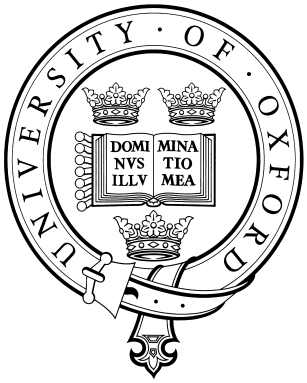
Causal Inference with Instruments and Other Supplementary Variables

Roland R. Ramsahai
St Catherine's College
University of Oxford

Michaelmas Term 2008



Thesis submitted in partial fulfilment of the requirements for the degree of DPhil in Statistics in the Department of Statistics at the University of Oxford.

Title: Causal Inference with Instruments and Other Supplementary Variables
Name: Roland R. Ramsahai
College: St Catherine's College
Degree: Doctor of Philosophy in Statistics
Date: Michaelmas Term 2008

Abstract

Instrumental variables have been used for a long time in the econometrics literature for the identification of the causal effect of one random variable, B, on another, C, in the presence of unobserved confounders. In the classical continuous linear model, the causal effect can be point identified by studying the regression of C on A and B on A, where A is the instrument. An instrument is an instance of a supplementary variable which is not of interest in itself but aids identification of causal effects. The method of instrumental variables is extended here to generalised linear models, for which only bounds on the causal effect can be computed. For the discrete instrumental variable model, bounds have been derived in the literature for the causal effect of B on C in terms of the joint distribution of (A, B, C). Using an approach based on convex polytopes, bounds are computed here in terms of the pairwise (A, B) and (A, C) distributions, in direct analogy to the classic use but without the linearity assumption. The bounding technique is also adapted to instrumental models with stronger and weaker assumptions. The computation produces constraints which can be used to invalidate the model. In the literature, constraints of this type are usually tested by checking whether the relative frequencies satisfy them. This is unsatisfactory from a statistical point of view as it ignores the sampling uncertainty of the data. Given the constraints for a model, a proper likelihood analysis is conducted to develop a significance test for the validity of the instrumental model and a bootstrap algorithm for computing confidence intervals for the causal effect. Applications are presented to illustrate the methods and the advantage of a rigorous statistical approach. The use of covariates and intermediate variables for improving the efficiency of causal estimators is also discussed.

Acknowledgements

This research was funded by a scholarship from the Government of Trinidad and Tobago.

Many thanks to Steffen Lauritzen for his insightful supervision and expert advice throughout and to Sir David Cox for detailed discussions on previous versions of this thesis. I am grateful to Phil Dawid, Thomas Richardson, Vanessa Didelez and Ib Skovgaard for valuable comments and interactions. I have also benefited from contact with Mathias Drton and Peter Clifford.

I must mention the numerous internet websites and online forums which I consulted for help with LaTeX syntax.

Contents

Chapter 1

Introduction

Causal inference from observational studies has long been recognised as a fundamental problem in Statistics. Much success has been achieved in the literature using various notions of causality. This thesis attempts to enhance the ammunition available to a researcher wanting to make inference about causal relations without being able to manipulate the allocation of treatments in a study. In particular, techniques for causal inference with supplementary variables are investigated. Two classical methods are the use of instrumental variables for identification and covariates for improving precision via analysis of covariance. These provide the basis for much of the discussion. An overview of this thesis is given in §1.1.

1.1 Overview

Chapter 1 presents an amateur touch on the philosophical analysis of the concept of causality while the rest of the chapter is devoted to formalising it within a statistical framework. The need to formalise causality within a

statistical framework is stressed and some notation is introduced.

Using the idea that causality is the study of interventions, various frameworks which have been used in the literature to discuss causality are given in Chapter 2. The concept of probabilistic causality is the preferred choice here and, together with influence diagrams, is used for the technical analyses of the thesis.

In Chapter 3, the use of continuous instrumental variables for causal inference is described. For the discrete case, Chapter 4 outlines results for causal bounds on the trivariate distribution in the literature. Significance tests for the instrumental variable model and confidence intervals for causal effects are discussed in Chapter 5. Chapter 6 is about the issues involved in the choice of supplementary variables to record for the improvement of the precision of causal estimators. The specific contributions of this thesis include

- The derivation of causal bounds in a continuous instrumental variable model where the classical linear model is relaxed to a generalised linear model. The link functions considered are concave, convex or logit.
- The illustration of a technique for computing constraints on the distributions represented by a model. It is applied to find observable constraints and causal bounds in a discrete instrumental variable model in terms of the pairwise marginal distributions.
- The bounding technique is also given some novel adjustments to facilitate sensitivity analysis. The derivation of non-trivial bounds without randomisation or exclusion restriction and with the monotonicity assumption is given.
- A significance test for the validity of the instrumental variable model

and its power function are derived.

- For the instrumental model, a method of calculating the maximum likelihood estimate of parameters using convex optimisation is derived. Together with the significance test, it is used for computing confidence intervals for the causal effect.
- The use of intermediate variables and covariates for improving the efficiency of a causal estimator is investigated for a multivariate Gaussian model.

Future possible extensions are discussed in Chapter 7.

1.2 Some thoughts on causality

Causality pervades many research areas. Researchers in Philosophy, Artifical Intelligence and Statistics, to name a few, have been both intrigued and humbled by the concept. Philosophers have struggled with the question, "*What is cause and effect?*" from the days of Aristotle till modern times (Russell, 1913; Holland, 1986). David Hume suggested that causal connections are the result of the observation of the constant conjunction of events (Hume, 1748). There are numerous inaccuracies with such a definition, as will be obvious from comments later on. Perhaps it is too ambitious to attempt to answer such a question or hope that any standard definition can ever be achieved. Despite the endless discussions on the nature of causality, some comments on the topic are given here. The ideas may serve as a baby step towards understanding the depth of the mystery surrounding the concept, as well as being an attempt to justify the role of Statistics as a vital tool towards

elucidation.

What is causality?

Such a question is extremely difficult. Just the fact that it has been debated over millennia or that it seems impossible to imagine a world without words or thoughts about 'cause and effect' demonstrates this. Insight has been gained by defining causality with ideas from the study of logic (Mill, 1843) and intelligence (Pearl, 2000). The latter fits more appropriately into this discussion. Intelligence (which is arguably just as elusive as causality) can be thought of, for the purposes here, as the ability to learn about Nature and use the knowledge gained to exploit her. Here learning about Nature means experiencing her mechanisms and replicating them as desired. An intermediate process would be assimilating all previous experience with any new one gained. If it is experienced that when the wind blows a tree branch shakes, unless previous experience dictates otherwise, a natural statement or thought would be

blowing of the wind causes the tree branch to shake.

It appears that the way the mind stores beliefs is by labelling cause-effect relationships. A relationship can only be branded as causal if it is believed that any observed relationship holds under intervention. One way of achieving this is by actually intervening and observing the outcome. Association can be treated as a term for incomplete causal information, where the knowledge about Nature that is stored is limited to no interventions. Basically, cause is the label given to an event to be conjured to produce a desired ef-

fect. Causal information is subjective so causal connections may or may not accurately portray existing mechanisms in Nature (Williamson, 2005).

Statistical causality

Apart from the cause-effect label given in the mind, a degree of uncertainty of the relationship is also assigned. This adheres to the belief that Nature cannot be told what to do but humbly nudged in the desired direction with the hope that her uncertain laws produce the desired outcome. Whether the uncertainty is a result of the ignorance of the complete physical and metaphysical mechanisms of Nature (Laplace, 1814) is of no consequence to this discussion.

The inherent uncertainty is not obvious in the example with the wind and the tree branch because little uncertainty is associated with it, it is almost deterministic. Man has evolved to automatically store the causal information gained from petty situations within plain sight. In everyday life, it is a trivial task for most circumstances encountered. It is extremely unlikely, based on the observation of the wind shaking a tree branch, to imagine that conjuring blowing wind would not shake the tree branch. The causal information is extracted instinctively. Throughout history, causal information has even been extracted from non-trivial scientific studies, sometimes without any concern about statistical causality (Aalen and Frigessi, 2007). In more complex situations, such as clinical trials, the task is far from simple and large datasets are just the start of the difficulties. Some issues which arise are

- Multiple causes, effects and interactions: there may be multiple events which are producing multiple effects simultaneously. A web of relation-

ships would obviously be difficult to assess mentally. There is also the potential for unobserved variables which distort the relationship.

- Indeterministic degree of cause-effect relationship: many relationships of interest are not black and white. Assuming no other variables affect the relationship, it would be difficult for the mind to compute a measure of the causal information provided by data.

It is with such situations that Statistics becomes handy, with its ability to handle uncertainy about multiple relationships in a formal framework. Statistics has had a great impact on investigating causal relations, notably clinical trials. This is despite the scepticism of some of the most influential founders of Statistics, such as Karl Pearson (Aalen and Frigessi, 2007). A fact that cannot be over emphasised is that there is still a lot to be done. Current practice accepts the causal conclusions from randomised experiments but, from observational studies they remain controversial. Allegations of invalid causal conclusions are not unjustified because of the inconvenient potential existence of unobserved confounders.

The causal thinking of the mind naturally leads to causal language in communication. It is because causality is so fundamental to our thinking that the word 'cause' has become irreversibly tangled in everyday language without any care having been taken in placing it there. To rid causal statements of the possibility of misinterpretation, causality must be formally redefined from everyday language, just as has been done with words such as 'consistency' and 'expectation'. Various notions of causality within Statistics are reviewed in Cox and Wermuth (2004) but for the purposes here it is thought of as the study of the effect of interventions or manipulations. This approach

is in line with the work of Robins (1986), Spirtes, Glymour and Scheines (1993), Pearl (1995a), Lauritzen (2001), Dawid (2002), Didelez, Dawid and Geneletti (2006) and others.

1.3 Intervention and observation

The observation of any association between events in a system is not necessarily related to the effect of external interventions. Such association may be the result of the common influence of a disturbance within the system. According to the Yule-Simpson paradox (Yule, 1903; Simpson, 1951), unobserved common influences make it possible for interventions to have an opposite effect to any observed effects.

It is necessary to make statements regarding the distribution of variables when certain variables are 'set' or forced to take some value by intervention. To represent interventions in a probabilistic framework, the notation $(\cdot \,||\, \cdot)$ (Lauritzen, 2001) is used for *intervention conditioning*, as opposed to the usual $(\cdot \,|\, \cdot)$ for *observation conditioning*. The '||' symbol is equivalent to the 'do(·)' notation, first used in Goldszmidt and Pearl (1992), and the '$\mathbb{P}_{man(\cdot)}$' notation of Spirtes, Glymour and Scheines (1993). For two random variables B and C, $\mathbb{P}(C \,||\, B = b)$ is the probability of C given that B is actively forced by an external mechanism to take the value b whereas $\mathbb{P}(C \,|\, B = b)$ is the probability of C given that it is passively observed that B takes the value b naturally. With observation conditioning it does not matter how the value of the variable conditioned on arises. Various other types of strategies for intervention on variables in a system are described in §2.1.2. In statistical

models it is often assumed that certain conditional independence relations hold regardless of the type of conditioning.

Randomised clinical trials have been very often used for causal inference. The most important aspect of such experiments is that the treatments are assigned to units randomly. Provided that the assignment mechanism is actually random, randomisation guarantees that the value set is not influenced by any variable in the study. Therefore it statistically mimics an external intervention so that causal effects can be observed.

1.4 Causal models

Cause and effect are merely constructs of the mind to store information about the blueprint of Nature based on experience and the level of detail of the blueprint information depends on the experience. Statements about causes only make sense when augmented with information about the subsystem to which they apply. This is because the level of refinement in the definition of a subsystem affects the causal relations that hold. In everyday language the scope of a causal statement is frequently omitted. Perhaps in the evolution of language for efficient communication brevity is of higher priority and the scope may be obvious or too complicated to specify. The impropriety of this practice can be seen from the fact that the modified statement,

blowing of the wind causes the tree branch to shake if the greenhouse door is open,

reflects a different causal relation than the previous one. A greenhouse is also included so the subsystem of Nature about which information is being

conveyed is different. In the same way causal conclusions are only relevant for a particular model, although there is a natural inclination by the mind to extrapolate knowledge when confronted with new circumstances. This must be emphasised for statistical models because no causal statement can truly be known to be universally applicable. The issue of whether a model, and hence its conclusions, are applicable to a practical situation is a separate one that can be justified by expert knowledge. A model can be thought of as defining what Nature consists of.

The qualitative component of a causal model specifies which factors are deemed to exist in Nature or be relevant and which of them influence one another. Consider the statements relating smoking habit, amount of tar in the lungs, a hypothetical 'smoking gene' and lung cancer,

(1) Smoking habit affects lung cancer.

(2) Lung cancer does not affect smoking habit.

(3) Smoking habit affects the amount of tar in the lungs.

(4) The amount of tar in the lungs does not affect smoking habit.

(5) The amount of tar in the lungs affects the occurrence of lung cancer.

(6) The occurrence of lung cancer does not affect the amount of tar in the lungs.

(7) Smoking habit does not affect the occurrence of the smoking gene.

(8) The occurrence of the smoking gene affects smoking habit.

(9) The occurrence of the smoking gene affects the occurrence of lung cancer.

(10) The occurrence of lung cancer does not affect the occurrence of the smoking gene.

Various models can be defined with various combinations of these statements. Consider three models,

- Model I: (1) - (2).
- Model II: (1), (2) and (7) - (10).
- Model III: (3) - (6).

The models can also be represented more easily by path diagrams (Wright, 1921), Fig.1.1, but the graphical representation of models is disliked by some (Rubin, 2004).

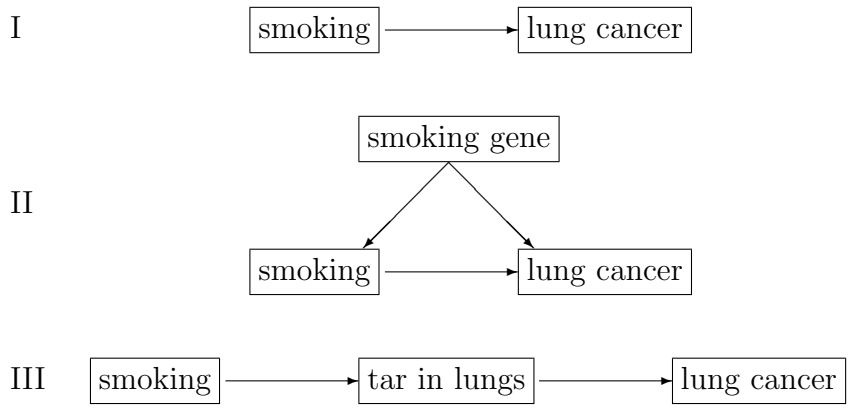


Figure 1.1: Path diagrams representing various models for the relationships between smoking habit, the amount of tar in the lungs, a hypothetical 'smoking gene' and lung cancer.

The definition of Nature in model I consists of only smoking habit and lung cancer. Model II and III additionally include the smoking gene and the amount of tar in the lungs respectively, so they can be considered as refinements of model I. Each of the arrows represents some underlying collection of mechanisms which results in the variable at the tail influencing the distribution of the variable at the head. When a variable is referred to as a cause,

it is the event of the variable changing which is the cause. Each of the arrows represents a sub-model, e.g. model I is a sub-model of model II.

1.5 Causal quantity of interest

It is necessary to develop some measure of the influence one variable has on another, the *causal effect*. Such a quantity is useful to compare the various tools for a task or the various treatments for a desired response. The context of a study can determine which measure of a causal effect is appropriate. It may be that inference is required for theoretical or practical purposes.

Consider the summary of data from a hypothetical study given in Fig.1.2, where the death rate is the proportion of people in the population who die from a particular disease. Analysis of data from population 1 shows that the ratio of deaths due to lung cancer for smokers to non-smokers is 6 whereas for coronary thrombosis it is only 1.5. However the difference in death rates between the two groups is much larger for coronary thrombosis. It is the differences in death rates which matter for public health concerns since concerns are about the number of lives that can be saved. Public health policy makers would define the causal quantity of interest as a difference between parameters of the distributions for smokers and non-smokers. For population 2 the contrast in differences for the diseases is not as pronounced.

Comparing population 2 with population 1, the death ratios are constant for each disease for both populations. Thus, from a scientific point of view, the ratios of the death rates are the relevant aspects of the data. They may be useful in forming theories about the biological mechanisms which smoking

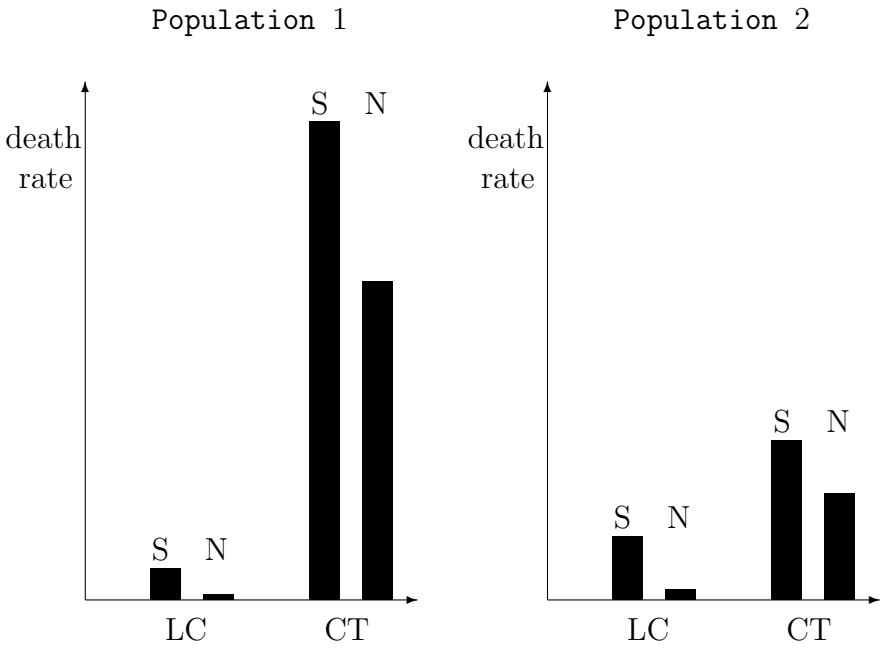


Figure 1.2: Data from a hypothetical study comparing death rates due to lung cancer (LC) and coronary thrombosis (CT) for smokers (S) and non-smokers (N).

affects. The scientific investigator would define the causal quantity of interest as a ratio. Berkson (1958) discusses such issues with regard to studies investigating smoking as a cause of lung cancer. Note also that the large ratio of deaths attributed to lung cancer can give rise to suspicion that the contrast is not merely due to some unobserved confounder.

Another factor which affects whether a ratio or difference is of interest is the form of a statistical model. A measure of a causal effect is useful if it summarises contrast in the distribution of the effect variable when the value of a cause variable is changed. Consider a model in which

$$g\{\mathbb{E}(C \,||\, B)\} = \beta B,$$

where $g(\cdot)$ is a continuous monotone function and (B, C) are continuous.

If $g(\cdot)$ is the identity function and a researcher is interested in the expected value of C when B is set to some value then the value of β, a difference, summarises the interesting component of the model. This is because $\mathbb{E}(C \,||\, B)$ can be obtained if β is known.

Let

$$\mathrm{CE}(B \to C) = \frac{\mathbb{E}(C \,||\, B + \Delta B) - \mathbb{E}(C \,||\, B)}{\Delta B}.$$

Since

$$\mathrm{CE}(B \to C) = \beta,$$

then $\mathrm{CE}(B \to C)$ is a useful measure of the causal effect of B on C. If $g(\cdot)$ is instead the $\log(\cdot)$ function,

$$\mathrm{CE}(B \to C) \neq \beta,$$

and no longer provides an appropriate definition of the causal effect. However β, which represents a ratio, will still be appropriate.

If instead it is the variance $\mathbb{V}(C \,||\, B)$ which is of interest then β will lose its value as a summary of the causal effect. It is only $\mathbb{P}(C \,||\, B)$ which stands its ground as being always important for quantifying a causal effect and any sufficient summary of it is model specific. The entire distribution of $\mathbb{P}(C \,||\, B)$ may sometimes be needed without any reduction in dimension of the measure being possible.

In this thesis, the *average causal effect* (ACE) of a random variable, B, on

another random variable, C, is defined as

$$\mathrm{ACE}(B \to C) = \frac{\mathbb{E}(C \,||\, B + \Delta B) - \mathbb{E}(C \,||\, B)}{\Delta B}, \tag{1.1}$$

if B is a discrete variable and

$$\mathrm{ACE}(B \to C) = \frac{\partial \mathbb{E}(C \,||\, B)}{\partial B}, \tag{1.2}$$

if B is a continuous variable.

The *individual causal effect* (ICE) of a random variable, B, on another random variable, C, in the presence of U, is defined as

$$\mathrm{ICE}(B \to C; U) = \frac{\mathbb{E}(C \,||\, B + \Delta B \,|\, U) - \mathbb{E}(C \,||\, B \,|\, U)}{\Delta B}, \tag{1.3}$$

if B is a discrete variable and

$$\mathrm{ICE}(B \to C; U) = \frac{\partial \mathbb{E}(C \,||\, B \,|\, U)}{\partial B}, \tag{1.4}$$

if B is a continuous variable, where U is the collection of unobserved confounders. This is not the usual definition of $\mathrm{ICE}(\cdot\,; U)$ and U may include other variables which define individual characteristics. For terms involving multiple types of conditioning, the convention adopted is that all operations are performed from right to left (Lauritzen, 2001). $\mathrm{ICE}(\cdot\,; U)$ is the causal effect for a fixed value of the unobserved confounders or for a particular unit which is identified by or classified according to U. In general, $\mathrm{ACE}(B \to C)$ and $\mathrm{ICE}(B \to C; U)$ may not be constant or equal to each other.

Chapter 2

Causal frameworks and influence diagrams

The frameworks can be broadly classified into probabilistic and deterministic causality. Probabilistic causality was popularised by Suppes (1970) and deterministic causality is based on the philosophy of causal determinism of Laplace (1814) (Gillies, 2000). Both approaches include the use of influence diagrams for causal modelling, which is described in detail in Dawid (2002). The models of §2.3 and §2.4 are also based on the concept of probabilistic causality but are not used subsequently.

2.1 Probabilistic causality

2.1.1 Probabilistic semantics of DAGs

Formal *directed acyclic graphs* (DAG) are used because of their natural causal interpretations, partly due to their resemblance to path diagrams (cf. §1.4). They have been successfully used for specifying statistical assumptions in a

model and are thoroughly discussed in Lauritzen (1996). A purely probabilistic DAG consists of a set of *vertices* or *nodes*, $\mathcal{N}$, and a set of *directed edges*, $\mathcal{E}$. If $\{\alpha_1, \alpha_2\} \in \mathcal{N}$ and $(\alpha_1, \alpha_2) \in \mathcal{E}$ then $(\alpha_2, \alpha_1) \notin \mathcal{E}$. It is said that there is a directed edge from α_1 to α_2 and this is written as $\alpha_1 \rightarrow \alpha_2$. The assumption of *stability* (Pearl, 2000) is made throughout. Stability requires that the conditional independence relations represented on the graph be identical to those exhibited by the probability distribution. In a DAG which represents the probability distribution of a set of random variables, $\mathcal{X}$, every $\alpha \in \mathcal{N}$ corresponds to a random variable $\mathcal{X}_\alpha \in \mathcal{X}$. The probability distribution function has the form

$$\mathbb{P}(\mathcal{X}) = \prod_{\alpha \in \mathcal{N}} \mathbb{P}\{\mathcal{X}_\alpha \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}\}, \tag{2.1}$$

where 'pa(α)' is the set of 'parents' of the node α. This factorisation property is equivalent to the collection of conditional independence relations

$$\mathcal{X}_\alpha \perp\!\!\!\perp \mathcal{X}_{\mathrm{nd}(\alpha)} \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)} \qquad \forall \alpha \in \mathcal{N}, \tag{2.2}$$

where 'nd(α)' is the set of 'non-descendents' of the node α. Other conditional independence properties are implied by Eq.(2.2), from the axioms of conditional independence (Dawid, 1979). However, all conditional independence relations can be derived directly from the DAGs using the concepts of 'd-separation' (Verma and Pearl, 1988) and a 'moral graph' (Lauritzen et al., 1990). Let $\mathrm{An}(A \cup B \cup S)$ be the smallest ancestral set containing the set of nodes $\{A \cup B \cup S\}$.

Definition 2.1. (d-separation) *A and B are d-separated by S if any path from A to B contains a vertex α such that either*

- *$\alpha \in S$ and arrows do not meet head to head on this path or*
- *α and its descendents are not in S and arrows meet head to head.*

Definition 2.2. (Moral graph) *The moral graph of a DAG is obtained by*

- *adding undirected edges between all pairs of vertices which have common children and are not already joined,*
- *then removing all arrowheads to form an undirected graph.*

All of the conditional independence properties of a graph can then be read directly using the d-separation or moralisation criterion.

Theorem 2.3. (d-separation criterion) *If A and B are d-separated by S in the original DAG then $A \perp\!\!\!\perp B \mid S$.*

Theorem 2.4. (Moralisation criterion) *If A and B are separated by S in the moral graph of $\mathrm{An}(A \cup B \cup S)$ then $A \perp\!\!\!\perp B \mid S$.*

Both criteria are equivalent according to the following theorem of Lauritzen (1996), which was proved in a more general setting by Richardson (2003).

Theorem 2.5. (Equivalence of separation criteria) *A and B are d-separated by S in the original DAG if and only if A and B are separated by S in the moral graph of $\mathrm{An}(A \cup B \cup S)$.*

Consider the example of a DAG in Fig.2.1, which represents the distribution of four random variables A, B, C and U.

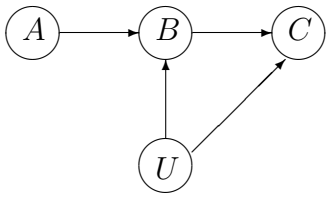


Figure 2.1: Example of a DAG representing a model involving four random variables A, B, C and U.

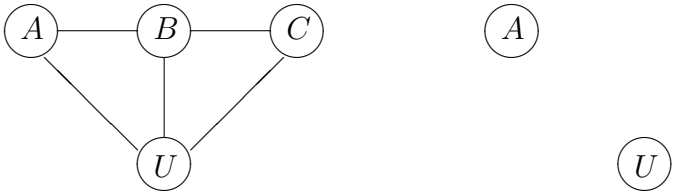


Figure 2.2: Moral graphs of the set of variables in $\mathrm{An}(A, C)$ (left) and $\mathrm{An}(A, U)$ (right) for DAG in Fig.2.1.

The moral graphs of various sets of variables for the DAG in Fig.2.1 is given in Fig.2.2. The relations $C \perp\!\!\!\perp A \,|\, (B, U)$ and $A \perp\!\!\!\perp U$ can be derived from Fig.2.1 and Fig.2.2 using either criterion, from Theorem 2.5.

2.1.2 Causal DAGs and augmented DAGs

The probabilistic semantics of a DAG for conditional independence are well defined but what causal assumptions are implied by its structure? The answer is none. The path diagrams of Fig.1.1 were seen to have a natural causal interpretation and they have no non-trivial visible difference to DAGs. To properly exploit this natural causal interpretation of a DAG, causal semantics must be developed. The probabilistic approach here treats causal inference as a decision problem but there are other approaches called 'probabilistic causality' which do not (Suppes, 1970). The quantity of interest is some measure of contrast in the probability distribution of some effect, Y, when a cause variable, X, is 'set' to different values. The consequence (loss) is

some function of Y and the decisions to choose from are the various values in the sample space of X. In the same way that the difference in the consequences of various decisions is of interest, the difference in distribution of Y for various set values of X is of interest. Decisions are used to model external interventions because their values are set by some mechanism external to the model. It is also in direct analogy to the modelling of the effect of decisions which, as mentioned before, is a main aim of causal analyses.

The deterministic causality approach makes the further assumption that there is a deterministic relationship between a decision and consequence but the particular relationship which applies is random. That approach will be discussed in §2.2.

In addition to modelling the joint distribution of a system of variables, it is now necessary to model the joint distribution when some variable in the system is 'set' by an external mechanism. Pearl (1995a) extends the semantics by which a DAG is to be interpreted to simultaneously model the joint distribution of variables under various scenarios or *regimes*. Various regimes are determined by the strategy which is used to set a variable in the system and the *observational regime* simply models the situation in which no intervention is performed. A strategy may involve previously observed variables, be random or simply be to set a variable to a fixed value (Didelez, Dawid and Geneletti, 2006). For an intervention in which the collection of variables corresponding to nodes in $\mathcal{W} \subseteq \mathcal{N}$, $\mathcal{X}_{\mathcal{W}}$, is set to a fixed value $\mathcal{X}_{\mathcal{W}'}$, it is assumed that the joint distribution conditional on the intervention is

$$\mathbb{P}(\mathcal{X} \,||\, \mathcal{X}_{\mathcal{W}} = \mathcal{X}_{\mathcal{W}'}) = \prod_{\alpha \in \mathcal{N} \setminus \mathcal{W}} \mathbb{P}\{\mathcal{X}_\alpha \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}\}\Big|_{\mathcal{X}_{\mathcal{W}} = \mathcal{X}_{\mathcal{W}'}}. \tag{2.3}$$

A DAG equipped with these modified semantics (assumption in Eq.(2.3)) is called a *causal DAG* or *intervention DAG* (Dawid, 2002; Lauritzen, 2001). In Fig.2.3, both DAGs represent the same conditional independence properties since they factorise as

$$\mathbb{P}(A, B) = \mathbb{P}(B \,|\, A)\mathbb{P}(A) = \mathbb{P}(A \,|\, B)\mathbb{P}(B).$$

The direction of the arrow makes no difference but this is not the case for DAGs in general.

Figure 2.3: DAG representing the joint distribution of (A, B).

However, if the DAGs in Fig.2.3 are interpreted as causal DAGs, the DAG on the left and right represent the assumptions

$$\mathbb{P}(A \,||\, B) = \mathbb{P}(A), \qquad \mathbb{P}(B \,||\, A) = \mathbb{P}(B),$$

respectively. It is said that the DAGs in Fig.2.3 are observationally equivalent but differ as causal DAGs.

Since the joint distribution under various regimes is being modelled, a *regime indicator* decision variable, F_α, is used to represent the various ways in which the value of $\mathcal{X}_\alpha$ arises. Examples of types of strategies are discussed in Didelez, Dawid and Geneletti (2006) and are

- *observational* ($F_\alpha =$ '*obs*'): do not intervene so that the value of $\mathcal{X}_\alpha$ arises naturally,

$$\mathbb{P}\{\mathcal{X}_\alpha \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}, F_\alpha = \text{'}obs\text{'}\}.$$

- *atomic intervention* ($F_\alpha = \alpha'$): force the value of $\mathcal{X}_\alpha$ to be fixed at some constant value α',

$$\begin{aligned}\mathbb{P}\{\mathcal{X}_\alpha \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}, F_\alpha = \alpha'\} &= \mathbb{P}(\mathcal{X}_\alpha \,|\, F_\alpha = \alpha') \\ &= \mathbb{P}(\mathcal{X}_\alpha \,||\, \mathcal{X}_\alpha = \mathcal{X}_{\alpha'}) \\ &= \begin{cases} 1 & \text{if } \mathcal{X}_\alpha = \mathcal{X}_{\alpha'} \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

- *conditional intervention* ($F_\alpha = d_{\{\mathrm{pa}(\alpha)\}}$): force the value of $\mathcal{X}_\alpha$ to arise from some distribution, $\tilde{\mathbb{P}}(\cdot)$, possibly depending on a subset of $\mathcal{X}_{\mathrm{pa}(\alpha)}$. The specified distribution may depend on the value of $\mathcal{X}_{\mathrm{pa}(\alpha)}$ in a random or deterministic way,

$$\begin{aligned}\mathbb{P}[\mathcal{X}_\alpha \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}, F_\alpha = d_{\{\mathrm{pa}(\alpha)\}}] &= \mathbb{P}[\mathcal{X}_\alpha \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}, F_\alpha = d_{\{\mathrm{pa}(\alpha)\}}] \\ &= \tilde{\mathbb{P}}\{\mathcal{X}_\alpha; \mathcal{X}_{\mathrm{pa}(\alpha)}\}\end{aligned}$$

Strategies describe how the value of the variables corresponding to a single node arises. If it is necessary to specify how the variables corresponding to a collection of nodes arise, strategies are simply specified for each node in the collection. Conditional interventions are particularly useful in defining *direct* and *indirect effects* (Pearl, 2001; Didelez, Dawid and Geneletti, 2006). A direct effect is a measure of the portion of a causal effect which is not conveyed by mediating variables and an indirect effect is the remainder of the causal effect. Such concepts would however depend on the definition of a causal effect, as discussed in §1.5.

Therefore the probability distribution of $\mathcal{X}$ varies according to the regime so

Eq.(2.1) must be altered to

$$\mathbb{P}(\mathcal{X} \mid F_{\mathcal{X}}) = \prod_{\alpha \in \mathcal{N}} \mathbb{P}\{\mathcal{X}_\alpha \mid \mathcal{X}_{\mathrm{pa}(\alpha)}, F_\alpha\}, \tag{2.4}$$

which is a more general form of Eq.(2.3). It may be sensible to restrict $F_\alpha =$ '*obs*' if $\mathcal{X}_\alpha$ cannot be realistically intervened on. Comparing Eqs.(2.1) and (2.4), it is obvious that $\mathbb{P}(\mathcal{X} \mid F_{\mathcal{X}})$ factorises according to a probabilistic DAG in which each node $\alpha \in \mathcal{N}$ has the set of parents $\{\mathrm{pa}(\alpha), F_\alpha\}$. Therefore the joint distribution over various scenarios can be represented by an influence diagram which is formed by augmenting each node $\alpha \in \mathcal{N}$ in a probabilistic DAG with an extra square parent node F_α, e.g. Fig.2.4. *Augmented DAGs* are described in Pearl (1993), Dawid (2002) and Lauritzen (2001). The intervention nodes, F_α, are decision variables and do not have a marginal distribution so only probabilistic expressions which involve conditioning on a given value of F_α, i.e. marginals of $\mathbb{P}(\mathcal{X} \mid F_{\mathcal{X}})$, are meaningful. For ease of notation, if $F_\alpha =$ '*obs*' then no value of the regime indicator is specified and the intervention node is omitted from the augmented DAG.

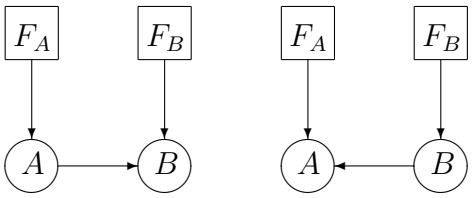


Figure 2.4: Augmented DAGs representing the joint distribution of (A, B).

