A Philosophical Analysis of Causality in Econometrics

Damien James Fennell

London School of Economics
and Political Science

Thesis submitted to the University of London
for the completion of the degree
of a Doctor of Philosophy

August 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

This thesis makes explicit, develops and critically discusses a concept of causality that is assumed in structural models in econometrics. The thesis begins with a development of Herbert Simon's (1953) treatment of causal order for linear deterministic, simultaneous systems of equations to provide a fully explicit mechanistic interpretation for these systems. Doing this allows important properties of the assumed causal reading to be discussed including: the invariance of mechanisms to intervention and the role of independence in interventions. This work is then extended to basic structural models actually used in econometrics, linear models with errors-in-the-equations. This part of the thesis provides a discussion of how error terms are to be interpreted and sets out a way to introduce probabilistic concepts into the mechanistic interpretation set out earlier. The resulting analysis is then critically compared with similar work by economists, Stephen LeRoy (1995) and Kevin Hoover (2001a) who both develop Simon's work on causal order in different ways. In the latter part of the thesis, the mechanistic interpretation set out at the beginning is used to interpret identification conditions. Typically, these are presented in econometrics as mathematical conditions for determining whether unknown parameters in equations can be measured from observation. In the thesis it is shown that the identification conditions imposed on sets of equations when interpreted mechanistically require a sparseness of causal structure that ensures that experiments are hypothetically possible of the causal structure. It also analyses the role of identifiability conditions in causal inference. The final part of the thesis shows that the mechanistic interpretation developed in the thesis succeeds, unlike Simon's own methods for analysing spurious correlation, in avoiding important criticisms by Nancy Cartwright (1989) whose own approach to inferring causal structure from observations is also critically analysed.

Contents

List of Tables and Figures

Acknowledgements

First thanks go to my parents and my brothers and sister, Jérôme, Claire and Vincent, for their constant support in spite of the often unclear explanations of what this PhD is all about. Hopefully now all will become clear! (At least I hope.)

I would especially like to thank Nancy Cartwright, not only for her excellent supervision, but also for her kindness. To have a good supervisor is a blessing but to have a supervisor who is your friend also, well. In addition, I would like to thank Stephen LeRoy, Mary Morgan and Marcel Boumans for their helpful and encouraging comments. Also, thanks to Jon Williamson for the greatly appreciated last-minute comments on the thesis.

Thanks to Antti Saaristo for the countless fun discussions over coffee on a whole range of philosophical topics, a key part of my philosophical education over the past five years. Also thanks to Philipp Beckmann, Gregor Betz, Philipp Dorstewitz, Vincent Guillin, Federica Russo, Georg Theiner and Bettina Woll for the lunchtime social support. I am also very grateful to everyone in the philosophy department at UCSD, I was made very welcome there; thanks to Andrew Hamilton, Anna Alexandrova and Belinda Aber. Thanks too to Chris Eads for being a crucial link to the real world beyond the academic bubble.

Lastly, I want to thank Hakan for listening to my thoughts, worries and generally putting up with me. Without you, everything would have been so much harder – *teşekkürler*.

Chapter 1

Introduction

1. Causality in Philosophy and Economics

Causal claims matter. So many actions are guided by beliefs about causes: from simple actions, such as taking vitamin C pills to avoid catching a cold, to institutional actions, such as tightening border controls to keep out illegal immigrants. To successfully navigate and control our environment, it helps to know the causes of things.[1] Economics is no exception. As shown by the recession that followed the previous government's hurried withdrawal from the Exchange Rate Mechanism in 1992, it is important to get economic decisions right. Crucial to this is understanding the causal relations in the economy.

It may seem a truism to say that to intervene effectively requires information about causal relations. But for most of the last century, in both philosophy and economics, explicit causal talk was frowned upon. In philosophy there were important influences that inhibited the explicit use of causal concepts. Perhaps most influential was (and remains) David Hume's (1739) analysis of causality in terms of time-ordered, contiguous and constantly conjoined events. This analysis has formed the basis of subsequent Humean attempts to explain causal concepts away using regularities.[2] Also influential at the time was Bertrand Russell's (1913) argument that causal concepts should be dropped in favour of functional relationships like those used in the physics. Russell argued that the concept of causality was ambiguous and confusing and, as seen by its absence in physics, unnecessary. Both Hume and Russell's work were highly influential on the logical positivism that dominated philosophy in the first half of the 20th century. Logical positivism restricted the attribution of truth to empirically verifiable claims. This left no place for metaphysics.[3] Following both Hume and Russell,

[1] This is enshrined in the motto of the London School of Economics: *rerum causas cognoscere* ('to know the causes of things').

