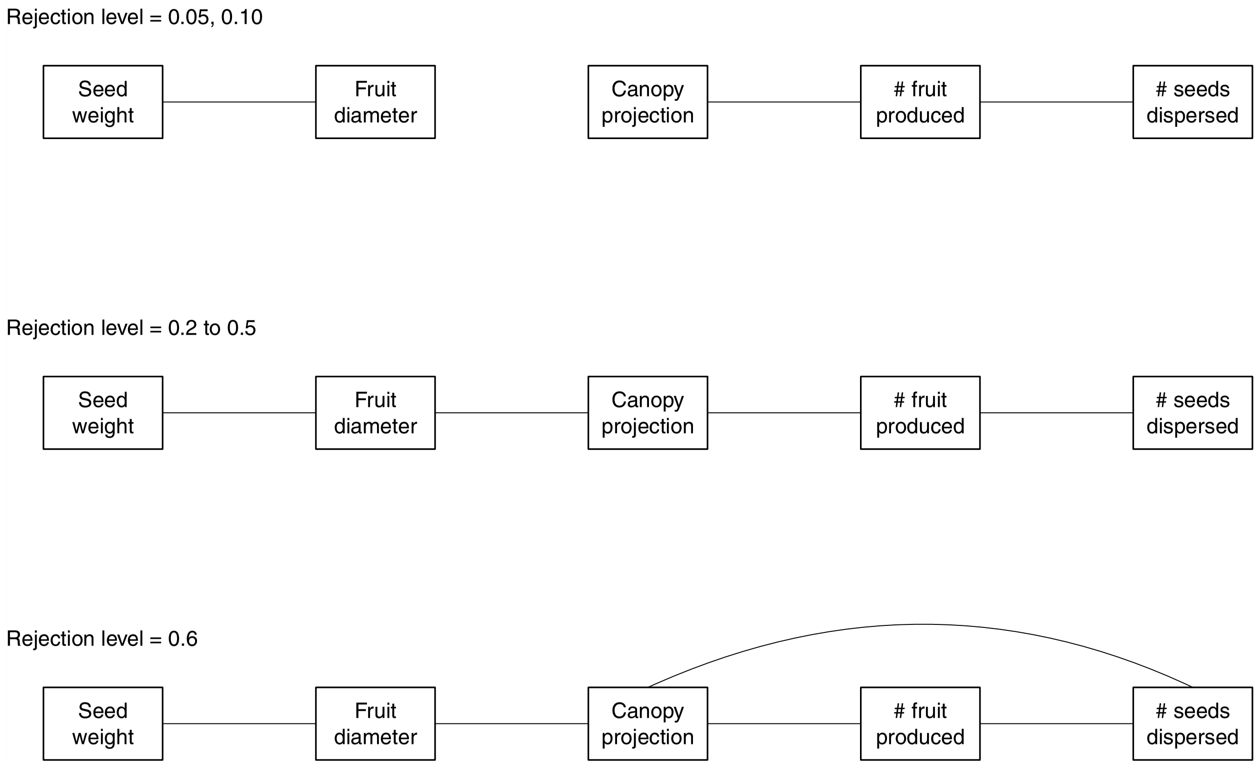


Figure 8.26. Output of the undirected dependency graph algorithm when applied to the empirical data of Jordano (1995) at various significance levels. #, number of.

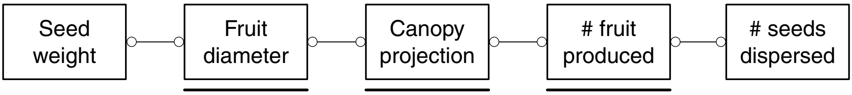


Figure 8.27. The final partially oriented graph that is produced based on the middle undirected graph of Figure 8.26. #, number of.

can be rejected without even orienting it. This is because it predicts that each of {seed weight, fruit diameter} is unconditionally d-separated from – and therefore independent of – each of {canopy projection, number of fruit produced, number of seeds dispersed}. Applying the d-sep test to only these independent statements yields a χ^2 value of 27.18 with 12 degrees of freedom ($p = 0.007$). If we then go to the second undirected graph, obtained using rejection levels of 0.2 to 0.5, we can apply the orientation phase of the algorithm. Until we get to a very high rejection level for this phase (0.4) we always find that each unshielded pattern is a definite non-collider. At a rejection level of 0.4 we are informed that fruit diameter is a collider. Figure 8.27 shows the partially oriented acyclic graph based on the middle undirected graph.

Despite the small sample size (60 observations) we have already discovered quite a lot of information about the possible causal relationships between these variables. There is no evidence that there are latent variables that are common causes of more than two observed variables; if there were then we would see three or more variables with a saturated pattern between them. Remember that we are in an exploratory mode. We are looking for possible models that accord with our available evidence about the correlational shadows but we also want our model to accord with any previous biological knowledge that we might possess. For instance, consider the relationship between the number of cherry fruits produced and the number of seeds dispersed by the birds (number of fruits produced o—o number of seeds dispersed). Since seeds can't be dispersed by birds before the fruit has been produced, we can exclude the orientation: (number of fruits produced ← number of seeds dispersed). It is possible that the orientation is: (number of fruits produced ↔ number of seeds dispersed), although it is difficult to conceive of a latent variable that determines both how many fruits the tree will produce and also how many of these fruits will be eaten by the birds. However, if we accept this orientation involving a latent variable then we must also exclude the orientation: (canopy projection o→ number of fruits produced), since this would produce a collider. This would therefore force us to accept the following orientation: (canopy projection←number of fruits produced). Such an orientation disagrees not only with much empirical evidence but also with the time ordering of the phenomenon, since the

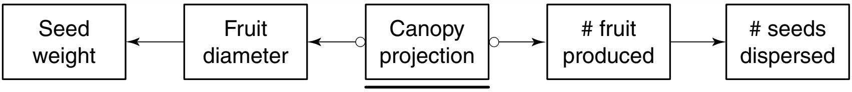


Figure 8.28. The partially oriented graph that is retained as most biologically plausible. #, number of.

canopy is produced before the fruits are made. If we begin with the biologically reasonable hypothesis that the total photosynthetic capital of the tree, of which the canopy projection area is a measure, determines both how many fruit will be produced and the average size of each fruit, then we are immediately led to the partially oriented directed acyclic graph in Figure 8.28.

Such a result, to me, is incredible. With five observed variables we had a little over 59000 possible directed acyclic models. This algorithm, combined with a few reasonable biological observations, reduced this huge number to a few reasonable models. Since I know that the statistical power to detect small non-zero correlations is not great with only 60 observations, I would not bet my salary on the accuracy of Figure 8.28, but I would feel much more confident about proposing a model derived from Figure 8.28 as a useful biological hypothesis to be tested with independent data.

This is the real strength of these discovery algorithms. Unless pre-existing theory is already quite solid, then proposing a complete causal model from such theory often degenerates into asking: ‘If I were God, and the world was a machine, then how would I construct it?’. Since few of us are gods and the world is not really a machine, such ‘hypothesis generation’ can easily mask unbridled speculation. The discovery algorithms first show us the correlational shadows that our data contain, which causal processes might reasonably have cast them, and which causal processes were unlikely to have cast these correlational shadows. This constrains our speculation, forces us to consider different alternative models, and also forces us to explicitly justify any causal process that appears to contradict what the data seem to say.