Since the factorisation property of Eq.(2.4) holds for an augmented DAG and the intervention nodes are source nodes, i.e. $\mathrm{pa}(F_\alpha) = \{\emptyset\}$ or $\mathrm{An}(F_\alpha) = \{\emptyset\}$ $\forall \alpha \in \mathcal{N}$, analogously to purely probabilistic DAGs, all conditional indepen-

dence relations can be derived directly from the augmented DAG using the moralisation or d-separation criterion. It is possible to derive conditional independence relations involving F_α but, as stated before, the only meaningful probabilistic expressions involve conditioning on a given value of F_α. The augmented DAG on the left in Fig.2.4 represents the assumption $A \perp\!\!\!\perp F_B$. Therefore

$$\mathbb{P}(A \,||\, B) = \mathbb{P}(A \,|\, F_B = B) = \mathbb{P}(A),$$

which is the same assumption represented in the causal DAG on the left in Fig.2.3. An augmented DAG can be used to represent causal assumptions by interpreting it according to the probabilistic semantics of a DAG to avoid extending the semantics of a probabilistic DAG. The only conditional independence relations not represented on the graph are

$$\mathcal{X}_\alpha \perp\!\!\!\perp \mathrm{pa}(\mathcal{X}_\alpha) \,|\, F_\alpha \quad \forall \alpha \in \mathcal{N}, \tag{2.5}$$

when the strategy indicated by F_α does not involve $\mathrm{pa}(\mathcal{X}_\alpha)$, e.g. some constant α'. In Fig.2.4, for the DAG on the left, using the semantics of an augmented DAG it can also be shown that

$$\begin{aligned}\mathbb{P}(B \,||\, A = A') &= \mathbb{P}(B \,|\, F_A = A') \\ &= \mathbb{P}(B \,|\, A = A', F_A = A') \\ &= \mathbb{P}(B \,|\, A = A'),\end{aligned}$$

since $B \perp\!\!\!\perp A \,|\, F_A = A'$ and $B \perp\!\!\!\perp F_A \,|\, A$. Therefore it is possible to derive expressions linking the joint distribution under various regimes. This is very

important for causal inference, where it is often necessary to compute intervention distributions when observational data alone is available.

2.2 Deterministic causality

2.2.1 Potential outcomes

The framework of *potential outcomes* (PO) is an attempt by Rubin (1974, 1978) to develop a causal framework, by extending the work of Neyman (1923) on experiments. One main idea is that the same notation used for randomised experiments can be adapted for use in observational studies. In order for this notation to be adapted for use in observational studies, the assignment mechanism needs to be explicitly modelled separately because assignments are no longer randomised. Thus the PO framework can be more appropriately thought of as "potential outcomes with assignment mechanism" (Rubin, 1990). The framework of Neyman (1923) is described and its extension to facilitate causal inference in observational studies is discussed.

Neyman's work on experiments

Consider the case where the causal effect of the variety of a crop, B, on the yield, C, is of interest. Neyman (1923) describes the design of a field experiment in which a field is split into n plots, each of which can potentially be exposed to one of m varieties. The random variable $C_{(b)i}$, for $b = 1, \ldots, m$ and $i = 1, \ldots, n$, is the "potential yield" or in other words, the potential value of C if plot i was exposed to variety b. For each member of a repeated

fixed sequence or ordering of the varieties, plots are randomly sampled without replacement. This assignment is stochastically identical to the concept of randomisation of Fisher (1925) but conceptually different (Rubin, 1990). The vector of potential outcomes for a plot i, is denoted by the random variable C_i^p and the entire array of potential outcomes for the n plots is denoted by $\mathbf{C^P}$, Fig.2.5.

B

	1	2		m
$C_1^p\{$	$C_{(1)1}$	$C_{(2)1}$		$C_{(m)1}$
$C_2^p\{$	$C_{(1)2}$	$C_{(2)2}$		$C_{(m)2}$
$C_n^p\{$	$C_{(1)n}$	$C_{(2)n}$		$C_{(m)n}$

Figure 2.5: Potential yields, $\mathbf{C^P}$, in the field experiment of Neyman (1923).

The plots are units in an experiment to which a variety (treatment) is randomly assigned, so it is only ever possible to observe the value of $C_{(b)i}$ for one value of b for each plot i. In terms of the notation introduced in §1.3, $\mathbb{P}\{C_{(b)}\} = \mathbb{P}(C \,||\, B = b)$. Within this framework the ICE of B on C is usually defined for a unit i (not conditioning on fixed characteristics U, cf. §1.5) as

$$\begin{aligned} \mathrm{ICE}(B \to C; C_i^p) &= \mathbb{E}\{C_{(b_2)i} - C_{(b_1)i} \,|\, i, \mathbf{C^P}\} \\ &= \mathbb{E}\{C_{(b_2)i} - C_{(b_1)i} \,|\, C_i^p, \mathbf{C^P}\} \\ &= C_{(b_2)i} - C_{(b_1)i}, \end{aligned}$$

since $C_{(b)}$ is a deterministic function of $\mathbf{C^p}$. The notation $\mathbb{E}(\cdot \mid \mathbf{C^p})$ is used to emphasize that the array of potential outcomes for the sample of units is fixed, although unknown (Rubin, 1990). Since the entire vector C_i^p can never be observed, $\text{ICE}(B \to C; C_i^p)$ is never observable for any unit. In this framework the average causal effect of B on C (cf. §1.5) is defined as

$$\begin{aligned}\text{ACE}(B \to C) &= \mathbb{E}_{C_i^p}\{\text{ICE}(B \to C; C_i^p) \mid \mathbf{C^p}\} \\ &= \mathbb{E}_{C_i^p}\{C_{(b_2)i} - C_{(b_1)i} \mid \mathbf{C^p}\} \\ &= \mathbb{E}_{C_i^p}\{C_{(b_2)i} \mid \mathbf{C^p}\} - \mathbb{E}\{C_{(b_1)i} \mid \mathbf{C^p}\},\end{aligned}$$

where the expectations are over i or C_i^p. This definition is convenient to the PO framework since it does not depend on the joint distribution of $\{C_{(b_1)i}, C_{(b_2)i}\}$ which, by definition, cannot be empirically verified. Let $\mathbf{B}$ be the vector of assignments of the random sample of n units. For a given set of data, $\mathbf{B}$ is fixed and the estimator

$$\bar{C}_{(b_2)} - \bar{C}_{(b_1)}, \tag{2.6}$$

where $\bar{C}_{(b)}$ is the average yield over all units actually assigned treatment b, is an unbiased estimator of $\text{ACE}(B \to C)$ over repeated randomised assignments $\mathbf{B}$ (Neyman, 1923), since

$$\mathbb{E}_{\mathbf{B}}\{\bar{C}_{(b_2)} - \bar{C}_{(b_1)} \mid \mathbf{C^p}\} = \mathbb{E}_i\{C_{(b_2)i} - C_{(b_1)i} \mid \mathbf{C^p}\}.$$

Therefore the estimator in Eq.(2.6) can be used to estimate $\text{ACE}(B \to C)$ but, as discussed in §1.5, there seems to be no theoretical reason why the

causal effect should always be a difference.

Rubin's extension to observational studies

Rubin (1974, 1978) attempts to extend the framework of Neyman for causal inference in observational studies. However, the estimator in Eq.(2.6) can be biased for observational studies since treatment assignment is not necessarily randomised,

$$\mathbb{P}(\mathbf{B} \mid \mathbf{C^p}) \neq \mathbb{P}(\mathbf{B}).$$

The definitions of ACE($B \to C$) and ICE($B \to C; C_i^p$) remain the same since the $\mathbf{C^p}$ structure is retained, but the estimator is different. The proposed estimator of ACE($B \to C$) is the average of ICE($B \to C; C_i^p$) in the sample. As noted before, ICE($B \to C; C_i^p$) cannot simply be observed because only the PO for one treatment can be observed for any unit.

Rubin formulates the problem as a missing data problem in which the PO for the treatment actually received is observed and the unobserved elements of $\mathbf{C^p}$ need to be estimated. This requires explicit modelling of the assignment mechanism, $\mathbb{P}(\mathbf{B} \mid \mathbf{C^p})$. One approach is to use the EM-algorithm (Dempster, Laird and Rubin, 1977) to estimate with missing data.

Certain assumptions can be made to facilitate identification of the causal effect. For example, assuming that there exists a set of units which have the same C_i^p but are actually assigned different values of B would allow ICE($B \to C \mid C_i^p$) to be estimated. A major problem though is that there is no way of empirically verifying assumptions about the properties of potential outcomes.

2.2.2 Functional DAGs

In this section a class of influence diagrams called *functional DAGs* (Dawid, 2002) is described. It has been proposed by Heckerman and Shachter (1995) for causal modelling. The framework is based on the decision theoretic framework of Savage (1954), in which the primitives are *act*, *consequence* and *possible state of the world* and can be used to create a graphical representation of a PO model.

Acts are determined by decisions and the consequences are anything that happens as a result of such decisions. The acts are deterministically mapped to consequences but the mapping depends on the possible state of the world, which is random and unknown. The treatment assignment is an act and the value of the outcome variable is a consequence.

Functional DAGs are extensions of probabilistic DAGs (cf. §2.1.2). Squares are used to represent a decision or act, in the same way as an augmented DAG (cf. §2.1.2), and double and single circles correspond to deterministic and random nodes respectively. In the framework, the only random nodes are the possible states of the world, *mapping variables*. Assuming (A, B, C) are discrete, consider the observably equivalent causal DAG and functional DAG in Fig.2.6.

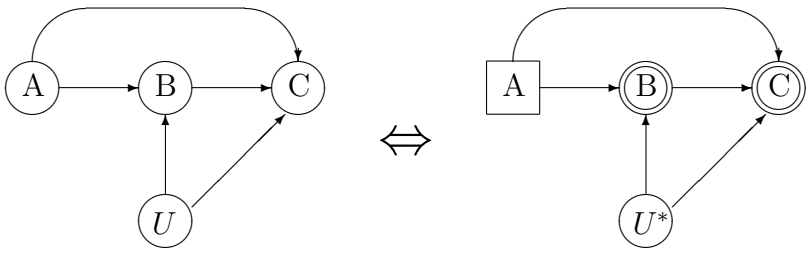


Figure 2.6: Example of a causal DAG and its observably equivalent functional DAG.

In Fig.2.6, the unobserved random variable U is replaced by U^*, which represents the mapping from A to (C, B). The node A is a decision node since it is set by an act and not influenced by any other variables in the model. The consequence (C, B) is a deterministic function of U^* and A, i.e.

$$(C, B) = g(A, U^*).$$

The specification of the relationship between (A, U) and (C, B), is replaced by a deterministic function,

$$\mathbb{P}(C, B \,||\, A, U^*) = \begin{cases} 1 & (C, B) = g(A, U^*) \\ 0 & \text{otherwise,} \end{cases}$$

and a marginal distribution is specified for U^*. From the conditional independence relations in Fig.2.6, it can be shown that $\mathbb{P}(C, B \,||\, A, U) = \mathbb{P}(C, B \mid A, U)$, which implies that the observational distribution is also specified by the deterministic function. In a certain sense, the randomness of $\mathbb{P}(C, B \mid A, U)$ has been 'collapsed' over U to form U^*, since U^* is the only random variable in the model. From the usual rules of probability,

$$\begin{aligned} \mathbb{P}(C, B \mid A) &= \textstyle\sum_u \mathbb{P}(C, B \mid A, U)\mathbb{P}(U) && \text{[probabilistic model]} \\ &= \textstyle\sum_{u^*} \mathbb{P}(C, B \,||\, A, U^*)\mathbb{P}(U^*) && \text{[PO model],} \end{aligned} \tag{2.7}$$

which implies that

$$\mathbb{P}(C, B \mid A) = \sum_{\{u^* : g(a, u^*) = (c, b)\}} \mathbb{P}(U^*). \tag{2.8}$$

Therefore the distribution of U^* can be chosen such that $\mathbb{P}(C, B \mid A)$ is equivalent to any probability distribution which is specified over the observable variables, $\mathbb{P}(C, B \mid A)$, provided that U^* has sufficiently many discrete states. At most $(|C| \times |B|)^{|A|}$ states are needed, where $|V|$ is the size of the state space of a discrete variable V (Heckerman and Shachter, 1995). Therefore, if U^* is defined as a discrete variable such that $|U^*| = (|C| \times |B|)^{|A|}$, an equivalent but not unique model can be found for any distribution $\mathbb{P}(C, B \mid A)$.
Consider the functional DAG of Fig.2.7, which is equivalent to the model in Fig.2.1. Since the probability distribution represented by functional DAGs factorise according to Eq.(2.1), the absence of the edge $A \to C$ in Fig.2.7 makes it possible to partition

$$U^* = (\sigma, \eta),$$

where σ and η are the mapping variables which represent the mapping from A to B and B to C respectively.

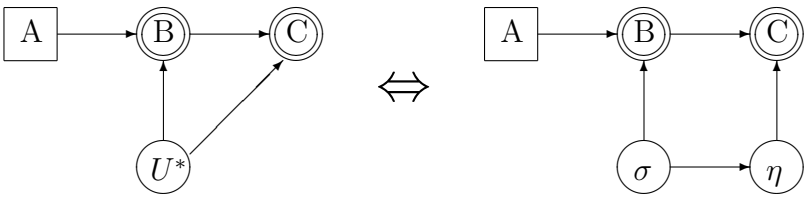


Figure 2.7: Functional DAG formed by removing $A \to C$ edge from the diagrams in Fig.2.6.

The sizes of the state spaces of the mapping variables are at most

$$|\sigma| = |B|^{|A|}, \quad |\eta| = |C|^{|B|}, \tag{2.9}$$

and the conditional probabilities are replaced by deterministic functions

$$\left\{ \begin{array}{c} \mathbb{P}(C \,|\, B, U) \\ \mathbb{P}(B \,|\, A, U) \end{array} \right\} = \left\{ \begin{array}{c} \mathbb{P}(C \,||\, B, U) \\ \mathbb{P}(B \,||\, A, U) \end{array} \right\} \\ \Downarrow \\ \begin{aligned} C &= g_1(B, U) = g_1(B, \eta) \\ B &= g_2(A, U) = g_2(A, \sigma). \end{aligned} \tag{2.10}$$

The concept of a mapping variable is identical to that of the PO for a unit since the vector of potential outcomes for a unit i, C_i^p, deterministically maps the value of a treatment variable to the value of the response.

In the terminology of §2.1.2, the regime indicators for the observable variables are set such that the factors in Eq.(2.4) are

$$\mathbb{P}[\mathcal{X}_\alpha \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}, F_\alpha = d_{\{\mathrm{pa}(\alpha)\}}] = \tilde{\mathbb{P}}\{\mathcal{X}_\alpha \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}\} = g_\alpha\{\mathrm{pa}(\alpha)\}, \tag{2.11}$$

where $\tilde{\mathbb{P}}(\cdot)$ is the modified probability measure under the intervention and $g_\alpha(\cdot)$ are deterministic functions, e.g. Eq.(2.10). The strategy for the observable variables is a conditional intervention, as described in §2.1.2. The strategies for unobservable variables are conditional interventions, set so that the joint distribution of $\mathcal{X}$ marginalises to a specific distribution over the observable variables, e.g. Eqs.(2.7) and (2.8). From Eq.(2.11), since $\mathcal{X}_\alpha \perp\!\!\!\perp F_{\mathrm{pa}(\alpha)} \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}$

$$\tilde{\mathbb{P}}\{\mathcal{X}_\alpha \,|\, \mathcal{X}_{\mathrm{pa}(\alpha)}\} = \tilde{\mathbb{P}}\{\mathcal{X}_\alpha \,||\, \mathcal{X}_{\mathrm{pa}(\alpha)}\} = g_\alpha\{\mathrm{pa}(\alpha)\}. \tag{2.12}$$

2.2.3 Structural equations

The functional models described in §2.2.1 and §2.2.2 are technically equivalent to *structural equation models* (SEM) (Goldberger, 1972; Bollen, 1989; Pearl, 2000). Thus they are all subject to the same weakness that they may not necessarily model the true underlying deterministic mechanisms of nature accurately. SEMs have played a vital role in statistical inference in Economics over the past few decades. Such models can be traced to the work of Wright (1921, 1934) and Haavelmo (1943). They rely on the specification of a system of equations which represent the relationship between variables in a system. Certain parameters express the stochastic element of the relationship. In a SEM, the equations are the same as the deterministic functions of Eq.(2.12) and the SEM is observably equivalent to its probabilistic counterpart. Consider the SEMs relating the variables A, B and C

$$\begin{aligned} C &= g_1(B, \epsilon_1) \\ B &= g_2(A, \epsilon_2), \end{aligned} \tag{2.13}$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are deterministic functions and the ϵ's are the unobservable "unexplained rest" (Haavelmo, 1943), not explained by the theory. They correspond to the unobservable variables of the probabilistic model. This can be seen easily since if $U := (\epsilon_1, \epsilon_2)$ then the models in Eqs.(2.13) and (2.10) are identical. Strotz and Wold (1960) clarify the meaning of structural equations as representing the effect of a manipulation of the causally dependent variables, and thus their equivalence to functional models. Similarly, Goldberger (1972) defines SEMs as, "stochastic models in which each

equation represents a causal link, rather than a mere empirical association."

2.3 Chain event graphs

Another type of diagram used for causal modelling is the chain event graph (CEG) of Smith and Anderson (2008). It is a generalisation of the probabilistic DAGs described in §2.1.1. CEGs explicitly model the sample space of variables in a model and are based on the event tree of Shafer (1996). An event tree represents how a process might unfold. Consider a model involving two binary variables A and B, each with sample space $\{0, 1\}$. Fig.2.8 is an example of an event tree.

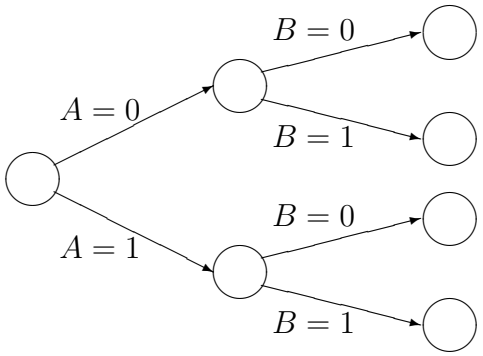


Figure 2.8: Event tree which represents a model of the joint distribution of the binary random variables A and B.

Events are represented by the paths on the tree from the root to a terminal vertex and the vertices correspond to *situations*. The framework is more expressive than a probabilistic DAG since asymmetric assumptions such as $\mathbb{P}(B = 0 \,|\, A = 1) = 1$ can be represented directly on the graph by removing the edge corresponding to $B = 1$ which emanates from the vertex corresponding to the situation in which A takes the value 1. On the other hand,

event trees may represent redundant relationships and are then unnecessarily complex. A CEG is a function of the event tree which collapses vertices with equivalent future unfoldings of the process. This requires all paths to have the same terminal vertex. For the model in Fig.2.8, if $A \perp\!\!\!\perp B$ or

$$\mathbb{P}(B = 1 \,|\, A = 1) = \mathbb{P}(B = 1 \,|\, A = 0)$$
$$\mathbb{P}(B = 0 \,|\, A = 1) = \mathbb{P}(B = 0 \,|\, A = 0),$$

then the CEG is given in Fig.2.9.

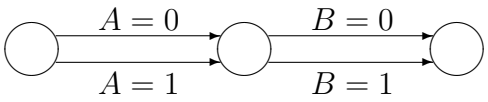


Figure 2.9: CEG corresponding to the model of Fig.2.8 with the additional assumption $A \perp\!\!\!\perp B$.

The CEG for event trees with certain symmetric properties are equivalent to probabilistic DAGs. Conditional independence relations can also be encoded in a CEG (Smith and Anderson, 2008).

From the discussion in §1.2, a cause can be intuitively defined as an event. Therefore, since CEGs model events, a natural application of such models is in causal modelling. Riccomagno and Smith (2007) gives full details on the extension of CEGs to causal CEGs. In a causal CEG, interventions force all paths to pass through certain situations or vertices.

2.4 Dynamic causality and mixed graphs

Thus far, only models of static causality have been considered. This section describes graphical models for the dynamic relationships among vari-

ables in a multivariate time series (Eichler, 2007). The diagrams use the concept of Granger causality (Granger, 1969) to define relationships. Let $\mathcal{X} = \{\mathcal{X}(t), t \in \mathbb{Z}\}$ be a weakly stationary multivariate time series.

Definition 2.6. (Granger-noncausality) *Let A and B be disjoint subsets of $S \subseteq \mathcal{X}$ and $I_{\mathcal{X}_S} = \{I_{\mathcal{X}_S}(t), t \in \mathbb{Z}\}$ be the sequence of closed subspaces generated by the subsets $\bar{\mathcal{X}}_S = \{\mathcal{X}_S(q), q \leq t\}$. Then the process $\mathcal{X}_A$ is* ***Granger-noncausal*** *for $\mathcal{X}_B$ with respect to the information set $I_{\mathcal{X}_S}$ (denoted $\mathcal{X}_A \not\to \mathcal{X}_B \ [I_{\mathcal{X}_S}]$) if*

$$\mathcal{X}_B(t+1) \perp\!\!\!\perp \bar{\mathcal{X}}_A(t) \,|\, \bar{\mathcal{X}}_{S\setminus A}(t) \quad \forall t \in \mathbb{Z}.$$

The processes $\mathcal{X}_A$ and $\mathcal{X}_B$ are ***contemporaneously independent*** *with respect to the information set $I_{\mathcal{X}_S}$ (denoted $\mathcal{X}_A \not\sim \mathcal{X}_B \ [I_{\mathcal{X}_S}]$) if*

$$\mathcal{X}_B(t+1) \perp\!\!\!\perp \mathcal{X}_A(t+1) \,|\, \bar{\mathcal{X}}_S(t) \quad \forall t \in \mathbb{Z}.$$

The path diagram associated with the time series $\mathcal{X}$ consists of a set of edges, $\mathcal{E}$, such that

- $A \to B \Leftrightarrow \mathcal{X}_A \not\to \mathcal{X}_B \ [I_{\mathcal{X}}]$),
- $A \text{---} B \Leftrightarrow \mathcal{X}_A \not\sim \mathcal{X}_B \ [I_{\mathcal{X}}]$).

Intuitively, the directed edges express movement in time but the dashed edges correspond to relationships within the same point in time. Similarly to DAGs, the path diagrams can be used to represent conditional independence relations. A vertex α is a **collider** on a path if the edges preceding and

succeeding α both have an arrowhead or a dashed tail at α (e.g. $\rightarrow \alpha \leftarrow$, $\text{---} \alpha \leftarrow$); otherwise it is a **non-collider** on the path.

Definition 2.7. (m-separation) *A path between A and B is said to be* ***m-connected*** *given a set S if*

- *every non-collider on the path is not in S, and*
- *every collider on the path is in S,*

otherwise it is ***m-blocked*** *given S. If all paths between A and B are m-blocked given S then A and B are* ***m-separated*** *given S.*

Theorem 2.8. (m-separation criterion) *If A and B are m-separated given S then $A \perp\!\!\!\perp B \mid S$.*

Definition 2.7 and Theorem 2.8 can be compared to Definition 2.1 and Theorem 2.3 for DAGs. The concept of m-separation is described here from Eichler (2007) but is equivalent to Richardson (2003), where '$\leftrightarrow$' edges are used instead of '---' edges. Richardson (2003) also describes an alternative criterion, for querying conditional independence relationships in mixed graphs, which is analagous to the moralisation criterion described in Definition 2.2 and Theorem 2.4. An example of a mixed graph is given in Fig.2.10.

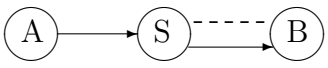


Figure 2.10: Example of a mixed graph which represents a model for the joint distribution of the random variables A, B and S.

In the mixed graph of Fig.2.10, A and B are not m-separated by any set. On the paths

- $A \rightarrow S \dashrightarrow B$: S is a collider,
- $A \rightarrow S \rightarrow B$: S is a non-collider.

The latter is m-blocked by S but the former is not. It should be noted that although Granger causality expresses an asymmetric dependence relationship between variables, it does not assume anything about the result of interventions. Therefore, thus far, there has been no mention of a causal (in the sense considered here) interpretation of the path diagrams.

Eichler and Didelez (2007) gives a causal interpretation of the graphs by specifying conditional independence relations involving $\mathcal{X}$ and regime indicators (cf. §2.1.2). They develop analogues of the 'back-door' and 'front-door' criteria (Pearl, 1993; Pearl, 1995a) for the identifiability of interventions. Continuous time versions of the path diagrams are described in Didelez (2006) and Didelez (2008) but it is emphasised that great care is needed when imposing any causal interpretation.

Chapter 3

Continuous instrumental variables

In §2.1.2, the idea of finding expressions for causal parameters in terms of the observational distribution was introduced. This is the fundamental idea which enables causal inference in observational studies. A major setback is that, in the presence of unobserved variables, such expressions are not always obtainable. The concept of an *instrumental variable* (IV) has been very useful in making causal inference in the presence of unobserved confounders, particularly in the econometrics literature (Durbin, 1954; Bowden and Turkington, 1984). A formal definition of an IV is given later but the intuition behind their use is that causal inference is possible by observing additional unconfounded variables, the instruments. The use of IVs can be traced back to at least Wald (1940), Reiersøl (1941, 1945) and Geary (1942, 1943).
Lauritzen (2003, 2004) describe the use of IVs for surrogates. In genetic epidemiology, the 'Mendelian instrument' represents a genotype and inference is required about the causal effect of a phenotype on a disease (Didelez and Sheehan, 2007). In studies involving partial compliance, IVs are also important (Angrist, Imbens and Rubin, 1996; Balke and Pearl, 1997). Although

IVs often provide a point estimator of a causal effect they don't always trivialise causal inference. Sometimes only bounds on the causal effect are attainable. Their classical use involves a linear model and is described in §3.1. The rest of the chapter proposes extensions to a generalised linear model with non-linear link functions. Details of models with discrete variables are deferred to Chapter 4.

3.1 Classical model for continuous instrumental variables

A special type of SEM (cf. §2.2.3) is a *linear SEM* in which the functions are linear. It is within these types of models that instrumental variables (Bowden and Turkington, 1984) have traditionally been used. In the classical case, (A, B, C) are continuous and the linear SEM is

$$C = \beta_{c|b}B + U, \tag{3.1}$$

where U is an unobservable random variable and, without loss of generality, $\mathbb{E}(U) = 0$, since expectations can be replaced by covariances.
If $\text{cov}(B, U) = 0$ then, by the law of large numbers, a consistent and unbiased estimator of $\beta_{c|b}$ is the usual regression estimator

$$\hat{\beta}_{c|b} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{C} = \frac{\sum B_i C_i}{\sum B_i^2},$$

where

$$\mathbf{C}^T = (C_1, \ldots, C_n),$$

is a vector of n i.i.d. random samples of C and similarly for B.

When $\text{cov}(B, U) \neq 0$, $\hat{\beta}_{c|b}$ is not unbiased and consistent. If there exists another random variable A then premultiplying throughout results in the expression

$$\mathbf{A}^T\mathbf{C} = \beta_{c|b}\mathbf{A}^T\mathbf{B} + \mathbf{A}^T\mathbf{U}.$$

If $\text{cov}(A, U) = 0$ then

$$\mathbb{E}(\mathbf{A}^T\mathbf{C}) = \beta_{c|b}\mathbb{E}(\mathbf{A}^T\mathbf{B}).$$

Assuming $\text{cov}(A, B) \neq 0$, as $n \to \infty$, by the law of large numbers, a consistent but biased estimator of β is

$$\hat{\beta}^{\,iv} = (\mathbf{A}^T\mathbf{B})^{-1}\mathbf{A}^T\mathbf{C} = \frac{\frac{1}{n}\sum A_i C_i}{\frac{1}{n}\sum A_i B_i} \to \beta_{c|b}, \tag{3.2}$$

which is called the IV estimator (Durbin, 1954). The *exclusion restriction* assumption, $C \perp\!\!\!\perp A \,|\, (B, U)$, is implicit since A does not appear in the SEM for C in Eq.(3.1). The IV estimator can be interpreted as

$$\hat{\beta}^{\,iv} = \{(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{B}\}^{-1}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C} = \frac{\hat{\beta}^{\,ls}_{c|a}}{\hat{\beta}^{\,ls}_{b|a}}, \tag{3.3}$$

where $\hat{\beta}^{\,ls}_{c|a}$ is the least squares estimate of the regression coefficient of C on A. Another interpretation of the IV estimator is as a generalised method of moments estimator (Hansen, 1982). Note that, for $\hat{\beta}^{\,iv}_{c|b}$ to be consistent it is unnecessary for the SEM relating B and A to be linear.

The statistical assumptions of the classical linear SEM of Eq.(3.1) can be

relaxed to

$$\mathbb{E}(C \mid B, U) = \beta_{c|b} B + U, \tag{3.4}$$

$C \perp\!\!\!\perp A \mid (B, U)$, $A \perp\!\!\!\perp U$ and the assumption $\mathbb{E}(U) = 0$, which is trivial since expectations can be replaced by covariances. Assuming stability (Pearl, 2000; cf. §2.1.1), the DAG in Fig.2.1 represents the condition that A is an instrument (cf. §2.1.1). Stability is not required if $A \not\perp\!\!\!\perp B$. The assumption $C \perp\!\!\!\perp A \mid (B, U)$ can also be referred to as the assumption of *zero direct effect* of A on C but this interpretation relies on the additional assumption that A can be intervened in.

Under the assumptions of the IV model

$$\begin{aligned} \mathbb{E}(AC) &= \mathbb{E}\{\mathbb{E}(AC \mid B, U)\} \\ &= \mathbb{E}[\mathbb{E}\{A(\beta_{c|b} B + U) \mid B, U\}] \\ &= \beta_{c|b} \mathbb{E}(AB), \end{aligned}$$

which implies that, by the law of large numbers,

$$\hat{\beta}^{\,iv} \xrightarrow{p} \frac{\mathbb{E}(AC)}{\mathbb{E}(AB)} = \beta_{c|b},$$

from Eq.(3.2). Therefore $\hat{\beta}^{\,iv}$ is a consistent estimator of $\beta_{c|b}$ as $n \to \infty$. Assuming linearity and additivity holds between B and (A, U),

$$\frac{\beta_{c|a}}{\beta_{b|a}} = \beta_{c|b}. \tag{3.5}$$

If (A, B, C, U) follows a multivariate normal distribution then linearity and additivity automatically hold and the least squares interpretation of the IV estimator in Eq.(3.3) implies that $\hat{\beta}^{iv}$ is the maximum likelihood estimator of $\beta_{c|b}$. To give a probabilistic causal intepretation of the IV estimator, the DAG of Fig.2.1 is augmented with regime indicators to form Fig.3.1.

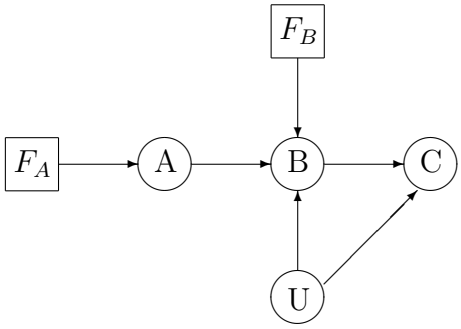


Figure 3.1: Augmented DAG representing instrumental variable model. The intervention variable F_A is not required for most of the results, only certain interpretations.

Since $C \perp\!\!\!\perp B \,|\, F_B = B, U$, from Eq.(2.5), and $C \perp\!\!\!\perp F_B \,|\, B, U$ and $U \perp\!\!\!\perp F_B$, from Fig.3.1,

$$\begin{aligned} \mathbb{E}(C \,||\, B) &= \mathbb{E}_U\{\mathbb{E}(C \,|\, B, U) \,|\, F_B\} \\ &= \mathbb{E}_U\{\mathbb{E}(C \,|\, B, U)\} \\ &= \beta_{c|b} B. \end{aligned}$$

Therefore

$$\mathrm{ACE}(B \to C) = \beta_{c|b},$$

from Eq.(1.2), and it can be said that $\hat{\beta}^{iv}$ is an estimate of the average causal effect of B on C. Following similar arguments for $\beta_{c|a}$ and $\beta_{b|a}$, from Eqs.(1.2) and (3.5),

$$\mathrm{ACE}(B \to C) = \frac{\mathrm{ACE}(A \to C)}{\mathrm{ACE}(A \to B)}. \tag{3.6}$$

With linearity and additivity throughout, the ACEs in Eq.(3.6) are equal to their corresponding ICEs. For certain models in which linearity and additivity do not hold between C and (B, U) but hold between B and (A, U), Eq.(3.6) is still valid. An example is given in Chen, Geng and Jia (2007).

3.2 Generalised linear instrumental models

Generalised linear models (Nelder and Wedderburn, 1972) have proven a worthwhile extension of linear models. The ability to adjust the link function by which the parameter space is transformed drastically improves the flexibility of such models. The success of such models readily spawns the idea that the IV technique can be used for many more systems than those which seem to appropriately follow the linear model of Eq.(3.4).

Assuming B and the instrument A are continuous, consider the alteration of Eq.(3.4) to the generalised linear model

$$g\{\mathbb{E}(C \mid B, U)\} = \beta B + U, \tag{3.7}$$

where $\mathbb{E}(A \mid B, U) \geq 0$, $\mathbb{E}(AB) \geq 0$, $g(\cdot)$ is a smooth, increasing link function and $\mathbb{E}(U) = 0$ as before, which can be relaxed for certain results.

Since any deterministic function of A satisfies the same conditional independence relations as A, then A can be re-defined as a function of itself to ensure that it is positive over its entire sample space. If the sample space of A has a finite minimum, a linear transformation can be used else some non-linear alternative, e.g. 0.5^A. Hence the assumption $E(A \mid B, U) \geq 0$ is trivial.

Apart from the technical issues for statistical inference which a non-linear link function gives rise to, there is also the philosophical difficulty of defining the causal effect for such a model. In light of the comments of §1.5, possible candidates are $\text{ACE}(B \to C)$, $\text{ICE}(B \to C; U)$, β and $\mathbb{E}(C \,||\, B)$. The suitability of each will be discussed.

For concave and convex link functions, the derivations of causal bounds are given in §3.2.1 and §3.2.2. The ideas are adapted to a generalised linear model with the logit$(\cdot)$ link function in §3.2.3. It is important to note that Eq.(3.7) is not the same as replacing C with $g(C)$ in Eq.(3.4). A model in which either C or B in Eq.(3.4) is transformed is not interesting because the data can simply be transformed and the classic IV technique used.

3.2.1 Concave link function

The fundamental idea behind the derivation of the causal bounds presented here is the exploitation of the relationship between a curved link function and its linear approximation. In this section only link functions with constant convexity are discussed.

A function $h(\cdot)$ is convex if for any two points x_1 and x_2 in its domain and any $t \in [0, 1]$,

$$h\{tx_1 + (1-t)x_2\} \leq th(x_1) + (1-t)h(x_2).$$

If $h'(\cdot)$ exists everywhere

$$h(x) \geq h(x_0) + h'(x_0)(x - x_0), \tag{3.8}$$

where x_0 is some arbitrary point in the domain of $h(\cdot)$. If $h''(\cdot)$ exists everywhere, by Taylor's theorem

$$h(x) = h(x_0) + h'(x_0)(x - x_0) + h''(x^*)\frac{(x - x_0)^2}{2},$$

where $x^* \in (x, x_0)$ and $h''(\cdot) \geq 0$. The inequalities are reversed if $h(\cdot)$ is concave.