[2] Hume's influential idea is that the necessity in the relationship between a cause and its effect cannot be observed, all that can be observed is the constant conjunction of a cause and its effect, that is, repeated observations of the cause and effect happening together. Humean views of causality take this to heart and try to reduce causal relations to regular associations between types of events.

[3] See, for example, Carnap (1932).

the logical positivists viewed talk of causality as unduly metaphysical, ambiguous and to be avoided.

Economics and econometrics were not immune to this reluctance to engage in causal talk. [4] In a recent paper entitled 'Lost Causes', Kevin Hoover (2004) provides historical evidence and a discussion of the phenomenon.[5] Interestingly, the dip in causal talk is particularly pronounced in the period *after* econometrics developed (1930-40's). Despite its development during the period when logical positivism was dominant, the founders of econometrics analysed the problem of how to distinguish causal and non-causal relations.[6] Though not always explicitly put in causal terms, one of their aims was to develop a method for identifying causal relations between factors of interest from economic data, which could be exploited for policy purposes.[7]

Indeed for a short time at the beginning of the 1950's causal language was particularly explicit in some papers on econometric method. This is clear in Koopmans (1950), Orcutt (1952) and Simon (1952; 1953; 1954). In these papers the word 'cause' is freely used. Of these papers, Herbert Simon's (1953) paper is perhaps the most interesting. This is because in it he attempts to present a formal definition of causal order. [8] However, his work is not antithetical to the strict empiricism of the time. Though Simon sees causal order as a useful concept in science (unlike Russell), he attempts to make it empirically respectable by operationalising the concept. This attempt to bring together explicit causal talk and the empiricism of the time, however, in fact preceded the drop in causal talk discussed in Hoover (2004). This, according to Hoover (2001a, pp.147-149), was in part due to Simon's assuming an equivalence between his definition of causal

[4] Perhaps the most striking example of the use of language that hides causal content is the continued use of 'structural' rather than 'causal' to characterise relations that are, in common sense terms, causal.
[5] Kevin Hoover has nice graphs showing the decline of causal language in economic and econometric papers (2004, pp.152 -153).
[6] See Morgan (1990) for a relevant history of the development of econometrics.
[7] For examples of relevant early econometric work, see Tinbergen (1939) and Frisch (1938).
[8] 'Causal order' is Simon's term for causal structure, I also use it in this way throughout the thesis.

order and conditions for identification.[9] This encouraged subsequent econometric analysis to take the non-causal identification conditions as an acceptable substitute for causal discussion. As Hoover puts it, after the Simon paper '[c]ausal language simply faded away.' (2001a, p.147).

Happily for those who think causal language eases discussion of intervention, talk of causes is no longer taboo in philosophy. Logical positivism is no longer dominant and philosophers, such as Patrick Suppes (1970) and Nancy Cartwright (1983) have attacked Russell's dismissal of causality.[10] In addition, Suppes, Cartwright and other philosophers have developed diverse analyses of causality.[11] Since then, the philosophy of causality has blossomed. One area where work has progressed particularly rapidly is in the Bayesian Network analysis of causal relations. This work, developed by Peter Spirtes, Clark Glymour, Richard Scheines, Judea Pearl and others,[12] assigns a causal interpretation to a Bayesian network.[13] Other important developments have included the growth of analysis on the relationship between causal relations and counterfactuals.[14]

This resurgence in interest in causality is not confined to philosophy. Econometrics too has recently begun to discuss causality more openly. One concept of causality which has been accepted by the econometric mainstream is that of Granger causality, see Granger (1980; 1988). This concept, closely related to Suppes's theory of probabilistic causality, is Humean in flavour.[15] It is an approach where, like Hume, causes are assumed to precede their effects in time.

---

[9] Identification conditions for a set of structural equations with unknown coefficients ensure that the unknown values of the coefficients can be deduced from observations and knowledge about the form of the equations.