The next empirical example shows that we shouldn’t accept the output of these discovery algorithms blindly. Their purpose is to help us to develop useful causal hypotheses, not to replace the scientist with a computer algorithm. In Chapters 3 and 4 I presented a path model relating specific leaf area, leaf nitrogen content, stomatal conductance, net photosynthetic rate, and the CO_2 concentration within the leaf. This model (Figure 8.29A) was based on the pre-existing model of stomatal regulation produced by Cowan and Farquhar (1977). When I apply the Causal

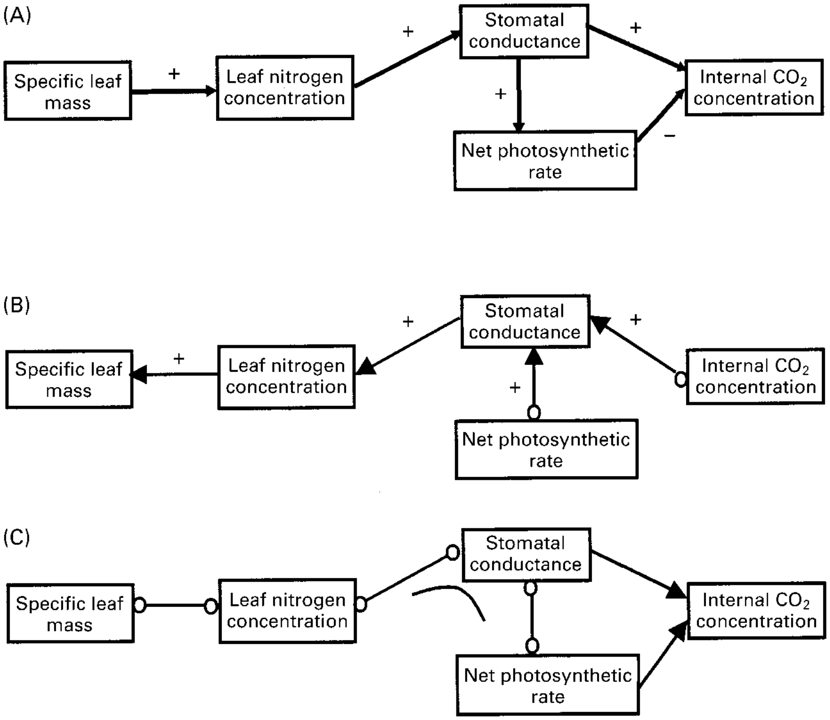


Figure 8.29. Model (A) is the one proposed in Chapter 3 based on biological arguments. The partially oriented graph (B) is the output of the Causal Inference algorithm when no constraints are placed on it. The partially oriented graph (C) is the output of the Causal Inference algorithm when simple constraints are placed on it based on well-known physical laws.

Inference algorithm to the empirical data[32] the resulting partially oriented graphs make no biological sense. At low rejection levels (0.2 and lower) none of the suggested graphs fit the data. At higher rejection level (0.2 to 0.5) a graph (Figure 8.29B) is suggested that does produce a path model with a non-significant MLX^2 value but this graph contradicts some well-established biological knowledge of leaf gas exchange. Note that in this graph the net photosynthetic rate is causally independent of the CO_2 concentration within the leaf, even though this is physically impossible. The amount of CO_2 within the leaf is determined by the rate at which it is diffusing into the leaf across the concentration gradient from the higher outside

[32] I use only the 35 species that have a C_3 photosynthetic system, and each variable is transformed to its natural logarithm to ensure multivariate normality.

concentration to the lower concentration within the leaf (i.e. stomatal conductance) and the rate at which it is being removed from the intercellular air by photosynthesis. The concentration within the leaf is therefore determined by the net rate at which it is being fixed by photosynthesis (by definition, the net photosynthetic rate) and the rate at which gases are diffusing across the stomates (measured by the stomatal conductance).

Why would the Causal Inference algorithm produce such an erroneous output? The reason is almost surely because this biological process violates one of the assumptions of the algorithm; namely that the probability distribution of these variables is faithful to the causal process that generated it[33]. This means that independence or partial independence relationships are assumed to be due to the way in which the variables are causally linked together rather than on special numerical values of the strengths of the direct causal relationships that manage to cancel each other. Imagine that you glance out of the window and see a single person walking down the lane. One hypothesis for this observation might be that there are really two people but that one is positioned behind the other in such a way that she is perfectly hidden by the person in front. A simpler, and more parsimonious, hypothesis is that there is really only one person coming down the lane. Both hypotheses are possible but the first requires that you are witnessing a very special juxtaposition of distances, shapes and sizes of people. The illusion of a single person would disappear as soon as any of those conditions change. We could say that such special conditions are *unfaithful* to our general expectations and so we would reject the hypothesis unless we had very good independent reasons to believe that someone might be hiding due to such special conditions. In the same way, these discovery algorithms assume that if we observe an observational independence between two variables, then this means that the two are causally independent. It is always possible that the two variables only *appear* to be independent because positive and negative direct and indirect relationships cancel each other out, but this would require a very special balancing of causal effects. Just as with the example of two people appearing to be a single person, unless we have good reasons for suspecting such a curious observation, we would choose the more parsimonious explanation.

In fact, the correlation coefficient between the ln-transformed net photosynthetic rates and the ln-transformed internal CO_2 concentrations in these data was only 0.051 ($p=0.77$). Therefore the line between these two variables would be immediately removed when we are constructing the undirected dependency graph. We know the net photosynthetic rate must

[33] Unfaithfulness only properly applies to the population probability distribution.

be a cause of the amount of CO_2 within the leaf and yet there appears to be no relationship between the two variables in these data. The reason for this apparent contradiction can be found in the Cowan and Farquhar (1977) model of stomatal regulation upon which the path model in Figure 8.24A is based. According to this theory, the stomates are regulated in order to maintain the internal CO_2 concentration at the 'break point'; that point at which carbon fixation is limited equally by the regeneration of Rubisco due to ATP production from the light reaction of photosynthesis and the amount of Rubisco available in the dark reaction of photosynthesis. In other words, the overall correlation between net photosynthetic rate and internal CO_2 concentration is determined by two different causal paths. One path is the direct effect of net photosynthetic rate in reducing the internal CO_2 concentration (net photosynthesis→internal CO_2). The other path is the trek from stomatal conductance that is a common cause of both net photosynthetic rate and internal CO_2 concentration (net photosynthesis←stomatal conductance→internal CO_2). Increasing the stomatal conductance increases the amount of CO_2 that enters the leaf, thus increasing both the photosynthetic rate and the internal CO_2 concentration. Furthermore, in order to maintain a constant internal CO_2 concentration, the stomates must ensure that the increase in internal CO_2 due to diffusion through the stomates is just enough to counter the decrease in CO_2 that is caused by the resulting increase in the net photosynthetic rate. By balancing these positive and negative effects, such homeostatic control maintains a constant internal CO_2 concentration but also produces an unfaithful probability distribution.

Since the operational definition of net photosynthetic rate is the rate at which CO_2 is being removed from the air within the leaf, we have good independent reasons to suspect that net photosynthetic rate will exert a direct negative effect on the internal CO_2 concentration. We can now apply the Causal Inference algorithm again but add the constraint that stomatal conductance and net photosynthetic rates must each remain as direct causes of the internal CO_2 concentration. This constraint is justified by simple physical laws of passive diffusion of gases across a concentration gradient. The resulting graph is shown in Figure 8.29C. The Causal Inference algorithm has suggested a partially oriented graph that is statistically equivalent to the path model that was proposed based on the Cowan–Farquhar theory of stomatal regulation. We have only to note that it is biologically more reasonable to suppose that the leaf nitrogen concentration is caused by specific leaf mass rather than the inverse[34] and we recover the path model in its entirety.

[34] The concentration of nitrogen does not vary strongly through the depth of the leaf. Therefore, if there is more leaf biomass per leaf area (i.e. the leaf is thicker) then there will be more nitrogen per unit leaf area.

The assumption of faithfulness is really based on a parsimony argument. It says that, if the only causal information available is that obtained from the observational data at hand and we have different possible causal structures that are exactly equivalent in their predictions of (partial) independence, then it is preferable to assume a causal structure whose independence predictions are robust rather than to assume a causal structure whose independence predictions require a special balancing of direct and indirect causal effects. When we have good causal information exterior to the data at hand then such information should be used. With small samples this is especially important because low statistical power means that real, but weak, effects might be incorrectly interpreted as independence.