For a model in which the link function $g(\cdot)$ is concave, i.e. $h(\cdot) = g^{-1}(\cdot)$ is convex, from Eqs.(3.7) and (3.8),

$$\begin{aligned}
\mathbb{E}(AC \,|\, B, U) &= \mathbb{E}(A \,|\, B, U)\mathbb{E}(C \,|\, B, U) \\
&= \mathbb{E}(A \,|\, B, U)h(\beta B + U) \\
\mathbb{E}(AC \,|\, B, U) &\geq \mathbb{E}(A \,|\, B, U)\{h(\phi) + h'(\phi)(\beta B + U - \phi)\} \\
\mathbb{E}(AC \,|\, B, U) &\geq \mathbb{E}[A\{h(\phi) + h'(\phi)(\beta B + U - \phi)\} \,|\, B, U],
\end{aligned}$$

since $A \perp\!\!\!\perp C \,|\, (B, U)$, where ϕ is any point in the domain of $h(\cdot)$, i.e. the range of $g(\cdot)$. Moving terms around and taking expectations over (B, U)

$$\begin{aligned}
\mathbb{E}[A\{C - h(\phi) + \phi h'(\phi)\}] &\geq h'(\phi)\beta\mathbb{E}(AB) \\
\beta^u_\phi &\geq \beta\mu_{ab},
\end{aligned}$$

since $A \perp\!\!\!\perp U$ and $\mathbb{E}(U) = 0$, where $\mu_{ab} = \mathbb{E}(AB)$ and

$$\beta^u_\phi = \mu_a\phi + \{\mu_{ac} - \mu_a h(\phi)\}/h'(\phi). \tag{3.9}$$

The bound can be tightened by minimising β_ϕ^u over the domain of $h(\cdot)$,

$$\frac{\partial \beta_\phi^u}{\partial \phi} = \frac{-h''(\phi)}{\{h'(\phi)\}^2}\{\mu_{ac} - \mu_a h(\phi)\} = 0,$$

at $\phi = g(\frac{\mu_{ac}}{\mu_a})$ and

$$\left.\frac{\partial^2 \beta_\phi^u}{\partial \phi^2}\right|_{\phi = g(\frac{\mu_{ac}}{\mu_a})} = \mu_a \left\{\frac{h''(\phi)}{h'(\phi)}\right\} + \{\mu_{ac} - \mu_a h(\phi)\}\frac{\partial}{\partial \phi}\left[\frac{-h''(\phi)}{\{h'(\phi)\}^2}\right] \geq 0,$$

since $h''(\phi) > 0$,

$$\begin{aligned} \mathbb{E}(A \,|\, B, U) \geq 0 \quad &\Rightarrow \quad \mu_a \geq 0 \\ g'(\cdot) \geq 0 \quad &\Rightarrow \quad h'(\cdot) \geq 0. \end{aligned}$$

Therefore $\phi = g(\frac{\mu_{ac}}{\mu_a})$ produces the tightest upper bound on β.

For a lower bound, by Jensen's inequality,

$$\begin{aligned} \mathbb{E}\{Ag(C) \,|\, B, U\} &= \mathbb{E}\{A \,|\, B, U\}\mathbb{E}\{g(C) \,|\, B, U\} \\ &\leq \mathbb{E}(A \,|\, B, U) g\{\mathbb{E}(C \,|\, B, U)\}. \end{aligned}$$

since $A \perp\!\!\!\perp C \,|\, (B, U)$ and $g(\cdot)$ is concave. Therefore by taking expectations over B and U

$$\mathbb{E}\{Ag(C)\} \leq \mathbb{E}\{\mathbb{E}(A \,|\, B, U)(\beta B + U)\} = \beta \mathbb{E}(AB),$$

since $A \perp\!\!\!\perp U$ and $E(U) = 0$. It follows that

$$\beta \in \mathcal{B}_1 = [\eta_1, \eta_2], \tag{3.10}$$

where

$$\eta_1 = \frac{\mathbb{E}\{Ag(C)\}}{\mathbb{E}(AB)}, \qquad \eta_2 = \frac{\mathbb{E}(A).g(\frac{\mu_{ac}}{\mu_a})}{\mathbb{E}(AB)}, \tag{3.11}$$

since $E(AB) \geq 0$. Although β is neither equal to $\text{ACE}(B \to C)$ nor $\text{ICE}(B \to C; U)$ it is a measure of an interesting component of the model. The magnitude of β can be used as a measure of the strength of the relationship between B and C under intervention, in a similar way to a typical generalised linear model. This is because

$$\mathbb{E}(C \,||\, B \,|\, U) = \mathbb{E}(C \,||\, B \,|\, B, U) = \mathbb{E}(C \,|\, B, U), \tag{3.12}$$

since $C \perp\!\!\!\perp B \,|\, F_B = B, U$ from Eq.(2.5) and $C \perp\!\!\!\perp F_B \,|\, B, U$ from Fig.3.1. Much can also be deduced from the sign of β.

Theorem 3.1. (same sign for measures of causal effect) *The quantities β, $ICE(B \to C; U)$ and $ACE(B \to C)$ all have the same sign for almost all U.*

Proof of theorem 3.1. From Eqs.(1.4), (3.7) and (3.12),

$$\beta = \frac{\partial g\{\mathbb{E}(C \,||\, B \,|\, U)\}}{\partial B} = \frac{\partial g\{\mathbb{E}(C \,||\, B \,|\, U)\}}{\partial \mathbb{E}(C \,||\, B \,|\, U)} \times \text{ICE}(B \to C; U).$$

Since $g(\cdot)$ is increasing then β and $\text{ICE}(B \to C; U)$ have the same sign. The equivalence of sign of β and $\text{ICE}(B \to C; U)$ also follows from the expression

$$\text{ICE}(B \to C; U) = \beta \times \frac{\partial h(\beta B + U)}{\partial(\beta B + U)}, \tag{3.13}$$

since $h(\cdot)$ is also increasing. It is true that

$$\begin{aligned}\mathbb{E}(C \,||\, B) &= \int \mathbb{E}(C \,|\, F_B = B, U) p(U \,|\, F_B = B)\, dU \\ &= \int \mathbb{E}(C \,||\, B \,|\, U) p(U)\, dU,\end{aligned}$$

since $U \perp\!\!\!\perp F_B$, from Fig.3.1. Similarly to Eq.(3) of Cox and Wermuth (2003),

$$\frac{\partial \mathbb{E}(C \,||\, B)}{\partial B} = \int \left\{ \frac{\partial \mathbb{E}(C \,||\, B \,|\, U)}{\partial B} p(U) + 0 \right\} dU, \tag{3.14}$$

for regular distributions, which implies that

$$\mathrm{ACE}(B \to C) = \mathbb{E}_u\{\mathrm{ICE}(B \to C; U)\}.$$

□

From Theorem 3.1, the sign of $\mathrm{ICE}(B \to C; U)$ and $\mathrm{ACE}(B \to C)$ are constant and will be obtainable when the sign of β is computable from the bounds. Estimates of η_1 and η_2 are

$$\hat{\eta}_1 = (\mathbf{A}^T\mathbf{B})^{-1}\mathbf{A}^T g(\mathbf{C}), \qquad \hat{\eta}_2 = (\mathbf{A}^T\mathbf{B})^{-1}\mathbf{A}^T g\left(\frac{\sum A_i C_i}{\sum A_i}\right), \tag{3.15}$$

where

$$\{g(\mathbf{C})\}^T = \{g(C_1), \ldots, g(C_n)\},$$

since $\hat{\eta}_1 \xrightarrow{p} \eta_1$ and $\hat{\eta}_2 \xrightarrow{p} \eta_2$ by the law of large numbers. The estimator $\hat{\eta}_1$ in Eq.(3.15) does not apply if $g(C)$ is infinite. Since the estimators in Eq.(3.15) can be computed from observational data then it is possible to partially identify the causal parameter β from an observational study.

Depending on the aim of a study, the interval $\mathcal{B}_1$ itself can be interesting, but it is also possible to obtain some non-trivial inference about $E(C \,||\, B)$. This can be obtained by use of the knowledge of the interval in which β lies. Since $h(\cdot)$ is convex, and $h'(\cdot) \geq 0$,

$$\begin{aligned}
\mathbb{E}(C \,||\, B) &= \mathbb{E}_u\{\mathbb{E}(C \,|\, B, U)\} \\
&= \mathbb{E}_u\{h(\beta B + U)\} \\
\mathbb{E}(C \,||\, B) &\geq h\{\mathbb{E}_u(\beta B + U)\} \\
\mathbb{E}(C \,||\, B) &\geq h(\beta B) \geq \min_{\beta \in \mathcal{B}_1} h(\beta B) = h\{\min_{\beta \in \mathcal{B}_1}(\beta B)\},
\end{aligned}$$

by Jensen's inequality, where $\min_{\beta \in \mathcal{B}_1}(\beta B)$ depends on the sign of B. Another approach which obtains the same result is

$$\begin{aligned}
g\{\mathbb{E}(C \,|\, B, U)\} &= \beta B + U \\
g(\theta) + g'(\theta)\{\mathbb{E}(C \,|\, B, U) - \theta\} &\geq \beta B + U \\
g(\theta) + g'(\theta)\{\mathbb{E}(C \,||\, B) - \theta\} &\geq \beta B \geq \min_{\beta \in \mathcal{B}_1}(\beta B) \\
\mathbb{E}(C \,||\, B) &\geq \mu^l_{c||b}(\theta),
\end{aligned}$$

since $\mathbb{E}(U) = 0$, where

$$\mu^l_{c||b}(\theta) = \theta + \frac{1}{g'(\theta)} \left\{ \min_{\beta \in \mathcal{B}_1}(\beta B) - g(\theta) \right\},$$

and θ is any point in the domain of $g(\cdot)$. The bound can be tightened by maximising $\mu^l_{c||b}(\theta)$,

$$\frac{\partial \mu^l_{c||b}}{\partial \theta} = \frac{-g''(\theta)}{\{g'(\theta)\}^2} \left\{ \min_{\beta \in \mathcal{B}_1}(\beta B) - g(\theta) \right\} = 0,$$

at $\theta_{min} = h\{\min_{\beta\in\mathcal{B}_1}(\beta B)\}$ and

$$\left.\frac{\partial^2 \mu^l_{c||b}}{\partial\theta^2}\right|_{\theta=\theta_{min}} = \frac{g''(\theta)}{g'(\theta)} + \left\{\min_{\beta\in\mathcal{B}_1}(\beta B) - g(\theta)\right\}\frac{\partial}{\partial\theta}\left[\frac{-g''(\theta)}{\{g'(\theta)\}^2}\right] \leq 0,$$

since $g''(\theta) \leq 0$ and $g'(\theta) \geq 0$. Therefore $\theta = h\{\min_{\beta\in\mathcal{B}_1}(\beta B)\}$ is the maximum point of $\mu^l_{c||b}(\theta)$ and

$$\mu^l_{c||b}(\theta) \leq h\left\{\min_{\beta\in\mathcal{B}_1}(\beta B)\right\},$$

which implies that

$$\mathbb{E}(C \,||\, B) \geq h\left\{\min_{\beta\in\mathcal{B}_1}(\beta B)\right\}. \tag{3.16}$$

The lower bound of Eq.(3.16) can be estimated by using $[\hat{\eta}_1, \hat{\eta}_2]$ of Eq.(3.15) as the estimate of $\mathcal{B}_1$. It is also true that

$$\begin{aligned}\mathbb{E}\{g(C) \,||\, B\} &= \mathbb{E}_u[\mathbb{E}\{g(C) \,|\, B, U\}] \\ &\leq \mathbb{E}_u[g\{\mathbb{E}(C \,|\, B, U)\}] \\ &\leq \mathbb{E}_u[\beta B + U] = \beta B \leq \max_{\beta\in\mathcal{B}_1}(\beta B),\end{aligned} \tag{3.17}$$

which produces an upper bound on $\mathbb{E}\{g(C) \,||\, B\}$, but is not as useful as an upper bound on $\mathbb{E}(C \,||\, B)$. Both Eqs.(3.16) and (3.17) enable inference about the intervention distribution from observational data.

3.2.2 Convex link function

Analagous results to §3.2.1 also hold for a convex link function. For a model in which $g(\cdot)$ in Eq.(3.7) is convex, since it is assumed that $E(AB) \geq 0$,

$$\beta \in \mathcal{B}_2 = [\eta_2, \eta_1], \tag{3.18}$$

where η_2 and η_1 are defined in Eq.(3.11), and

$$\mathbb{E}(C \,||\, B) \leq h\left\{\max_{\beta \in \mathcal{B}_2}(\beta B)\right\}. \tag{3.19}$$

Also

$$\mathbb{E}\{g(C) \,||\, B\} \geq \min_{\beta \in \mathcal{B}_2}(\beta B). \tag{3.20}$$

Estimates of the bounds are obtained similarly to the concave link function in §3.2.1 from Eq.(3.15).

If $g(\cdot)$ is the identity link function, the bounds on β in Eqs.(3.10) and (3.18) collapse to a point and the expressions in Eqs.(3.16), (3.17), (3.19) and (3.20) become equalities. The estimators in Eq.(3.15) also collapse to the IV estimator of Eq.(3.2). Since $\mathbb{E}(C \,||\, B) = \beta B$, $\mathbb{E}(C \,||\, B)$ can then be point estimated. For $\mathbb{E}(AB) < 0$, the expressions for $\mathcal{B}_1$ and $\mathcal{B}_2$ are exchanged.

3.2.3 Logit Link Function

Thus far, only link functions with constant convexity were considered. The logit(x) function is an important link function but does not satisfy this re-

quirement since it is concave for $x < \frac{1}{2}$ and convex for $x \geq \frac{1}{2}$.

In applications involving the logit$(\cdot)$ function the sign of $\beta B + U$ may be fixed and known, e.g. for a rare disease. If $\beta B + U \leq 0$, the link function is concave and the bounds of Eqs.(3.10), (3.16) and (3.17) are valid. Similarly, if $\beta B + U \geq 0$, the link function is convex and the bounds of Eqs.(3.18), (3.19) and (3.20) are valid.

If the sign of $\beta B + U$ is unknown and $g(\cdot)$ in Eq.(3.7) is the logit link function

$$\begin{aligned}
\mathbb{E}\{Ag(C) \,|\, C \geq \tfrac{1}{2}, B, U\} &= \mathbb{E}(A \,|\, B, U) E\{g(C) \,|\, C \geq \tfrac{1}{2}, B, U\} \\
&\geq \mathbb{E}(A \,|\, B, U) g\{E(C \,|\, C \geq \tfrac{1}{2}, B, U)\} \\
&\geq \mathbb{E}(A \,|\, B, U) g\{E(C \,|\, B, U)\} \\
\mathbb{E}\{Ag(C) \,|\, C \geq \tfrac{1}{2}\} &\geq \beta \mathbb{E}(AB \,|\, C \geq \tfrac{1}{2}) + \mathbb{E}(AU \,|\, C \geq \tfrac{1}{2}).
\end{aligned}$$

by Jensen's inequality, since

$$\mathbb{P}\{A \,|\, C \geq \tfrac{1}{2}, g(C), B, U\} = \mathbb{P}\{A \,|\, g(C), B, U\},$$

$A \perp\!\!\!\perp C \,|\, (B, U)$, $\mathbb{E}(A \,|\, B, U) \geq 0$ and $g(C)$ is convex for $C \geq \frac{1}{2}$. Similarly, since $g(C)$ is concave for $C < \frac{1}{2}$,

$$\mathbb{E}\{Ag(C) \,|\, C < \tfrac{1}{2}\} < \beta \mathbb{E}(AB \,|\, C < \tfrac{1}{2}) + \mathbb{E}(AU \,|\, C < \tfrac{1}{2}).$$

To develop bounds in terms of observational parameters, assume $A \perp\!\!\!\perp U \,|\, C \geq \frac{1}{2}$. Since $A \perp\!\!\!\perp U$

$$A \perp\!\!\!\perp U \,|\, C \geq \tfrac{1}{2} \quad \Leftrightarrow \quad A \perp\!\!\!\perp U \,|\, C < \tfrac{1}{2} \cdot$$

Therefore the bounds

$$\beta\mathbb{E}(AB \,|\, C \geq \tfrac{1}{2}) \leq \mathbb{E}\{Ag(C) \,|\, B, C \geq \tfrac{1}{2}\}$$
$$\beta\mathbb{E}(AB \,|\, C < \tfrac{1}{2}) > \mathbb{E}\{Ag(C) \,|\, B, C < \tfrac{1}{2}\},$$

are obtained by taking expectations over U. Estimators for the bounds in terms of observational data can be obtained similarly to Eq.(3.15). Therefore it is possible to partially identify the parameter β from observational data with the extra assumption $A \perp\!\!\!\perp U \,|\, C \geq \frac{1}{2}$. The extra assumption is required because of the variable convexity of the logit$(\cdot)$ function.

3.2.4 Illustrations for specific link functions

Log link function

To illustrate the use of the bounds, consider a hypothetical national health survey whose aim is to assess the effect of the average price of a pack of cigarettes (B) on the number of cigarettes smoked per day (C). The relationship between B and C may be confounded by the amount of funds that are spent on advertising. Let U be a measure of the budget for advertising cigarettes. Assuming that government policy on the tobacco industry is solely based on concerns about population health, a possible IV, A, is the level of taxes imposed on the cigarette companies. It is assumed that the effect of taxes on the smoking habit of the population is only via the price of cigarettes and tax levels are set independently of any confounders. Data

are simulated from the following distributions

$$A \sim \text{Uniform}[5, 10]$$
$$U \sim \text{Uniform}[-1, 1]$$
$$B \,|\, A, U \sim \text{Uniform}[2A + U - 1, 2A + U + 1]$$
$$C \,|\, B, U \sim \text{Po}(e^{0.5B+U}),$$

where $\beta = 0.5$. The summary statistics for 100000 simulated samples are given in Table 3.1.

$\frac{1}{n}\sum A$	$\frac{1}{n}\sum AB$	$\frac{1}{n}\sum AC$	$\frac{1}{n}\sum Ag(C)$
7.505	116.826	58878.86	58.41285

Table 3.1: Summary statistics for a simulation of 100000 samples.

From Table 3.1 and Eqs.(3.10) and (3.15), $\beta \in [0.5000, 0.5761]$. This demonstrates that non-trivial bounds can be obtained and are close to the true value of β. Since the bounds imply that β is positive then it can be deduced from Theorem 3.1 that $\text{ICE}(B \to C; U)$ and $\text{ACE}(B \to C)$ are positive. Using these bounds, from Eqs.(3.16) and (3.17),

$$\mathbb{E}(C \,||\, B) \geq e^{0.5B}, \qquad E\{\ln(C) \,||\, B\} \leq 0.5761B,$$

respectively. The sampling variation of the data was ignored here but will be considered in a different setting in Chapter 5.

Chapter 4

Discrete instrumental variables

The focus of Chapter 3 was on IV models with continuous variables. Here attention is diverted to discrete IV models. In a discrete IV model the observable variables A, B and C are discrete but no assumptions are made about the state space of U. For the discrete IV model, constraints on the observable trivariate distribution are given in §4.1.

However, problems arise when only bivariate data are available, which is often the case when exploiting 'Mendelian randomisation' in genetic epidemiology (Didelez and Sheehan, 2007). In Mendelian randomisation, A is the genotype, B the phenotype and C is the occurrence of a disease of interest. It is often the case that only genotype-phenotype and genotype-disease data are available. Bounds on the causal effect of B on C, in terms of the $(C \mid A)$ and $(B \mid A)$ distributions, and constraints, which must be satisfied by the bivariate distribution if the model is valid, are derived in §4.3. This is an important example as it is in direct analogy to the classical instrumental variable approach, only involving data on $(C \mid A)$ and $(B \mid A)$.

Throughout the chapter binary variables are predominantly discussed for

clarity and ease of notation. As with the models involving continuous variables, often only causal bounds can be computed. The approach to deriving constraints on the observable distribution is given in §4.2 and extended in §4.3 to bound the causal effect. Applications to data are also provided. In recognition of the fact that the IV model involves assumptions that may not always be plausible, §4.4 discusses the derivation of causal bounds under various assumptions. Additional assumptions may also be deemed appropriate in certain studies by expert knowledge. The incorporation of some such assumptions into the technique is shown.

4.1 Constraints on trivariate distribution

There have been numerous advances in the use of IVs with discrete variables when the linearity and additivity assumptions in Eq.(3.4) do not hold. Pearl (1995b) derives a falsifiable condition for A to be an instrumental variable for the causal effect of B on C or for Figs.2.1 and 2.7 to hold. It is the 'instrumental inequality'

$$\max_B \sum_C \left[\max_A \mathbb{P}(C, B \mid A)\right] \leq 1, \tag{4.1}$$

which represents a set of constraints on the joint distribution of (A, B, C). When A, B and C are binary, Robins (1989) and Manski (1990) derive non-parametric bounds for

$$\text{ACE}(B \to C) = \mathbb{P}(C = 1 \,||\, B = 1) - \mathbb{P}(C = 1 \,||\, B = 0), \tag{4.2}$$

in terms of the observable joint observational distribution of (A, B, C). The definition of ACE($B \to C$) in Eq.(4.2) follows from Eq.(1.1). Balke and Pearl (1997) improve the bounds by formulating the problem as a linear programming problem in the PO framework (cf. §2.2.1). Within the probabilistic framework, using the conditional independence relations represented in Figs.2.1 and 3.1, Dawid (2003) derives the same bounds. Angrist, Imbens and Rubin (1996) have also provided a causal interpretation of the IV estimand for binary variables,

$$\frac{\text{cov}(C, A)}{\text{cov}(B, A)} = \frac{\mathbb{E}\left[C_i\{B_{(1)i}, 1\} - C_i\{B_{(0)i}, 0\}\right]}{\mathbb{E}\left\{B_{(1)i} - B_{(0)i}\right\}},$$

within the PO framework, where $C_i\{B_{(1)i}, 1\}$ is the value of C for unit i when A is set to 1. The interpretation relies on the untestable monotonicity assumption about the joint distribution of the POs of units, $B_{(1)i} \geq B_{(0)i}$.
In §4.2, a method for computing constraints on a statistical model is described, in analogy to the instrumental inequality of Eq.(4.1). Extra statistical and causal assumptions are then added in §4.3 to extend the method to derive bounds on ACE($B \to C$) for the IV model. In §4.2 and §4.3, only the classical assumptions are used but various assumptions are removed and added to the classical set in §4.4 for sensitivity analysis.

4.2 Computing constraints on distributions

Consider a model in which there are 4 random variables, A, B and C with sample spaces $\{1, 2\}$, $\{0, 1\}$ and $\{0, 1\}$ respectively and U with an unknown

sample space. The state space of A is labelled differently since A is considered as a variable which represents different types of treatments and not the presence or absence of treatment. Initially, no statistical assumptions are made except for the sample spaces of the observable variables. U is unobservable by definition so no assumption is made about it.

The probability distribution of $(C, B, A \,|\, U)$ can be represented by a vector

$$\vec{v} = (\xi^*_{001}, \xi^*_{011}, \xi^*_{101}, \xi^*_{111}, \xi^*_{002}, \xi^*_{012}, \xi^*_{102}, \xi^*_{112}),$$

where $\xi^*_{cba} = \mathbb{P}(C, B, A \,|\, U)$. The random variable U can be treated as a parameter of the distribution since $\vec{v}$ varies as U varies. Without any assumptions, by the axioms of probability, $\vec{v}$ is a point in a 7 dimensional subspace of $[0,1]^8$ or a hyperplane in $[0,1]^8$ since

$$\textstyle\sum_c \sum_b \sum_a \xi^*_{c,b,a} = 1. \tag{4.3}$$

Let the hyperplane represented by Eq.(4.3) be $\mathcal{Z}$. In other words, $\mathcal{Z}$ is the set of all vectors in $[0,1]^8$ that represent probability distributions. Therefore $\vec{v} \in \mathcal{Z}$. The factorisation of any joint probability distribution yields

$$\mathbb{P}(C, B, A \,|\, U) = \mathbb{P}(C \,|\, B, A, U)\mathbb{P}(B \,|\, A, U)\mathbb{P}(A \,|\, U). \tag{4.4}$$

Therefore the vector

$$\vec{\tau} = (\eta_{01}, \eta_{11}, \eta_{02}, \eta_{12}, \delta_1, \delta_2, \psi),$$

$\vec{\tau} \in [0,1]^7$, where

$$\begin{aligned} \eta_{ba} &= \mathbb{P}(C = 1 \,|\, B, A, U) \\ \delta_a &= \mathbb{P}(B = 1 \,|\, A, U) \\ \psi &= \mathbb{P}(A = 2 \,|\, U), \end{aligned} \tag{4.5}$$

can also be used to represent any probability distribution. The mapping is possible since

$$\sum_C \mathbb{P}(C \,|\, B, A, U) = \sum_B \mathbb{P}(B \,|\, A, U) = \sum_A \mathbb{P}(A \,|\, U) = 1,$$

for all B and A. Both $[0,1]^7$ and $\mathcal{Z}$ are 7 dimensional. The relation in Eq.(4.4) together with the codes in Eq.(4.5) define a mapping

$$\Xi : \vec{\tau} \in [0,1]^7 \rightarrow \vec{v} \in \mathcal{Z}.$$

For the complete graph, i.e. without any conditional independence relations,

$$\Xi([0,1]^7) = \mathcal{Z}.$$

Next consider the model with the additional assumption $C \perp\!\!\!\perp A \,|\, (B, U)$. This implies that only values of $\vec{\tau}$ and $\vec{v}$ which represent distributions that exhibit this property are possible under the model. Define $\vec{u}$ as

$$\vec{u} = (\xi_{001}, \xi_{011}, \xi_{101}, \xi_{111}, \xi_{002}, \xi_{012}, \xi_{102}, \xi_{112}),$$

where $\xi_{cba} = \mathbb{P}(C, B, A)$. In order to derive falsifiable constraints for the model, it is necessary to determine the set of possible $\vec{u}$. Since $\xi_{cba} = \mathbb{E}_u[\xi^*_{cba}]$, the set of possible $\vec{u}$ lies in the convex hull of the set of possible $\vec{v}$. However, the set of possible $\vec{v}$ for the model is not obvious but the set of possible $\vec{\tau}$ is simply the intersection of the hyperplanes

$$\eta_{01} = \eta_{02}, \qquad \eta_{11} = \eta_{12}. \tag{4.6}$$

Let $\mathcal{T}$ be the set of $\vec{\tau}$ which satisfy the model restrictions. Therefore $\mathcal{T}$ is the set of $\vec{\tau}$ which satisfy Eq.(4.6),

$$\mathcal{T} = \{\vec{\tau} \in [0,1]^7 : \eta_{01} = \eta_{02}, \eta_{11} = \eta_{12}\} \subset [0,1]^7,$$

$\dim(\mathcal{T}) = 5$ and $\Xi(\mathcal{T}) = \mathcal{V} \subseteq \mathcal{Z}$, where $\mathcal{V}$ is the set of possible $\vec{v}$ for the model. The set of possible $\vec{v}$ can be found by transforming $\mathcal{T}$. Only 5 independent components of $\vec{\tau}$ are required to transform $\mathcal{T}$.
Consider first the transformation of the extreme vertices of $\mathcal{T}$, $\hat{\mathcal{T}}$. Define $\hat{\mathcal{V}} = \Xi(\hat{\mathcal{T}})$. The extreme vertices are listed and the transformation is represented in Fig.4.1, where

$$\eta_0 = \eta_{01} = \eta_{02}, \qquad \eta_1 = \eta_{11} = \eta_{12}.$$

Let $\mathcal{H}$ and $\hat{\mathcal{H}}$ be the convex hull of $\mathcal{V}$ and $\hat{\mathcal{V}}$ respectively. Since $\mathcal{H}$ is the convex hull of the set of possible $\vec{v}$ then it is the set of possible $\vec{u}$. The vector $\vec{u}$ must satisfy the inequalities which define $\hat{\mathcal{H}}$ to fit the model, i.e. to satisfy $C \perp\!\!\!\perp A \,|\, (B, U)$, since $\mathcal{H} = \hat{\mathcal{H}}$ from Theorem 4.1.

η_0	η_1	δ_1	δ_2	ψ		ξ^*_{001}	ξ^*_{011}	ξ^*_{101}	ξ^*_{111}	ξ^*_{002}	ξ^*_{012}	ξ^*_{102}	ξ^*_{112}
0	0	0	0	0		1	0	0	0	0	0	0	0
0	0	0	1	0		1	0	0	0	0	0	0	0
0	0	1	0	0		0	1	0	0	0	0	0	0
0	0	1	1	0		0	1	0	0	0	0	0	0
0	1	0	0	0		1	0	0	0	0	0	0	0
0	1	0	1	0		1	0	0	0	0	0	0	0
0	1	1	0	0		0	0	0	1	0	0	0	0
0	1	1	1	0		0	0	0	1	0	0	0	0
1	0	0	0	0		0	0	1	0	0	0	0	0
1	0	0	1	0		0	0	1	0	0	0	0	0
1	0	1	0	0		0	1	0	0	0	0	0	0
1	0	1	1	0		0	1	0	0	0	0	0	0
1	1	0	0	0		0	0	1	0	0	0	0	0
1	1	0	1	0		0	0	1	0	0	0	0	0
1	1	1	0	0		0	0	0	1	0	0	0	0
1	1	1	1	0	$\rightarrow$	0	0	0	1	0	0	0	0
0	0	0	0	1		0	0	0	0	1	0	0	0
0	0	0	1	1		0	0	0	0	0	1	0	0
0	0	1	0	1		0	0	0	0	1	0	0	0
0	0	1	1	1		0	0	0	0	0	1	0	0
0	1	0	0	1		0	0	0	0	1	0	0	0
0	1	0	1	1		0	0	0	0	0	0	0	1
0	1	1	0	1		0	0	0	0	1	0	0	0
0	1	1	1	1		0	0	0	0	0	0	0	1
1	0	0	0	1		0	0	0	0	0	0	1	0
1	0	0	1	1		0	0	0	0	0	1	0	0
1	0	1	0	1		0	0	0	0	0	0	1	0
1	0	1	1	1		0	0	0	0	0	1	0	0
1	1	0	0	1		0	0	0	0	0	0	1	0
1	1	0	1	1		0	0	0	0	0	0	0	1
1	1	1	0	1		0	0	0	0	0	0	1	0
1	1	1	1	1		0	0	0	0	0	0	0	1

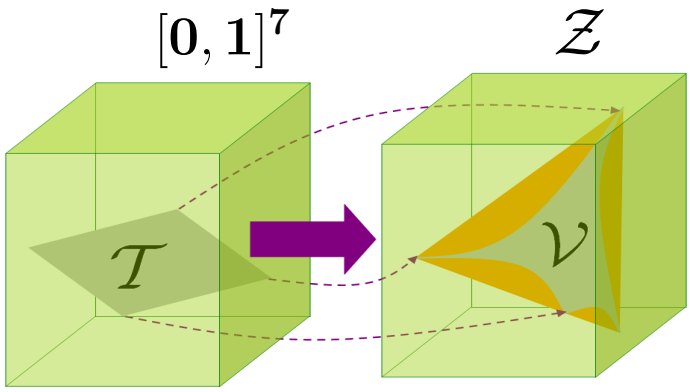


Figure 4.1: Transformation of polytope represented by extreme vertices (above) and diagrammatically (below).

Theorem 4.1. $\mathcal{H} = \hat{\mathcal{H}}$.

Proof of theorem 4.1. Following Dawid (2003), since $\hat{\mathcal{V}} \subseteq \mathcal{V}$ and $\hat{\mathcal{H}}$ is the minimal convex set containing $\hat{\mathcal{V}}$ then $\hat{\mathcal{H}} \subseteq \mathcal{H}$.

Let $m(\vec{v})$ be an affine function of $\vec{v}$ for $\vec{v} \in \hat{\mathcal{V}}$. Consider the inequality (i.e. closed half space in $[0,1]^8$) $m(\vec{v}) \geq 0$ or $m\{\Xi(\vec{\tau})\} \geq 0$ for $\vec{\tau} \in \hat{\mathcal{T}}$. From Eqs.(4.4) and (4.5), $m\{\Xi(\vec{\tau})\}$ is a monotonic function of any component of $\vec{\tau}$ when the other three are fixed. Therefore the minimum of $m\{\Xi(\vec{\tau})\}$ over $\mathcal{T}$ is attained for some $\vec{\tau} \in \hat{\mathcal{T}}$. Therefore

$$m\{\Xi(\vec{\tau})\} \geq 0 \text{ for all } \vec{\tau} \in \hat{\mathcal{T}} \Rightarrow \; m\{\Xi(\vec{\tau})\} \geq 0 \text{ for all } \vec{\tau} \in \mathcal{T}.$$

This means that any half space containing $\hat{\mathcal{V}}$ also contains $\mathcal{V}$. Since $\hat{\mathcal{H}}$ is the intersection of all half spaces containing $\hat{\mathcal{V}}$ then $\mathcal{V} \subseteq \hat{\mathcal{H}}$. Since $\hat{\mathcal{H}}$ is convex and $\mathcal{H}$ is the minimal convex set containing $\mathcal{V}$ then $\mathcal{H} \subseteq \hat{\mathcal{H}}$. □

Note that the proof of Theorem 4.1 does not use the specific form of $\Xi(\cdot)$, only its monotonicity in each coordinate.

A program such as *Polymake* (Gawrilow and Joswig, 2004) can be used to find the representation of $\hat{\mathcal{H}}$ in terms of its facets or inequalities. The method of computation of the constraints is described in Appendix A.

It is clear that the inequalities here are all trivial since $\hat{\mathcal{H}} = \mathcal{Z}$ from Fig.4.1 because $\hat{\mathcal{V}}$ is the set of extreme vertices of $[0,1]^8$. This does not necessarily imply that $\mathcal{V} = \mathcal{Z}$. The method described here determines the constraints on ξ_{cba} for $C \perp\!\!\!\perp A \,|\, (B, U)$ but can easily be adjusted to find the constraints that quantities such as $\mathbb{P}(C \,|\, A)$, $\mathbb{P}(B \,|\, A)$, $\mathbb{P}(C \,|\, B)$ etc., must satisfy. This

is done by transforming $\mathcal{T}$ to a vector of quantities related to these terms instead of a vector of $\mathbb{P}(C, B, A \,|\, U)$, as will be shown later.

4.3 Causal bounds and constraints

In this section bounds are computed for the average causal effect of B on C, as defined by Eq.(4.2), for the model from §4.2, in which $C \perp\!\!\!\perp A \,|\, (B, U)$. The additional assumption that $A \perp\!\!\!\perp U$ is added so that the model being considered is the IV model of Fig.2.1.

4.3.1 Pairwise marginals

To derive causal bounds, the assumptions represented in Fig.4.2 are considered, which include those of Fig.2.1 and extra causal assumptions.

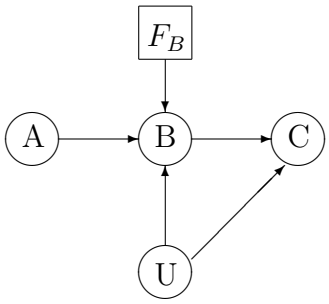


Figure 4.2: Augmented DAG for instrumental variable model which includes causal assumptions.