[10] Suppes (1970, pp.6-7) argues that Russell's arguments based on physics no longer apply, since modern physics does use causal concepts. While Cartwright (1983, chapter 1) argues that Russell is wrong to see functional relationships as an adequate substitute for causal concepts, since causal concepts are required to account for the difference between effective and ineffective strategies.

[11] Suppes (1970) is a classic text setting out his theory of probabilistic causality. Cartwright's can be found in her books (1983), (1989) and (1999).

[12] See, for example, Spirtes, Glymour and Scheines (1993) and Pearl (2000).

[13] A Bayesian network is a convenient representation of a joint probability distribution using a directed acyclic graph. The connection with causality comes by relating conditional probabilities with causal relations, as is standard in theories of probabilistic causality. See Williamson (2004) for an overview of Bayesian Networks and their relationship with causality.

[14] See Collins *et al.* (2004) for a selection of recent papers on the topic.

[15] Roughly, one earlier random variable is a 'Granger-cause' of another later variable, if the history the earlier variable improves predictions of the later variable given the later variable's and other relevant variables' history.

The Granger approach contrasts with the approach of the Cowles Commission in the 1950's where causes could be simultaneous with their effects.[16] This latter approach is motivated by a desire to model equilibrium relationships using simultaneous equation systems.[17] It is also a key characteristic of Simon's treatment of causal order (1953) that two or more variables can be (simultaneously) co-determined. Recently, there has also been renewed interest in Simon's work in econometrics. Economists such as James Heckman, Kevin Hoover and Stephen LeRoy[18] have developed formal definitions of causal relations based to a greater or lesser extent on Simon's 1950's work.

The general resurgence of causal discussion in econometrics is also evident in a recent special issue in the journal of econometrics. The issue is built around a paper by Adams *et al.*(2003) that presents a mammoth study of the causal relationships between socioeconomic status and health in elderly Americans. The paper is very rich and breaks new ground by supplementing the Granger approach to causality with invariance tests typically associated with the structural (Simon-like) approach to causality. Unsurprisingly, the paper generates a lot of comment which the rest of the journal presents. The commentaries can be broken into two camps: those that discuss the hypotheses Adams *et al.* test (e.g. Adda *et al.*(2003), Poterba (2003)); and those that discuss the methodology, that is, definitions of causal relations and methods used for finding out about them (e.g. Florens (2003), Geweke (2003), Granger (2003), Hausman (2003), Heckman (2003), Hoover (2003), Mealli and Rubin (2003), and Robins (2003)). The fact that the disproportionate number of comments fall in the second, methodological camp shows the importance to econometricians today of obtaining clear answers to the following two questions.

(i) What is meant by a causal relation?

(ii) How does one find out about causal relations?

---

[16] During the 40's and 50's there was a methodological debate between those who wanted to restrict modelling to time-ordered, dynamic systems and those who wanted to admit simultaneous equation models. See Morgan (1991).

[17] This motivation can be understood in part by the overwhelming emphasis of economic theory on modelling systems in equilibrium.

[18] See, for example, Heckman (2000), Hoover (2001a; 2001b) and LeRoy (1995; 2004).

This thesis addresses these two questions for a particular treatment of causality: the influential position set out in Simon (1953). In this way, it aims ultimately to contribute to the growing discussion of causality in econometrics.

Note also that these two questions matter to anyone affected by economic policy. This is because the way results of studies like Adams *et al.* (2003) are interpreted and used depends vitally on answers given to (i) and (ii). For instance, Adams *et al.* claim that their study supports the hypothesis that there is no causal impact from socioeconomic status to health in the elderly population they study. Such claims might encourage significant policy changes, such as cuts in government spending on pensions. Policy changes like these are everyone's concern. Therefore, the answers given to questions (i) and (ii) are important, not only to philosophers and econometricians, but ultimately to all those affected by economic policy decisions. This is in part what motivates this thesis.

2. Overview of Thesis

This thesis aims to develop a clear, explicit presentation of a particular treatment of causality. The type of causal relations analysed are those that are implicitly assumed in simultaneous structural equation models in econometrics. Such models are used widely throughout economics and econometrics, wherever equilibrium relations are modelled as static relations.[19] Since modelling equilibrium relations in this way is central to economic theory, focussing on these systems gives the work a wide relevance.