8.18 Orienting edges in the undirected dependency graph without assuming an acyclic causal structure

A recurring theme in science fiction stories is the Universal Translator; a device that can infallibly translate back and forth between any set of languages. In the case of acyclic causal structures we had an imperfect, but still quite serviceable, translation device. d-separation, applied to the directed acyclic graph, could infallibly translate from the language of causality to the language of probability distributions. We could not use it to translate infallibly backwards from a probability distribution to the causal graph because different causal structures can generate the same joint probability distribution. This is why the discovery algorithms output a partially oriented acyclic graph rather than a single DAG. Still, this is quite useful in reducing hypothesis space down to a manageable set of possible DAGs. When we move on to search algorithms for (possibly) cyclic causal processes then the problem gets even more difficult because our translation device, d-separation, can't be generally applied to non-linear cyclic causal processes. None the less, Richardson (1996b) has produced an algorithm that is provably correct for cyclic causal structures (given population measures of association and faithfulness) under the assumption that the functional relationships between the variables are linear and that there are no latent variables generating associations between more than two observed variables.

As you might have already feared, this algorithm is both more complicated and requires some new definitions and notational conventions. Taken one at a time, each part of the algorithm is still intuitively comprehensible. The algorithm is based on the notion of a *partial ancestral graph* (PAG). A PAG is an extension of the partially oriented inducing path graph of acyclic models that was used in the Causal Inference algorithm. Here are the conditions for a graph to be a PAG:

1. There is an edge between two variables, A and B, if and only if A and B are d-connected given any subset of other observed variables in the graph; i.e. if and only if there is an inducing path between A and B. This is the same as for the graphs output from the Causal Inference algorithm.
2. If there is an edge between A and B that is out of A with the notation $A\text{—}\star\mathrm{B}$ (but not necessarily into B), then A is an ancestor of B.
3. If there is an edge between A and B that is into B with the notation $A\star\!\!\rightarrow B$ (but not necessarily out of A), then B is *not* an ancestor of A.
4. If there is an underlining at the middle variable of a triplet with notation $A\star\text{—}\underline{\star B\star}\text{—}\star C$ then the edges *do not* collide at B. Therefore B is an ancestor of either A or B but not of both.
5. If there is an edge from A to B and from C to B ($A\rightarrow B\leftarrow C$) but B is *not* a descendant of a common child of A and B, then B is doubly underlined[35]; thus $A\rightarrow\underline{\underline{B}}\leftarrow C$. This is the first big difference from the POIPG graphs output from the Causal Inference algorithm. In fact, such a condition is impossible in an acyclic causal structure as will be explained in more detail below.
6. Any edge endpoint not marked in one of the above ways is left with a small circle; thus $A\mathrm{o}\text{—}$ means that the mark after the A is unknown and could be any of $A\leftarrow$, $A\text{—}$ or $A\star\text{—}$.

If you go over these points slowly then you will see that the only major difference is in point 5. Let's look more closely at it. To begin, let's look at what would happen if we applied the Causal Inference algorithm to the causal structure, shown in Figure 8.30, containing a feedback loop[36].

Figure 8.31 shows the development of the partially oriented graph. The undirected dependency graph has no line between A and D because, applying d-separation to the true graph in Figure 8.30, A and D are unconditionally d-separated. However, there is a line between A and C and between D and B even though none appears in the true graph. This is not a mistake. Remember that a line in the undirected dependency graph simply means that the two variables joined by the line are d-connected given any subset of other variables (i.e. that there is an inducing path between them).

[35] Richardson (1996b) used a dotted underline rather than a double underline.

[36] In Chapter 2 I described how such cyclic graphs can result from a dynamic process in which the same variables are measured at different times and included in the same data set without any explicit indexing of the time dimension. Richardson (1996b) has an extended discussion of the interpretation of cyclic patterns in causal graphs.

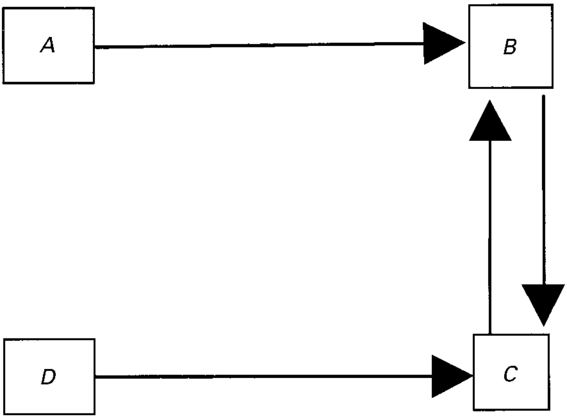


Figure 8.30. A directed cyclic graph with a feedback relationship between variables *B* and *C*.

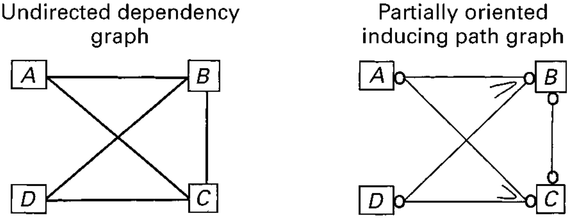

Undirected dependency graph
Partially oriented inducing path graph
A
B
C
D
A
B
C
D

Figure 8.31. The graph on the left is the undirected dependency graph that results from the directed cyclic graph in Figure 8.30. The graph on the right is the partially oriented inducing path graph.

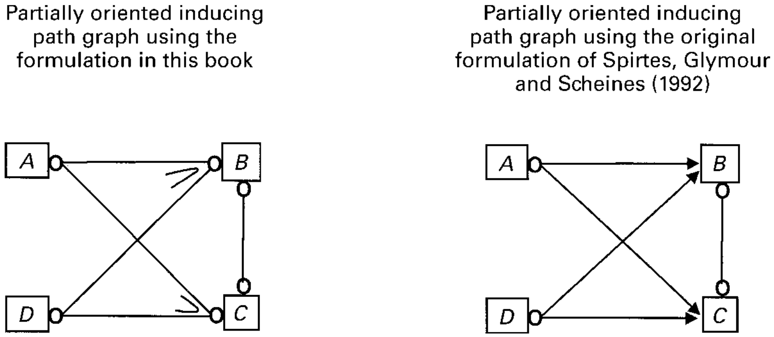
A is always d-connected to *C* because there are two undirected paths given the feedback relationship (*A*→*B*→*C* and *A*→*B*←*C*), and so conditioning on *B* (or {*B*,*D*}) will always leave one path open. The line between *D* and *B* occurs for the same reason. Now, when we apply the orientation phase of the Causal Inference algorithm to the undirected dependency graph, we find that the two unshielded patterns (*A*—*C*—*D* and *D*—*B*—*A*) are non-colliders. To see this, write out the two undirected paths between *A* and *D*: *A*→*B*→*C*←*D* and *A*→*B*←*C*←*D*. The unshielded pattern *A*—*C*—*D* is a collider (according the Causal Inference algorithm) if *A* and *D* are never d-separated given *C* plus any possible subset of other observed variables (i.e. {*C*} and {*B*,*C*}). Now, *A* and *C* are never d-separated given {*C*} because of the undirected path *A*→*B*→*C*←*D*. However, given {*B*,*C*} then both undirected paths are d-separated and so the Causal Inference algorithm would declare the unshielded pattern (*A*—*C*—*D*) to be a definite non-collider. The same result would occur for the unshielded pattern (*A*—*C*—*D*).

This mistake made by the Causal Inference algorithm occurs because of the feedback relationship between B and C. It could never occur in an acyclic causal process. Yet this mistake is actually very informative because it suggests a way of detecting feedback relationships. To see this we have to go back to the algorithm used to construct the undirected dependency graph.