From the DAG in Fig.4.2,

$$\mathbb{P}(C \,|\, A) = \textstyle\sum_u \mathbb{P}(C \,|\, A, U)\mathbb{P}(U)$$
$$\mathbb{P}(B \,|\, A) = \textstyle\sum_u \mathbb{P}(B \,|\, A, U)\mathbb{P}(U),$$

since $U \perp\!\!\!\perp A$. The relation $C \perp\!\!\!\perp B \,|\, (F_B = B, U)$ also holds for the model in Fig.4.2 but is not represented by the augmented DAG (cf. §2.1.2). It follows from the fact that B is a deterministic function of F_B under the regime $F_B = B$, as in Eq.(2.5). Therefore, since $C \perp\!\!\!\perp F_B \,|\, (B, U)$ and $U \perp\!\!\!\perp F_B$ from Fig.4.2 then

$$\begin{aligned}
\mathbb{P}(C \,||\, B) &= \mathbb{P}(C \,|\, F_B = B) \\
&= \textstyle\sum_u \mathbb{P}(C \,|\, U, F_B = B)\mathbb{P}(U \,|\, F_B) \\
&= \textstyle\sum_u \mathbb{P}(C \,|\, U, F_B = B, B)\mathbb{P}(U \,|\, F_B) \\
&= \textstyle\sum_u \mathbb{P}(C \,|\, U, B)\mathbb{P}(U).
\end{aligned}$$

Let

$$\begin{aligned}
\vec{v} &= (\gamma^*_{01}, \gamma^*_{11}, \gamma^*_{02}, \gamma^*_{12}, \theta^*_{01}, \theta^*_{11}, \theta^*_{02}, \theta^*_{12}, \alpha^*) \\
\vec{u} &= (\gamma_{01}, \gamma_{11}, \gamma_{02}, \gamma_{12}, \theta_{01}, \theta_{11}, \theta_{02}, \theta_{12}, \alpha),
\end{aligned}$$

where

$$\begin{aligned}
\gamma^*_{ca} &= \mathbb{P}(C \,|\, A, U) \\
\gamma_{ca} &= \mathbb{P}(C \,|\, A) \\
\theta^*_{ba} &= \mathbb{P}(B \,|\, A, U) \\
\theta_{ba} &= \mathbb{P}(B \,|\, A) \\
\alpha^* &= \mathbb{P}(C = 1 \,|\, B = 1, U) - \mathbb{P}(C = 1 \,|\, B = 0, U) \\
\alpha &= \mathrm{ACE}(B \to C).
\end{aligned}$$

Similarly to §4.2, any possible $\vec{u}$ lies in the convex hull of the set of possible $\vec{v}$ since

$$\sum \mathbb{P}(U) = 1, \qquad \mathbb{P}(U) \geq 0 \quad \forall\, U.$$

The technique of §4.2 is then applied to find linear constraints involving the γ's, θ's and α. If the α^* and α components are omitted from $\vec{v}$ and $\vec{u}$ respectively, no causal assumptions are necessary and constraints on $(\vec{\gamma}, \vec{\theta})$ are produced for the model of Fig.2.1.

The mapping $\Xi(\cdot)$ can be expressed as

$$\begin{aligned}
\gamma^*_{01} &= (1-\eta_0)(1-\delta_1) + (1-\eta_1)\delta_1 \\
\gamma^*_{11} &= \eta_0(1-\delta_1) + \eta_1\delta_1 \\
\gamma^*_{02} &= (1-\eta_0)(1-\delta_2) + (1-\eta_1)\delta_2 \\
\gamma^*_{12} &= \eta_0(1-\delta_2) + \eta_1\delta_2 \\
\theta^*_{01} &= 1-\delta_1 \\
\theta^*_{11} &= \delta_1 \\
\theta^*_{02} &= 1-\delta_2 \\
\theta^*_{12} &= \delta_2 \\
\alpha^* &= \eta_1 - \eta_0.
\end{aligned}$$

The transformation of the extreme vertices with or without α^* and α can be seen in Fig.4.3. Only half of the set of extreme vertices are considered since $\mathbb{P}(A \,|\, U)$ is irrelevant for the transformation. In such a case $\dim(\mathcal{V}) = 4$ but $\dim(\mathcal{T}) = 5$ because Ξ is a many to one mapping.

Without any causal considerations in the analysis, the constraints

$$\begin{aligned}
\theta_{01} + \theta_{02} &\geq \gamma_{01} - \gamma_{02} \\
\theta_{01} + \theta_{02} &\geq \gamma_{02} - \gamma_{01} \\
\theta_{11} + \theta_{12} &\geq \gamma_{01} - \gamma_{02} \\
\theta_{11} + \theta_{12} &\geq \gamma_{02} - \gamma_{01},
\end{aligned} \tag{4.7}$$

η_0	η_1	δ_1	δ_2	ψ		γ^*_{01}	γ^*_{11}	γ^*_{02}	γ^*_{12}	θ^*_{01}	θ^*_{11}	θ^*_{02}	θ^*_{12}	α^*
0	0	0	0	0		1	0	1	0	1	0	1	0	0
0	0	0	1	0		1	0	1	0	1	0	0	1	0
0	0	1	0	0		1	0	1	0	0	1	1	0	0
0	0	1	1	0		1	0	1	0	0	1	0	1	0
0	1	0	0	0		1	0	1	0	1	0	1	0	1
0	1	0	1	0		1	0	0	1	1	0	0	1	1
0	1	1	0	0		0	1	1	0	0	1	1	0	1
0	1	1	1	0	$\rightarrow$	0	1	0	1	0	1	0	1	1
1	0	0	0	0		0	1	0	1	1	0	1	0	-1
1	0	0	1	0		0	1	1	0	1	0	0	1	-1
1	0	1	0	0		1	0	0	1	0	1	1	0	-1
1	0	1	1	0		1	0	1	0	0	1	0	1	-1
1	1	0	0	0		0	1	0	1	1	0	1	0	0
1	1	0	1	0		0	1	0	1	1	0	0	1	0
1	1	1	0	0		0	1	0	1	0	1	1	0	0
1	1	1	1	0		0	1	0	1	0	1	0	1	0

Figure 4.3: Transformation to the extreme vertices corresponding to the polytope which represents the IV model in terms of the pairwise marginals.

or

$$|\gamma_{01} - \gamma_{02}| \leq \theta_{01} + \theta_{02} \leq 2 - |\gamma_{01} - \gamma_{02}|,$$

are obtained, in addition to the trivial constraints $\vec{\gamma} \geq 0$, $\vec{\theta} \geq 0$ and $\sum_c \gamma_{ca} = \sum_b \theta_{ba} = 1 \; \forall \; a$. The constraints of Eq.(4.7) can also be expressed as

$$0 \leq \mathbb{P}(B = 0 \,|\, A = 1) + \mathbb{P}(B = 0 \,|\, A = 2) + |\mathrm{ACE}(A \rightarrow C)| \leq 2,$$

since

$$\mathrm{ACE}(A \rightarrow C) = \gamma_{01} - \gamma_{02}. \tag{4.8}$$

Thus far it has not been assumed that interventions in A are possible, as in Fig.4.2. However Eq.(4.8) relies on the additional assumption that interventions in A are possible.

The inequalities of Eq.(4.7) represent constraints on the observable distribution (observable component of the model). It is true that $\mathcal{V} \subset \mathcal{H}$ (where $\mathcal{V}$ and $\mathcal{H}$ are defined analagously to §4.2) for these constraints and those of §4.3.2. Therefore it is possible for the observable constraints to be satisfied but the distribution conditional on U (unobservable component of the model) to not fit the model. This occurs when $(\vec{\gamma}^*, \vec{\theta}^*) \in \mathcal{V}^c \cap \mathcal{H}$. Therefore Eq.(4.7) cannot be used to determine if a distribution actually fits the model but may be able to determine if a distribution does not fit the model. In other words, a subset of those distributions which do not fit the model can be detected. Eq.(4.7) can be considered as a version of the 'instrumental inequality' of Eq.(4.1) for the joint $(C \,|\, A)$ and $(B \,|\, A)$ distributions.

With causal assumptions, various constraints are obtained. The constraints involving observables only are the same as Eq.(4.7) but the constraints involving α are

$$\mathrm{ACE}(B \to C) \geq \max \left\{ \begin{array}{c} 2\gamma_{01} - \gamma_{02} + 2\theta_{01} - 3 \\ \gamma_{01} + \theta_{01} - 2 \\ \gamma_{02} + \theta_{02} - 2 \\ -\gamma_{01} + 2\gamma_{02} + 2\theta_{02} - 3 \\ -\gamma_{01} + \gamma_{02} - \theta_{01} + \theta_{02} - 1 \\ -\gamma_{01} - \theta_{01} \\ -\gamma_{02} - \theta_{02} \\ \gamma_{01} - 2\gamma_{02} - 2\theta_{02} \\ -2\gamma_{01} + \gamma_{02} - 2\theta_{01} \\ \gamma_{01} - \gamma_{02} + \theta_{01} - \theta_{02} - 1 \end{array} \right\} \tag{4.9}$$

$$\mathrm{ACE}(B \to C) \leq \min \left\{ \begin{array}{c} -2\gamma_{01} + \gamma_{02} + 2\theta_{01} + 1 \\ \gamma_{01} - 2\gamma_{02} + 2\theta_{02} + 1 \\ 2\gamma_{01} - \gamma_{02} - 2\theta_{01} + 2 \\ -\gamma_{01} + 2\gamma_{02} - 2\theta_{02} + 2 \\ \gamma_{01} - \gamma_{02} - \theta_{01} + \theta_{02} + 1 \\ -\gamma_{02} + \theta_{02} + 1 \\ \gamma_{01} - \theta_{01} + 1 \\ \gamma_{02} - \theta_{02} + 1 \\ -\gamma_{01} + \theta_{01} + 1 \\ -\gamma_{01} + \gamma_{02} + \theta_{01} - \theta_{02} + 1 \end{array} \right\}. \tag{4.10}$$

The α's are not estimable from observational data. Therefore, without experimental data, all of the constraints cannot be used to determine whether a distribution is invalid under the model. Instead, Eq.(4.7) is used to check whether a distribution fits and if it is not invalid then the constraints involving α are used to bound α. Eq.(4.7) has to be checked separately, ahead of bounding. This is because the bounds do not automatically satisfy Eq.(4.7) once they are non-empty, just like the cases considered in Balke and Pearl (1997) and Dawid (2003).

Provided that Eq.(4.7) holds, it is assumed (but not necessarily true, as mentioned before) that the model is valid and therefore the inequalities involving the α's will be satisfied. Therefore the validity of the bounds relies on this assumption but since α is not estimable there is no way to check it. A formal significance test for instrumental inequalities will be discussed in §5.2.

From Eqs.(4.8), (4.9) and (4.10),

$$\mathrm{ACE}(B \to C) \geq \max \left\{ \begin{array}{c} -\mathrm{ACE}(A \to C) - \mathrm{ACE}(A \to B) - 1 \\ \mathrm{ACE}(A \to C) + \mathrm{ACE}(A \to B) - 1 \end{array} \right\}$$

$$\mathrm{ACE}(B \to C) \leq \min \left\{ \begin{array}{c} \mathrm{ACE}(A \to C) - \mathrm{ACE}(A \to B) + 1 \\ -\mathrm{ACE}(A \to C) + \mathrm{ACE}(A \to B) + 1 \end{array} \right\},$$

since

$$\mathrm{ACE}(A \to B) = \theta_{01} - \theta_{02}, \tag{4.11}$$

which additionally assumes intervention in A is possible, but is not necessary for the derivation of Eqs.(4.9) and (4.10). It follows that a sufficient condition for at least one tight bound is that $|\mathrm{ACE}(A \to C)|$ and $|\mathrm{ACE}(A \to B)|$ be near 1. If their signs are the same then the lower bound is tight else the upper bound is tight. Thus if the magnitude of the causal effect of A on both B and C is large then the bounds are useful.

4.3.2 Trivariate distribution

It is possible to find bounds in terms of $\mathbb{P}(C, B \mid A)$ for the model in Fig.4.2. Simply transform $\vec{\tau}$ to a vector with components $\mathbb{P}(C, B \mid A, U)$ and α^*. Since $\mathbb{P}(C, B \mid A) = \mathbb{E}_u\{\mathbb{P}(C, B \mid A, U)\}$, the vector of $\mathbb{P}(C, B \mid A)$ and α lies in the convex hull of the set of vectors with components $\mathbb{P}(C, B \mid A, U)$ and α^*. The bounds and constraints produced are

$$\begin{aligned} \zeta_{00.1} + \zeta_{10.2} &\leq 1 \\ \zeta_{10.1} + \zeta_{00.2} &\leq 1 \\ \zeta_{11.1} + \zeta_{01.2} &\leq 1 \\ \zeta_{01.1} + \zeta_{11.2} &\leq 1, \end{aligned} \tag{4.12}$$

$\vec{\zeta} \geq 0$, $\sum_c \sum_b \zeta_{cb.a} = 1 \; \forall \; a$ and

$$\text{ACE}(B \to C) \geq \max \left\{ \begin{array}{c} \zeta_{00.1} + \zeta_{11.2} - 1 \\ \zeta_{11.1} + \zeta_{00.2} - 1 \\ -\zeta_{01.1} - \zeta_{10.1} + \zeta_{11.1} - \zeta_{10.2} - \zeta_{11.2} \\ -\zeta_{10.1} - \zeta_{11.1} - \zeta_{01.2} - \zeta_{10.2} + \zeta_{11.2} \\ -\zeta_{01.1} - \zeta_{10.1} \\ -\zeta_{01.2} - \zeta_{10.2} \\ -\zeta_{00.1} - \zeta_{01.1} + \zeta_{00.2} - \zeta_{01.2} - \zeta_{10.2} \\ \zeta_{00.1} - \zeta_{01.1} - \zeta_{10.1} - \zeta_{00.2} - \zeta_{01.2} \end{array} \right\} \tag{4.13}$$

$$\text{ACE}(B \to C) \leq \min \left\{ \begin{array}{c} 1 - \zeta_{10.1} - \zeta_{01.2} \\ 1 - \zeta_{01.1} - \zeta_{10.2} \\ \zeta_{00.1} - \zeta_{01.1} + \zeta_{11.1} + \zeta_{00.2} + \zeta_{01.2} \\ \zeta_{00.1} + \zeta_{01.1} - \zeta_{01.2} + \zeta_{00.2} + \zeta_{11.2} \\ \zeta_{00.1} + \zeta_{11.1} \\ \zeta_{00.2} + \zeta_{11.2} \\ \zeta_{10.1} + \zeta_{11.1} + \zeta_{00.2} + \zeta_{11.2} - \zeta_{10.2} \\ \zeta_{00.1} - \zeta_{10.1} + \zeta_{11.1} + \zeta_{10.2} + \zeta_{11.2} \end{array} \right\}, \tag{4.14}$$

where $\zeta_{cb.a} = \mathbb{P}(C, B \,|\, A)$. Eqs.(4.13) and (4.14) are the same as those in Balke and Pearl (1997) and Dawid (2003) for the model in which A is an instrument for the effect of B on C. The method employed here is exactly that used in Dawid (2003).

4.3.3 Data analysis with causal bounds

Lipid Research Clinics coronary data

Consider the Lipid Research Coronary Primary Prevention Trial (Lipid Research Clinic Program, 1984), which was analysed by Efron and Feldman (1991) and Balke and Pearl (1997). Subjects were randomised into two groups, 172 men were given the placebo and 165 were given the treatment,

a	$\hat{\zeta}_{00.a}$	$\hat{\zeta}_{01.a}$	$\hat{\zeta}_{10.a}$	$\hat{\zeta}_{11.a}$
1	0.919	0	0.081	0
2	0.315	0.139	0.073	0.473

a	$\hat{\theta}_{0a}$	$\hat{\theta}_{1a}$
1	1	0
2	0.388	0.612

a	$\hat{\gamma}_{0a}$	$\hat{\gamma}_{1a}$
1	0.919	0.081
2	0.454	0.546

Table 4.1: Relative frequencies derived from Lipid Research Clinics Coronary Primary Prevention Trial (1984).

and the subjects' cholesterol levels were measured. There was partial compliance of the subjects with the treatment assigned. The relative frequencies are given in Table 4.1. In this trial, the relative frequencies are the maximum likelihood estimates of the parameters $\vec{\zeta}$ (cf. §5.1). The uncertainty in the data will be ignored here but properly taken into consideration in Chapter 5.

From Eqs.(4.13) and (4.14) the bounds using the $(C, B \mid A)$ distribution are

$$\mathrm{ACE}(B \to C) \geq \max \left\{ \begin{array}{c} 0.392, -0.685, -0.627, 0.18, -0.081, \\ -0.212, -0.816, 0.384 \end{array} \right\}$$

$$\mathrm{ACE}(B \to C) \leq \min \left\{ \begin{array}{c} 0.78, 0.927, 1.373, 1.568, 0.919, \\ 0.788, 0.796, 1.384 \end{array} \right\}$$

$$\Rightarrow \qquad 0.392 \leq \alpha \leq 0.780,$$

as in Balke and Pearl (1997) and Dawid (2003). Using only the $(C \mid A)$ and $(B \mid A)$ distributions, from Eqs.(4.9) and (4.10),

$$\text{ACE}(B \to C) \geq \max \left\{ \begin{array}{c} 0.384, -0.081, -1.158, -2.235, -2.077, \\ -1.919, -0.842, -0.765, -3.384, 0.077 \end{array} \right\}$$

$$\text{ACE}(B \to C) \leq \min \left\{ \begin{array}{c} 1.616, 1.787, 1.384, 1.213, 0.853, \\ 0.934, 0.919, 1.066, 1.081, 1.147 \end{array} \right\}$$

$$\Rightarrow \qquad 0.384 \leq \text{ACE}(B \to C) \leq 0.853.$$

Remarkably, the width of the bounds for the bivariate and trivariate distributions are not much different. Therefore it is still possible to make useful inference with the less informative bivariate data.

Vitamin A Supplementation

Another example of partial compliance is the study of Vitamin A supplementation in northern Sumatra, described by Sommer and Zeger (1991). The study consisted of children in 450 villages, 11588 children (221 villages) were assigned to the control group and 12094 (229 villages) to the treatment group. The relative frequencies, which are the maximum likelihood estimates, are given in Table 4.2. Those assigned to the control group were not given a placebo because of government policy. The causal effects were also analysed by Balke and Pearl (1997).

From Eqs.(4.13) and (4.14) the bounds using the $(C, B \mid A)$ distribution are

$$\text{ACE}(B \to C) \geq \max \left\{ \begin{array}{c} -0.1946, -0.9972, -1.9898, -0.3928, -0.9936, \\ -0.1982, -0.2018, -0.991 \end{array} \right\}$$

$$\text{ACE}(B \to C) \leq \min \left\{ \begin{array}{c} 0.0054, 0.8028, 0.0102, 0.8072, \\ 0.0064, 0.8018, 1.5982, 0.009 \end{array} \right\}$$

a	$\hat{\zeta}_{00.a}$	$\hat{\zeta}_{01.a}$	$\hat{\zeta}_{10.a}$	$\hat{\zeta}_{11.a}$
1	0.0064	0	0.9936	0
2	0.0028	0.0010	0.1972	0.7990

a	$\hat{\theta}_{0a}$	$\hat{\theta}_{1a}$
1	1	0
2	0.2	0.8

a	$\hat{\gamma}_{0a}$	$\hat{\gamma}_{1a}$
1	0.0064	0.9936
2	0.0038	0.9962

Table 4.2: Relative frequencies derived from vitamin A data of Sommer and Zeger (1991).

$$\Rightarrow \qquad -0.1946 \leq \alpha \leq 0.0054,$$

as in Balke and Pearl (1997) and Dawid (2003). Using only the $(C \,|\, A)$ and $(B \,|\, A)$ distributions, from Eqs.(4.9) and (4.10),

$$\mathrm{ACE}(B \to C) \geq \max \left\{ \begin{array}{c} -0.991, -0.9936, -1.7962, -2.5988, -1.8026, \\ -1.0064, -0.2038, -0.4012, -2.009, -0.1974 \end{array} \right\}$$

$$\mathrm{ACE}(B \to C) \leq \min \left\{ \begin{array}{c} 2.991, 1.3988, 0.009, 1.6012, 0.2026, \\ 1.1962, 0.0064, 0.8038, 1.9936, 1.7974 \end{array} \right\}$$

$$\Rightarrow \qquad -0.1974 \leq \mathrm{ACE}(B \to C) \leq 0.0064.$$

Here again the bounds are very similar. The similarity of the bounds when given the bivariate and trivariate data for these two examples may not necessarily hold in general. For each of the examples the data from all of the tables were derived from the same study but, when ignoring sampling uncertainty, it is irrelevant whether each individual table was obtained from a different

study. However, when sampling variation is taken into account, various issues may arise when using data from multiple studies. Also in both cases it was checked separately that the estimated parameters satisfied Eqs.(4.7) and (4.12) before the bounds were calculated. It is the condition that the constraints are satisfied by the relative frequencies which makes them likelihood estimates (cf. §5.1). The sampling uncertainty in the data was ignored in the above analyses but techniques to quantify the probability that the causal effect is within specific bounds will be discussed in Chapter 5.

4.4 Sensitivity analysis of causal bounds

In §4.1, §4.2 and §4.3, various bounds on the observable distribution were described when the exclusion restriction (zero direct effect) and randomisation assumptions hold. However those assumptions may not always be appropriate so it is necessary to assess the sensitivity of the bounds to the various assumptions.

In this section, the technique of §4.3 is extended for the derivation of causal bounds in models where the assumptions of randomisation and no direct effect do not hold. The augmented DAGs for the models without the assumptions are given in Fig.4.4. The *monotonicity* assumption (Imbens and Angrist, 1994) is also considered in §4.4.3.

4.4.1 Randomisation assumption

In §4.2, the technique for computing constraints was described for the model without randomisation, $U \not\perp\!\!\!\perp A$. The example was solely for illustrative pur-

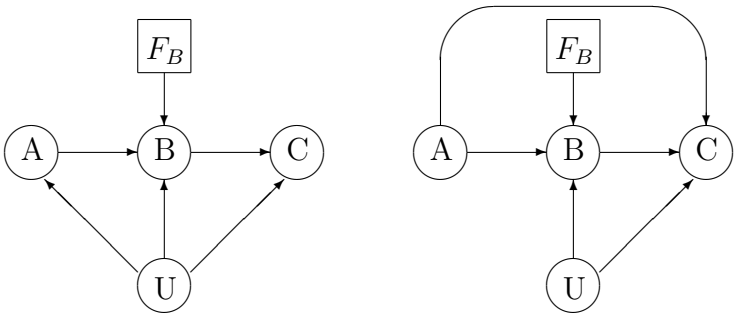


Figure 4.4: Augmented DAGs which represent the causal IV model without randomisation (left) and without exclusion restriction (right).

poses and only trivial constraints were derived.

Let the state space of A be $\{1, 2, \ldots, l\}$ and (B, C) remain the same. For all of the observable binary variables, $l = 2$. Consider the modification of the technique in §4.2 and §4.3. Let $\psi_a = \mathbb{P}(A \,|\, U)$ and redefine

$$\vec{\tau} = (\eta_{01}, \eta_{11}, \ldots, \eta_{0l}, \eta_{1l}, \delta_1, \ldots, \delta_l, \psi_1, \ldots, \psi_{l-1}).$$

It is true that $\vec{\tau} \in [0,1]^{4l-1}$ and fully parameterises $\mathbb{P}(A, B, C \,|\, U)$. Since $C \perp\!\!\!\perp A \,|\, (B, U)$, the subspace of $[0,1]^{4l-1}$ which corresponds to the set of $\vec{\tau}$ valid for the model, $\mathcal{T}$, is

$$\mathcal{T} = \{\vec{\tau} \in [0,1]^{4l-1} : \eta_{01} = \eta_{02} = \ldots = \eta_{0l}, \eta_{11} = \eta_{12} = \ldots = \eta_{1l}\}.$$

Let

$$\vec{v}_i = (\gamma^*_{0i}, \gamma^*_{1i}, \theta^*_{0i}, \theta^*_{1i}, \alpha^*),$$

and

$$\eta_0 = \eta_{0i}, \qquad \eta_1 = \eta_{1i}, \qquad \text{for } i = 1, \ldots, l.$$

Since $\Xi(\cdot)$ is the function which maps $\vec{\tau}$ to $\vec{v}_i$, the set of $\vec{v}_i$ which are valid for the model is $\Xi(\mathcal{T})$ and $\Xi(\cdot)$ can be expressed as

$$\begin{aligned}
\gamma^*_{0i} &= (1-\eta_0)(1-\delta_i) + (1-\eta_1)\delta_i \\
\gamma^*_{1i} &= \eta_0(1-\delta_i) + \eta_1\delta_i \\
\theta^*_{0i} &= 1-\delta_i \\
\theta^*_{1i} &= \delta_i \\
\alpha^* &= \eta_1 - \eta_0.
\end{aligned}$$

Similarly to §4.2, the convex hull of $\Xi(\hat{\mathcal{T}})$ is the same as the convex hull of $\Xi(\mathcal{T})$, where $\hat{\mathcal{T}}$ is the set of extreme points of $\mathcal{T}$. Thus it is only necessary to determine the convex hull of $\Xi(\hat{\mathcal{T}})$. Consider $\hat{\mathcal{T}}$ and its transformation, $\Xi(\hat{\mathcal{T}})$, which are given in Fig.4.5.

η_0	η_1	δ_i		γ^*_{0i}	γ^*_{1i}	θ^*_{0i}	θ^*_{1i}	α^*
0	0	0		1	0	1	0	0
0	0	1		1	0	0	1	0
0	1	0		1	0	1	0	1
0	1	1	$\rightarrow$	0	1	0	1	1
1	0	0		0	1	1	0	-1
1	0	1		1	0	0	1	-1
1	1	0		0	1	1	0	0
1	1	1		0	1	0	1	0

Figure 4.5: Transformation to the extreme vertices corresponding to the polytope which represents the IV model without randomisation in terms of the pairwise marginals.

Only those components of $\vec{\tau}$ which are relevant to the transformation are given in the table in Fig.4.5. The inequalities which define the convex hull of the transformed space are found using Polymake, which outputs the halfspace representation of the convex polytope, i.e. inequality constraints on $\vec{v}_i$.

Since $C \perp\!\!\!\perp B \,|\, (U, F_B = B)$ from Eq.(2.5) and $C \perp\!\!\!\perp F_B \,|\, (B, U)$, $U \perp\!\!\!\perp F_B$ and $C \perp\!\!\!\perp A \,|\, (B, U)$ from the DAG in Fig.4.4,

$$\begin{aligned} \mathbb{P}(C \,||\, B) &= \textstyle\sum_a \sum_u \mathbb{P}(C \,|\, B, U)\mathbb{P}(U \,|\, A)\mathbb{P}(A) \\ &= \mathbb{E}_a(\alpha'_a), \end{aligned} \tag{4.15}$$

where

$$\alpha'_a = \sum_u \mathbb{P}(C \,|\, B, U)\mathbb{P}(U \,|\, A = a).$$

It is also true in general that

$$\begin{aligned} \mathbb{P}(C \,|\, A) &= \textstyle\sum_u \mathbb{P}(C \,|\, A, U)\mathbb{P}(U \,|\, A) \\ \mathbb{P}(B \,|\, A) &= \textstyle\sum_u \mathbb{P}(B \,|\, A, U)\mathbb{P}(U \,|\, A). \end{aligned}$$

Therefore

$$\vec{w}_i = (\gamma_{0i}, \gamma_{1i}, \theta_{0i}, \theta_{1i}, \alpha'_i),$$

is a weighted sum of $\vec{v}_i$, with weights $\mathbb{P}(U \,|\, A = i)$. This implies that any possible $\vec{w}_i$ lies in the convex hull of the set of possible $\vec{v}_i$ since

$$\sum \mathbb{P}(U \,|\, A) = 1, \qquad \mathbb{P}(U \,|\, A) \geq 0 \quad \forall\, U,$$

and $\vec{w}_i$ must satisfy the same constraints as $\vec{v}_i$. The following constraints are obtained

$$\begin{aligned} 0 &\leq \gamma_{0i} + 2\gamma_{1i} - \theta_{0i} + \alpha'_i \\ 0 &\leq \gamma_{0i} + \theta_{0i} + \alpha'_i \\ 0 &\leq \gamma_{1i} + \theta_{0i} - \alpha'_i \\ 0 &\leq 2\gamma_{0i} + \gamma_{1i} - \theta_{0i} - \alpha'_i. \end{aligned}$$

Therefore, for $i = 1, \ldots, l$,

$$\max \left\{ \begin{array}{c} \gamma_{0i} + \theta_{0i} - 2 \\ -\gamma_{0i} - \theta_{0i} \end{array} \right\} \leq \alpha'_i \leq \min \left\{ \begin{array}{c} -\gamma_{0i} + \theta_{0i} + 1 \\ \gamma_{0i} - \theta_{0i} + 1 \end{array} \right\}. \tag{4.16}$$

Since $\alpha = \mathbb{E}_a(\alpha'_a)$ from Eq.(4.15) then

$$\text{ACE}(B \to C) \geq \min_{i=1,\ldots,l} \left[\max \left\{ \begin{array}{c} \gamma_{0i} + \theta_{0i} - 2 \\ -\gamma_{0i} - \theta_{0i} \end{array} \right\} \right]$$

$$\text{ACE}(B \to C) \leq \max_{i=1,\ldots,l} \left[\min \left\{ \begin{array}{c} -\gamma_{0i} + \theta_{0i} + 1 \\ \gamma_{0i} - \theta_{0i} + 1 \end{array} \right\} \right].$$

However, if marginal A data is available, the bounds can be improved to

$$\text{ACE}(B \to C) \geq \sum_{i=1}^{l} \left[\max \left\{ \begin{array}{c} \gamma_{0i} + \theta_{0i} - 2 \\ -\gamma_{0i} - \theta_{0i} \end{array} \right\} \mathbb{P}(A = i) \right]$$

$$\text{ACE}(B \to C) \leq \sum_{i=1}^{l} \left[\min \left\{ \begin{array}{c} -\gamma_{0i} + \theta_{0i} + 1 \\ \gamma_{0i} - \theta_{0i} + 1 \end{array} \right\} \mathbb{P}(A = i) \right],$$

or

$$-1 + \mathbb{E}_a(|\gamma_{1a} - \theta_{0a}|) \leq \text{ACE}(B \to C) \leq 1 - \mathbb{E}_a(|\gamma_{0a} - \theta_{0a}|). \tag{4.17}$$

Eqs.(4.9) and (4.10) contain the terms in Eq.(4.16). Therefore the bounds with the randomisation restriction are at least as narrow as those without, as expected. Although Eq.(4.17) bounds the unobservable causal effect, there are no falsifiable constraints to invalidate the model since all constraints involve α'_i. Any analysis via this approach assumes that the model is correct. The main difference between the case considered here and that of §4.2 and §4.3 is that the vector $\vec{\tau}$ is mapped to a vector $\vec{v}_i$ for each $A = i$ here instead

of $\vec{v}$. The results presented for the model with no randomisation apply for all finite values of l.

For the trivariate distribution, bounds in terms of $\mathbb{P}(C, B \mid A)$ can be found. Redefine $\vec{v}_i = (\zeta^*_{00.i}, \zeta^*_{01.i}, \zeta^*_{10.i}, \zeta^*_{11.i}, \alpha^*)$ where $\zeta^*_{cb.a} = \mathbb{P}(C, B \mid A, U)$. Since

$$\mathbb{P}(C, B \mid A) = \textstyle\sum_u \mathbb{P}(C, B \mid A, U)\mathbb{P}(U \mid A),$$

then $\vec{w}_i = (\zeta_{00.i}, \zeta_{01.i}, \zeta_{10.i}, \zeta_{11.i}, \alpha'_i)$ lies in the convex hull of the set of $\vec{v}_i$. The set of valid $\vec{v}_i$ can again be found by transformation of $\hat{\mathcal{T}}$, as in §4.3. The transformation of the extreme vertices of the polytope for $A = i$ is given in Fig.4.6.

η_0	η_1	δ_i		$\zeta^*_{00.i}$	$\zeta^*_{01.i}$	$\zeta^*_{10.i}$	$\zeta^*_{11.i}$	α^*
0	0	0		1	0	0	0	0
0	0	1		0	1	0	0	0
0	1	0		1	0	0	0	1
0	1	1	$\rightarrow$	0	0	0	1	1
1	0	0		0	0	1	0	-1
1	0	1		0	1	0	0	-1
1	1	0		0	0	1	0	0
1	1	1		0	0	0	1	0

Figure 4.6: Transformation to the extreme vertices corresponding to the polytope which represents the IV model without randomisation in terms of the trivariate distribution.

Using Polymake to find the inequalities which define the convex hull of the set of valid $\vec{v}_i$, the following constraints on $\vec{w}_i$ are obtained

$$-\zeta_{01.i} - \zeta_{10.i} \leq \alpha'_i \leq \zeta_{00.i} + \zeta_{11.i}.$$

Since $\alpha = \mathbb{E}_a(\alpha'_a)$ from Eq.(4.15),

$$\min_{i=1,\ldots,l}\{-\zeta_{01.i} - \zeta_{10.i}\} \leq \text{ACE}(B \to C) \leq \max_{i=1,\ldots,l}\{\zeta_{00.i} + \zeta_{11.i}\},$$

and

$$-\zeta_{01} - \zeta_{10} \leq \text{ACE}(B \to C) \leq \zeta_{00} + \zeta_{11}, \tag{4.18}$$

where $\zeta_{cb} = \mathbb{P}(C, B)$. The causal effect is bounded by the joint (C, B) distribution only. Eq.(4.18) is the no-assumption bound of Manski (1990) but here it is derived without the use of counterfactuals. This derivation requires no assumptions also since a variable A which satisfies the condition $C \perp\!\!\!\perp A \,|\, (B, U)$ trivially exists. Simply construct a variable $A = B$ and this satisfies the required criterion. Here again the bounds with the randomisation restriction, Eqs.(4.13) and (4.14), are at least as narrow as those without, Eq.(4.18). The terms in the expression for the bounds without randomisation correspond exactly to terms in the expression for the bounds with randomisation.

4.4.2 Non-zero direct effect

The exclusion restriction assumption is fairly strong and may not be satisfied in practice. For this reason the IV model is inappropriate for many studies. A very useful alternative is a model in which there is a weaker assumption on the direct effect of the instrument on the response. Consider a binary model

with the weaker assumption

$$0 < |\mathbb{P}(C \mid B, A = 1, U) - \mathbb{P}(C \mid B, A = 2, U)| \leq \epsilon,$$

where $0 < \epsilon \leq 1$, and the larger the value of ϵ the weaker the assumption. Constraints under this weaker IV model are important and the use of the same general approach to derive constraints is demonstrated. The augmented DAG is given in Fig.4.7, where the $A \rightarrow C$ edge represents a weaker exclusion restriction, defined as

$$\mathbb{P}(C \mid B, A = 1, U) \neq \mathbb{P}(C \mid B, A = 2, U).$$

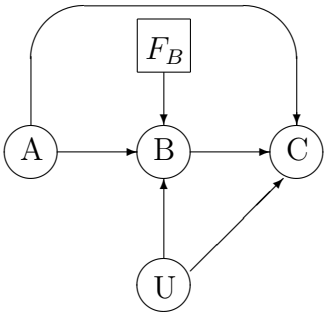


Figure 4.7: Augmented DAG representing the instrumental variable model with a weaker exclusion restriction assumption.

The adaptation of the technique is considered for the case where $\epsilon = 0.5$ and A, B and C are binary. Therefore $\mathcal{T}$, the set of valid $\vec{\tau}$ for the model, is the subspace of $[0, 1]^7$ for which

$$|\eta_{i1} - \eta_{i2}| \leq 0.5, \quad \text{for } i = 0, 1.$$

η_{01}	η_{02}	η_{11}	η_{12}	δ_1	δ_2
0	0	0	0	0	0
0	0	0	0.5	0	1
0	0	0.5	0	1	0
0	0	0.5	0.5	1	1
0	0	0.5	1	0	0
0	0	1	0.5	0	1
0	0	1	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮

$$\Downarrow$$

$\zeta^*_{00.1}$	$\zeta^*_{01.1}$	$\zeta^*_{10.1}$	$\zeta^*_{11.1}$	$\zeta^*_{00.2}$	$\zeta^*_{01.2}$	$\zeta^*_{10.2}$	$\zeta^*_{11.2}$
1	0	0	0	1	0	0	0
1	0	0	0	0	0.5	0	0.5
0	0.5	0	0.5	1	0	0	0
0	0.5	0	0.5	0	0.5	0	0.5
1	0	0	0	1	0	0	0
1	0	0	0	0	0.5	0	0.5
0	0	0	1	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 4.8: Transformation to the extreme vertices corresponding to the polytope which represents the IV model with a weaker exclusion restriction in terms of the trivariate distribution.