It is important to note that the particular formalisation of causal relations – which this thesis presents, develops and critically compares with that of others – is not intended as a universal theory of causal relations. I make no claim that this approach to causality can be used everywhere and anywhere. Instead, the aim is to make explicit one particular way of attributing causal content to one kind of mathematical model used in econometrics. Ultimately this should help in understanding how econometrics results, like Adams *et al.*'s (2003), can be

[19] Strictly speaking the approach here is general in that it can also be applied to time-ordered relations. What is distinctive about the approach is that it is tailor-made to fit simultaneous equation systems. However, this does not prevent it being applied to non-simultaneous systems.

understood and used.[20] Nevertheless, for systems where the assumptions of the formal system of causal relations presented in this thesis are met, then this treatment of causal relations is applicable. However, in cases where the assumptions are not met, one must be cautious. There may well be other, possibly inconsistent views of causal relations that may be more appropriate for understanding such cases.[21]

To give a brief outline of the thesis, chapter two develops a strong reading of Herbert Simon's (1953) formal treatment of causal order. In doing this, it aims to unpack the causal content attributed to simple sets of equations of the kind Simon analyses. The resulting strong reading is an explicit formalisation of causal concepts that can be attributed to such systems of equations. Unlike Simon's treatment however, it is not assumed that an equivalence holds between identification conditions and causal order.[22] The chapter finishes with an exploration of some of the important properties of the causal concepts in this formalisation.

Chapter three continues the work of chapter two by extending the strong reading to apply to more general systems of equations. The aim is to move from the extremely simple deterministic, linear systems of equations, for which a causal interpretation is provided in chapter two, to slightly more complex systems like the simplest models actually used in econometrics. Importantly, this includes introducing error terms and random variables. The chapter also takes advantage of one proposed extension to present a brief exploration of different types of interventions.

The next chapter looks in some detail at work by Stephen LeRoy (2004) and Kevin Hoover (2001a). Their work is particularly relevant since both hold

[20] Though the thesis develops a causal interpretation for simple systems of equations like those used in econometrics (see chapter three), the highly complex models used in studies like Adams *et al.* (2003) are beyond the scope of the thesis. The aim expressed here remains to be fulfilled through further work.

[21] A strong advocate of having different theories of causality for different situations is Nancy Cartwright. See, for example, Cartwright (2003c).

[22] This makes the treatment here more general than Simon's since it can be applied to unidentifiable systems of equations.

positions on causal order that are developed from Simon's work on causal order. The chapter begins with a presentation of LeRoy's position and raises some relevant criticisms. The chapter then looks at Hoover's treatment and concludes that Hoover and LeRoy's positions are both unduly restrictive when compared with the strong reading.

Chapter five analyses the relationship between identification conditions and causal order. As mentioned above, this is important given Simon's equivalence claim between causal order and identification conditions. In this chapter, Simon's position is set out clearly and criticised. The next part of the chapter analyses, from the perspective of the strong reading, just what imposing identification conditions on structural equations requires of the causal order denoted by that system. The chapter ends with a brief look at the epistemic role of the identifiability conditions, both for measuring the values of structural coefficients (those that represent the strength of causal influence between factors) and for performing limited inferences about the causal order (i.e. structure) between factors.

The final chapter moves more deeply into the questions of how causal relations can be inferred from observation, by looking in some depth at work by Simon (1954), which controversially claims to show that causal relations can be deduced from observed correlations. It considers and develops Cartwright's (1989) criticism of this claim, concluding that the strong reading provides a better approach for setting out how causal relations can be inferred from observation in the way Simon would like. It then critically considers Cartwright's own method for inferring causal order. The chapter finishes with a brief comparison of Cartwright's method for inferring causal order with that of the strong reading.

Finally, the thesis ends with some brief comments on how the work in the thesis might be developed in the future in relation to both econometrics and the philosophy of science.

Chapter 2

Mathematical Equations and their Causal Interpretations: The Strong Reading of Herbert Simon's Concept of Causal Order

1. Introduction

This chapter explores one way sets of linear equations can be used to represent causal relations. It focuses and builds on Herbert Simon's influential 1953 paper 'Causal Ordering and Identifiability' in which Simon defines a causal order for sets of equations. Simon's work is focused on because it presents one of the best attempts to set out explicitly the causal content of deterministic linear simultaneous structural equation models. This is particularly relevant to econometrics since these models are simpler versions of the models actually used in econometrics to model equilibrium relations.