In the algorithm for the undirected dependency graph we choose two variables (X and Y) and then look for a conditioning subset of other observed variables that renders X and Y independent (thus d-separated). We begin with the smallest such conditioning subset – the null subset containing no other variables – and test for independence of X and Y given this null subset (i.e. unconditional independence). If this does not occur then we test the first-order conditioning subsets, and so on. As soon as we find a conditioning subset that renders X and Y independent then we remove the line between them and stop. Let's call the conditioning subset that renders X and Y independent during the algorithm the '*Separation set of X and Y*', or **Sepset**(X,Y). Now, in an acyclic causal process in which the DAG has a collider, for example $X{\star}{\rightarrow}Z{\leftarrow}{\star}Y$, then Z will *never* be a member of **Sepset**(X,Y) because any conditioning set that contains Z will make X and Y d-connected. Furthermore, in an acyclic causal process of this type in which a given variable is a non-collider, for example $X{\star}\text{—}\underline{oZo}\text{—}Y$, then Z will *always* be a member of **Sepset**(X,Y). Therefore, given an unshielded pattern (X—Z—Y) and the assumption of no cyclic causal relationships, we can orient this unshielded pattern simply by determining whether Z is a member of **Sepset**(X,Y). If Z is a member of **Sepset**(X,Y) then Z is a non-collider in the unshielded pattern; if not, then Z is a collider. Since **Sepset**(X,Y) is the smallest subset that d-separates X and Y there can be other subsets that also d-separate X and Y. However, if Z is a collider along X—Z—Y then neither **Sepset**(X,Y) nor any other subset that d-separates X and Y will ever contain Z in an acyclic graph. However, in a cyclic causal process this is not true. Because the causal influence goes both ways in a cyclic graph it is possible for Z to be absent from **Sepset**(X,Y) given the unshielded pattern X—Z—Y even though Z plus some other variables can d-separate X and Y.

For instance, let's look again Figure 8.31. There is an unshielded pattern (A—B—D) and the conditioning sets that d-separate A and D are: {null} – since A and D are unconditionally d-separated – and $\{B,C\}$[37]. **Sepset**(A,D) is {null} because this is the smallest conditioning set. Since B

[37] There are two undirected paths between A and D: $A{\rightarrow}B{\rightarrow}C{\leftarrow}D$ and $A{\rightarrow}B{\leftarrow}C{\leftarrow}D$. The first undirected path is d-separated given $\{B,C\}$ because B is blocked. The second undirected path is d-separated given $\{B,C\}$ because C is blocked.



Partially oriented inducing path graph using the formulation in this book

Partially oriented inducing path graph using the original formulation of Spirtes, Glymour and Scheines (1992)

Figure 8.32. The partially oriented inducing path graphs that result from the two different formulations of the Causal Inference algorithm.

is not in **Sepset**(*A*,*D*) this would normally mean that *B* is a collider if there were no cycles. If there were no cycles then this would also mean that no other set that d-separates *A* and *D* could contain *B*, yet {*B*,*C*} does d-separate *A* and *D*. The discovery algorithm for cyclic causal structures uses this fact. If, given an unshielded pattern *X*—*Z*—*Y*, *Z* appears in some subsets of other variables that d-separate *X* and *Y* but not in others, then there is a cyclic orientation.

Before presenting the full algorithm I should head off some potential confusion that could result if people compared my description of the algorithms for the undirected dependency graph and the Causal Inference algorithm with those published by Spirtes, Glymour and Scheines (1993). In my descriptions I did not refer to Sepsets but these play an important role in the original descriptions. I did this because, for acyclic causal processes, the result is the same in the population, although the use of Sepsets reduces the computational cost by one's not having to retest for d-separation during the orientation phases. If applied to causal processes having cyclic structures, the result can be different. Figure 8.32 shows the resulting partially oriented inducing path graphs that are obtained from the causal process shown in Figure 8.30 when applying the Causal Inference algorithm as I have presented it, and as Spirtes, Glymour and Scheines (1993) originally presented it.

Notice that the unshielded patterns (*A*—*B*—*D* and *A*—*C*—*D*) are oriented as non-colliders in my formulation and colliders in the original formulation. This is because **Sepset**(*A*,*D*) is empty (a null set). In the unshielded pattern *A*o—o*B*o—o*D*, *B* is not in **Sepset**(*A*,*D*) and so *B* is therefore considered a collider in the original formulation. In my formulation, given *A*o—o*B*o—o*D*, one must find that *A* and *D* are conditionally associated given *B* plus any other subset of observed variables, not just

Subset(*A*,*D*). Since *A* and *D* are d-separated given {*B*,*C*}, then *A*o—o*B*o—o*D* is oriented as *A*o—$\underline{\text{o}B\text{o}}$—o*D*. It is precisely when these two formulations disagree that we have evidence for a feedback relationship. Now, I give the Cyclic Causal Discovery (CCD) algorithm of Richardson.

8.19 The Cyclic Causal Discovery algorithm

1. Form the undirected dependency graph. As soon as a pair of variables (*X*,*Y*) are found to be independent given a conditioning set **Q** then let **Sepset**(*X*,*Y*) = **Sepset**(*Y*,*X*) = **Q** and go on to another pair of variables. Orient each edge (*X*—*Y*) between two variables in the undirected dependency graph as (*X*o—o*Y*).
2. For each unshielded pattern (*X*o—o*Z*o—o*Y*) orient as (*X*o→*Z*←o*Y*) if *Z* is not in **Sepset**(*X*,*Y*) and orient as (*X*o—$\underline{\text{o}Z\text{o}}$—o*Y*) if *Z* is in **Sepset**(*X*,*Y*).
3. For each triplet of variables (*A*,*X*,*Y*) such that *X* and *Y* are adjacent (i.e. *X*o—o*Y*) and *A* is not adjacent to either *X* or *Y*, then (i) if **Sepset**(*A*,*Y*) is not a subset[38] of **Sepset**(*A*,*X*) then orient *X*o—★*Y* as *X*←*Y* else (ii) if **Sepset**(*A*,*X*) is not a subset of **Sepset**(*A*,*Y*) but *A* and *X* are d-connected given **Sepset**(*A*,*Y*) then orient *X*o—★*Y* as *X*←*Y*.
4. Find each triplet of variables that are now oriented as (*A*→*B*←*C*). Now find those variables *V* that are *local* to *A*; the set of all such variables is called **Local**(*A*). A variable *V* is *local* to *A* if either *A*★—★*V* or *A*→*Y*←*V*. Next form a new set (***T***) consisting of all variables in **Local**(*A*) except for *B*, *C* or those also in **Sepset**(*A*,*C*). If any subset of ***T*** d-separates *A* and *C* then orient as (*A*→$\underline{\underline{B}}$←*C*) and record ***T*** plus **Sepset**(*A*,*C*) plus *B* in new sets called **SupSepset**(*A*,*B*,*C*) and **SupSepset**(*C*,*B*,*A*). The double underline means that *X* and *Y* do not both collide at *Z*[39].
5. Find a quadruple of variables (*A*,*B*,*C*,*D*) such that *B* and *D* are adjacent and the following patterns exist: *A*→$\underline{\underline{B}}$←*C* and either *A*→*D*←*C* or *A*→$\underline{\underline{D}}$←*C*. If such patterns exist then orient o—o*D*

[38] A set is a subset of another set if every element in the first set is also in the second set. For instance ***A*** = {*X*,*Y*,*Z*} is a subset of ***B*** = {*W*,*X*,*Y*,*Z*}.

[39] The original formulation of this algorithm uses a dotted underline rather that a double underline. Although this section appears quite complicated, the complications arise in an effort to save computational time. Basically, this step simply looks for a pattern (*A*★→*B*←★*C*) such that *B* is not in **Sepset**(*A*,*C*) but there is at least one subset of other variables that includes *B* and that d-separated *A* and *C*. This is the clue needed to detect a feedback relationship.

or B—oD as B→D if D is not in **SupSepset**(A,B,C) and orient as B*—D if D is in **SupSepset**(A,B,C).