The set of extreme vertices, $\hat{\mathcal{T}}$, will be subject to this constraint. The derivation of falsifiable constraints is discussed for this particular model since minor adjustments to it readily produce causal bounds and the aim here is primarily demonstrative. Consider the mapping of $\vec{\tau}$ to

$$\vec{v} = (\zeta^*_{00.1}, \zeta^*_{01.1}, \zeta^*_{10.1}, \zeta^*_{11.1}, \zeta^*_{00.2}, \zeta^*_{01.2}, \zeta^*_{10.2}, \zeta^*_{11.2}),$$

where the transformation of some of the extreme vertices are given in Fig.4.8. The resulting non-trivial constraints on the observable distribution are

$$\zeta_{00.1} + \zeta_{10.2} - \zeta_{10.1} - \zeta_{00.2} \leq 1$$
$$\zeta_{10.1} + \zeta_{00.2} - \zeta_{00.1} - \zeta_{10.2} \leq 1$$
$$\zeta_{11.1} + \zeta_{01.2} - \zeta_{01.1} - \zeta_{11.2} \leq 1$$
$$\zeta_{01.1} + \zeta_{11.2} - \zeta_{11.1} - \zeta_{01.2} \leq 1,$$

which is a weaker version of the instrumental inequality of Eq.(4.12), as expected. By redefining $\vec{v}$ with an additional component α^*, causal bounds can be derived.

4.4.3 Montonicity assumption

Assume that all observable variables are binary. Here, monotonicity refers to the assumption that increasing A increases the probability of a larger value of B and can be expressed mathematically by

$$\mathbb{P}(B = 1 \,|\, A = 2, U) \geq \mathbb{P}(B = 1 \,|\, A = 1, U) \quad \text{or} \quad \delta_2 \geq \delta_1,$$

a weaker version of that described in Imbens and Angrist (1994) and further discussed in Angrist, Imbens and Rubin (1996). It is also applied to bounds on causal effects in Balke and Pearl (1997). An example where the monotonicity assumption can be justified is in the education of teenagers about the dangers of smoking. The variable A represents whether a teenager is educated ($A = 2$) or not ($A = 1$) and B whether the subject develops a smoking habit ($B = 0$) or not ($B = 1$). This is a weaker assumption than assuming that those who are educated never smoke. Under this assumption the subject who is educated may still actually take up smoking but the assumption is

valid as long as they are more likely to not smoke when educated. Using the terminology of Shafer (1996), conditional on U, the event of A increasing is a *positive sign* for a larger value of B since

$$\begin{aligned}&\mathbb{P}(B=1\,|\,A=2,U)-\mathbb{P}(B=0\,|\,A=2,U)\\&\quad\geq\mathbb{P}(B=1\,|\,A=1,U)-\mathbb{P}(B=0\,|\,A=1,U).\end{aligned}$$

The monotonicity assumption restricts the space of the vector of probabilities so that the bounds produced are those which apply in the restricted space. This assumption may reduce or have no effect on the width of the bounds. The monotonicity assumption cannot be represented on a DAG since it represents assumptions about the sample space of the random variables. The chain event graphs of Smith and Anderson (2008) can be used to represent it (cf. §2.3).

To demonstrate the effect of the monotonicity assumption, consider a case where only trivial bounds are produced without it. Suppose it is necessary to make inference on the causal effect of A on B given data on the effect of A on C. When the technique of §4.3 is bluntly applied to the IV model only trivial bounds are obtained.

Let

$$\varphi=\mathrm{ACE}(A\rightarrow B).$$

It is obvious from Eq.(4.7) that there are no constraints on $(\theta_{01}-\theta_{02})$ and thus, from Eq.(4.11), no constraints on $\mathrm{ACE}(A\rightarrow B)$. Further assumptions are necessary to produce useful bounds.

Let

$$\varphi^* = \mathbb{P}(B = 1 \,|\, A = 2, U) - \mathbb{P}(B = 1 \,|\, A = 1, U).$$

Additionally assuming monotonicity, all of the vertices with $\delta_2 < \delta_1$ are removed. The omitted edges transform to the edges with $\varphi^* = -1$ since the assumption basically constrains $\varphi^* \geq 0$. The new polytope transformation is given in Fig.4.9.

η_0	η_1	δ_1	δ_2		γ^*_{01}	γ^*_{11}	γ^*_{02}	γ^*_{12}	φ^*
0	0	0	0		1	0	1	0	0
0	0	0	1		1	0	1	0	1
0	0	1	0		–	–	–	–	–
0	0	1	1		1	0	1	0	0
0	1	0	0		1	0	1	0	0
0	1	0	1		1	0	0	1	1
0	1	1	0		–	–	–	–	–
0	1	1	1	→	0	1	0	1	0
1	0	0	0		0	1	0	1	0
1	0	0	1		0	1	1	0	1
1	0	1	0		–	–	–	–	–
1	0	1	1		1	0	1	0	0
1	1	0	0		0	1	0	1	0
1	1	0	1		0	1	0	1	1
1	1	1	0		–	–	–	–	–
1	1	1	1		0	1	0	1	0

Figure 4.9: Transformation to the extreme vertices corresponding to the polytope which represents the IV model with the monotonicity assumption in terms of the pairwise marginals.

The resulting constraints are $\vec{\gamma} \geq 0$, $\sum_c \gamma_{ca} = 1$ and

$$\max \left\{ \begin{array}{l} \gamma_{01} - \gamma_{02} \\ -\gamma_{01} + \gamma_{02} \end{array} \right\} \leq \varphi \leq 1 \quad \text{or} \quad |\gamma_{01} - \gamma_{02}| \leq \varphi \leq 1. \tag{4.19}$$

The transformed polytope is represented in Fig.4.10. The •'s are the vertices

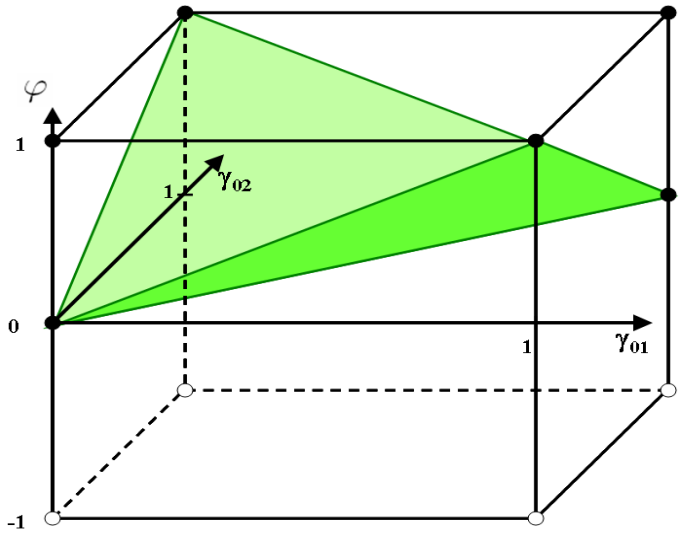


Figure 4.10: Transformed polytope, restricted and unrestricted by the monotonicity assumption. The extreme vertices which satisfy montonicity are the •'s the ∘'s are those which do not.

which are not removed after assuming monotonicity and the ∘'s are the vertices which are removed. It is obvious from Fig.4.10 that the convex hull of the vertices before assuming monotonicity is the entire cube whereas, after monotonicity is assumed, it is just the intersection of the regions in the cube which are above the planes represented by $\varphi = |\gamma_{01} - \gamma_{02}|$ or

$$\text{ACE}(A \to B) \geq |\gamma_{01} - \gamma_{02}| = |\text{ACE}(A \to C)|.$$

Intuitively, the bound expresses the fact that, under the assumption of monotonicity, the causal effect of A on B is greater than or equal to the magnitude of the causal effect of A on C. This makes sense since it is assumed that $\varphi^* \geq 0$ and B lies on the causal pathway from A to C.

Another interesting exercise is to determine the effect of the monotonicity assumption on the constraints of §4.3. The technique is easily adjusted, as

in the derivation of Eq.(4.19), by removing those vertices which do not exist under monotonicity. For bivariate data, analagously to Eqs.(4.7), (4.9) and (4.10), the constraints under monotonicity are $\vec{\gamma}$, θ_{02}, $\theta_{11} \geq 0$,

$$\theta_{01} - \theta_{02} \geq |\gamma_{01} - \gamma_{02}|, \tag{4.20}$$

and the causal bounds

$$\text{ACE}(B \to C) \geq \max \left\{ \begin{array}{c} 2\gamma_{01} - \gamma_{02} + \theta_{01} - 2 \\ \gamma_{01} - 2\gamma_{02} - \theta_{02} \\ \gamma_{01} + \theta_{01} - 2 \\ -\gamma_{02} - \theta_{02} \\ \gamma_{01} - \gamma_{02} + \theta_{01} - \theta_{02} - 1 \end{array} \right\} \tag{4.21}$$

$$\text{ACE}(B \to C) \leq \min \left\{ \begin{array}{c} 2\gamma_{01} - \gamma_{02} - \theta_{01} + 1 \\ \gamma_{01} - 2\gamma_{02} + \theta_{02} + 1 \\ \gamma_{01} - \theta_{01} + 1 \\ -\gamma_{02} + \theta_{02} + 1 \\ \gamma_{01} - \gamma_{02} - \theta_{01} + \theta_{02} + 1 \end{array} \right\}, \tag{4.22}$$

which imply that $\vec{\theta} \geq 0$. Eq.(4.20) is actually the causal bound of Eq.(4.19) but here it is a testable constraint because (C, A) and (B, A) data is available and can therefore be used to reject models which do not adhere to the assumptions of exclusion restriction, randomisation and monotonicity. When Eq.(4.20) holds, the bounds in Eqs.(4.21) and (4.22) are at least as narrow as those of Eqs.(4.9) and (4.10), which means that the monotonicity assumption has improved the bounds.

For trivariate data, analagously to Eqs.(4.12), (4.13) and (4.14), the con-

straints are $\zeta_{00.2}$, $\zeta_{01.1}$, $\zeta_{10.2}$, $\zeta_{11.1} \geq 0$,

$$\begin{aligned}\zeta_{00.1} - \zeta_{00.2} &\geq 0\\ \zeta_{01.2} - \zeta_{01.1} &\geq 0\\ \zeta_{10.1} - \zeta_{10.2} &\geq 0\\ \zeta_{11.2} - \zeta_{11.1} &\geq 0,\end{aligned}$$

and

$$\zeta_{00.1} - \zeta_{00.2} - \zeta_{01.2} - \zeta_{10.2} \leq \text{ACE}(B \to C) \leq \zeta_{00.1} + \zeta_{01.1} + \zeta_{11.1} - \zeta_{01.2},$$

or

$$\text{ACE}(A \to C) - \zeta_{01.1} - \zeta_{10.2} \leq \text{ACE}(B \to C) \leq \text{ACE}(A \to C) + \zeta_{11.1} + \zeta_{00.2},$$

which imply that $\vec{\zeta} \geq 0$. The constraints correspond to those derived by Robins (1989) and Manski (1990) and are equivalent to the first two terms each in Eqs.(4.13) and (4.14) (Balke and Pearl, 1997; Pearl, 2000).

Chapter 5

Sampling variability in instrumental variable model

Thus far, the issue of identifiability of a causal effect has occupied centre stage. Only a small portion of the large literature on IVs in the field of Artifical Intelligence has been surveyed in §4.1 but much of it is with little regard for the uncertain nature of the data involved in any study. However there are some exceptions, e.g. Cheng and Small (2006), who discuss confidence intervals for bounded causal effects, and Cai et al. (2008), who find statistical estimators for causal bounds.

It is the aim here to address this problem by developing the analysis of an IV model with a firm statistical grounding. Only the constraints on the trivariate distribution will be considered to illustrate the ideas and all observable variables are assumed to be binary throughout the chapter except in §5.6, which elaborates on the more general case.

Although the causal bounds of Eqs.(4.13) and (4.14) provide a means of partially identifying the causal effect, they are merely parameters of the statistical model and need to be estimated. If the m.l.e. of $\vec{\zeta}$ is available it can

be plugged into Eqs.(4.13) and (4.14) to find the m.l.e. of the bounds. The first problem of interest is therefore to compute the m.l.e. of $\vec{\zeta}$.

Since U is unobserved, one approach is to convert the model to a PO model (cf. §2.2.1) and use the EM algorithm (Dempster, Laird and Rubin, 1977). If the set of relative frequencies is used for estimation of $\vec{\zeta}$ they may not be appropriate since the parameter space of the joint distribution of (A, B, C) is constrained by Eq.(4.12). Alternatively, as will be done here, the m.l.e. can be directly calculated subject to the validity constraints. A technique for the estimation of parameters is presented in §5.1.

The validity of the estimate of the causal bounds relies on the validity of the IV model so a significance test of the IV model is required to add credibility to any conclusions. Although the constraints of Eq.(4.12) do not provide such a test, it is the first step towards such a goal.

In §5.2 a statistical test for the model with binary observable variables is described. Issues involved in the power of the test are also discussed. Such a test is not merely a significance test of conditional independence since the marginal model involving only the observable variables do not imply any such relations.

The results of §5.2 are compared to various simulations in §5.3 to assess their merit. The final addition to the ensemble of statistical tools is a method for finding confidence intervals for α or the causal bounds. A non-parameteric bootstrap approach is described in §5.4 to compute confidence intervals. It utilises the results of §5.1 and §5.2 for the binary case.

5.1 Estimation of parameters

In this section the computation of the m.l.e.'s is considered for binary observable variables but the ideas apply in general. The constraints on the parameters $\vec{\zeta}$ are the inequalities of Eq.(4.12). Let $\mathcal{Z}_0$ be the polytope in $\mathcal{Z}$ which satisfies the constraints in Eq.(4.12), where the simplex $\mathcal{Z}$ is given as

$$\mathcal{Z} = \{\vec{\zeta} : \sum \zeta_{cb.a} = 1, \zeta_{cb.a} \geq 0 \quad \forall a, b, c\}, \tag{5.1}$$

similarly to Eq.(4.3). Therefore

$$\dim(\mathcal{Z}) = \dim(\mathcal{Z}_0).$$

Since $\vec{\zeta} \in \mathcal{Z}_0$ is a necessary and sufficient condition for the distribution of the observable variables $\vec{\zeta}$ to fit the IV model, the m.l.e. of $\vec{\zeta}$, if the model is true, is given by

$$\vec{\pi}^{\,0} = \underset{\vec{\zeta}^* \in \mathcal{Z}_0}{\text{argmax}} \ \mathcal{L}(\vec{\zeta}^*).$$

where $\mathcal{L}(\cdot)$ is the likelihood function. Therefore it is necessary to maximise

$$l(\vec{n} \,|\, \vec{\zeta}) = \sum_{abc} n_{abc} \log P(A, B, C) \propto \sum_{abc} n_{abc} \log \zeta_{cb.a}, \tag{5.2}$$

for $\vec{\zeta} \in \mathcal{Z}_0$, where $l(\cdot) = \log \mathcal{L}(\cdot)$ and n_{abc} is the number of units for which $(A, B, C) = (a, b, c)$. Since $l(\vec{n} \,|\, \vec{\zeta})$ is the sum of $\log(\cdot)$ functions, which are concave, then it is concave. Therefore it is a convex optimisation problem subject to the linear constraints of Eqs.(4.12) and (5.1). The m.l.e. of $\vec{\zeta}$ over

$\mathcal{Z}$ are the relative frequencies,

$$\pi_{cb.a} = \frac{n_{abc}}{n_a} \qquad \forall a, b, c.$$

Therefore, if $\vec{\pi} \in \mathcal{Z}_0$, where

$$\vec{\pi} = (\pi_{00.1}, \pi_{01.1}, \pi_{10.1}, \pi_{11.1}, \pi_{00.2}, \pi_{01.2}, \pi_{10.2}, \pi_{11.2}),$$

the constrained maximum likelihood estimate of $\vec{\zeta}$ is simply the vector of relative frequencies $\vec{\pi}$. However, if $\vec{\pi} \notin \mathcal{Z}_0$ then convex optimisation is needed. The probability of $\vec{\pi} \in \mathcal{Z}_0$ depends on the parameter $\vec{\zeta}$ and is studied in further detail in §5.2.

Using the approach of this section, it is possible to directly compute the m.l.e. of $\vec{\zeta}$. It also follows that in many circumstances, although relative frequencies are used as estimators when ignoring sampling variability, they are valid statistical estimators since they are equivalent to the m.l.e. In cases where the relative frequencies are not the m.l.e., the method can be implemented in standard optimisation software.

5.2 Significance test for the binary instrumental variable model

As mentioned in §4.3, the validity of the causal bounds relies on the validity of the IV model. Hence, before trusting the m.l.e., a significance test of the constraints in Eq.(4.12) is appropriate.

Consider the likelihood ratio test for null and alternative hypotheses

$$H_0 : \vec{\zeta} \in \mathcal{Z}_0$$
$$H_1 : \vec{\zeta} \in \mathcal{Z} \backslash \mathcal{Z}_0.$$

Ideas can be borrowed from the literature on order restricted statistical inference (El Barmi and Dykstra, 1995) since the fundamental theme remains the same, to test convex constraints on the parameter space. The classical null distribution of the log-likelihood ratio test statistic (Wilks, 1938) and the extension by Wald (1943), when the true parameter lies in the alternative hypothesis, does not directly apply here. This is because their null hypotheses are hyperplanes in the parameter space and $\mathcal{Z}_0$ is not. Chernoff (1954) derives the null distribution when the null hypothesis is not necessarily a hyperplane and Feder (1968) extends the work to compute the distribution of the test statistic under the alternative. Some of the results of these papers are relevant as approximations to the distribution of the test statistic.

5.2.1 p-value of test

Let $\vec{\pi} = (\vec{\pi}_1, \vec{\pi}_2)$ and $\vec{\zeta} = (\vec{\zeta}_1, \vec{\zeta}_2)$, where

$$\vec{\pi}_i = (\pi_{00.i}, \pi_{01.i}, \pi_{10.i}, \pi_{11.i})$$
$$\vec{\zeta}_i = (\zeta_{00.i}, \zeta_{01.i}, \zeta_{10.i}, \zeta_{11.i}),$$

and $\textsc{lr}(\vec{\pi}; \vec{\zeta})$ be the log-likelihood ratio between $\vec{\pi}$ and $\vec{\zeta}$. Since

$$\textsc{lr}(\vec{\pi}; \cdot) = -2\{l(\cdot) - l(\vec{\pi})\} = 2\{n_1 \, \textsc{kl}(\vec{\pi}_1; \cdot) + n_2 \, \textsc{kl}(\vec{\pi}_2; \cdot)\},$$

then, from Eq.(5.2),

$$\vec{\pi}^{\,0} = \underset{\vec{\zeta}^{*}\in\mathcal{Z}_0}{\operatorname{argmin}}\; 2\{n_1\,\mathrm{KL}(\vec{\pi}_1;\vec{\zeta}_1^{*}) + n_2\,\mathrm{KL}(\vec{\pi}_2;\vec{\zeta}_2^{*})\} = \underset{\vec{\zeta}^{*}\in\mathcal{Z}_0}{\operatorname{argmax}}\; l(\vec{n}\,|\,\vec{\zeta}^{*}),$$

where

$$\mathrm{KL}(\vec{p};\vec{q}) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right),$$

is the Kullback-Leibler divergence of $\vec{q}$ from $\vec{p}$ and

$$\vec{p} = (p_1,\ldots,p_k), \qquad \vec{q} = (q_1,\ldots,q_k),$$

are probability distributions for a discrete random variable. $\mathrm{KL}(\cdot)$ is a non-symmetric measure of the distance of one point from another in the parameter space or the space of distributions and is not in general equal to the log-likelihood ratio. Since $\vec{\pi}$ and $\vec{\pi}^{\,0}$ are the m.l.e.'s of $\vec{\zeta}$ for $\vec{\zeta}\in\mathcal{Z}$ and $\vec{\zeta}\in\mathcal{Z}_0$ respectively, $\Lambda(\vec{\pi}) = \mathrm{LR}(\vec{\pi};\vec{\pi}^{\,0})$, where $\Lambda(\vec{\pi})$ is the log-likelihood ratio statistic for the test of H_0 vs H_1. To derive the p-value for the test, first consider Theorem 5.1.

Theorem 5.1. (distribution of LR for binary IV model under $\vec{\pi}^{\,0}$) *For the likelihood ratio test of H_0 vs H_1, assuming $\pi_{cb.a} > 0\ \forall\ a,b,c$, the asymptotic distribution of $\Lambda(\vec{\pi})$ under $\vec{\pi}^{\,0}$ is*

$$\lim_{n\to\infty} \mathbb{P}_{\vec{\pi}^{\,0}}\{\Lambda(\vec{\pi}) \geq t\} = \ \tfrac{1}{2}\mathbb{P}(\chi_1^2 \geq t),$$

where t is the observed value of the log-likelihood ratio test statistic and $t = 0$ for $\vec{\pi}\in\mathcal{Z}_0$.

To prove Theorem 5.1, some theorems and definitions are required. Lemma 5.2 shows that $\mathcal{Z}_0$ possesses a convenient property which makes it relatively statistically simple. A face of codimension k of a polytope is defined by changing k of the inequalities which define the polytope to equalities and leaving the complement unaltered.

Lemma 5.2. (convenient property of $\mathcal{Z}_0$) *Assuming $\pi_{cb.a} > 0 \ \forall \ a, b, c$, if $\vec{\pi} \notin \mathcal{Z}_0$ then $\mathcal{Q}(\vec{\pi})$ is a facet (face of codimension 1) of $\mathcal{Z}_0$, where $\mathcal{Q}(\vec{\pi})$ is the face of $\mathcal{Z}_0$ in which $\vec{\pi}^0$ lies.*

Proof of lemma 5.2. If $\Delta_{\text{Q}}^{\text{z}} > 2$ then $\sum_{c,b} \pi_{cb.a}^0 > 1$ for some $a \in \{1, 2\}$, where $\Delta_{\text{Q}}^{\text{z}}$ is the codimension of $\mathcal{Q}(\vec{\pi})$. Therefore since $\mathcal{Q}(\vec{\pi}) \subseteq \mathcal{Z}_0 \subseteq \mathcal{Z}$

$$\Delta_{\text{Q}}^{\text{z}} \in \{0, 1, 2\},$$

If $\Delta_{\text{Q}}^{\text{z}} = 2$ then $\exists \ (a, b, c) \ \pi_{cb.a}^0 = 0$, from Eq.(4.12). Also if $\pi_{cb.a} \neq 0$ and $\pi_{cb.a}^0 = 0$ then $l(\vec{n} \,|\, \vec{\pi}^0) = -\infty$ and $\exists \, \vec{\zeta}^*$ such that $\zeta_{cb.a}^* \neq 0$ and $l(\vec{n} \,|\, \vec{\zeta}^*) > -\infty$. Therefore $\pi_{cb.a}^0 = 0 \Rightarrow \pi_{cb.a} = 0$ (Lauritzen, 1996), which implies that $\Delta_{\text{Q}}^{\text{z}} \neq 2$. Since $\text{LR}(\vec{\pi}, \cdot)$ is a continuous monotone function, $\Delta_{\text{Q}}^{\text{z}} = 0$ only if $\vec{\pi} \in \mathcal{Z}_0$,

$$\Delta_{\text{Q}}^{\text{z}} = 0 \Rightarrow \vec{\pi} \in \mathcal{Z}_0.$$

Therefore

$$\Delta_{\text{Q}}^{\text{z}} \neq 0 \Leftarrow \vec{\pi} \notin \mathcal{Z}_0.$$

□

The regularity conditions needed to derive the asymptotic distribution of the

log-likelihood ratio test statistic are given in Definition 5.3 (van der Vaart, 1998; Drton, 2007). Recall that a statistical model $\mathcal{P}_{\mathcal{Z}}$ is differentiable in quadratic mean at $\vec{\zeta} \in \mathcal{Z}$ if there exists a measurable vector valued function $\dot{l}(\cdot)$ such that as $\vec{\zeta}^* \to \vec{\zeta}$,

$$\int \left\{ \sqrt{p_{\vec{\zeta}^*}(\vec{n})} - \sqrt{p_{\vec{\zeta}}(\vec{n})} - \frac{1}{2}(\vec{\zeta}^* - \vec{\zeta})^T \dot{l}(\vec{\zeta}) \sqrt{p_{\vec{\zeta}}(\vec{n})} \right\}^2 d\nu(\vec{n}) = o(||\vec{\zeta}^* - \vec{\zeta}||^2).$$

Definition 5.3. (regular statistical model) *A statistical model $\mathcal{P}_{\mathcal{Z}}$ is regular at $\vec{\zeta} \in \mathcal{Z} \subseteq \mathbb{R}^q$ if*

- *$\vec{\zeta}$ is an inner point of $\mathcal{Z}$*
- *the model $\mathcal{P}_{\vec{\zeta}}$ is differentiable in quadratic mean at $\vec{\zeta}$ with a nonsingular Fisher information matrix*
- *for every $\vec{\zeta}^1$ and $\vec{\zeta}^2$ in a neighbourhood of $\vec{\zeta}$ and for a measurable function $\dot{l}(\cdot)$ such that $\int \dot{l}(\vec{n})^2 d\mathcal{P}_{\vec{\zeta}}(\vec{n}) < \infty$,*

$$|\log p_{\vec{\zeta}^1}(\vec{n}) - \log p_{\vec{\zeta}^2}(\vec{n})| \leq \dot{l}(\vec{n}) ||\vec{\zeta}^1 - \vec{\zeta}^2||.$$

The result in Theorem 5.4 (Chernoff, 1954; Thm 16.7, van der Vaart, 1998) is used to derive the asymptotic distribution of the log-likelihood ratio test statistic. For simplicity it is assumed that $n_1 = a_1 n$ and $n_2 = a_2 n$ as $n \to \infty$, where a_1 and a_2 are constants.

Theorem 5.4. (asymptotic behaviour of likelihood ratio) *Let the model $(\mathcal{P}_{\vec{\zeta}^*} : \vec{\zeta}^* \in \mathcal{Z} \subseteq \mathbb{R}^q)$ be regular at $\vec{\zeta}$ (cf. Definition 5.3). If the m.l.e.'s $\vec{\pi}^0$ and $\vec{\pi}$ are consistent under $\vec{\zeta}$ and the set $\sqrt{n}(\mathcal{Z}_0 - \vec{\zeta})$ converges (cf. p. 101, van der Vaart, 1998) to the set $\tilde{\mathcal{Z}}_0$ then, as $n \to \infty$, $\Lambda(\vec{\pi})$ converges*

under $\vec{\zeta}_n = \vec{\zeta} + \frac{\lambda}{\sqrt{n}}$ *in distribution to the squared Mahalanobis distance*

$$\inf_{h \in \tilde{\mathcal{Z}}_0} (X-h) I(\vec{\zeta})(X-h),$$

which has the same distribution as the squared Euclidean distance between X *and the linearly transformed set* $I(\vec{\zeta})^{1/2}\tilde{\mathcal{Z}}_0$,

$$\inf_{h \in \tilde{\mathcal{Z}}_0} ||I(\vec{\zeta})^{1/2}X - I(\vec{\zeta})^{1/2}h||^2,$$

where $X \sim N\{\lambda, I(\vec{\zeta})^{-1}\}$ *or*

$$\inf_{h \in \tilde{\mathcal{Z}}_0} ||Z + I(\vec{\zeta})^{1/2}\lambda - I(\vec{\zeta})^{1/2}h||^2,$$

where $Z \sim N(0, I)$ *and* I *is the identity matrix.*

Theorem 5.4 basically states that the test of H_0 vs H_1 is asymptotically equivalent to

$$\begin{aligned} \tilde{H}_0(\vec{\zeta}) &: \lambda \in \tilde{\mathcal{Z}}_0 \\ \tilde{H}_1(\vec{\zeta}) &: \lambda \in \mathbb{R}^q \backslash \tilde{\mathcal{Z}}_0, \end{aligned}$$

based on the observation X from a normal distribution. The hypotheses are functions of $\vec{\zeta}$ since $\tilde{\mathcal{Z}}_0$ depends on $\vec{\zeta}$. To determine what the set $\tilde{\mathcal{Z}}_0$ in Theorem 5.4 is, consider the following definitions from Drton (2007) which are based on Rockafellar and Wets (1998) and Geyer (1994).

Definition 5.5. (tangent cone) *The tangent cone* $T_{\mathcal{Z}}(\vec{\zeta})$ *of the set* $\mathcal{Z} \subseteq \mathbb{R}^q$ *at the point* $\vec{\zeta} \in \mathbb{R}^q$ *is the set of vectors in* $\mathbb{R}^q$ *that are limits of sequences* $k_n(\vec{\zeta}_n - \vec{\zeta})$, *where* k_n *are positive reals and* $\vec{\zeta}_n \in \mathcal{Z}$ *converge to* $\vec{\zeta}$.

Definition 5.6. (Chernoff-regular) *The set $\mathcal{Z} \subseteq \mathbb{R}^q$ is Chernoff regular at $\vec{\zeta}$ if for every vector τ in the tangent cone $T_{\mathcal{Z}}(\vec{\zeta})$ there exist $\epsilon > 0$ and a map $\alpha : [0, \epsilon) \to \mathcal{Z}$ with $\alpha(0) = \vec{\zeta}$ such that $\tau = \lim_{t \to 0+}[\alpha(t) - \alpha(0)]/t$.*

Proof of theorem 5.1. From Eq.(4.12), $\mathcal{Z}_0$ is a semi-algebraic set. Every semi-algebraic set $\mathcal{Z}_0 \subseteq \mathbb{R}^q$ is everywhere Chernoff-regular (Lem. 3.3, Drton, 2007). Therefore, from Geyer (1994), the set $\sqrt{n}(\mathcal{Z}_0 - \vec{\zeta})$ converges under $\vec{\zeta}$ to $T_{\mathcal{Z}_0}(\vec{\zeta})$ in the sense of van der Vaart (1998) and Theorem 5.4. It follows that in Theorem 5.4 $\tilde{\mathcal{Z}}_0 \equiv T_{\mathcal{Z}_0}(\vec{\zeta})$. From Lemma 5.2, $T_{\mathcal{Z}_0}(\vec{\pi}^0)$ is the closed half-space with boundary line through the origin in $\mathbb{R}^8$, even if $\vec{\pi} = \vec{\pi}^0$. The asymptotic behaviour of the parameter space near $\vec{\pi}^0$ is represented in Fig.5.1.

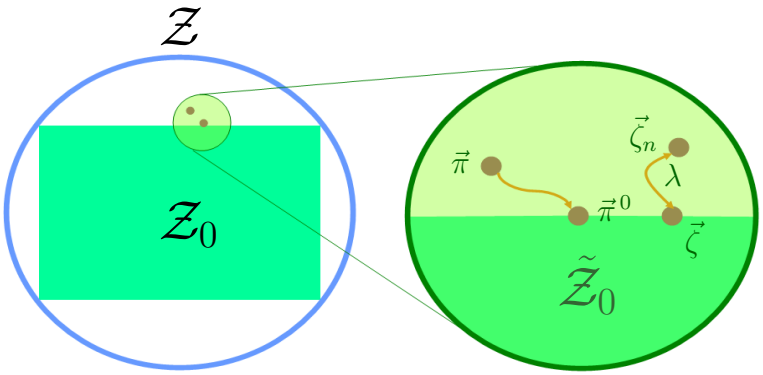


Figure 5.1: Geometry of the local parameter space near $\vec{\pi}^0$. The figure on the right is a scaled view of the region near $\vec{\pi}^0$.

Since $\vec{\pi}^0$ lies on the hyperplane which is the boundary of the tangent cone $T_{\mathcal{Z}_0}(\vec{\pi}^0)$ and the regularity conditions of Definition 5.3 are satisfied,

$$\tilde{\Lambda}(\vec{\pi}) \xrightarrow{d} \inf_{h \in T_{\mathcal{Z}_0}(\vec{\pi}^0)} ||Z - I(\vec{\zeta})^{1/2} h||^2,$$

from Theorem 5.4, where $\tilde{\Lambda}(\vec{\pi})$ is the log-likelihood ratio statistic for the test

$\tilde{H}_0(\vec{\zeta})$ vs $\tilde{H}_1(\vec{\zeta})$, $Z \sim N(0, I)$ and I is the identity matrix. Therefore $\tilde{\Lambda}(\vec{\pi})$ converges in distribution to the squared Euclidean distance between a draw from $N(0, I)$ and $T_{\mathcal{Z}_0}(\vec{\pi}^0)$. Since a standard normal vector is rotationally symmetric, the distribution of the distance to the half-space does not depend on the orientation of the half-space (cf. Ex. 16.9, van der Vaart, 1998). Therefore

$$\tilde{\Lambda}(\vec{\pi}) \xrightarrow{d} ||Z \vee 0||^2,$$

and

$$\lim_{n\to\infty} \mathbb{P}_{\vec{\pi}^0}\{\Lambda(\vec{\pi}) \geq t\} = \lim_{n\to\infty} \mathbb{P}_{\vec{\pi}^0}\{\tilde{\Lambda}(\vec{\pi}) \geq t > 0\} = \tfrac{1}{2}\mathbb{P}(\chi_1^2 \geq t), \tag{5.3}$$

from Chernoff (1954) and van der Vaart (1998), since $\mathbb{P}(||Z \vee 0||^2 > t) = \frac{1}{2}\mathbb{P}(\chi_1^2 \geq t)$ for every $t > 0$. □

Note that the distribution in Eq.(5.3) does not depend on the actual value of $\vec{\pi}^0$ as long as it lies on the hyperplane. If $\vec{\pi} \in \mathcal{Z}_0$ then the p-value of the test of H_0 vs H_1 is simply

$$\sup_{\vec{\zeta}^* \in \mathcal{Z}_0} \mathbb{P}_{\vec{\zeta}^*}\{\Lambda(\vec{\pi}) \geq 0\} = 1, \tag{5.4}$$

but if $\vec{\pi} \notin \mathcal{Z}_0$, the p-value is approximated here by

$$\sup_{\vec{\zeta}^* \in \mathcal{Z}_0} \mathbb{P}_{\vec{\zeta}^*}\{\Lambda(\vec{\pi}) \geq t\} \approx \mathbb{P}_{\vec{\pi}^0}\{\Lambda(\vec{\pi}) \geq t\}, \tag{5.5}$$

since $\vec{\pi}^0$ is asymptotically near the closest point in $\mathcal{Z}_0$ to the true parameter.

Therefore

$$\sup_{\vec{\zeta}^* \in \mathcal{Z}_0} \mathbb{P}_{\vec{\zeta}^*}\{\Lambda(\vec{\pi}) \geq t\} \approx \ \tfrac{1}{2}\mathbb{P}(\chi^2_1 \geq t), \tag{5.6}$$

from Theorem 5.1. In §5.3, simulations are carried out to demonstrate the validity of the results in Eqs.(5.4) and (5.6). The accuracy of the informal but sensible approximation of Eq.(5.5) is of particular concern.

5.2.2 Power function

Previously the distribution of $\Lambda(\vec{\pi})$ was only considered when the true parameter $\vec{\zeta}$ was exactly on the boundary of $\mathcal{Z}_0$, $\bar{\mathcal{Z}}_0$. Here the power function throughout the entire parameter space is of interest and is given in Theorem 5.7.