Herbert Simon's paper presents a detailed formal definition of causal order for sets of equations, but gives only a sketchy discussion of how this formal order is to be interpreted. Therefore, this chapter attempts to extend Simon's interpretative discussion to present a more complete and explicit picture of how Simon's formally defined 'causal' order for sets of equations can be interpreted. It also aims to make explicit the properties of causal relations that the resulting interpretation assumes. Importantly, the position set out in this chapter differs from Simon's in that it is more general. Unlike Simon's treatment of the causal order in his 1953 paper, here it is *not* required that causally ordered systems be identifiable.[1] Instead, the chapter focuses on the first part of Simon's paper, before he introduces identification, and builds on the analysis there to set out a distinctive 'strong reading' of Simon's formal order for sets of equations.

The strong reading presents a formalisation of causal concepts. However, it is important to note that I do not claim that this formalisation of causal concepts holds of, or applies to causal systems in general. The aim is simply to set out

[1] An equation is identifiable in a set of equations if its coefficients can be deduced from knowledge of the form of all of the equations and from observations of the variables in the system of equations. In chapter five I discuss identification and how Simon ties his concept of causal order to it in more detail.

clearly the causal content that can be attributed to simple mathematical models, like those used in econometrics. Ultimately, the hope is that this reading can be developed to provide a causal semantics for structural models actually used in econometrics.

The chapter is structured as follows. It starts by presenting a problem faced when trying to represent causal relations with equations, the problem that mathematically equivalent sets of equations can have different causal interpretations. This 'conceptual equivalence problem' needs to be overcome if sets of equations are to explicitly represent causal relations. The next section presents Simon's formal definition of causal order and shows how it helps with, but does not suffice to solve the conceptual equivalence problem. The problem is that it lacks an explicit causal interpretation, so the next section fills this in so that the conceptual equivalence problem is avoided. This is done by building on Simon's comments on how the formal order is to be understood. The result is a clear interpretation of sets of equations and a definition of the causal order that such sets of equations represent. With this in place, the chapter identifies and discusses important properties of the now explicit causal relations. Finally, the chapter reconsiders how the conceptual equivalence problem is solved to emphasise the importance of directly controllable factors and mechanisms in the reading of sets of equations. It then proposes a formal change to the syntax of equations to make formally explicit the strong reading, the causal interpretation of equations developed in this chapter.

2. A Challenge to Representing Causal Relations with Mathematical Equations

Mathematical models are widely used in economics and in many other disciplines. Of their different uses one of the most significant is to present idealized hypotheses of the causal relations that obtain between different factors of interest to the modeller. For example, consider

$$p = \alpha$$
$$q = \beta p + \gamma, (\beta < 0)$$

as presented these equations can be interpreted using a purely *mathematical interpretation*, that is, there are two variables p and q and three coefficients α, β, γ and these satisfy the two linear equations and the inequality above. Now suppose

that these equations are intended to act as a mathematical model of 'something else' which is done by attributing the equations with some further content. Suppose that p denotes the price of a good and q denotes quantity demanded of that good. Suppose also that the coefficients α, β, γ denote factors that are not caused by those denoted by p and q. Finally, suppose that factors denoted on the right hand side of an equation are causes, while the factor denoted on the left is the effect of the causes denoted on the right. This additional content provides an alternative *model interpretation* for the equations: the equations denote a causal model of demand in which increases/decreases in price cause decreases/increases in demand.

This simple mathematical model of demand highlights a general point about mathematical models: the functional relations of a mathematical model can be read in two ways. The first way is as a piece of pure mathematics, that is, the equations in virtue of being mathematical equations have a mathematical reading under which they can be manipulated according to the rules of algebra (or whatever calculus is appropriate). However, when a set of equations is used as a model of 'something else' then in virtue of being about something else, there will be a distinct 'model' interpretation of the set of equations. Typically, the model interpretation is much richer than the mathematical interpretation since it draws on a wealth of background knowledge and theory about what is to be modelled using the set of equations. The mathematical interpretation, on the other hand, is much simpler since it relies solely on the highly abstract semantics of pure mathematics.