6. Find a quadruple of variables (A,B,C,D) such that B and D are adjacent, D is not adjacent to both A and C, and $A\rightarrow\underline{\underline{B}}\leftarrow C$. If such conditions exist then orient Bo—oD or B—oD as B→D if A and C are not d-separated given **SupSepset**(A,B,C) plus D.

This algorithm probably seems overwhelming. Since it is incorporated into the TETRAD III program you don't have to understand it sufficiently to actually program it, only well enough to have an intuitive knowledge of what it does. The most important part is to be able to interpret it. I'll go over each section of the algorithm and provide an intuitive explanation. Note, however, that this algorithm is the most general algorithm of all those presented so far[40]. If the causal process is acyclic then this algorithm will give the same output as the Causal Inference algorithm even if the functional relationships are non-linear.

Section 1: This is simply the algorithm for constructing the undirected dependency graph. If there is an edge between two variables (X,Y) then there is an inducing path between them and no other variable, or set of these other variables, can d-separate X and Y. The reason for constructing **Sepset**(X,Y) and **Sepset**(Y,X) is simply so that we don't have to keep conducting the independence tests in the other sections of the algorithm. We could have done this in the Causal Inference algorithm as well. In fact, the original formulation of the Causal Inference algorithm, as implemented in TETRAD II and TETRAD III does use separation sets. The separation sets provide useful information because every variable in **Sepset**(X,Y) is an ancestor of either X or Y.

Section 2: This section is simply the algorithm for determining whether variable Y in an unshielded pattern (Xo—oZo—oY) is a collider (thus Xo→Z←oY) or a non-collider (thus $X\text{o—}\underline{\text{o}Z\text{o}}\text{—}Y$). Now that we have **Sepset**(X,Y) we don't have to re-do all of the (conditional) independence tests. If we see that Z is in **Sepset**(X,Y) then Z is a non-collider and if Z isn't in **Sepset**(X,Y) then it is a collider. This uses the fact (above) that if Z is in **Sepset**(X,Y) then Z is an ancestor of either X or Y. Therefore, the orientation can't be Xo→Z←oY because this would imply that Z was a descendant of both X and Y. This also explains why causal processes having feedback relationships like that shown in Figure 8.25 produce different results when we apply my version of the Causal Inference algorithm and the original version that uses separation sets. In such feedback processes a vari-

[40] There are a few 'propagation' rules that can be added after the algorithm is finished. For instance, $X\text{o}\rightarrow\underline{Y}\text{o—o}Z$ implies Xo→Y—oZ.

able can be *both* an ancestor and a descendant at the same time. This is not possible in an acyclic causal structure.

Section 3: We know that X and Y are adjacent (Xo—★Y) and that A is not adjacent to either X or Y by looking at the partially oriented graph. We also know that those variables that d-separate A and Y are not a subset of those variables that d-separate A and X; i.e. that A and X are still d-connected given **Sepset**(A,Y). Therefore X is not an ancestor of Y and we can orient Xo—★Y as X←★Y.

Section 4: This section begins by looking for a triplet of variables that has already been oriented as $A \rightarrow B \leftarrow C$ in section 2 of the algorithm. However, we have already seen that if there are feedback loops then a variable can be both an ancestor and a descendant of another variable. Therefore, this section tries to find some set of variables that d-separates A and C while including B. Remember that if we see $A \rightarrow B \leftarrow C$ then this means that, in section 2, we had found A, B and C to form an unshielded pattern and that B *wasn't* a member of the separation set that d-separated A and C. So we will have found two separation sets, one with B and one without B, that d-separate A and C. This is the signal for a feedback loop. Since this section looks for the smallest set that includes B and **Sepset**(A,C) – i.e. **Supsepset**(A,B,C) – this means that every variable in **Supsepset** (A,B,C) is an ancestor of A, B, or C. The double underline that is added to B means that both A and C can't both collide at B; some equivalent graphs have $A \rightarrow B$ and C is not adjacent to B, while other equivalent graphs have $C \rightarrow B$ and A is not adjacent to B.

Sections 5 and 6: Since every member of **Supsepset**(A,B,C) is an ancestor of A, B, or C we can now use this information to orient Bo—oD.

The proof of the correctness of each section of this algorithm, given the assumptions, is provided by Richardson (1996a,b). Let's apply this algorithm to the causal structure shown in Figure 8.30, reproduced as Figure 8.33.

Assuming that we have a very large sample size, so that we can ignore errors in determining probabilistic independence due to sampling variations, the undirected dependency graph, obtained after section 1, is shown in Figure 8.34.

There are only two unshielded patterns (Ao—oBo—oD and Ao—oCo—oD). Since A and D are unconditionally d-separated the **Sepset** (A,D) is empty; i.e. **Sepset**(A,D) = {null}. Therefore we orient these two unshielded patterns as $A \rightarrow B \leftarrow D$ and $A \rightarrow C \leftarrow D$. Figure 8.35 shows the partially oriented ancestral graph after this step.

No changes are made after applying section 3 because the necessary patterns don't exist. When we apply section 4 we find that **Sepset**(A,D) =

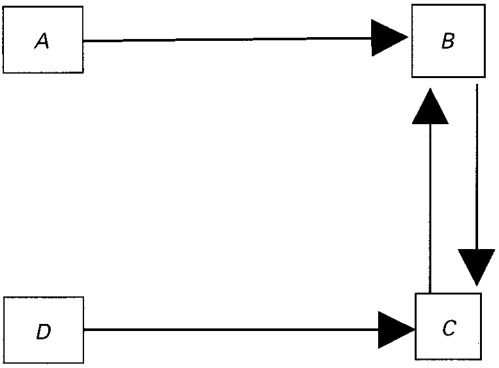


Figure 8.33. A directed cyclic graph with a feedback relationship between variables *B* and *C*.

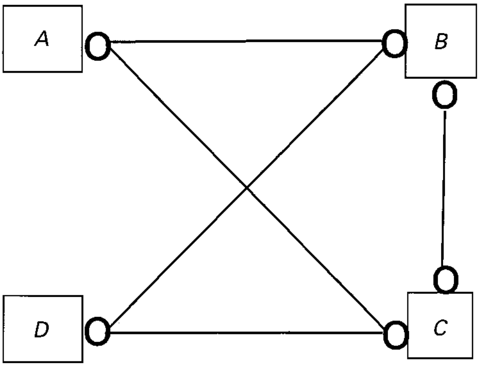


Figure 8.34. The undirected dependency graph, obtained after section A of the Cyclic Causal Discovery algorithm.

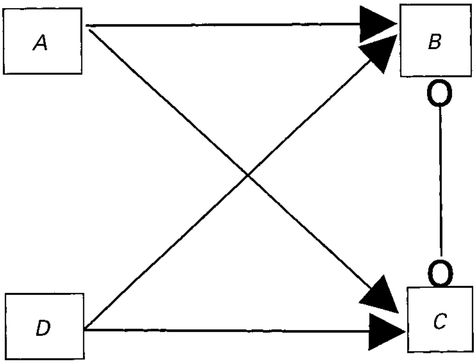


Figure 8.35. The partially oriented graph that is obtained after orienting based on the unshielded colliders.

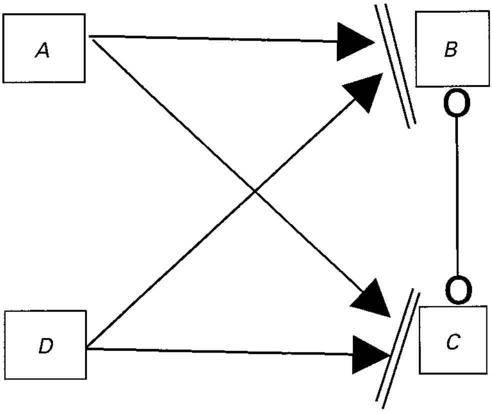
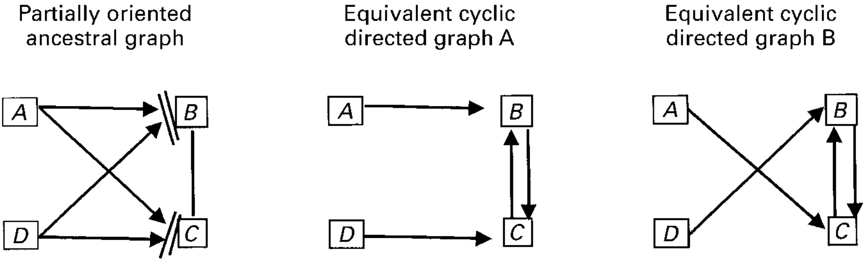


Figure 8.36. The partially oriented graph, obtained after section D of the Cyclic Causal Discovery algorithm.