Theorem 5.7. (power function for LR test for binary IV model) *Let $d(\vec{\zeta};\cdot)$ be the minimum Euclidean divergence from $\vec{\zeta}$ to a set and $d(\vec{\zeta};\bar{\mathcal{Z}}_0) \leq 0$ for $\vec{\zeta} \in \mathcal{Z}_0$. For the likelihood ratio test of H_0 vs H_1, assuming $\pi_{cb.a} > 0\ \forall\ a, b, c$, the asymptotic power function at $\vec{\zeta}_n$ (as defined in Theorem 5.4) is*

- $d(\vec{\zeta};\bar{\mathcal{Z}}_0) < -O(n^{-\frac{1}{2}})$: $\lim_{n\to\infty} \mathbb{P}\{\Lambda(\vec{\pi}) \geq t_w\} \approx 0$
- $d(\vec{\zeta};\bar{\mathcal{Z}}_0) > O(n^{-\frac{1}{2}})$: $\lim_{n\to\infty} \mathbb{P}\{\Lambda(\vec{\pi}) \geq t_w\} \approx 1$
- $d(\vec{\zeta};\bar{\mathcal{Z}}_0) = O(n^{-\frac{1}{2}})$,

$$\lim_{n\to\infty} \mathbb{P}_{\vec{\zeta}_n}\{\Lambda(\vec{\pi}) \geq t_w\} = \begin{cases} 1 - \Phi\{\sqrt[+]{t_w} - \sqrt[\pm]{\lambda}\} & \text{for } t_w > 0 \\ 1 & \text{for } t_w = 0, \end{cases}$$

where $t > t_w$ is the critical region for a test of size $w\%$,

$$\lambda = n_1 d(\vec{\zeta_1}; \bar{\mathcal{Z}}_0)^2 + n_2 d(\vec{\zeta_2}; \bar{\mathcal{Z}}_0)^2,$$

$\sqrt{\lambda}$ is positive for $\vec{\zeta_n} \notin \mathcal{Z}_0$ and negative for $\vec{\zeta_n} \in \mathcal{Z}_0$.

Proof of theorem 5.7. Since $\zeta_{cb.a} > 0 \; \forall \; a, b, c$, from Lemma 5.2 and Theorem 5.4,

- $\vec{\zeta} \in \mathcal{Z}_0 \backslash \bar{\mathcal{Z}}_0$: $\tilde{\mathcal{Z}}_0 \equiv T_{\mathcal{Z}_0}(\vec{\zeta}) \equiv \mathbb{R}^8$
- $\vec{\zeta} \in \mathcal{Z} \backslash \mathcal{Z}_0$: $\tilde{\mathcal{Z}}_0 \equiv T_{\mathcal{Z}_0}(\vec{\zeta}) \equiv \emptyset$
- $\vec{\zeta} \in \bar{\mathcal{Z}}_0$: $\tilde{\mathcal{Z}}_0 \equiv T_{\mathcal{Z}_0}(\vec{\zeta})$ is the closed half-space with boundary line through the origin.

From Theorem 5.4, since the regularity conditions of Definition 5.3 are satisfied, the power function for the test H_0 vs H_1 is asymptotically equal to the power function of $\tilde{H}_0(\vec{\zeta})$ vs $\tilde{H}_1(\vec{\zeta})$ at $\vec{\zeta_n}$ when $\vec{\zeta_n}$ is within a local distance of $O(n^{-\frac{1}{2}})$ of $\vec{\zeta}$,

$$\sqrt{n}\,||\vec{\zeta_n} - \vec{\zeta}\,|| = O(1).$$

Therefore, from the rotational invariance of the normal distribution,

$$\tilde{\Lambda}(\vec{\pi}) \xrightarrow{d} \Lambda = ||(Z + \sqrt[+]{\lambda}\,) \vee 0||^2,$$

where $\sqrt[+]{\lambda}$ is the positive square root of λ and for large n

$$\begin{aligned} \lambda &= \min_{\vec{\zeta^*} \in \bar{\mathcal{Z}}_0} 2\{n_1 \,\mathrm{KL}(\vec{\zeta_1}; \vec{\zeta_1^*}) + n_2 \,\mathrm{KL}(\vec{\zeta_2}; \vec{\zeta_2^*})\} \\ &\approx n_1 d(\vec{\zeta_1}; \bar{\mathcal{Z}}_0)^2 + n_2 d(\vec{\zeta_2}; \bar{\mathcal{Z}}_0)^2. \end{aligned}$$

Λ is the minimum squared distance from the standard normal vector Z to the affine subspace $-I(\vec{\zeta})^{1/2}\lambda + I(\vec{\zeta})^{1/2}\tilde{\mathcal{Z}}_0$. Therefore, for $\vec{\zeta} \in \mathcal{Z}\backslash\mathcal{Z}_0$,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\{\Lambda(\vec{\pi}) \geq t_w\} &= \lim_{n\to\infty} \mathbb{P}\{\tilde{\Lambda}(\vec{\pi}) \geq t_w \mid \mathcal{Q}(\vec{\pi})\} \\
&= \lim_{n\to\infty} \mathbb{P}\left\{ \sqrt{\tilde{\Lambda}(\vec{\pi})} \geq \sqrt[+]{t_w} \;\middle|\; \mathcal{Q}(\vec{\pi}) \right\} \\
&= \begin{cases} 1 - \Phi(\sqrt[+]{t_w} - \sqrt[+]{\lambda}) & \text{for } t_w > 0 \\ 1 & \text{for } t_w = 0, \end{cases}
\end{aligned} \tag{5.7}$$

where $\Phi(\cdot)$ is the distribution function of a standard normal random variable. By similar reasoning for $\vec{\zeta} \in \mathcal{Z}_0$,

$$\mathbb{P}_{\vec{\zeta}\in\mathcal{Z}_0}\{\Lambda(\vec{\pi}) \geq t_w\} \approx \begin{cases} 1 - \Phi(\sqrt[+]{t_w} - \sqrt[-]{\lambda}) & \text{for } t_w > 0 \\ 1 & \text{for } t_w = 0. \end{cases} \tag{5.8}$$

□

When $\vec{\zeta} \in \bar{\mathcal{Z}}_0$, $\lambda = 0$ and Eq.(5.8) is equivalent to Eq.(5.3). In §5.3, simulations are used to illustrate the validity of the results in Theorem 5.7.

5.3 Simulations and plots

The distribution of the log-likelihood ratio statistic for the test of the binary IV model constraints was derived in §5.2. At many points in the derivation, approximations were made based on asymptotic properties and a few, hopefully innocent, simplifying approximations. Therefore a vital step in the development of this test is to check the quality of the approximate distribution of the underlying test statistic by simulation. Potential issues are

- size of the sample at which asymptotic results become acceptable
- power of the test.

Samples were simulated from the distribution corresponding to the value of $\vec{\zeta} \in \mathcal{Z}_0$ given in Table 5.1 and for various values of $\vec{\zeta} \in \mathcal{Z} \backslash \mathcal{Z}_0$, for various sample sizes, which are also given in Table 5.1. More simulations were performed for $\vec{\zeta} \in \mathcal{Z} \backslash \mathcal{Z}_0$ since that region of the power function is of more interest. The number of samples for each simulation is `n` and $n_1 = n_2 = \frac{n}{2}$.

n	$\vec{\zeta} \in \mathcal{Z}_0$								λ
	$\zeta_{00.1}$	$\zeta_{01.1}$	$\zeta_{10.1}$	$\zeta_{11.1}$	$\zeta_{00.2}$	$\zeta_{01.2}$	$\zeta_{10.2}$	$\zeta_{11.2}$	
3360	0.1899	0.7030	0.0375	0.0696	0.5649	0.0631	0.0845	0.2875	0.3685

n	$\vec{\zeta} \in \mathcal{Z} \backslash \mathcal{Z}_0$								λ
	$\zeta_{00.1}$	$\zeta_{01.1}$	$\zeta_{10.1}$	$\zeta_{11.1}$	$\zeta_{00.2}$	$\zeta_{01.2}$	$\zeta_{10.2}$	$\zeta_{11.2}$	
3360	0.6786	0.1012	0.1369	0.0833	0.4167	0.0357	0.3273	0.2202	0.1477
336	0.1012	0.1964	0.6012	0.1012	0.5	0.1012	0.2976	0.1012	3.482
3360	0.3095	0.3155	0.1369	0.2381	0.8988	0.0595	0.0119	0.0298	10.25

Table 5.1: Summary of parameters for simulations with various sample sizes.

For each value of $\vec{\zeta}$, the empirical distribution was computed over all simulated samples and the approximate non-null distribution function was computed according to Theorem 5.7. The plots are given in Figs.5.2 and 5.3. Fortunately all of the plots seem to show that the approximation fits fairly well. However, the second simulation in Fig.5.3 is not as convincing as rest, on account of its much smaller sample size.

From Eqs.(5.7) and (5.8),

$$\lim_{n\to\infty} \mathbb{P}_{\vec{\zeta}}\{\Lambda(\vec{\pi}) = 0\} \approx \begin{cases} \Phi(0 - \sqrt[+]{\lambda}) & \text{for } \vec{\zeta} \in \mathcal{Z} \backslash \mathcal{Z}_0 \\ \Phi(0 - \sqrt[-]{\lambda}) & \text{for } \vec{\zeta} \in \mathcal{Z}_0. \end{cases} \tag{5.9}$$

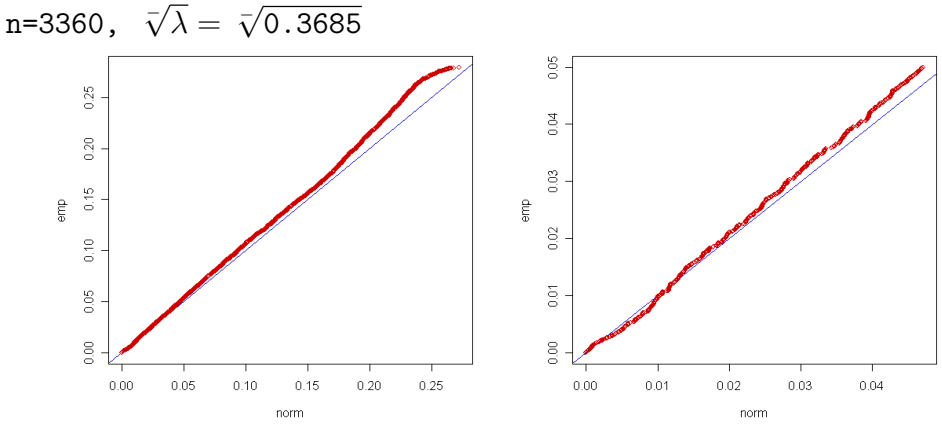


Figure 5.2: Plots of empirical vs approximate distribution function of the square root of the LR statistic. For the plot 10000 simulations, each with sample size n, are used. The plots are for all of the simulated samples (left) and for those with an empirical distribution less than 0.05 (right).

To confirm the validity of Eq.(5.9), random values of $\vec{\zeta} \in \mathcal{Z}$ were chosen and used to simulate 10000 samples, each with n= 336. The empirical estimate, proportion of samples for which $\vec{\pi} \in \mathcal{Z}_0$ or $\Lambda(\vec{\pi}) = 0$, was compared to the approximation in Eq.(5.9). Fig.5.4 shows the plots, where "o" and "*" represent points which correspond to $\vec{\zeta}$ being inside and outside of the polytope respectively. The plot on the left includes all simulated points whereas on the right only those points for which the expected frequencies are all greater than 5 are plotted.

At a glance the plots imply that the asymptotic approximation applies for expected frequencies greater than 5. At the same time the left side plot of Fig.5.4 does contain some reasonable evidence in support of the approximation at low values of the expected frequencies. Overall the results in Eqs.(5.4) and (5.6) and Theorem 5.7 are supported by the simulations.

To address concerns about whether the power of such a test for a sample of

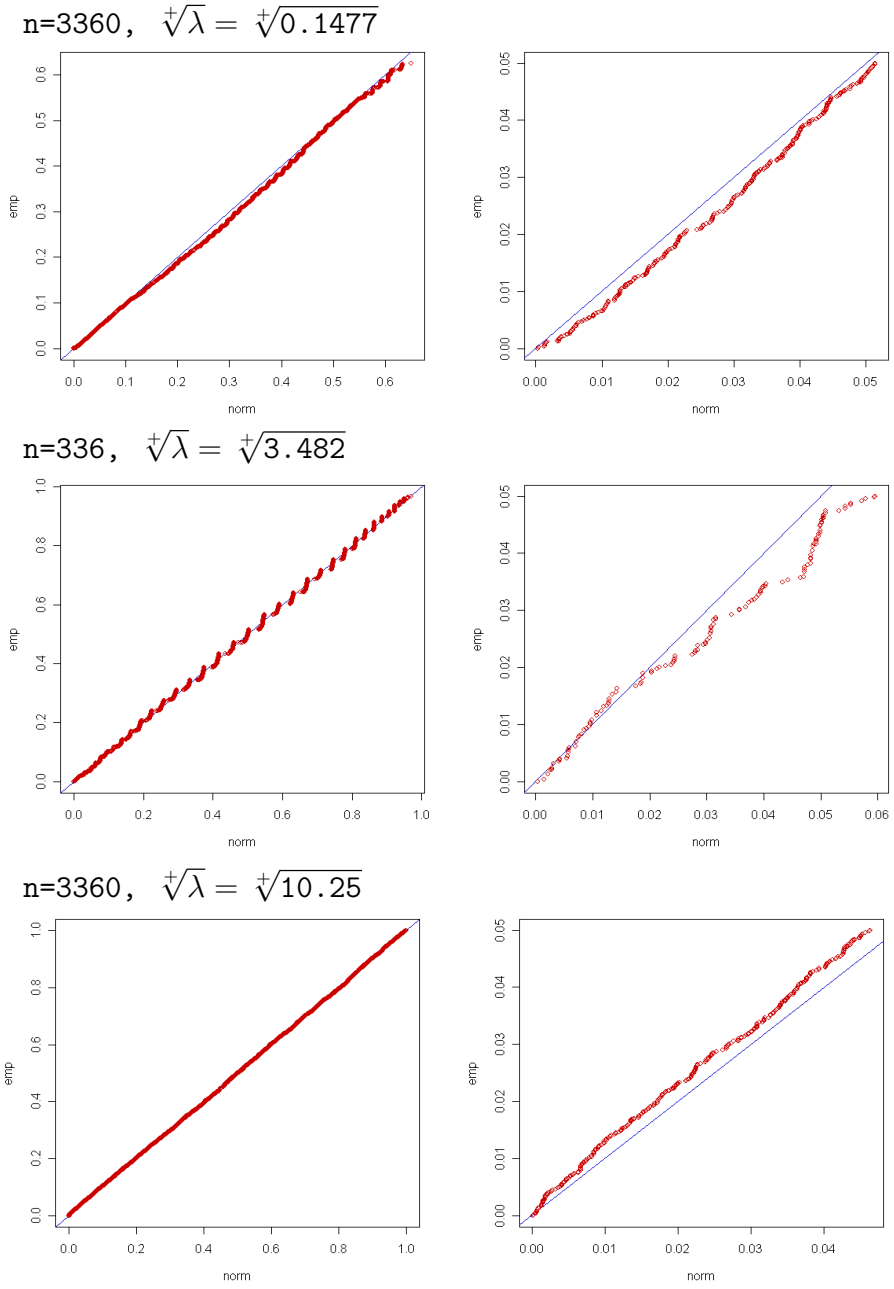


Figure 5.3: Plots of empirical vs approximate distribution function of the square root of the LR statistic. For each plot 10000 simulations, each with sample size `n`, are used.

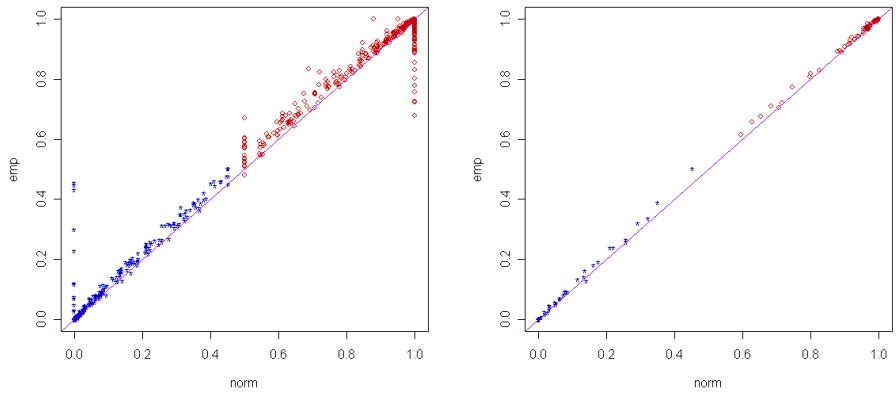


Figure 5.4: Plots of the probability of being inside the polytope, computed empirically and by the approximate normal distribution. Plot on right excludes samples with minimum cell frequency less than 5.

realistic size is ever high enough to be of any use, Fig.5.5 contains plots of the power function, based on the approximation in Theorem 5.7, vs $\sqrt{\lambda}$ for critical regions of tests of various sizes, $\mathtt{w}$.

As it should be, the value of the power function is low when the true value of the parameter $\vec{\zeta} \in \mathcal{Z}_0$, i.e. $\sqrt{\lambda} < 0$. The largest value of $\sqrt{\lambda}$ attained among all of the values of $\vec{\zeta}$ which were used to plot Fig.5.4 is $\sqrt{\lambda}_{\max} = 26.22$. Therefore a large power is attainable for $n = 336$ since the power at $\sqrt{\lambda}_{\max}$ for any size of test in Fig.5.5 is close to 1. Also, increasing the sample size increases $\lambda_{\max}$.

From the plots in Fig.5.5 or Theorem 5.7, it is clear that

$$\mathtt{w} = \sup_{\vec{\zeta}^* \in \mathcal{Z}_0} \mathbb{P}_{\vec{\zeta}^*}\{\Lambda(\vec{\pi}) \geq t_w\} \leq \inf_{\vec{\zeta}^* \in \mathcal{Z} \setminus \mathcal{Z}_0} \mathbb{P}_{\vec{\zeta}^*}\{\Lambda(\vec{\pi}) \geq t_w\},$$

which means that the test is unbiased.

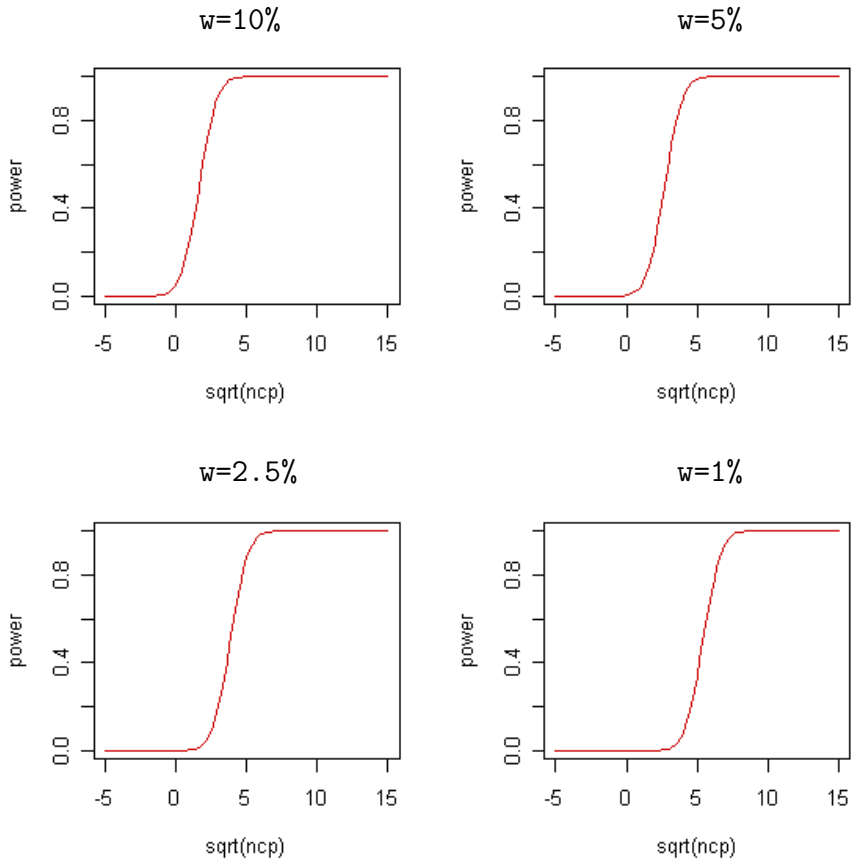


Figure 5.5: Plots of power function as a function of the square root of the minimum KL distance to polytope which defines the null hypothesis, for various test sizes `w`.

5.4 Confidence intervals for causal effects

For a given set of data, if the IV model cannot be rejected by the statistical test, ideally, the estimate of the causal bound would be augmented with a confidence interval (C.I.). In this section a method for producing confidence intervals for the causal bounds is discussed. The method relies on the techniques for estimation and testing discussed in §5.1 and §5.2. The algorithm is:

1. Simulate a large number of bootstrap samples.
2. For each sample, test IV model,
   if reject IV model
   set bounds to $[-1, 1]$,
   else
   compute m.l.e. of $\vec{\zeta}$ and use to compute causal bounds.
3. Order the upper bounds and ignore the highest $\frac{1}{2}w\%$.
4. Order the lower bounds and ignore the lowest $\frac{1}{2}w\%$.
5. The $(1-w)\%$ C.I. is the range of the remaining values.

The IV model is tested since the bounds only apply if the IV constraints hold. The C.I. is for the causal bounds and is thus a conservative C.I. for α.

5.5 Data analysis for binary IV model

The analysis of the data from the Lipid Research Clinics (LRC) coronary data and Vitamin A Supplementation (VAS) study described in §4.3.3 was carried out ignoring the uncertainty in the data. Here the data sets are analysed while taking into account sampling variability, in the usual statistical style.

Following the classical approach, firstly, the test statistic and the p-value are computed, according to Eqs.(5.4) and (5.6). Estimates and confidence intervals are computed for the data sets. For both studies the m.l.e. of ζ is the vector of relative frequencies since they satisfy the instrumental inequalities. The confidence intervals are calculated by the boostrap algorithm of §5.4, with 1000 bootstrap samples. The results of the testing and estimation are

	n	$\vec{n}$							
		$n_{00.1}$	$n_{01.1}$	$n_{10.1}$	$n_{11.1}$	$n_{00.2}$	$n_{01.2}$	$n_{10.2}$	$n_{11.2}$
LRC	337	158	0	14	0	52	23	12	78
VAS	23682	74	0	11514	0	34	12	2385	9663

Table 5.2: Data from Lipid Research Clinics (LRC) coronary data and Vitamin A Supplementation (VAS) study described in §4.3.3. These are the count data which correspond to the relative frequencies in Tables 4.1 and 4.2.

given in Table 5.3, where $\hat{\alpha}^l$ and $\hat{\alpha}^u$ are the estimates of the lower and upper bound on α respectively. From Table 5.3, non-trivial C.I.'s were obtained.

	t_{obs}	p-value	$[\hat{\alpha}^l, \hat{\alpha}^u]$	95% C.I. for α
LRC	0	1	$[0.392, 0.780]$	$[0.307, 0.834]$
VAS	0	1	$[-0.1946, 0.0054]$	$[-0.2016, 0.0070]$

Table 5.3: Results of statistical analysis of Lipid Research Clinics (LRC) coronary data and Vitamin A Supplementation (VAS) study.

In the simulation from the LRC data, 370 samples did not satisfy the instrumental inequalities but only 21 were rejected at a 5% level. However, for the VAS data all of the bootstrap samples of data satisfied the instrumental inequalities. This is because the Kullback-Leibler divergence of the boundary of the polytope from the unconstrained m.l.e. (relative frequencies) is larger for the VAS data than the LRC data.

Both the data from the Lipid Research Clinics Program and the Vitamin A Supplementation study contain relative frequencies less than 5, which can affect the validity of the asymptotic results used. However all of the results still hold as long as the number of components of $\vec{n}$ which are zero is less than four. Some sets of artificial data which do not possess this weakness

are given in Table 5.4.

	n	$\vec{n}$							
		$n_{00.1}$	$n_{01.1}$	$n_{10.1}$	$n_{11.1}$	$n_{00.2}$	$n_{01.2}$	$n_{10.2}$	$n_{11.2}$
(i)	3360	1140	170	230	140	700	60	550	370
(ii)	3360	1509	35	51	85	1210	148	93	229
(iii)	3360	176	299	629	576	1346	138	44	152

Table 5.4: Hypothetical data from multiple studies similar to the studies described in §4.3.3.

The data is analysed and the results are given in Table 5.5. The m.l.e. of ζ is the vector of relative frequencies for (ii) since they satisfy the instrumental inequalities but estimation for (i) uses the convex optimisation approach of §5.1.

	t_{obs}	p-value	$[\hat{\alpha}^l, \hat{\alpha}^u]$	95% C.I. for α
(i)	$\sqrt[+]{0.1477}$	0.535	$[-0.0189, 0.5378]$	$[-1, 1]$
(ii)	0	1	$[0.0387, 0.8565]$	$[0.0161, 0.8732]$
(iii)	$\sqrt[+]{128.2}$	7.66×10^{-4}	$*$	$*$

Table 5.5: Results of statistical analysis of hypothetical data sets.

Estimation is not carried out for data set (iii) since the statistical significance test for the binary IV model rejects it at a 5% level. Although data set (i) only produces a trivial 95% confidence interval for α, it produces an 82% confidence interval of $[-0.0970, 0.5976]$. This may be useful if inference is based on accumulated evidence from multiple studies.

Data set (i) is also a very interesting example since the relative frequencies do not satisfy the instrumental inequalities but yet the model cannot be rejected, because of sample variability. It illustrates the value of a proper

statistical analysis of the IV model.

5.6 Significance test for non-binary instrumental variable model

In §5.2, a statistical test was described for the constraints in Eq.(4.12). Here, those results are extended for non-binary models in which the constraints on the parameters define a polytope in the parameter space.

Consider Fig.5.6, where the large circle and the polytope in the centre represent $\mathcal{Z}$ and $\mathcal{Z}_0$ respectively. If $\vec{\pi}$ lies in the subspace of $\mathcal{Z}$ represented by a dotted line, then the m.l.e. under H_0 is the point of intersection of the line and $\mathcal{Z}_0$, $\vec{\pi}^0$. If $\vec{\pi}$ lies in a particular partition then the m.l.e. under H_0 lies in the facet of the polytope corresponding to the partition.

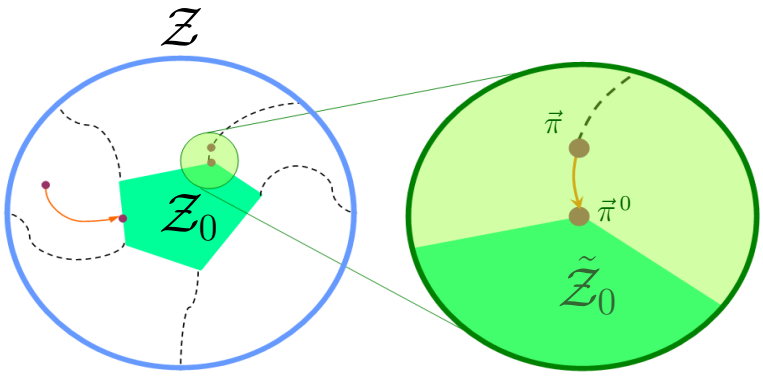


Figure 5.6: Local geometry of parameter space near $\vec{\pi}^0$. Points along each dotted line are nearest to the point of the polytope where the line meets it. The figure on the right is a scaled view of the region near $\vec{\pi}^0$. The figure on the left shows two points and the head of each arrow represents the nearest point in the polytope to them.

Similarly to the binary case, the asymptotic distribution of $\Lambda(\vec{\pi})$ is equivalent

to that of $\tilde{\Lambda}(\vec{\pi})$ and is determined by the tangent cone at $\vec{\pi}^0$, $T_{\mathcal{Z}_0}(\vec{\pi}^0)$. Unlike the binary model, $\mathcal{Q}(\vec{\pi})$ is not necessarily a hyperplane in $\mathcal{Z}$ since Lemma 5.2 does not apply, so $T_{\mathcal{Z}_0}(\vec{\pi}^0)$ is not necessarily a half-space through the origin. However it will be defined by a collection of linear inequalities through the origin.

Without loss of generality, the polytopes $\mathcal{Z}_0$ and $\mathcal{Q}(\vec{\pi})$ are of the form

$$\mathcal{Z}_0 = \left[\vec{\zeta} \in \mathcal{Z} : h_j(\vec{\zeta}) \leq 0 \text{ for } j \in \{1, \ldots, s\}\right],$$

and

$$\mathcal{Q}(\vec{\pi}) = \left[\vec{\zeta} \in \mathcal{Z} : \begin{cases} h_j(\vec{\zeta}) = 0 \text{ for } j \in \{1, \ldots, q\} \\ h_j(\vec{\zeta}) \leq 0 \text{ for } j \in \{q+1, \ldots, s\} \end{cases}\right],$$

where

$$h_j(\vec{\zeta}) = \sum_{abc} \zeta_{cb.a}\, \beta_{[j,cb.a]},$$

and the β's are known constants. It should be noted that the constraints, including Eq.(4.12), can always be expressed in this form but this is not necessary for the theory to apply since the tangent cone will pass through the origin regardless. It follows that the null hypothesis is

$$T_{\mathcal{Z}_0}(\vec{\pi}^0) = \left[\vec{\zeta} \in \mathcal{Z} : h_j(\vec{\zeta}) \leq 0 \text{ for } j \in \{1, \ldots, q\}\right].$$

To derive the distribution of $\tilde{\Lambda}(\vec{\pi})$, consider Lemma 5.8 (Thm 4.2, El Barmi and Dykstra, 1995).

Lemma 5.8. (distribution of LR) *Define the following hypotheses*

$$\begin{aligned} H_e &: h_j(\vec{p}\,) = 0 \quad j = 1, \ldots, q \\ H_\omega &: h_j(\vec{p}\,) \leq 0 \quad j = 1, \ldots, q \\ H_\Omega &: \vec{p} \in \mathcal{P}, \end{aligned}$$

where

$$h_j(\vec{p}\,) = \textstyle\sum_{i=1}^{k} p_i \beta_{ji},$$

$\vec{p} = (p_1, \ldots, p_k)$ *is the vector of parameters for a multinomial distribution, the* β*'s are known constants,* $q \leq k-1$ *and* $\mathcal{P}$ *is the set of possible* $\vec{p}$ *in* $[0,1]^k$*. If* $\vec{p}^{\,0} \in H_e$ *and* T_{12} *is the log-likelihood ratio test statistic for the test of* H_ω *vs* $H_\Omega - H_\omega$ *for any real* t*,*

$$\lim_{n\to\infty} \mathbb{P}_{\vec{p}^{\,0}}(T_{12} \geq t) = \sum_{j=0}^{q} w_j(\vec{p}^{\,0}) \mathbb{P}(\chi_j^2 \geq t),$$

where $\chi_0^2 \equiv 0$*. Let* $\mathbf{H}$ *be the* $\{(k-1)\times q\}$ *matrix with* (i,j)*th element* $(\beta_{ji} - \beta_{jk})$ *and* $\mathbf{B}$ *be the* $\{(k-1)\times(k-1)\}$ *matrix with* (i,j)*th element* $\{p_i^0(\delta_{ij} - p_j^0)\}$*, where* δ_{ij} *is the Kronecker delta function.*
For $1 \leq j \leq q-1$ *and* $\vartheta \in \Theta$*,*

$$w_j(\vec{p}^{\,0}) = \sum_{\vartheta : card(\vartheta) = j} \mathbb{P}\left[\mathbf{N}\left\{\mathbf{0}, \mathbf{\Sigma}_1(\vartheta)\right\} \geq \mathbf{0}\right] \mathbb{P}\left[\mathbf{N}\left\{\mathbf{0}, \mathbf{\Sigma}(\vartheta)\right\} \geq \mathbf{0}\right],$$

where Θ *is the class of all subsets of* $\{1, \ldots, q\}$*,* $\vartheta^c = \{1, \ldots, q\} \backslash \vartheta$*,* $\mathbf{H}(\vartheta)$ *is*

the submatrix of $\mathbf{H}$ *with columns determined by the indices in* ϑ *and*

$$\begin{aligned}\boldsymbol{\Sigma}_1(\vartheta) &= \{\mathbf{H}^T(\vartheta)\mathbf{B}\mathbf{H}(\vartheta)\}^{-1} \\ \boldsymbol{\Sigma}(\vartheta) &= \mathbf{H}^T(\vartheta^c)\{\mathbf{B} - \mathbf{B}\mathbf{H}(\vartheta)\boldsymbol{\Sigma}_1(\vartheta)\mathbf{H}^T(\vartheta)\}\mathbf{H}^T(\vartheta^c).\end{aligned}$$

Also,

$$\begin{aligned}w_0(\vec{p}^{\,0}) &= \mathbb{P}\{\mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_0) \geq \mathbf{0}\} \\ w_q(\vec{p}^{\,0}) &= 1 - \textstyle\sum_{j=0}^{q-1} w_j(\vec{p}^{\,0}),\end{aligned}$$

where

$$\boldsymbol{\Sigma}_0 = \mathbf{H}^T\mathbf{B}\mathbf{H}.$$

The distribution of $\tilde{\Lambda}(\vec{\pi})$ can be obtained from Lemma 5.8 by redefining the matrices $\mathbf{H}$ and $\mathbf{B}$ since $\vec{\zeta}$ is the vector of parameters for all of the multinomial distributions conditional on each value of A. Note that $\mathbf{H}$ and $\mathbf{B}$ only include the independent parameters. $\mathbf{B}$ is simply the expected information matrix at $\vec{\zeta}$, $\mathcal{I}(\vec{\zeta})$.

Let the sample space of A be $\{1, \ldots, m_a\}$ and similarly for B and C. Since

$$h_j(\vec{\zeta}\,) = \sum_{A=a}\left\{\sum_{\neq(m_b,m_c)} \zeta_{cb.a}\,\beta_{[j,cb.a]} + \beta_{[j,m_cm_b.a]}\left(1 - \sum_{\neq(m_b,m_c)} \zeta_{cb.a}\right)\right\},$$

$\mathbf{H}$ is the matrix of

$$\frac{\partial h_j(\vec{\zeta}\,)}{\partial \zeta_{cb.a}} = \beta_{[j,cb.a]} - \beta_{[j,m_cm_b.a]}$$

for $(c, b) \neq (m_b, m_c)$.

Chapter 6

Efficiency of causal estimators

Much attention has been given to IVs, which are basically supplementary variables that are recorded to achieve identifiability of a causal quantity, as mentioned in Chapter 3. This very useful idea can also be extended to improve the efficiency of estimators. The technique of analysis of covariance, which involves observing a covariate to improve the precision of an estimator, originates from Fisher (1932) and has been well developed throughout the literature (Cochran, 1957; Cox and McCullagh, 1982). The similar concept of observing intermediate variables, possibly of no practical importance, to increase the precision of estimators was introduced by Cox (1960). Cox (1960) considers a model which involves a treatment variable, T, an intermediate variable, L, and a response variable, R. The total regression of R on T is of interest and the model is represented by the DAG in Fig.6.1.

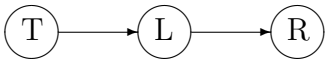


Figure 6.1: DAG for the model discussed in Cox (1960), which involves a treatment (T), intermediate (L) and response (R) variable.