One of the main reasons for mathematising a model is that, if done successfully, one can derive mathematical results from the equations that when interpreted from the model perspective (hopefully) provide new and interesting claims about the model the equations represent. This is beneficial for many reasons. Among other things, it can aid model development and can be used to generate predictions for testing hypotheses in the model. However, the mathematical and model interpretations of the equations must correspond in the right way for this to work. Specifically, benefits of mathematical modelling are put in jeopardy if there is a

divergence between the acceptable mathematical manipulations of the set of equations and the model interpretation.

To see how such problems can occur, reconsider my simple example above. Mathematically, one is perfectly entitled to manipulate the equations and redefine the coefficients to obtain the following mathematically equivalent set of equations.

$$q = \alpha'$$
$$p = \beta' q + \gamma', (\beta' < 0)$$
$$\text{where } \alpha' = \alpha\beta + \gamma, \beta' = (1/\beta) \,\&\, \gamma' = -(\gamma/\beta).$$

Being mathematically equivalent, these equations have identical meaning to the original set of equations under the mathematical interpretation. However, this is not true under the model interpretation. If one follows the method for reading equations used earlier (reading effects on the left hand side and causes on the right) this new set of equations has a radically different model interpretation to the first set of equations: an increase/decrease in demand causes a decrease/increase in price. Thus, by performing mathematically acceptable transformations one completely changes the earlier model interpretation of the equations in the original demand model. This is a serious problem because this derived set of equations no longer represents the original demand model. As such, it is of no use for developing that demand model, for testing it using predictions and so on.

I call this a 'conceptual equivalence' problem. The problem is that the two sets of equations above are equivalent under a mathematical interpretation but not under a model interpretation. It is a serious problem because it implies, if one uses the first set of equations as a model in the way set out earlier, that one is not free to mathematically manipulate the equations since this may unwittingly change the meaning of the set of equations under the model interpretation. More intuitively, the problem is that important aspects under the model perspective have no *explicit* counterpart under the mathematical perspective. In the example here, the mathematical form of the demand model did not make explicit the causal order assumed in the model interpretation of the equations. Since there was no mathematical feature of the equations that denoted this causal order, it was

possible to implicitly change this causal order when deriving an alternative, mathematically equivalent set of equations.

The most direct way to solve a conceptual equivalence problem is to impose conditions that ensure that the two sets of equations are mathematically equivalent if and only if they are equivalent under the model interpretation for those equations. One way to do this is to ensure that an isomorphism holds between the mathematical equations and a 'model reading' of those equations. Specifically, this requires that for each mathematical term there is a corresponding model term and that mathematical relations satisfied by the mathematical terms correspond to model relations satisfied by the model terms which correspond to the mathematical terms.[2] Essentially, this is to demand a clear set of translation rules for moving between the mathematical and model interpretation of the equations, constructed such that both mathematical and model reasoning respect these translation rules. If this is done, then however one reasons with the equations, be it mathematically or directly using the model concepts, one can be sure that if one translates the results of such reasoning, one obtains a meaningful and correct result in the other interpretation.

In this chapter the object of interest is the representation of causal relations using mathematical equations. Therefore the conceptual equivalence problem that I am concerned to solve is that of the example presented here: how can one ensure that one has a mathematical formalism which makes explicit the causal content in the model so that mathematical derivations from those equations respect that content? An obvious first step is to define in the mathematical domain something that can represent the causal content in the model. This is why in this chapter, I start from Herbert Simon's work because his formal order definition in his 1953 paper provides an ideal candidate for this. I now present his formal order and consider how it helps with the conceptual equivalence problem.

[2] More precisely, the isomorphism is between a set of mathematical terms and a *subset* of model terms. So the requirement is simply that formalism represent correctly a subset of the model concepts. In this way the requirement allows the model language to be richer in content than the mathematical language. This is necessary because typically the model language embodies on a lot of theory, other hypotheses, specifics about the situation being modelled and other influences which cannot be fully formalised. Indeed, as is discussed below, the model concept of 'experimenter/nature', used by Simon, is an example of a model term which is not represented explicitly by the mathematical formalism.

3. Simon's Formal Ordering Method

In this section, I set out Herbert Simon's formal methods for causally ordering variables in a set of equations. The section begins with a brief outline of two of Simon's ordering methods, one mathematical the other logical, before discussing the contribution it makes towards solving the conceptual equivalence problem.