Partially oriented ancestral graph

Equivalent cyclic directed graph A

Equivalent cyclic directed graph B

Figure 8.37. The partially oriented ancestral graph that results from the Cyclic Causal Discovery algorithm is shown on the left followed by the two equivalent cyclic graphs.

{null} but that A and D are d-separated given $\{B,C\}$. Therefore we re-orient $A \to B \leftarrow D$ to be $A \to \underline{\underline{B}} \leftarrow D$ and re-orient $A \to C \leftarrow D$ to be $A \to \underline{\underline{C}} \leftarrow D$. Next we construct **Supsepset**$(A,B,D) = \{B,C)$ and **Supsepset** $(A,C,D) = \{B,C\}$. The result is in Figure 8.36.

The double underlining means that A and D can't both be ancestors of either B or C at the same time. Upon arriving at section 5 we see that $A \to \underline{\underline{B}} \leftarrow D$ and $A \to \underline{\underline{C}} \leftarrow D$ and that B and C are adjacent. We can therefore go on and apply this section. Now **Supsepset**$(A,B,D) = \{B,C\}$ includes C and so the edge Bo—oC is oriented as B—oC and **Supsepset**(A,C,D) also includes B and so the edge B—oC is oriented as B—C, meaning that B and C are each the other's ancestor. The necessary conditions are not met in section 6 and so no further changes are made. The final partially oriented ancestral graph is shown in Figure 8.37 with the two directed cyclic models that are equivalent to it.

The last step is to devise a general discovery algorithm that is applicable to acyclic or cyclic causal processes with linear or non-linear functional relationships between the variables. In fact Richardson's (1996b) Ph.D. thesis provides just such an algorithm but, unfortunately, it is based on an unproven (as yet!) conjecture of Spirtes (1995). The study of causal models with feedback is an active area of research and, with luck, this chapter will soon be out of date.

8.20 In conclusion . . .

Given the dim view of causality that is adopted by most empiricists, it is ironic that the approach to causality taken in this book is almost, well . . . *empirical*. Rather than defining causality, one looks for those properties of relationships that scientists have deemed to call 'causal', and then develop a mathematical language that possesses such properties. In time perhaps this will lead to a comprehensive definition that can be accepted by everyone. For myself, I view 'causality' as a relationship between events or classes of events (i.e. variables) that possesses the properties of asymmetry, transitivity and the Markovian condition[41]. I expect that as our mathematical language of causality improves, we will be able to better express our scientific notions of causality using mathematics and this should lead to better tests of causal hypotheses as well as better discovery algorithms.

The various methods in this book all attempt to detect or test causal relationships using observational data. I have not intended my book to be an encyclopaedic treatment of the relationship between cause and correlation – there is certainly much more to be said – but I hope that it will be useful as you watch the correlational shadows dance across the screen of nature's Shadow Play.

Enjoy the show.

[41] I know. In Chapter 1 I promised not to give a definition of causality. I couldn't resist the temptation; I'm an academic and academics are drawn to definitions like young boys are drawn to puddles. We like to jump in and stir up the mud.

Appendix

The following is a list of software programs useful for path analysis, structural equations modelling (SEM) or exploration of causal structures. The inclusion of a program in this list is not an endorsement; I have not even used many of the listed programs. Since new versions of these programs are constantly being produced, you should visit the Internet websites before choosing.

AMOS (a commercial structural equations program)
SmallWaters Corporation
1507 E. 53rd Street, no. 452
Chicago, IL 60615, USA
http://www.smallwaters.com/amos

EQS (a commercial structural equations program)
Multivariate Software, Inc.
4924 Balboa Blvd, Encino, CA 91316, USA
http://www.mvsoft.com

LISREL and PRELIS
Scientific Software International
7383 N. Lincoln Avenue, Suite 100,
Lincolnwood, IL 60712-1704, USA
http://www.ssicentral.com

Mx (a free software program for structural equations modelling)

Mx is very complete but is not as user-friendly as the others and requires that the user enter command code. Both the program and the user's manual can be downloaded from http://views.usu.edu/mx/

Mplus (a commercial structural equations program)
11965 Venice Blvd, Suite 407
Los Angeles, CA 90066, USA
http://StatModel.com

CALIS – a procedure of SAS (a commercial statistical program of which CALIS is a procedure for SEM).
SAS Institute, Cary, NC 27513–2414, USA.

SEPATH – a module of Statistica (a commercial statistical program of which SEPATH is a module for SEM)
Statsoft,
2325 East 13th Street,
Tulsa, OK 74104, USA

TETRAD II (a program and manual for exploratory methods in causal modelling, some of which are described in Chapter 8)
Lawrence Erlbaum Associates, Inc.
10 Industrial Avenue,
Mahwah, NJ 07430-2262, USA
orders@Leahq.mhs.compuserve.com

(Note that TETRAD III is freely available from
http://hss.cmu.edu/HTML/departments/philosophy/TETRAD/tetrad.html)

TOOLBOX: this is a set of DOS-executable programs that I have written for various tasks in causal modelling that are not generally available in other programs. The price is very modest. These programs can be obtained directly from me (Bill Shipley, Department of Biology, University of Sherbrooke, Sherbrooke (Qc), J1K 2R1, Canada, bshipley@courrier.usherb.ca). To use them from within WINDOWS you have to first create a DOS window. These programs include:

- PARCOR: a program to calculate partial correlations (Pearson, Spearman) of various orders and associated probabilities;
- PERMR: a program to calculate partial correlations of various orders and associated probabilities using permutation methods.
- DGRAPH: a program to test path models without latent variables using the d-sep test described in Chapter 3.
- EPA2: a program that performs the exploratory methods without cyclic relationships, as described in Chapter 8, and also as described by Shipley (1997).

- TESTTET: an exploratory program to determine whether the observed correlations or covariances between a set of four variables provides evidence for an unmeasured common cause. This program is based on the Vanishing Tetrad algorithm of Chapter 8.
- IDEN: a program to determine whether a measurement model is identified based on Davis (1993), as described in Chapter 6.
- MULTICOV: a program to calculate covariance matrices from hierarchical data of up to five levels, as described in Chapter 7. You can use these covariance matrices to test multilevel models from within commercial SEM programs.

Finally, you might want to join an Internet discussion group (SEMNET) devoted to SEM. Many of the top statistical developers of SEM methods are members and are always willing to answer questions about the mechanics, statistics or philosophy of SEM. To join, visit http://bama.ua.edu/archives/semnet.html

References

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science* **10**: 364–376.

Bentler, P.M. (1995). *EQS structural equations program manual, version 3.0.* Los Angeles, BMDP Statistical Software.

Bentler, P.M. and Bonnett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* **88**: 588–606.

Beveridge, W.I.B. (1957). *The art of scientific investigation.* New York, Random House.

Blalock, H.M. (1961). Correlation and causality: the multivariate case. *Social Forces* **39**: 246–251.

Blalock, H.M. (1964). *Causal inferences in nonexperimental research.* Chapel Hill, NC, University of North Carolina.

Bollen, K.A. (1989). *Structural equations with latent variables.* New York, John Wiley and Sons.

Bollen, K.A. and Long, J.S., (1993). *Testing structural equation models.* Newbury Park, CA, Sage.