Consider linear regressions of the form

$$R \,|\, L = \gamma_{rl} L + \epsilon_{r|l}$$
$$L \,|\, T = \gamma_{lt} T + \epsilon_{l|t},$$

where

$$\epsilon_{a|b} \sim N(0, \sigma_{aa|b}),$$

and $\sigma_{ab|c} = \mathrm{cov}(A, B \,|\, C)$. This can be equivalently stated as

$$\begin{aligned} R \,|\, L &\sim N(\gamma_{rl} L, \sigma_{rr|l}) \\ L \,|\, T &\sim N(\gamma_{lt} T, \sigma_{ll|t}), \end{aligned} \tag{6.1}$$

where γ_{ab} is the regression coefficient of B in the regression of A on B. Assuming Fig.6.1 is a causal DAG, $\mathbb{E}(R \,|\, L) = \mathbb{E}(R \,||\, L)$ and $\mathbb{E}(L \,|\, T) = \mathbb{E}(L \,||\, T)$ but it is not necessary to assume that interventions in L and C are possible, only in T. Therefore from Eqs.(6.1)

$$\mathrm{ACE}(T \to R) = \gamma_{rt} = \gamma_{rl}.\gamma_{lt}, \tag{6.2}$$

where $\mathrm{ACE}(\cdot)$ is defined in Eq.(1.2). Cox (1960) compared the estimators based on (R, T) and (R, L, T) data and showed that the m.l.e.'s always satisfy the relationship

$$\mathbb{V}_n^{\infty}(\hat{\gamma}_{rt}) \geq \mathbb{V}_n^{\infty}(\hat{\gamma}_{rl}.\hat{\gamma}_{lt}),$$

where $\mathbb{V}_n^{\infty}(\cdot)$ is the asymptotic variance sequence and $\hat{\gamma}$ is the m.l.e. of γ, and that appreciable additional precision can be potentially obtained from observing the intermediate variable L, in addition to (R, T). Consider the DAG in

Fig.6.2 for a multivariate Gaussian model with variables $(X_1, \ldots, X_k)$, which is a generalisation of Fig.6.1.

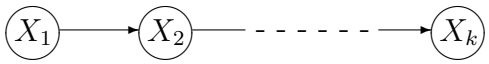


Figure 6.2: DAG for a more general version of the model discussed in Cox (1960).

The following expression holds for Fig.6.2 when all of the X's are recorded,

$$\frac{1}{\mathbb{I}_n(\gamma_{k,1})} = \sum_{i=1}^{k-1} \frac{1}{\mathbb{I}_n(\gamma_{i+1,i})}, \tag{6.3}$$

where

$$\mathbb{I}_n(\gamma) = \gamma^2 \frac{I_n(\gamma)}{n} = \left\{ \frac{n\mathbb{V}_n^{\infty}(\hat{\gamma})}{\gamma^2} \right\}^{-1} = \frac{\rho^2}{1-\rho^2}, \tag{6.4}$$

$I_n(\cdot)$ is the information matrix, ρ is the correlation coefficient and $\gamma_{j,i}$ is the regression coefficient of X_i in the regression of X_j on X_i, with proof in Appendix B.1.

$\mathbb{I}_n(\cdot)$ is the inverse of the square of the asymptotic standard error of the m.l.e. per unit value of the parameter (standardised asymptotic variance) or the standardised Fisher information. It represents the ratio of the proportion of the variation in the dependent variable explained to unexplained by the independent variable. $\mathbb{I}_n(\cdot)$ is equivalent to '*Cohen's f*' (Cohen, 1988), which is a dimensionless measure of effect size commonly used in the behavioural sciences.

Eq.(6.3) demonstrates additivity of the standardised variation in the estimators and will be used later to provide an intuitive interpretation of expressions

for relative efficiency.

Although the discussion of Cox (1960) was aimed at regression coefficient estimators, assuming the DAGs are causal, the parameters are equivalent to causal effects and will be interpreted as the latter. In this chapter, the variance of multiple estimators of causal effects are investigated. In models where multiple expressions for causal effects exist, criteria must be specified to determine which expression to use. The choice of expression can affect the efficiency of the estimator. Analagously to Cox (1960), it is demonstrated here that increased precision can be obtained by observing supplementary variables, whether intermediate or covariate.

Since the cost of data collection is always an important issue there may not always be the option to observe additional variables. In certain studies there may be restrictions on the number of variables recorded, e.g. funding may only be available to measure covariates or intermediate variables but not both. Such circumstances are also considered.

In §6.1 a model in which there exist multiple expressions for a particular causal effect is described and the expressions are derived. The asymptotic variances of the estimators for each expression is given in §6.2 and compared in §6.3.

6.1 Multiple expressions for causal effect

Consider the model represented by the DAG in Fig.6.3, where C is a confounder (or covariate) and the joint distribution of (R, L, T, C) is multivariate normal. It is assumed throughout that there exists a known intermediate

variable, L, and a known covariate, or concomitant variable, C, which satisfy the assumptions of the model in Fig.6.3.

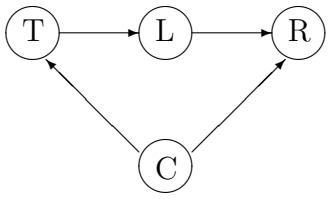


Figure 6.3: Causal DAG for a model with an intermediate (L) and covariate (C) variable.

Let

$$\begin{aligned} R \,|\, L, C &\sim N(\gamma_{rl|c} L + \gamma_{rc|l} C, \sigma_{rr|l,c}) \\ L \,|\, T &\sim N(\gamma_{lt} T, \sigma_{ll|t}) \\ T \,|\, C &\sim N(\gamma_{tc} C, \sigma_{tt|c}), \end{aligned} \tag{6.5}$$

denote the relationships between the nodes and their parents in Fig.6.3 and, without loss of generality, assume $\mathbb{E}(C) = 0$.

It should be noted that Eqs.(6.5) also represent the intervention distributions since

$$\begin{aligned} \mathbb{P}(R \,|\, L, C) &= \mathbb{P}(R \,||\, L, C) \\ \mathbb{P}(L \,|\, T) &= \mathbb{P}(L \,||\, T) \\ \mathbb{P}(T \,|\, C) &= \mathbb{P}(T \,||\, C), \end{aligned}$$

from the causal DAG in Fig.6.3.

Using the 'back-door' formula (Pearl, 1993), the 'front-door' formula (Pearl, 1995a) and the 'extended back-door' formula (Lauritzen, 2001), multiple expressions for ACE($T \to R$) can be obtained in terms of observational parameters.

Let $\mu_{a|c} = \mathbb{E}(A \,|\, C)$ and $\gamma_{ab|c}$ be the partial regression coefficient of B in the

regression of A on B and C. It follows that

- Back-door (R, C, T distribution):

$$\begin{aligned}\mathbb{E}(R \,||\, T^*) &= \mathbb{E}_c\{\mathbb{E}(R \,|\, C, T^*)\} \\ &= \gamma_{rc|t}\mu_c + \gamma_{rt|c}T^* \\ \Rightarrow \text{ACE}(T \to R) &= \gamma_{rt|c},\end{aligned}$$

- Front-door (R, L, T distribution):

$$\begin{aligned}\mathbb{E}(R \,||\, T^*) &= \mathbb{E}_l[\mathbb{E}_t\{\mathbb{E}(R \,|\, L, T)\} \,|\, T^*] \\ &= \mathbb{E}_l[\gamma_{rl|t}L + \gamma_{rt|l}\mu_t \,|\, T^*] \\ &= \gamma_{rl|t}\mathbb{E}(L \,|\, T^*) + \gamma_{rt|l}\mu_t \\ &= \gamma_{rl|t}\gamma_{lt}T^* + \gamma_{rt|l}\mu_t \\ \Rightarrow \text{ACE}(T \to R) &= \gamma_{rl|t}\gamma_{lt},\end{aligned}$$

- Extended Back-door (R, L, C, T distribution):

$$\begin{aligned}\mathbb{E}(R \,||\, T^*) &= \mathbb{E}_l[\mathbb{E}_c\{\mathbb{E}(R \,|\, L, C)\} \,|\, T^*] \\ &= \mathbb{E}_l[\gamma_{rl|c}L + \gamma_{rc|l}\mu_c \,|\, T^*] \\ &= \gamma_{rl|c}\mathbb{E}(L \,|\, T^*) + \gamma_{rc|l}\mu_c \\ &= \gamma_{rl|c}\gamma_{lt}T^* + \gamma_{rc|l}\mu_c \\ \Rightarrow \text{ACE}(T \to R) &= \gamma_{rl|c}\gamma_{lt}.\end{aligned}$$

Hence, in particular, when $\gamma_{lt} \neq 0$, $\gamma_{rl|c} = \gamma_{rl|t}$. It can be easily deduced that, unlike the model in Fig.6.1, Eq.(6.2) does not hold because of the confounding by C. In fact, there is no expression for $\text{ACE}(T \to R)$ in terms of the joint

observational distribution of R and T only so the problem of interest here is what factors affect the efficiency of the estimator when L, C or both are recorded in addition to the (R, T) data. Similarly to Cox (1960), Lauritzen (2001) suggests a comparison of the efficiency of the estimators based on the multiple expressions for the causal effect

$$\gamma_{rt|c} = \gamma_{rl|t}.\gamma_{lt} = \gamma_{rl|c}.\gamma_{lt}, \tag{6.6}$$

which are all path coefficients (Wright, 1921) for the $T \to R$ path.

6.2 Asymptotic variance of causal estimators

The asymptotic variance sequence of the estimators of ACE($T \to R$) in Eq.(6.6) are given in this section. They are

- R, C and T recorded

$$\mathbb{V}_n^{\infty}(\hat{\gamma}_{rt|c}) = \frac{1}{n}\left(\frac{\sigma_{rr|l,c}}{\sigma_{tt|c}} + \gamma_{rl|c}^2\frac{\sigma_{ll|t}}{\sigma_{tt|c}}\right), \tag{6.7}$$

- R, L and T recorded

$$\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) = \frac{1}{n}\left(\gamma_{lt}^2\frac{\sigma_{rr|l,t}}{\sigma_{ll|t}} + \gamma_{rl|t}^2\frac{\sigma_{ll|t}}{\sigma_{tt}}\right), \tag{6.8}$$

- R, L, C and T recorded

$$\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}) = \frac{1}{n}\left(\gamma_{lt}^2\frac{\sigma_{rr|l,c}}{\sigma_{ll|c}} + \gamma_{rl|c}^2\frac{\sigma_{ll|t}}{\sigma_{tt}}\right), \tag{6.9}$$

with proofs in Appendix B.2. These expressions can be compared to determine the conditions under which certain variables should be recorded to obtain the most efficient estimator. To facilitate a more intuitive comparison, the asymptotic variances are expressed in terms of correlation coefficients,

- R, C and T recorded

$$n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rt|c}) = \gamma_{rl|c}^2.\gamma_{lt}^2 \left[\frac{1-\rho_{rl|c}^2}{\rho_{rl|c}^2} \left\{ 1 + \frac{1-\rho_{lt}^2}{\rho_{lt}^2(1-\rho_{tc}^2)} \right\} + \frac{1-\rho_{lt}^2}{\rho_{lt}^2(1-\rho_{tc}^2)} \right], \quad (6.10)$$

- R, L and T recorded

$$n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) = \gamma_{rl|t}^2.\gamma_{lt}^2 \left(\frac{1-\rho_{rl|t}^2}{\rho_{rl|t}^2} + \frac{1-\rho_{lt}^2}{\rho_{lt}^2} \right), \quad (6.11)$$

- R, L, C and T recorded

$$n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}) = \gamma_{rl|c}^2.\gamma_{lt}^2 \left(\frac{1-\rho_{rl|c}^2}{\rho_{rl|c}^2} + \frac{1-\rho_{lt}^2}{\rho_{lt}^2} \right), \quad (6.12)$$

where $\rho_{ab|c}$ is the correlation between A and B conditioning on a fixed value of C. The derivations of Eqs.(6.10), (6.11) and (6.12) are given in Appendix B.3 and a comparison in §6.3.

6.3 Comparison of causal estimators

In this section the asymptotic variance of the various estimators in Eqs.(6.10), (6.11) and (6.12) are compared to determine which variables should ideally

be recorded in addition to R and T. For two estimators $\hat{\gamma}_1$ and $\hat{\gamma}_2$, define

$$\nabla\mathbb{I}(\hat{\gamma}_1\,;\,\hat{\gamma}_2) = \mathbb{I}_n(\gamma_1)/\mathbb{I}_n(\gamma_2),$$

and

$$\Delta\mathbb{I}(\hat{\gamma}_1\,;\,\hat{\gamma}_2) = \frac{1}{\mathbb{I}_n(\gamma_2)} - \frac{1}{\mathbb{I}_n(\gamma_1)},$$

where $\nabla\mathbb{I}(\cdot)$ is the relative information and $\Delta\mathbb{I}(\cdot)$ is the difference in the inverse of the Fisher information. Both are dimensionless measures of contrast in effect size and will be used for formal comparisons of the efficiency of two estimators. The particular measure used will be chosen according to which has a more intuitive interpretation. If $\gamma_1 = \gamma_2$ then $\nabla\mathbb{I}(\hat{\gamma}_1\,;\,\hat{\gamma}_2)$ is the relative asymptotic efficiency of $\hat{\gamma}_1$ to $\hat{\gamma}_2$ and

$$\nabla\mathbb{I}(\hat{\gamma}_1\,;\,\hat{\gamma}_2) \geq 1 \quad \Leftrightarrow \quad \Delta\mathbb{I}(\hat{\gamma}_1\,;\,\hat{\gamma}_2) \geq 0,$$

implies that $\hat{\gamma}_1$ is more asymptotically efficient. The larger the value of both $\nabla\mathbb{I}(\hat{\gamma}_1\,;\,\hat{\gamma}_2)$ and $\Delta\mathbb{I}(\hat{\gamma}_1\,;\,\hat{\gamma}_2)$ the more efficient $\hat{\gamma}_1$ is compared to $\hat{\gamma}_2$.

6.3.1 Supplement C with L or not

In a study in which R, C and T are observed, or recorded, the purpose of additionally measuring L is to remove variation entering the system after T has exerted its effect. However, if an investigator suspects that L contains little information about $\text{ACE}(T \to R)$ that is not already available in R, C and T then they would opt to omit it.

Formal criteria for the omission of L are developed by comparing Eqs.(6.7)

and (6.9). Measures of the importance of L are the relative asymptotic efficiency of the estimators

$$\nabla\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c}) = \frac{\frac{1}{\mathbb{I}^*_n(\gamma_{rl|c})} + \frac{1}{\mathbb{I}^*_n(\gamma_{lt})}}{\frac{1}{\mathbb{I}_n(\gamma_{rl|c})} + \frac{1}{\mathbb{I}_n(\gamma_{lt})}}, \tag{6.13}$$

and

$$\Delta\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c}) = \frac{1}{\mathbb{I}_n(\gamma_{rl|c})}\left\{\frac{1+\mathbb{I}_n(\gamma_{tc})}{\mathbb{I}_n(\gamma_{lt})}\right\} + \frac{\mathbb{I}_n(\gamma_{tc})}{\mathbb{I}_n(\gamma_{lt})}, \tag{6.14}$$

from Eqs.(6.10) and (6.12), where $\mathbb{I}_n(\cdot)$ is defined in Eq.(6.4),

$$\mathbb{I}^*_n(\gamma_{rl|c}) = \frac{\mathbb{I}_n(\gamma_{rl|c})}{1+1/\mathbb{I}^*_n(\gamma_{lt})}, \qquad \mathbb{I}^*_n(\gamma_{lt}) = \frac{\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}.$$

Eq.(6.14) immediately follows from Eqs.(6.10) and (6.12) but an alternative proof is given in Appendix B.4. From Eqs.(6.13) and (6.14), since $\mathbb{I}_n(\cdot) \geq 0$,

$$\nabla\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c}) \geq 1, \quad \Delta\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c}) \geq 0,$$

which agrees with intuition since it means that the omission of an intermediate variable does not increase the efficiency of the estimator.

Intuitively, the denominator in Eq.(6.13) follows the additive formula of Eq.(6.3) but the corresponding quantities in the numerator are inflated by the omission of L, i.e. less information about the $T \to L$ and $L \to R$ edges are available. As expected, the inclusion of L in analyses cannot worsen the efficiency of the estimator because L removes variation entering the system after treatment allocation.

The larger the value of $\Delta\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c})$ the more beneficial it is to observe L. Therefore, from Eq.(6.14), as $\mathbb{I}_n(\gamma_{rl|c})$ or $\mathbb{I}_n(\gamma_{lt})$ increases the benefit of L decreases. This is because L contains less additional information than what is already available in T and R.

When

$$\rho_{tc} = 0,$$

and

$$\rho_{rl|c} = \rho_{rl} \quad \Leftrightarrow \quad \rho_{rc|l} = 0,$$

i.e. there is no confounding, Eq.(6.13) is equivalent to Eq.(18) of Cox (1960).

6.3.2 Supplement L with C or not

In a study in which R, L and T are recorded, additionally measuring C serves the purpose of removing variation entering the system before T has exerted its effect, just as in analysis of covariance.

In addition to increasing the precision of treatment contrasts, analysis of covariance can also serve to remove bias (Cochran, 1957; Cox and McCullagh, 1982) but observing C is not necessary to remove bias here because L alone is sufficient for that. If L was not recorded then C would also be needed for the bias adjustment. It may be decided to not record C if it does not improve the precision of the estimator of ACE($T \to R$) much. In a model without an intermediate variable the effect of the strengths of the relationships between C and R and T on the precision is analysed for the logistic model by Robinson and Jewell (1991).

The importance of C is quantified by comparing Eqs.(6.8) and (6.9) via

$$\nabla\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) = \frac{\frac{1}{\mathbb{I}_n(\gamma_{rl|t})} + \frac{1}{\mathbb{I}_n(\gamma_{lt})}}{\frac{1}{\mathbb{I}_n(\gamma_{rl|c})} + \frac{1}{\mathbb{I}_n(\gamma_{lt})}}, \tag{6.15}$$

from Eqs.(6.11) and (6.12), and

$$\Delta\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) = \frac{1}{\mathbb{I}_n(\gamma_{rl|c})}\left\{\frac{\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}\right\} + \frac{\mathbb{I}_n(\gamma_{rc|l})}{\mathbb{I}_n(\gamma_{rl|c})}\left\{\frac{1+\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}\right\}, \tag{6.16}$$

with proof in Appendix B.4. From Eqs.(6.15) and (6.16), since $\rho^2_{rl|c} \geq \rho^2_{rl|t}$ (proof in Appendix B.5), which implies that $\mathbb{I}_n(\gamma_{rl|c}) \geq \mathbb{I}_n(\gamma_{rl|t})$,

$$\nabla\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) \geq 1, \quad \Delta\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) \geq 0.$$

This agrees with intuition since it means that omission of a covariate cannot increase the efficiency of the estimator. Recording C removes variation entering the system before the allocation of treatment.

From Eq.(6.16), the benefit of recording C increases as $\mathbb{I}_n(\gamma_{rc|l})$ increases and decreases as $\mathbb{I}_n(\gamma_{tc})$ increases. The importance of C reflects the two competing effects and agrees with classical results (cf. Robinson and Jewell, 1991; §9.5 of Jewell, 2003). The effect of the strength of the $C \to T$ relationship exists because the increasing collinearity increases the redundant information or reduces the information which C contains that is not already available in T. Thus, similarly to L, the less redundant information contained in the supplementary variable the greater the benefit of recording it. The effect of the strength of the $C \to R$ relationship occurs because the stronger the relationship the greater the proportion of residual variation, in the regression

of R on L, which is removed by conditioning on C.

6.3.3 Replace C with L or not

For the scenarios in §6.3.1 and §6.3.2 the more supplementary variables recorded the greater the efficiency of the estimator. Thus a study with all variables measured represents the gold standard but such studies may be expensive or impossible to conduct. Therefore it is sometimes necessary to choose whether to record L or C only, in addition to R and T.

The relative efficiency of the corresponding estimators is very important and is useful to assess the trade off between cost and efficiency of the estimator. Such an analysis provides criteria for judging whether it is more beneficial for covariates or intermediate variables to be recorded in a particular study. From Eqs.(6.10) and (6.11)

$$\nabla\mathbb{I}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c}) = \frac{\frac{1}{\mathbb{I}^*_n(\gamma_{rl|c})} + \frac{1}{\mathbb{I}^*_n(\gamma_{lt})}}{\frac{1}{\mathbb{I}_n(\gamma_{rl|t})} + \frac{1}{\mathbb{I}_n(\gamma_{lt})}}. \tag{6.17}$$

To determine formal conditions for choosing whether to record L or C, consider the expression in Eq.(6.17). It is more beneficial to record L when $\nabla\mathbb{I}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c}) > 1$, which is equivalent to $\Delta\mathbb{I}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c}) > 0$, and $\nabla\mathbb{I}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c})$ can take values anywhere on the positive real line. The comparison is simpler by considering the quantity

$$\begin{aligned}&\Delta\mathbb{I}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c})\\ &= \frac{1}{\mathbb{I}_n(\gamma_{rl|c})}\left\{\frac{1+\mathbb{I}_n(\gamma_{tc})}{\mathbb{I}_n(\gamma_{lt})} - \frac{\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})} + \mathbb{I}_n(\gamma_{rl|c})\frac{\mathbb{I}_n(\gamma_{tc})}{\mathbb{I}_n(\gamma_{lt})} - \mathbb{I}_n(\gamma_{rc|l})\frac{1+\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}\right\},\end{aligned} \tag{6.18}$$

with proof in Appendix B.4. Based on Eq.(6.18), sufficient conditions to choose to record L instead of C are

$$\mathbb{I}_n(\gamma_{lt}) < \mathbb{I}_n(\gamma_{tc}), \quad \mathbb{I}_n(\gamma_{rc|l}) < \mathbb{I}_n(\gamma_{rl|c}).$$

Thus it is more beneficial to record L if the strength of each of the $T \to L$ and $C \to R$ relationships is smaller than the $C \to T$ and $L \to R$ relationships respectively. In other words, it is better to record L when

- L contains less redundant information than C from T,
- L contains more information than C about R.

These conditions are represented in Fig.6.4.

Figure 6.4: DAGs representing conditions for choosing to record L instead of C (left) and vice versa (right). The thicker arrow represents a stronger relationship, as measured by $\mathbb{I}_n(\cdot)$.

Interestingly the reverse conditions are not sufficient for C to be more useful. Sufficient conditions for recording C are actually

$$\mathbb{I}_n(\gamma_{lt}) > \mathbb{I}_n(\gamma_{tc}) + 1, \quad \mathbb{I}_n(\gamma_{rc|l}) > \mathbb{I}_n(\gamma_{rl|c}).$$

They are similar, but not exactly, opposite to those required for recording L instead and are represented in Fig.6.4. Therefore it is possible for the

observation of C to be less beneficial if it contains only slightly less redundant information than L from T,

$$\mathbb{I}_n(\gamma_{tc}) < \mathbb{I}_n(\gamma_{lt}) < \mathbb{I}_n(\gamma_{tc}) + 1,$$

even though

$$\mathbb{I}_n(\gamma_{rc|l}) > \mathbb{I}_n(\gamma_{rl|c}).$$

In such borderline cases the behaviour of $\Delta\mathbb{I}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c})$ is determined by $\mathbb{I}_n(\gamma_{rl|c})$ and $\mathbb{I}_n(\gamma_{rc|l})$.

Chapter 7

Discussion

Various techniques have been presented for causal inference using supplementary variables. Many of the ideas have a long history in the literature but are extended here. Nevertheless they are still limited in certain ways and can be further developed. In this chapter, such limitations are discussed and future directions of development are proposed.

Identification with continuous instrumental variables

The use of IVs for point identification stems from the econometric literature, where they were used for continuous linear structural models. Their use has been extended for the identification of causal parameters in a generalised linear IV model in Chapter 3. The approach relies heavily on the constant convexity of the link function in the model and only produces bounds on the causal parameters. The fact that only bounds can be derived does not discredit the technique but attests to the difficulty of any causal inference at all in such models. Additional assumptions, which may not always be justifiable, were introduced for the logit function because of its variable convexity.

The question of whether bounds can be found using only the convexity or other special properties of the logit function does deserve further attention. The link functions in models in many practical problems do fit in the class of constant convexity and so the bounds have a wide scope for applications. However there are always uses for bounds in models with link functions not considered here. It would also be worthy to develop the analysis of the bounds by taking into consideration the sampling uncertainty of the data.

Computation of constraints for discrete model

For the discrete IV model, a technique for deriving constraints on the parameters of the model was given in Chapter 4. The approach is via an analysis of the convex polytopes which define the DAG. The constraints which involve the causal parameter of interest can be used to bound it and the rest to develop a statistical test for the model. Only the binary IV model was considered but the method can be conceptually easily extended to more complex DAGs with non-binary variables. However difficulties arise with computations involving non-binary variables and there is no guarantee that non-trivial constraints can be derived for a particular DAG. Specific issues are

- **Specifying extreme vertices of polytope**. The extreme vertices of the polytope which defines a DAG, such as in Fig.4.3, correspond exactly to the maps in the functional DAG representation. From the factorisation of a DAG or Eq.(2.9), the number of vertices are $|C|^{|B|} \times |B|^{|A|}$. Therefore the binary case only involves 16 vertices but the simple extension to trivariate variables involves 729 vertices. Thus it can be seen that the number increases rapidly and computations by hand are impractical. This is easily

overcome by programming a two dimensional array (cf. Appendix A) but the maximum size of any array which can be used still limits the use of the approach.

- **Converting representation of polytope**. The conversion of the algebraic representation of the polytope in terms of extreme vertices to hyperplanes is currently performed by Polymake. Not much emphasis was placed on testing its limits since the computations presented here and much more complex ones were seamlessly handled by the web demo of Polymake. A study of its internal algorithms with regard to the specific types of polytopes encountered in such applications may be a worthwhile direction of research but not necessarily in keeping with a hard-core statistical theme. It would definitely be useful to set up an automatic interface between the spreadsheet, with the polytope representation to be converted, and the software used for the conversion. There might also be other software which perform the computation more efficiently.
- **Evaluating constraints and causal bounds**. As expected, the number of facets increases as the sizes of the state space of the variables increase. For a model with $|A| = |B| = 3$ and $|C| = 2$ the polytope is defined by 438 facets, including the causal bounds. Again the evaluation of so many constraints cannot be done manually and requires the use of an array but computing resources are of concern.

The technique only produces linear constraints and will only derive trivial constraints in a model which may possess non-trivial non-linear constraints. Perhaps it can be developed to deal with non-linear constraints.

Testing instrumental model

Prior to any estimation of parameters in the IV model, an investigator must first assess whether the model is valid. Although the derived constraints determine which observable distributions fit and not which unobservable ones fit (cf. §4.3.1), they can be used in the development of a statistical test for the instrumental model. Based on the formulation of Sir R. A. Fisher, the constraints only need to be used to determine whether the test is significant. A statistical test checks that the parameters lie in the polytope which defines the model. The polytope for the binary IV model was shown to possess some properties which simplified the derivation of the distribution of the test statistic and the power function.

Similarly to the computation of the constraints, the test becomes much more complex for non-binary models. However a worthwhile direction of research is the derivation of the power function for the test for non-binary IV models, possibly using the results of El Barmi and Dykstra (1995).

Future work would also involve the derivation of a proof for the expression for the approximate p-value in Eq.(5.5) or of a more accurate approximation.

Estimation in the instrumental model

If a model cannot be rejected then the next step is usually estimation of parameters. Estimation in the IV model is done via convex optimisation, simply maximise the likelihood subject to the constraints which specify which distributions are valid for the model. It is an alternative to the EM-algorithm and can be extended to DAGs in general, as long as constraints can be computed

for the DAG. Since only linear constraints are derived, the method of deriving constraints is very optimisation-friendly. This approach to estimation avoids the use of counterfactuals, which has some philosophical and computational advantages (Dawid, 2000) but a thorough assessment of its worth requires a comparison of its computational complexity to the EM-algorithm.

Note that, even though non-trivial causal bounds are available, the estimates of the parameters may be such that the estimated causal bounds are trivial. Furthermore, even if non-trivial estimates are produced, confidence intervals can still be trivial. Further research could involve investigating the adequacy of the bootstrap method used for computing confidence intervals.

Supplementary variables to improve precision

The other major use of supplementary variables which was considered is for increasing the precision of estimation. This includes analysis of covariance and the recording of intermediate variables to improve the precision of causal estimators. In a multivariate Gaussian model, conditions were given for choosing between covariates and intermediate variables when only one could be recorded.

The aim of such analyses is to develop criteria for choosing variables to record when costs are of concern. The process may be made more rigorous by using the tools of decision theory. Such a task would require formal specification of the costs involved. The analysis can also be extended to non-Gaussian models as in Robinson and Jewell (1991) but may yield intuitively contrasting results.

Appendix A

Method of computation of constraints

The method used for computing the algebraic expressions for the constraints of a DAG is described here. The algorithm is:

1. `Enter original polytope into software.`

   The two dimensional array of the extreme vertices of the polytope, which represents the factorised distributions, is input into the software, e.g. Excel, R. The extreme vertices are entered in the same format as in Fig.4.1, with each column corresponding to a coordinate.

2. `Compute the extreme vertices of transformed polytope.`

   This is done by programming the relevant map such that the format of the transformed polytope is the same.

3. `Copy and paste representation into Polymake web demo.`

   The format of Fig.4.1 is exactly that required for inputing "`POINTS`" into Polymake.

4. `Use Polymake to convert representation.`

   The extreme vertices representation (`POINTS`) is converted to the half-

space representation (`FACETS`). Polymake outputs inequalities in the same format, where each number corresponds to the coefficient of the component its column represents in the inequality.

5. `Evaluation by multiplying matrices.`

   To evaluate bounds and constraints on the parameters, the array (matrix) output from Polymake is multiplied by the matrix of estimated parameter values. If there are many constraints then this approach can be adapted to typeset the algebraic expressions.

For much of the computations in this thesis, Microsoft Excel 2003 was used but the algorithms are not software specific and can be implemented in other platforms. Statistical computations were either performed or double checked in R.

Appendix B

Expressions for asymptotic variance of causal estimators

B.1 Estimator given all X's recorded

Theorem B.1. *The relation*

$$\frac{1}{\mathbb{I}_n(\gamma_{k,1})} = \sum_{i=1}^{k-1} \frac{1}{\mathbb{I}_n(\gamma_{i+1,i})},$$

is true for the DAG in Fig.6.2 when all of the variables are recorded.

Proof of theorem B.1. When $k = 2$ the result is trivially true. Assume it is true for $k - 1$. By the delta method

$$n\mathbb{V}_n^{\infty}(\hat{\gamma}_{k,1}) \approx \gamma_{k,k-1}^2 n\mathbb{V}_n^{\infty}(\hat{\gamma}_{k-1,1}) + \gamma_{k-1,1}^2 n\mathbb{V}_n^{\infty}(\hat{\gamma}_{k,k-1})$$

since $\hat{\gamma}_{i+1,1} = \hat{\gamma}_{i+1,i} \cdots \hat{\gamma}_{2,1}$ for $i = 1, \ldots, n-1$ and

$$\mathrm{cov}_n^{\infty}(\hat{\gamma}_{i+1,i}, \hat{\gamma}_{j+1,j}) = 0 \text{ for } i \neq j,$$

from the factorisation of the likelihood function. From the inductive hypothesis

$$\begin{aligned} n\mathbb{V}_n^\infty(\hat{\gamma}_{k,1}) &= \gamma_{k,k-1}^2 n\mathbb{V}_n^\infty(\hat{\gamma}_{k-1,1}) + \gamma_{k-1,1}^2 n\mathbb{V}_n^\infty(\hat{\gamma}_{k,k-1}) \\ \gamma_{k,1}^{-2} n\mathbb{V}_n^\infty(\hat{\gamma}_{k,1}) &= \gamma_{k-1,1}^{-2} n\mathbb{V}_n^\infty(\hat{\gamma}_{k-1,1}) + \gamma_{k,k-1}^{-2} n\mathbb{V}_n^\infty(\hat{\gamma}_{k,k-1}) \\ \tfrac{1}{\mathbb{I}_n(\gamma_{k,1})} &= \textstyle\sum_{i=1}^{k-2} \tfrac{1}{\mathbb{I}_n(\gamma_{i+1,i})} + \tfrac{1}{\mathbb{I}_n(\gamma_{k,k-1})}, \end{aligned}$$

since $\gamma_{i+1,1} = \gamma_{i+1,i} \cdots \gamma_{2,1}$ for $i = 1, \ldots, n-1$. □

B.2 Derivations of asymptotic variances

This section gives the derivations of the expressions for the asymptotic variance of the estimators of ACE($T \to R$) in §6.2.

Estimator given R, C and T recorded

Let $f(\cdot)$ be the probability density function, then the likelihood function for $\gamma_{rt|c}$ is

$$\mathcal{L}(\gamma_{rt|c}) = f(R \,|\, T, C) f(T \,|\, C) f(C) \propto f(R \,|\, T, C).$$

Therefore the log-likelihood function is

$$l(\gamma_{rt|c}) \propto \ln f(R \,|\, T, C) \propto -\frac{n}{2} \ln(2\pi\sigma_{rr|t,c}) - \sum_{i=1}^{n} \frac{(r_i - \gamma_{rt|c} t_i - \gamma_{rc|t} c_i)^2}{2\sigma_{rr|t,c}},$$

where $\sigma_{ab|t,c} = \operatorname{cov}(A, B \,|\, T, C)$, and the information matrix of $(\gamma_{rt|c}, \gamma_{rc|t})$ is

$$I_n(\gamma_{rt|c}, \gamma_{rc|t}) = \begin{pmatrix} \frac{\partial^2 l}{\partial \gamma_{rt|c}^2} & \frac{\partial^2 l}{\partial \gamma_{rt|c} \partial \gamma_{rc|t}} \\ \frac{\partial^2 l}{\partial \gamma_{rt|c} \partial \gamma_{rc|t}} & \frac{\partial^2 l}{\partial \gamma_{rc|t}^2} \end{pmatrix} = \frac{1}{\sigma_{rr|t,c}} \begin{pmatrix} -\sum t_i^2 & -\sum t_i c_i \\ -\sum t_i c_i & -\sum c_i^2 \end{pmatrix}.$$

The information matrix of the vector $(\gamma_{rt|c}, \gamma_{rc|t})$ is considered because of the factorisation of the likelihood function. Since

$$T \,|\, C \sim N(\gamma_{tc}C, \sigma_{tt|c}),$$

and $\sigma_{tt|c}$ is functionally independent of C,

$$\begin{aligned} \sigma_{tt} &= \mathbb{E}_c(\sigma_{tt|c}) + \mathbb{V}_c(\gamma_{tc}C) \\ &= \sigma_{tt|c} + \gamma_{tc}^2\sigma_{cc}. \end{aligned}$$

Therefore, since $\gamma_{tc} = \sigma_{tc}/\sigma_{cc}$,

$$\sigma_{tt|c} = \frac{\sigma_{tt}\sigma_{cc} - \sigma_{tc}^2}{\sigma_{cc}} = \sigma_{tt}(1 - \rho_{tc}^2). \tag{B.1}$$

Eq.(B.1) can also be obtained from the properties of the concentration matrix of a multivariate Gaussian distribution. It follows that

$$I_n(\gamma_{rt|c}, \gamma_{rc|t}) = \frac{n}{\sigma_{rr|t,c}} \begin{pmatrix} \sigma_{tt} & \sigma_{tc} \\ \sigma_{tc} & \sigma_{cc} \end{pmatrix},$$

and

$$n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rt|c}) = nI_n(\gamma_{rt|c}, \gamma_{rc|t})_{tt}^{-1} = \frac{\sigma_{rr|t,c}}{\sigma_{tt|c}},$$

from Eq.(B.1). Alternatively

$$\begin{aligned} \mathbb{V}(\hat{\gamma}_{rt|c}) &= \mathbb{E}_{t,c}\{\mathbb{V}(\hat{\gamma}_{rt|c} \,|\, T, C)\} + \mathbb{V}_{t,c}\{\mathbb{E}(\hat{\gamma}_{rt|c} \,|\, T, C)\} \\ &= \mathbb{E}_{t,c}\{\sigma_{rr|t,c}(X^TX)_{tt}^{-1}\} + \mathbb{V}_{t,c}(\gamma_{rt|c}) \\ &= \sigma_{rr|t,c}.\sigma_{cc}/\{(n-3)(\sigma_{cc}\sigma_{tt} - \sigma_{tc}^2)\} \\ &= \sigma_{rr|t,c}/\{\sigma_{tt|c}(n-3)\}, \end{aligned}$$

from Eq.(B.1), where X is the design matrix for the regression of R on T and C, which implies that

$$\lim_{n\to\infty} n\mathbb{V}(\hat{\gamma}_{rt|c}) = \frac{\sigma_{rr|t,c}}{\sigma_{tt|c}}.$$

It is true that

$$\begin{aligned}\sigma_{rr|t,c} &= \mathbb{E}_{l|t,c}(\sigma_{rr|l,t,c} \,|\, T, C) + \mathbb{V}_{l|t,c}(\mu_{r|l,t,c} \,|\, T, C) \\ &= \sigma_{rr|l,t,c} + \mathbb{V}_{l|t,c}(\gamma_{rl|c}L + \gamma_{rc|l}C \,|\, T, C) \\ &= \sigma_{rr|l,c} + \gamma^2_{rl|c}\sigma_{ll|t}.\end{aligned}$$

since $R \perp\!\!\!\perp T \,|\, (L, C)$, $L \perp\!\!\!\perp C \,|\, T$ and $\sigma_{rr|l,t,c}$ is functionally independent of L. Therefore

$$n\mathbb{V}^{\infty}_{n}(\hat{\gamma}_{rt|c}) = \frac{\sigma_{rr|l,c}}{\sigma_{tt|c}} + \gamma^2_{rl|c}\frac{\sigma_{ll|t}}{\sigma_{tt|c}}.$$

This is the expression in Eq.(6.10).