Bollen, K.A. and Stine, R.A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In *Testing structural equation models*, ed. Bollen, K.A. and Long, J.S., pages 111–134. Newbury Park, CA, Sage.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* **37**: 62–83.

Browne, M.W. and Cudeck, R. (1993). Alternative ways of assessing model fit. *Testing structural equation models*, ed. Bollen, K.A. and Long, J.S., pages 136–162. Newbury Park, CA, Sage.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical linear models.* Newbury Park, CA, Sage.

Bumpus, H.C. (1899). The elimination of the unfit as illustrated by the introduced sparrow. *Biological Lectures of the Woods Hole Marine Biological Station* **6**: 209–226.

Burke, J. (1996). *The pinball effect and other journeys through knowledge.* Boston, MA, Little, Brown and Company.

Cleveland, W.S. and Devlin, S.J. (1988). Locally-weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**: 596–610.

Cleveland, W.S., Devlin, S.J. and Grosse, E. (1988). Regression by local fitting. *Journal of Econometrics* **37**: 87–114.

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1992). Local regression models. In *Statistical models in S*, ed. Chambers, J.M. and Hastie, T.J., pages 309–376. Pacific Grove, CA, Wadworth and Brooks.

Conover, W.J. and Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician* **35**: 124–129.

Cowan, I.R. and Farquhar, G.D. (1977). Stomatal function in relation to leaf metabolism environment. In *Integration of activity in the higher plant*, ed. Jennings, D. H., pages 471–505. Cambridge, Cambridge University Press.

Cowles, M. and Davis, C. (1982a). Is the .05 level subjectively reasonable? *Canadian Journal of Behavioural Sciences* **14**: 248–232.

Cowles, M. and Davis, C. (1982b). On the origins of the .05 level of statistical significance. *American Psychologist* **37**: 553–558.

D'Agostino, R.B., Belanger, A. and D'Agostino, R.B.J. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician* **44**: 316–321.

Davenport, C.B. (1917). Inheritance of stature. *Genetics* **2**: 313–389.

Davis, W.R. (1993). The FC1 rule of identification for confirmatory factor analysis. *Sociological Methods and Research* **21**: 403–437.

De Robertis, E.D.P. and De Robertis, E.M.F. (1980). *Cell and molecular biology*. Philadelphia, Saunders College.

DeCarlo, L.T. (1997). On the meaning and use of kurtosis. *Psychological Methods* **2**: 292–307.

Duhem, P. (1914). *La théorie physique: son objet, sa structure.* Paris, Rivière.

Dunn, G., Everitt, B. and Pickles, A. (1993). *Modelling covariances and latent variables using EQS.* London, Chapman & Hall.

Eliason, S.R. (1993). *Maximum likelihood estimation. Logic and practice.* Newbury Park, CA, Sage.

Epstein, R.J. (1987). *A history of econometrics.* New York, Elsevier Science Publishing.

Farebrother, R. (1987). Algorithm AS 231: the distribution of a noncentral chi-squared variable with nonnegative degrees of freedom. *Applied Statistics* **36**: 402–405.

Feiblman, J.K. (1972). *Scientific method.* The Hague, Martinus Nijhoff.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist* **125**: 1–15.

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

Fisher, R.A. (1925). *Statistical methods for research workers,* 1st edition. Edinburgh, Oliver & Boyd.

Fisher, R.A. (1926). *The design of experiments,* 1st edition. Edinburgh, Oliver and Boyd.

Fisher, R.A. (1950). *Contributions to mathematical statistics*. New York, Wiley.

Fisher, R.A. (1959). *Smoking. The cancer controversy*. Edinburgh, Oliver & Boyd.

Fisher, R.A. (1970). *The design of experiments,* 8th edition. New York, Hafner.

Forrest, D.W. (1974). *Francis Galton: The life and work of a Victorian genius*. New York, Taplinger.

Galton, F. (1869). *Hereditary genius: an inquiry into its laws and consequences*. London, Macmillan.

Geiger, D., Verma, T. and Pearl, J. (1990). Identifying independence in Bayesian networks. *networks* **20**: 507–534.

Glymour, G., Scheines, R., Spirtes, R. and Kelly, K. (1987). *Discovering causal structure. Artificial intelligence, philosophy of science, and statistical modeling*. Orlando, FL, Academic Press.

Goldberger, A.S. (1972). Structural equation methods in the social sciences. *Econometrica* **40**: 979–1002.

Goldstein, H. (1995). *Multilevel statistical models*. London, Academic Press.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, M. (1998). *A user's guide to MLwiN*. Bath, Multilevel Models Project.

Good, P. (1993). *Permutation tests. A practical guide to resampling methods for testing hypotheses,* 1st edition. New York, Springer-Verlag.

Good, P. (1994). *Permutation tests. A practical guide to resampling methods for testing hypotheses,* 2nd edition. New York, Springer-Verlag.

Griliches, Z. (1974). Errors in variables and other unobservables. *Econometrica* **42**: 971–998.

Grime, J. P. (1979). *Plant strategies and vegetation processes*. New York, John Wiley and Sons.

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

Harvey, P.H. and Pagel, M.D. (1991). *The comparative method in evolutionary biology*. Oxford, Oxford University Press.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. London, Chapman & Hall.

Heise, D. (1975). *Causal analysis*. New York, Wiley.

Hoogland, J.J. and Boomstra, A. (1998). Robustness studies in covariance structure modeling. An overview and a meta-analysis. *Sociological Methods and Research* **26**: 239–367.

Hottelling, H. (1953). New light on the correlation coefficient and its transformations. *Journal of the Royal Statistical Society, Series B* **15**: 193–232.

Howson, C. and Urbach, P. (1989). *Scientific reasoning. The Bayesian approach*. LaSalle, IL, Open Court.

Hox, J.J. (1993). Factor analysis of multilevel data. Gauging the Muthén model. In *Advances in longitudinal and multivariate analysis in the behavioural sciences*, ed. Oud, J.H.L. and van Blokland-Vogelesang, R.A.W., pages 141–156. Nijmegen, ITS.

Jobson, J.D. (1992). *Applied multivariate data analysis.* Volume I: *Regression and experimental design*. New York, Springer-Verlag.

Jordano, P. (1995). Frugivore-mediated selection on fruit and seed size: birds and St. Lucie's cherry. *Ecology* **76**: 2627–2639.

Jöreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32**: 443–482.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34**: 183–202.

Jöreskog, K.G. (1970). A general method for analysis of covariance structures. *Biometrika* **57**: 239–251.

Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system. In *Structural equation models in the social sciences*, ed. Goldberger, A.S. and Duncan, O.D., pages 85–112. New York, Academic Press.

Keesling, J.W. (1972). *Maximum likelihood approaches to causal analysis*. Chicago, Department of Education, University of Chicago.

Kempthorpe, O. (1979). *The design and analysis of experiments*. Huntington, NY, Robert E. Krieger.

Kendall, M. and Gibbons, J.D. (1990). *Rank correlation methods*. New York, Oxford University Press.

Kendall, M.G. and Stuart, A. (1983). *The advanced theory of statistics*. London, Charles Griffin & Company.

Korn, E.L. (1984). The ranges of limiting values of some partial correlations under conditional independence. *The American Statistician* **38**: 61–62.

Lande, R. and Arnold, S.J. (1983). The measurement of selection on correlated characters. *Evolution* **37**: 1210–1226.

Li, C.C. (1975). *Path analysis – a primer*. Pacific Grove, CA, Boxwood Press.

Longford, N.T. (1993). *Random coefficient models*. Oxford, Clarendon Press.

Mach, E. (1883). *The science of mechanics: A critical and historical account of its development*, 5th edition, with revisions through the 9th German editon. LaSalle, IL, Open Court.

Manly, B.F.J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology,* second edition. London, Chapman & Hall.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**: 519–530.

Mardia, K.V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya* **B36**: 115–128.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate analysis*. London, Academic Press.

Mayo, D.G. (1996). *Error and the growth of experimental knowledge*. Chicago, Chicago University Press.

McDonald, R.P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods and Research* **22**: 399–413.

Meziane, D. (1998). Étude de la variation interspécifique de la vitesse spécifique de croissance et modélisation de l'effet des attributs morphologiques, physiologiques et d'allocation de biomasse. PhD thesis, Université de Sherbrooke.

Mulaik, S.A. (1986). Toward a synthesis of deterministic and probabilistic formulations of causal relations by the functional relation concept. *Philosophy of Science* **53**: 313–332.

Muthén, B. (1990). Mean and covariance structure analysis of hierarchical data. Unpublished paper presented at Psychometric Society Meeting, Princeton, NJ.

Muthén, B. (1994a). Latent variable modeling of longitudinal and multilevel data. Unpublished paper presented at American Sociological Association, Section on Methodology, Showcase Session, Los Angeles, American Sociological Association.

Muthén, B. (1994b). Multilevel covariance structure analysis. *Sociological Methods and Research* **22**: 376–398.

Muthén, B.O. and Satorra, A. (1995). Complex sample data in structural equation modeling. In *Sociological methodology*, ed. Marsden, P. V., pages 267–316. Washington, DC, American Sociological Association.

Niles, H.E. (1922). Correlation, causation and Wright's theory of 'path coefficients'. *Genetics* **7**: 258–273.

Norton, B.J. (1975). Biology and philosophy: the methodological foundations of biometry. *Journal of the History of Biology* **8**: 85–93.

Passmore, J. (1966). *A hundred years of philosophy*. Harmondsworth, Middx, Penguin.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Mateo, CA, Morgan Kaufmann.

Pearl, J. (1997). The new challenge: from a century of statistics to an age of causation. *Computing Science and Statistics* **29**: 415–423.

Pearl, J. (2000). *Causality*. Cambridge, Cambridge University Press.

Pearl, J. and Dechter, R. (1996). Identifying independencies in causal graphs with feedback. In *Proceedings of the 12th conference on uncertainty in artificial intelligence*, pages 240–246. San Francisco, Morgan Kaufmann.

Pearson, E.S. and Kendall, M.G. (1970). *Studies in the history of statistics and probability*. London, Griffin.

Pearson, K. (1892). *The grammar of science,* 1st edition. London, Adam & Charles Black.

Pearson, K. (1911). *The grammar of science,* 3rd edition. London, Adam & Charles Black.

Peters, R.H. (1991). *A critique for ecology*. Cambridge, Cambridge University Press.

Pollack, J.L. (1986). *Contemporary theories of knowledge*. Totowa, Rowman & Littlefield.

Popper, K. (1980). *The logic of scientific discovery*. London, Hutchinson.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986).

Numerical recipes. The art of scientific computing. Cambridge, Cambridge University Press.

Provine, W. B. (1986). *Sewall Wright and evolutionary biology*. Chicago, University of Chicago Press.

Pugesek, B.H. and Tomer, A. (1996). The Bumpus house sparrow data: a reanalysis using structural equation models. *Evolutionary Ecology* **10**: 387–404.

Rao, M.M. (1984). *Probability theory with applications*. Orlando, FL, Academic Press.

Rapport, S. and Wright, T. (1963). *Science: method and meaning*. New York, New York University Press.

Richardson, T. (1996a). A discovery algorithm for directed cyclic graphs. In *Proceedings of the 12th conference on uncertainty in artificial intelligence*, pages 454–461. San Francisco, Morgan Kaufmann.

Richardson, T. (1996b). Models of feedback: interpretation and discovery. Ph.D. thesis, Carnegie–Mellon University, Pittsburgh.

Rigdon, E.E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research* **30**: 359–383.

Satorra, A. and Bentler, P.M. (1988). Scaling corrections for chi-squared statistics in covariance structure analysis. In *Proceedings of the American Statistical Association*, pages 308–313. Alexandria, VA, American Statistics Association.

Scott, A.J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* **77**: 848–854.

Shipley, B. (1995). Structured interspecific determinants of specific leaf area in 34 species of herbaceous angiosperms. *Functional Ecology* **9**: 312–319.

Shipley, B. (1997). Exploratory path analysis with applications in ecology and evolution. *American Naturalist* **149**: 1113–1138.

Shipley, B. (1999). Exploring hypothesis space: examples from organismal biology. In *Computation, causation and discovery*, ed. Glymour, C. and Cooper, G.F., pages 441–452. Menlo Park, CA, MIT/AAAI Press.

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

Shipley, B. and Hunt, R. (1996). Regression smoothers for estimating parameters of growth analyses. *Annals of Botany* **76**: 569–576.

Shipley, B. and Lechowicz, M.J. (2000). The functional coordination of leaf morphology and gas exchange in 40 wetland plant species. *Ecoscience*, **7**(2): 183–194.

Shipley, B. and Peters, R.H. (1990). A test of the Tilman model of plant strategies: relative growth rate and biomass partitioning. *The American Naturalist* **136**: 139–153.

Shirahata, S. (1980). Rank tests of partial correlation. *Bulletin of Mathematical Statistics* **19**: 9–18.

Simon, H. (1977). *Models of discovery*. Dordrecht, D. Reidel.

Sokal, R.R. and Rohlf, F.J. (1981). *Biometry*. New York, Freeman.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology* **15**: 201–293.

Spirtes, P. (1995). Directed cyclic graphical representation of feedback models. In *Proceedings of the 11th conference on uncertainty in artificial intelligence*, pages 491–498. San Mateo, CA., Morgan Kaufmann.

Spirtes, P., Glymour, C. and Scheines, R. (1990). Causality from probability. In *Evolving knowledge in natural science and artificial intelligence*, ed. McGee, G., pages 181–199. London, Pitman.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, prediction, and search.* New York, Springer-Verlag.

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

StatSci (1995). *S-PLUS guide to statistical and mathematical analysis, version 3.3.* Seattle, StatSci, A Division of MathSoft, Inc.

Steiger, J.H. (1989). *EzPATH: A supplementary manual for SYSTAT and SYGRAPH.* Evanston, IL., SYSTAT Inc.

Steiger, J.H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research* **25**: 173–180.

Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural equation models. In *Testing structural equation models.* Bollen, K.A. and Long, J.S., pages 10–39. Newbury Park, CA, Sage.

Van Hulst, R. (1979). On the dynamics of vegetation: Markov chains as models of succession. *Vegetatio* **40**: 3–14.

Verma, T. and Pearl, J. (1988). Causal networks: semantics and expressiveness. In *Proceedings of the 4th workshop on uncertainty in artificial intelligence*, Mountain View, CA, pages 352–359. Reprinted in Schachter, R., Levitt, T.S. and Kanal, L.N. (eds.) *Uncertainty in artificial intelligence*, vol. 4, pages 69–76. Amsterdam, Elsevier.

Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the 6th workshop on uncertainty in artifical intelligence*, Cambridge, MA, pp. 220–277. Reprinted in Bonissone, P., Henrion, M., Kanal, L.N. and Lemmer, J.F. (eds.) *Uncertainty in artificial intelligence*, vol. 6, pages 255–268. Amsterdam, Elsevier.

Wahba, G. (1991). *Spline models for observational data*. Philadelphia, SIAM Press.

Wishart, J. (1928). Sampling errors in the theory of two factors. *British Journal of Psychology* **19**: 180–187.

Wright, S. (1918). On the nature of size factors. *Genetics* **3**: 367–374.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. *Proceedings of the National Academy of Sciences* **6**: 320–332.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **10**: 557–585.

Wright, S. (1925). *Corn and hog correlations*. Washington, DC, US Department of Agriculture.

Wright, S. (1984). Diverse uses of path analysis. In *Human population genetics*, ed. Chakravarti, A., pages 1–34. New York, Van Nostrand Reinhold.

Index