Estimator given R, L and T recorded

The likelihood function for $(\gamma_{rl|t}, \gamma_{lt})$ is

$$\begin{aligned}\mathcal{L}(\gamma_{rl|t}, \gamma_{lt}) &= f(R \,|\, L, T) f(L \,|\, T) f(T) \\ &\propto f(R \,|\, L, T) f(L \,|\, T),\end{aligned}$$

and the log-likelihood function is

$$l(\gamma_{rl|t}, \gamma_{lt}) \propto -\frac{n}{2}\ln(4\pi^2\sigma_{rr|l,t}.\sigma_{ll|t}) - \sum_{i=1}^{n}\frac{(r_i - \gamma_{rl|t}l_i - \gamma_{rt|l}t_i)^2}{2\sigma_{rr|l,t}} - \sum_{i=1}^{n}\frac{(l_i - \gamma_{lt}t_i)^2}{2\sigma_{ll|t}}.$$

The information matrix of $(\gamma_{rl|t}, \gamma_{rt|l}, \gamma_{lt})$ is

$$I_n(\gamma_{rl|t}, \gamma_{rt|l}, \gamma_{lt}) = n \begin{pmatrix} \sigma_{ll}/\sigma_{rr|l,t} & \sigma_{lt}/\sigma_{rr|l,t} & 0 \\ \sigma_{lt}/\sigma_{rr|l,t} & \sigma_{tt}/\sigma_{rr|l,t} & 0 \\ 0 & 0 & \sigma_{tt}/\sigma_{ll|t} \end{pmatrix},$$

from which it follows by the delta method that

$$\begin{aligned} \mathbb{V}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) &\approx \mathbb{E}(\hat{\gamma}_{lt})^2\mathbb{V}(\hat{\gamma}_{rl|t}) + \mathbb{E}(\hat{\gamma}_{rl|t})^2\mathbb{V}(\hat{\gamma}_{lt}) \\ &\quad + 2\mathbb{E}(\hat{\gamma}_{rl|t})\mathbb{E}(\hat{\gamma}_{lt})\mathrm{cov}(\hat{\gamma}_{rl|t}, \hat{\gamma}_{lt}) \\ n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) &\approx \gamma_{lt}^2\{\tfrac{\sigma_{rr|l,t}.\sigma_{tt}}{\sigma_{ll}\sigma_{tt}-\sigma_{lt}^2}\} + \gamma_{rl|t}^2(\tfrac{\sigma_{ll|t}}{\sigma_{tt}}) \\ &\approx \gamma_{lt}^2(\tfrac{\sigma_{rr|l,t}}{\sigma_{ll|t}}) + \gamma_{rl|t}^2(\tfrac{\sigma_{ll|t}}{\sigma_{tt}}). \end{aligned}$$

Alternatively the Wishart (W) and inverse Wishart (W^{-1}) distributions can be used in the derivation. Since

$$\sum t^2 \sim W(n, \sigma_{tt}) \qquad \Leftrightarrow \qquad \frac{1}{\sigma_{tt}} \sum t^2 \sim \chi_n^2,$$

which implies that

$$\left(\sum t^2\right)^{-1} \sim W^{-1}(n, \sigma_{tt}^{-1}) \equiv \gamma^{-1}\left(\frac{n}{2}, \frac{\sigma_{tt}^{-1}}{2}\right),$$

where $\gamma^{-1}(\cdot)$ is the inverse gamma distribution, then

$$\begin{aligned} \mathbb{V}(\hat{\gamma}_{lt}) &= \mathbb{V}\{\mathbb{E}(\hat{\gamma}_{lt} \,|\, L, T)\} + \mathbb{E}\{\mathbb{V}(\hat{\gamma}_{lt} \,|\, L, T)\} \\ &= \mathbb{V}(\gamma_{lt}) + \mathbb{E}\left(\tfrac{\sigma_{ll|t}}{\sum t^2}\right) = \tfrac{\sigma_{ll|t}}{\sigma_{tt}}. \end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{V}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) &= \mathbb{E}\{\hat{\gamma}_{lt}^2.\mathbb{V}(\hat{\gamma}_{rl|t} \mid L,T)\} + \mathbb{V}\{\hat{\gamma}_{lt}\mathbb{E}(\hat{\gamma}_{rl|t} \mid L,T)\} \\ &= \mathbb{E}\{\hat{\gamma}_{lt}^2.\sigma_{rr|l,t}(Z^TZ)_{ll}^{-1}\} + \mathbb{V}(\hat{\gamma}_{lt}.\gamma_{rl|t}) \\ &\approx \sigma_{rr|l,t}\mathbb{E}\{\hat{\gamma}_{lt}^2\}\mathbb{E}\{(Z^TZ)_{ll}^{-1}\} + \gamma_{rl|t}^2\mathbb{V}(\hat{\gamma}_{lt}) \\ &\approx \tfrac{\sigma_{rr|l,t}}{\sigma_{ll|t}}(\tfrac{\sigma_{ll|t}}{n\sigma_{tt}} + \gamma_{lt}^2)\tfrac{1}{n-3} + \gamma_{rl|t}^2.\tfrac{\sigma_{ll|t}}{n\sigma_{tt}},\end{aligned}$$

since $K_{ll} = \frac{1}{\sigma_{ll|t}}$ and

$$Z^TZ \sim W\left\{n, \begin{pmatrix}\sigma_{ll} & \sigma_{lt}\\ \sigma_{lt} & \sigma_{tt}\end{pmatrix}\right\} \Rightarrow (Z^TZ)^{-1} \sim W^{-1}\left\{n, \begin{pmatrix}\sigma_{ll} & \sigma_{lt}\\ \sigma_{lt} & \sigma_{tt}\end{pmatrix}^{-1}\right\},$$

where K is the concentration matrix of the joint distribution of (L,T), Z is the design matrix for the regression of R on L and T and $\sigma_{rr|l,t}$ is the variance of the error term. Either approach gives the expression in Eq.(6.8), which is

$$n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) = \gamma_{lt}^2\frac{\sigma_{rr|l,t}}{\sigma_{ll|t}} + \gamma_{rl|t}^2\frac{\sigma_{ll|t}}{\sigma_{tt}}.$$

Estimator given R, L, C and T recorded

Similarly to the derivation of the expression for $\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt})$,

$$\mathbb{V}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}) \approx \mathbb{E}(\hat{\gamma}_{lt})^2\mathbb{V}(\hat{\gamma}_{rl|c}) + \mathbb{E}(\hat{\gamma}_{rl|c})^2\mathbb{V}(\hat{\gamma}_{lt}) + 2\mathbb{E}(\hat{\gamma}_{rl|c})\mathbb{E}(\hat{\gamma}_{lt})\mathrm{cov}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}),$$

by the delta method. One approach to evaluating this expression is to use $I_n(\gamma_{rl|c}, \gamma_{rc|l}, \gamma_{lt})$ to determine $\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|c})$, $\mathbb{V}_n^{\infty}(\hat{\gamma}_{lt})$ and $\mathrm{cov}_n^{\infty}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt})$. From the factorisation of the likelihood, $\mathrm{cov}_n^{\infty}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}) = 0$ so that $I_n(\gamma_{lt})$ can

be used to determine $\mathbb{V}_n^\infty(\hat{\gamma}_{lt})$ and $I_n(\gamma_{rl|c}, \gamma_{rc|l})$ can be used to determine $\mathbb{V}_n^\infty(\hat{\gamma}_{rl|c})$. Since $\hat{\gamma}_{lt}$ and $\hat{\gamma}_{rl|c}$ are asymptotically unbiased, the expressions for the variances can then be substituted in

$$\mathbb{V}_n^\infty(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}) \approx \gamma_{lt}^2 \mathbb{V}_n^\infty(\hat{\gamma}_{rl|c}) + \gamma_{rl|c}^2 \mathbb{V}_n^\infty(\hat{\gamma}_{lt}).$$

Alternatively $\mathbb{V}_n^\infty(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt})$ can be computed by using the inverse Wishart distribution, similarly to finding $\mathbb{V}_n^\infty(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt})$. Whichever method is used,

$$n\mathbb{V}_n^\infty(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}) = \gamma_{lt}^2 \frac{\sigma_{rr|l,c}}{\sigma_{ll|c}} + \gamma_{rl|c}^2 \frac{\sigma_{ll|t}}{\sigma_{tt}},$$

which is the expression in Eq.(6.9).

B.3 Proofs of asymptotic variances in terms of correlation coefficients

In this section the expressions for the asymptotic variances in terms of correlation coefficients in Eqs.(6.10), (6.11) and (6.12) are derived.

Estimator given R, C and T recorded

Since $L \perp\!\!\!\perp C \,|\, T$ from the causal DAG in Fig.6.3,

$$\begin{aligned}
\sigma_{ll|c} &= \mathbb{E}_{t|c}\{\mathbb{V}(L \,|\, T, C)\} + \mathbb{V}_{t|c}\{\mathbb{E}(L \,|\, T, C)\} \\
&= \sigma_{ll|t} + \gamma_{lt}^2 \sigma_{tt|c} \\
\frac{\sigma_{ll|c} - \sigma_{ll|t}}{\sigma_{ll|t}} &= \frac{\sigma_{lt}^2 \sigma_{tt}(1-\rho_{tc}^2)}{\sigma_{tt}^2 \sigma_{ll}(1-\rho_{lt}^2)},
\end{aligned}$$

from Eq.(B.1), which implies that

$$\frac{\gamma_{lt}^2\sigma_{tt|c}}{\sigma_{ll|t}} = \frac{\sigma_{ll|c} - \sigma_{ll|t}}{\sigma_{ll|t}} = \frac{\rho_{lt}^2}{1-\rho_{lt}^2}(1-\rho_{tc}^2) = \frac{\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}. \tag{B.2}$$

The following equation also holds for the model

$$\begin{aligned}\sigma_{rr|c} &= \mathbb{E}_{l|c}\{\mathbb{V}(R\,|\,L,C)\} + \mathbb{V}_{l|c}\{\mathbb{E}(R\,|\,L,C)\} \\ &= \sigma_{rr|l,c} + \gamma_{rl|c}^2\sigma_{ll|c},\end{aligned}$$

which implies that

$$\frac{\sigma_{rr|l,c}}{\gamma_{rl|c}^2\sigma_{ll|c}} = \frac{1-\rho_{rl|c}^2}{\rho_{rl|c}^2}. \tag{B.3}$$

From Eqs.(6.7), (B.2) and (B.3),

$$\begin{aligned}n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rt|c}) &= \tfrac{\sigma_{rr|l,c}}{\sigma_{tt|c}} + \gamma_{rl|c}^2\tfrac{\sigma_{ll|t}}{\sigma_{tt|c}} \\ &= \gamma_{rl|c}^2.\gamma_{lt}^2\left\{\tfrac{\sigma_{rr|l,c}}{\gamma_{rl|c}^2\sigma_{ll|c}}\left(\tfrac{\sigma_{ll|c}}{\gamma_{lt}^2\sigma_{tt|c}}\right) + \tfrac{\sigma_{ll|t}}{\gamma_{lt}^2\sigma_{tt|c}}\right\} \\ &= \gamma_{rl|c}^2.\gamma_{lt}^2\left[\tfrac{1-\rho_{rl|c}^2}{\rho_{rl|c}^2}\left\{1+\tfrac{1}{1-\rho_{tc}^2}\left(\tfrac{1-\rho_{lt}^2}{\rho_{lt}^2}\right)\right\} + \tfrac{1}{1-\rho_{tc}^2}\left(\tfrac{1-\rho_{lt}^2}{\rho_{lt}^2}\right)\right].\end{aligned}$$

which is the expression in Eq.(6.10).

Estimator given R, L and T recorded

Similarly to the derivation of Eq.(B.3),

$$\begin{aligned}\sigma_{rr|t} &= \mathbb{E}_{l|t}\{\mathbb{V}(R\,|\,L,T)\} + \mathbb{V}_{l|t}\{\mathbb{E}(R\,|\,L,T)\} \\ &= \sigma_{rr|l,t} + \gamma_{rl|t}^2\sigma_{ll|t},\end{aligned}$$

which implies that

$$\frac{\sigma_{rr|l,t}}{\gamma_{rl|t}^2\sigma_{ll|t}} = \frac{\sigma_{rr|l,t}}{\sigma_{rr|t} - \sigma_{rr|l,t}} = \frac{1-\rho_{rl|t}^2}{\rho_{rl|t}^2}. \tag{B.4}$$

From Eqs.(6.8) and (B.4),

$$\begin{aligned} n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) &= \gamma_{lt}^2\frac{\sigma_{rr|l,t}}{\sigma_{ll|t}} + \gamma_{rl|t}^2\frac{\sigma_{ll|t}}{\sigma_{tt}} \\ &= \gamma_{rl|t}^2.\gamma_{lt}^2\left(\frac{\sigma_{rr|l,t}}{\gamma_{rl|t}^2\sigma_{ll|t}} + \frac{\sigma_{ll|t}}{\gamma_{lt}^2\sigma_{tt}}\right) \\ &= \gamma_{rl|t}^2.\gamma_{lt}^2\left(\frac{1-\rho_{rl|t}^2}{\rho_{rl|t}^2} + \frac{1-\rho_{lt}^2}{\rho_{lt}^2}\right), \end{aligned}$$

which is the expression given in Eq.(6.11).

Estimator given R, L, C and T recorded

Also, from Eqs.(6.9), (B.1), (B.2) and (B.3),

$$\begin{aligned} n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}) &= \gamma_{lt}^2\frac{\sigma_{rr|l,c}}{\sigma_{ll|c}} + \gamma_{rl|c}^2\frac{\sigma_{ll|t}}{\sigma_{tt}} \\ &= \gamma_{rl|c}^2.\gamma_{lt}^2\left(\frac{\sigma_{rr|l,c}}{\gamma_{rl|c}^2\sigma_{ll|c}} + \frac{\sigma_{ll|t}}{\gamma_{lt}^2\sigma_{tt}}\right) \\ &= \gamma_{rl|c}^2.\gamma_{lt}^2\left(\frac{1-\rho_{rl|c}^2}{\rho_{rl|c}^2} + \frac{1-\rho_{lt}^2}{\rho_{lt}^2}\right), \end{aligned}$$

which is the expression given in Eq.(6.12).

B.4 Derivation of $\Delta\mathbb{I}$'s

In this section the expressions for the $\Delta\mathbb{I}$'s in Eqs.(6.14), (6.16) and (6.18) are derived.

From Eqs.(6.7), (6.9), (B.1) and (B.2)

$$\begin{aligned}
&n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rt|c}) - n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}) \\
&\quad = \tfrac{\sigma_{rr|l,c}}{\sigma_{tt|c}} + \gamma_{rl|c}^2\tfrac{\sigma_{ll|t}}{\sigma_{tt|c}} - \gamma_{lt}^2\tfrac{\sigma_{rr|l,c}}{\sigma_{ll|c}} - \gamma_{rl|c}^2\tfrac{\sigma_{ll|t}}{\sigma_{tt}} \\
&\quad = \gamma_{rl|c}^2.\gamma_{lt}^2\left(\tfrac{\sigma_{rr|l,c}}{\gamma_{rl|c}^2\sigma_{ll|c}}\right)\left(\tfrac{\sigma_{ll|c}}{\gamma_{lt}^2\sigma_{tt|c}} - 1\right) + \gamma_{rl|c}^2\tfrac{\sigma_{ll|t}}{\sigma_{tt}}\left(\tfrac{1}{1-\rho_{tc}^2} - 1\right) \\
&\quad = \gamma_{rl|c}^2.\gamma_{lt}^2\left\{\tfrac{1-\rho_{rl|c}^2}{\rho_{rl|c}^2}\left(\tfrac{\sigma_{ll|t}}{\gamma_{lt}^2\sigma_{tt|c}}\right) + \tfrac{\sigma_{ll|t}}{\gamma_{lt}^2\sigma_{tt}}\left(\tfrac{\rho_{tc}^2}{1-\rho_{tc}^2}\right)\right\} \\
&\quad = \gamma_{rl|c}^2.\gamma_{lt}^2\left(\tfrac{1-\rho_{lt}^2}{\rho_{lt}^2}\right)\left(\tfrac{\rho_{tc}^2}{1-\rho_{tc}^2}\right)\left\{\tfrac{1-\rho_{rl|c}^2}{\rho_{rl|c}^2}\left(\tfrac{1}{\rho_{tc}^2}\right) + 1\right\},
\end{aligned}$$

which implies that

$$\Delta\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c}) = \frac{1}{\mathbb{I}_n(\gamma_{rl|c})}\left\{\frac{1+\mathbb{I}_n(\gamma_{tc})}{\mathbb{I}_n(\gamma_{lt})}\right\} + \frac{\mathbb{I}_n(\gamma_{tc})}{\mathbb{I}_n(\gamma_{lt})}.$$

Since $R \perp\!\!\!\perp T \,|\, (L, C)$ and $\sigma_{rr|l,c}$ is functionally independent of C

$$\begin{aligned}
\sigma_{rr|l,t} &= \mathbb{E}_{c|l,t}(\sigma_{rr|l,t,c}) + \mathbb{V}_{c|l,t}(\mu_{r|l,t,c}) \\
&= \sigma_{rr|l,c} + \gamma_{rc|l}^2\sigma_{cc|t}.
\end{aligned} \tag{B.5}$$

Therefore from Eqs.(6.8), (6.9) and (B.5)

$$\begin{aligned}
&n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) - n\mathbb{V}_n^{\infty}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}) \\
&\quad = \gamma_{lt}^2\tfrac{\sigma_{rr|l,t}}{\sigma_{ll|t}} + \gamma_{rl|t}^2\tfrac{\sigma_{ll|t}}{\sigma_{tt}} - \gamma_{lt}^2\tfrac{\sigma_{rr|l,c}}{\sigma_{ll|c}} - \gamma_{rl|c}^2\tfrac{\sigma_{ll|t}}{\sigma_{tt}} \\
&\quad = \gamma_{lt}^2\left(\tfrac{\sigma_{rr|l,t}}{\sigma_{ll|t}} - \tfrac{\sigma_{rr|l,c}}{\sigma_{ll|c}}\right) \\
&\quad = \gamma_{lt}^2\left(\tfrac{\sigma_{rr|l,c}}{\sigma_{ll|t}} - \tfrac{\sigma_{rr|l,c}}{\sigma_{ll|c}}\right) + \gamma_{rl|c}^2.\gamma_{lt}^2\left\{\tfrac{\mathbb{I}_n(\gamma_{rc|l})}{\mathbb{I}_n(\gamma_{rl|c})}\right\}\tfrac{1+\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})} \\
&\quad = \gamma_{rl|c}^2.\gamma_{lt}^2\left[\tfrac{\sigma_{rr|l,c}}{\gamma_{rl|c}^2\sigma_{ll|c}}\left(\tfrac{\gamma_{lt}^2\sigma_{tt|c}}{\sigma_{ll|t}}\right) + \tfrac{\mathbb{I}_n(\gamma_{rc|l})}{\mathbb{I}_n(\gamma_{rl|c})}\left\{\tfrac{1+\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}\right\}\right],
\end{aligned}$$

since

$$\begin{aligned}\gamma_{lt}^2\frac{\gamma_{rc|l}^2\sigma_{cc|t}}{\sigma_{ll|t}} &= \gamma_{lt}^2\frac{\gamma_{rc|l}^2\sigma_{cc|l}}{\sigma_{rr|l,c}}\frac{\sigma_{rr|l,c}\sigma_{cc|t}}{\sigma_{ll|t}\sigma_{cc|l}}\\ &= \gamma_{rl|c}^2.\gamma_{lt}^2\frac{\gamma_{rc|l}^2\sigma_{cc|l}}{\sigma_{rr|l,c}}\frac{\sigma_{rr|l,c}}{\gamma_{rl|c}^2\sigma_{ll|c}}\frac{\sigma_{cc|t}\sigma_{ll|c}}{\sigma_{ll|t}\sigma_{cc|l}}\\ &= \gamma_{rl|c}^2.\gamma_{lt}^2\frac{\gamma_{rc|l}^2\sigma_{cc|l}}{\sigma_{rr|l,c}}\frac{\sigma_{rr|l,c}}{\gamma_{rl|c}^2\sigma_{ll|c}}\frac{1-\rho_{tc}^2}{1-\rho_{lt}^2}\\ &= \gamma_{rl|c}^2.\gamma_{lt}^2\left\{\frac{\mathbb{I}_n(\gamma_{rc|l})}{\mathbb{I}_n(\gamma_{rl|c})}\right\}\frac{1+\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}.\end{aligned}$$

It follows that

$$\Delta\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}) = \frac{1}{\mathbb{I}_n(\gamma_{rl|c})}\left\{\frac{\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}\right\} + \frac{\mathbb{I}_n(\gamma_{rc|l})}{\mathbb{I}_n(\gamma_{rl|c})}\left\{\frac{1+\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}\right\},$$

and

$$\begin{aligned}&\Delta\mathbb{I}(\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c})\\ &\quad= \Delta\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rt|c}) - \Delta\mathbb{I}(\hat{\gamma}_{rl|c}.\hat{\gamma}_{lt}\,;\,\hat{\gamma}_{rl|t}.\hat{\gamma}_{lt})\\ &\quad= \frac{1}{\mathbb{I}_n(\gamma_{rl|c})}\left\{\frac{1+\mathbb{I}_n(\gamma_{tc})}{\mathbb{I}_n(\gamma_{lt})} - \frac{\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})} + \mathbb{I}_n(\gamma_{rl|c})\frac{\mathbb{I}_n(\gamma_{tc})}{\mathbb{I}_n(\gamma_{lt})} - \mathbb{I}_n(\gamma_{rc|l})\frac{1+\mathbb{I}_n(\gamma_{lt})}{1+\mathbb{I}_n(\gamma_{tc})}\right\}.\end{aligned}$$

B.5 Relation between $\rho_{rl|t}^2$ and $\rho_{rl|c}^2$

Theorem B.2. $\rho_{rl|c}^2 \geq \rho_{rl|t}^2$.

Proof of theorem B.2. From Eqs.(6.6), (B.3), (B.4) and (B.5)

$$\begin{aligned}\frac{\sigma_{rr|l,t}}{\gamma_{rl|t}^2\gamma_{lt}^2\sigma_{ll|t}} &= \frac{\sigma_{rr|l,c}}{\gamma_{rl|c}^2\gamma_{lt}^2\sigma_{ll|t}} + \frac{\gamma_{rc|l}^2\sigma_{cc|t}}{\gamma_{rl|c}^2\gamma_{lt}^2\sigma_{ll|t}}\\ \frac{1-\rho_{rl|t}^2}{\rho_{rl|t}^2} &= \frac{\sigma_{rr|l,c}}{\gamma_{rl|c}^2\sigma_{ll|c}}\left(\frac{\sigma_{ll|c}}{\sigma_{ll|t}}\right)\left\{1+\frac{\gamma_{rc|l}^2\sigma_{cc|l}}{\sigma_{rr|l,c}}\left(\frac{\sigma_{cc|t}}{\sigma_{cc|l}}\right)\right\}\\ &= \frac{1-\rho_{rl|c}^2}{\rho_{rl|c}^2}\left(\frac{\sigma_{ll|c}}{\sigma_{ll|t}}\right)\left\{1+\frac{\rho_{rc|l}^2}{1-\rho_{rc|l}^2}\left(\frac{\sigma_{cc|t}}{\sigma_{cc|l}}\right)\right\}.\end{aligned}$$

It follows from Eq.(B.2) that

$$\frac{1-\rho_{rl|t}^2}{\rho_{rl|t}^2} \geq \frac{1-\rho_{rl|c}^2}{\rho_{rl|c}^2}.$$

□

References

Aalen, O.O. and Frigessi, A. (2007). What can statistics contribute to a causal understanding? *Scandinavian Journal of Statistics*. **34** 155-168.

Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*. **91** 444-455.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*. **92** 1171-1176.

Berkson, J. (1958). Smoking and lung cancer: some observations on two recent reports. *Journal of the American Statistical Association*. **53** 28-38.

Bollen, K.A. (1989). *Structural equations with latent variables*. John Wiley & Sons: New York.

Bowden, R.J. and Turkington, D.A. (1984). *Instrumental variables*. Cambridge University Press: Cambridge, MA.

Cai, Z., Kuroki, M., Pearl, J. and Tian, J. (2008). Bounds on direct effects in

the presence of confounded intermediate variables. *Biometrics*. **64** 695-701.

Cheng, J. and Small, D.S. (2006). Bounds on causal effects in three arm trials with non-compliance. *Journal of the Royal Statistical Society*, Ser. B. **68** 815-836.

Chen, H., Geng, Z. and Jia, J. (2007). Criteria for surrogate endpoints. *Journal of the Royal Statistical Society*, Ser. B. **69** 919-932.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*. **25** 573-578.

Cochran, W.G. (1957). Analysis of covariance: its nature and uses. *Biometrics*. **13** 261-281.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences,* 2^{nd} *edition*. Lawrence Erlbaum Associates: Hillsdale.

Cox, D.R. (1960). Regression analysis when there is prior information about supplementary variables. *Journal of the Royal Statistical Society*, Ser. B. **22** 172-176.

Cox, D.R. and McCullagh, P. (1982). Some aspects of analysis of covariance. *Biometrics*. **38** 541-561.

Cox, D.R. and Wermuth, N. (2003). A general condition for avoiding effect reversal after marginalization. *Journal of the Royal Statistical Society*, Ser. B. **65** 937-941.

Cox, D.R. and Wermuth, N. (2004). Causality: a statistical view. *International Statistical Review*. **72** 285-305.

Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society*, Ser. B. **41** 1-31.

Dawid, A.P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association*. **95** 407-448.

Dawid, A.P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*. **70** 161-189.

Dawid, A.P. (2003). Causal inference using influence diagrams: The problem of partial compliance (with discussion). In *Highly Structured Stochastic Systems*, (ed. P.J. Green, N.L. Hjort and S. Richardson). Oxford University Press: New York.

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

Didelez, V. (2006). Asymmetric separation for local independence graphs. In *Proceedings of the* 22^{nd} *Annual Conference on Uncertainty in Artificial Intelligence*. 130-137.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society*, Ser. B. **70** 245-264.

Didelez, V., Dawid, A.P. and Geneletti, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the* 22^{nd} *Annual Conference on Uncertainty in Artifical Intelligence*. 138-146.

Didelez, V. and Sheehan, N. (2007). Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*. **16** 309-330.

Drton, M. (2007). Likelihood ratio tests and singularities. *Annals of Statistics*. To appear.

Durbin, J. (1954). Errors in variables. *Review of the International Statistical Institute*. **22** 23-32.

Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*. **86** 9-26.

Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*. **137** 334-353.

Eichler, M. and Didelez, V. (2007). Causal reasoning in graphical time series models. In *Proceedings of the* 23^{rd} *Annual Conference on Uncertainty in Artificial Intelligence*. 109-116.

El Barmi, H. and Dykstra, R. (1995). Testing for and against a set of linear inequality constraints in a multinomial setting. *The Canadian Journal of Statistics*. **23** 131-143.

Feder, P.I. (1968). On the distribution of the log likelihood ratio test statistic

when the true parameter is near the boundaries of the hypothesis regions. *The Annals of Mathematical Statistics*. **39** 2044-2055.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh.

Fisher, R.A. (1932). *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh.

Gawrilow E. and Joswig M. (2004). Polymake web demo. Available online at URL: `http://www.math.tu-berlin.de/polymake/`.

Geary, R.C. (1942). Inherent relations between random variables. *Proceedings of the Royal Irish Academy*. **47** 63-76.

Geary, R.C. (1943). Relations between statistics: the general and the sampling problem. *Proceedings of the Royal Irish Academy*. **49** 177-196.

Geyer, C.J. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics*. **22** 1993-2010.

Gillies, D. (2000). *Philosophical theories of probability*. Routledge: London.

Goldberger, A.S. (1972). Structural equation methods in the social sciences. *Econometrica*. **40** 979-1001.

Goldszmidt, M. and Pearl, J. (1992). Rank based systems: a simple approach to belief revision, belief update, and reasoning about evidence and actions. In *Proceedings of the* 3^{rd} *International Conference on Knowledge Representation*

and Reasoning, (ed. B. Nebel, C. Rich and W. Swartout). Morgan Kaufmann: San Mateo, CA.

Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*. **37** 424-428.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*. **11** 1-12.

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*. **50** 1029-1054.

Heckerman, D. and Shachter, R. (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artifical Intelligence Research*. **3** 405-430.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*. **81** 945-960.

Hume, D. (1748). An enquiry concerning human understanding.

Imbens, G.W. and Angrist, J.D. (1994). Identification and estimation of local average treatment effects. *Econometrica*. **62** 467-475.

Jewell, N.P. (2003). *Statistics for Epidemiology*. Chapman & Hall/ CRC Press: London.

Laplace, P.S. (1814). *Essai philosophique sur les probabilities*. Courcier: Paris. Reprinted (1912) in English (F.W. Truscott and F.L. Emory, Trans.) by Wiley: New York.

Lauritzen, S.L. (1996). *Graphical models*. Oxford University Press: Clarendon, Oxford, UK.

Lauritzen, S.L. (2001). Causal inference from graphical models. In *Complex Stochastic Systems*, (ed. O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg). CRC Press: London.

Lauritzen, S.L. (2003). Graphical models for surrogates. *Bulletin of the International Statistical Institute*. **60** 144-147.

Lauritzen, S.L. (2004). Discussion on causality. *Scandinavian Journal of Statistics*. **31** 189-201.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N. and Leimer, H.G. (1990). Independence properties of directed Markov fields. *Networks*. **20** 491-505.

Lipid Research Clinic Program (1984). The lipid research clinics coronary primary prevention trial results, part I and II. *Journal of the American Medical Association*. **251** 351-374.

Manski, C.F. (1990). Non-parametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*. **80** 319-323.

Mill, J.S. (1843). A system of logic.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Ser. A. **135** 370-384.

Pearl, J. (1993). Comment: graphical models, causality and interventions. *Statistical Science*. **8** 266-269.

Pearl, J. (1995a). Causal diagrams for empirical research. *Biometrika*. **82** 669-710.

Pearl, J. (1995b). Causal inference from indirect experiments. *Artificial Intelligence in Medicine*. **7** 561-582.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press: Cambridge, UK.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th Annual Conference on Uncertainty in Artifical Intelligence*, 411-420.

Reiersøl, O. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*. **9** 1-24.

Reiersøl, O. (1945). Confluence analysis by means of instrumental sets of variables. *Arkiv för Matematik, Astronomi och Fysik*. **32A** 1-119.

Riccomagno, E.M. and Smith, J.Q. (2007). The causal manipulation of chain event graphs. Technical Report, University of Warwick.

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

Robins, J.M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modelling*. **7** 1393-1512.

Robins, J.M. (1989). The analysis of randomized and non-randomized AIDS

treatment trials using a new approach to causal inference in logitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, (ed. L. Sechrest, H. Freeman and A. Mulley). U.S. Public Health Service, Washington D.C.

Robinson, L.D. and Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review.* **59** 227-240.

Rockafellar, R.T. and Wets, R.J.-B. (1998). *Variational analysis.* Springer: New York.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology.* **66** 688-701.

Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics.* **6** 34-68.

Rubin, D.B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science.* **5** 472-480.

Rubin, D.B. (2004). Reply to discussion of "Direct and indirect causal effects via potential outcomes." *Scandinavian Journal of Statistics* **31** 196-198.

Russell, B. (1913). On the notion of cause. *Proceedings of the Aristotelian Society.* **13** 1-26.

Savage, L.J. (1954). *The foundations of statistics.* John Wiley and Sons: New York.

Shafer, G. (1996). *The art of causal conjecture*. MIT Press: Cambridge, MA.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, Ser. B. **13** 238-241.

Smith, J.Q. and Anderson, P.E. (2008). Conditional independence and chain event graphs. *Artificial Intelligence*. **172** 42-68.

Sommer, A. and Zeger, S.L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine*. **10** 45-52.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, prediction and search*. Springer-Verlag: New York.

Strotz, R.H. and Wold, H.O.A. (1960). Recursive vs nonrecursive systems: an attempt at synthesis (part I of a triptych on causal chain systems). *Econometrica*. **28** 417-427.

Suppes, P. (1970). *A probabilistic theory of causation*. North-Holland Publishing Company: Amsterdam.

van der Vaart, A.W. (1998). *Asymptotic statistics*. Cambridge University Press: New York.

Verma, T. and Pearl, J. (1988). Causal networks: semantics and expressiveness. In *Proceedings of the* 4^{th} *Workshop on Uncertainty in Artificial Intelligence* (Mountain View, CA), 352-359. Reprinted in *Uncertainty in Artificial Intelligence*, (ed. R. Shachter, T.S. Levitt and L.N. Kanal). **4** 69-76. Elsevier: Amsterdam.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*. **11** 284-300.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*. **54** 426-482.

Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*. **9** 60-62.

Williamson, J. (2005). *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press: New York.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*. **20** 557-585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*. **5** 161-215.

Yule, G.U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*. **2** 121-134.