3.1. The Formal 'Causal' Order for Sets of Equations

In his 1953 paper Simon presents a method for determining what he calls the 'causal order' for a set of equations.[3] The sets of equations for which formal orders are defined are special systems that meet certain conditions. First, the equations relate coefficients and variables (where Greek letters denote coefficients and Latin letters variables). The distinction between variables and coefficients is important as they are later interpreted in different ways by Simon. It is also assumed that the equations are linear in the variables and coefficients[4] and that they are linearly independent.[5] Simon also distinguishes between linear structures and linear models. A linear structure is a set of equations meeting the conditions above, where coefficients have specific, non-zero values. A linear model is the set of equations where coefficients can take any possible value; in other words it is the set of possible linear structures where all of these have the same functional form. A linear structure is called 'self-contained' if it is solvable for the variables in terms of the coefficients. Simon defines his formal 'causal' order for self-contained linear structures.[6] In simpler terms, Simon's formal order is defined for sets of linear equations that are solvable for the variables and which have no equations that are redundant for solving for the variables.

---

[3] In the discussion that follows I prefer to use 'formal order' rather than Simon's 'causal order'. The reason is that the order that Simon defines is itself merely an ordering that arises from where variables appear in equations. As it stands, there is nothing *causal* about this. This is why I prefer to use 'causal order' for the model reading of the formal order, where the intuitive causal content is clearer.

[4] Simon also extends his ordering method for non-linear systems that are solvable by sequential substitution in the way that solvable systems of linear equations are. However, I omit discussion of this here as it is not substantively different from the linear case.

[5] This means that no equation in the set can be derived using other equations in the set.

[6] Though it is immediate from the way that formal order is defined that all the self-contained linear structures in a self-contained linear model have the same formal order, see Simon (1953, p.15). So, one can equally take the definition of formal order as applying to self-contained linear models.

To find the formal order for a set of solvable linear equations using Simon's method, one begins by identifying the smallest subsets of equations which can be solved for the variables that appear in them: the complete subsets of 0^{th} order. One then solves for the variables in these subsets in terms of coefficients and substitutes these solved variables into any equations that remain outside of the complete subsets (provided there are some). These remaining equations are the derived set of equations. One then repeats the process, treating the derived set of equations as the whole set of equations was treated above. In other words, one identifies the complete subsets (now of 1^{st} order) for the set of equations, solves for the variables in these and then substitutes these into any remaining equations. Continuing this until no equations remain, the result is an ordered partition (the complete subsets) of the equations, in which a complete subset of equations of $n-1^{th}$ order 'directly precedes' a complete subset of n^{th} order if one of the variables solved for using the equations of the first complete set was substituted into some equation(s) of the second complete set when solving for that second set of equations. This sometimes branched order of sets of equations is what Simon calls the 'causal order'. From this order, Simon defines a variable as 'exogenous' relative to a complete subset of equations if it appears in that set of equations and in an earlier complete subset. A variable is 'endogenous' relative to a complete subset of equations if that complete subset is the first in which it appears.

In addition, Simon defines an alternative causal order over the variables that appear in the complete subsets of equations. In this case one orders the variables according to the order in which they are solved. For example, a set of variables that are solved for in a complete subset of 0^{th} order of equations, form a corresponding complete subset of variables of 0^{th} order. In this definition of causal order complete subsets of equations are replaced by the sets of variables that are solved using the complete subsets of equations. In this order, one complete subset of variables precedes a second subset if and only if a variable in the first is exogenous with respect to the equations that are used to solve for the variables in the second.

Despite the involved terminology, Simon's method is straightforward in practice. To see this, recall the earlier demand example (p and q are variables and α, β and γ are coefficients).

$$p = \alpha$$
$$q = \beta p + \gamma, (\beta < 0)$$

Following Simon's method, the first equation is solvable for p in terms of coefficients but the second equation is not solvable for q using that equation only. Thus the first equation forms the only complete subset of 0^{th} order. Solving for p and substituting its solution into the remaining equation gives us an equation for q which is solvable for q in terms of coefficients, so the second equation forms the only complete subset of 1^{st} order. At this point no equations remain so the process is complete. Since p was substituted into the second equation to solve for q, the complete subset of the first equation is causally ordered prior to the complete subset containing the second. It follows from this that p is endogenous relative to the first equation and exogenous relative to the second, while q is endogenous relative to the second equation.

To calculate the alternative formal order among the variables, note that p is solved for using the first equation alone so $\{p\}$ appears at the beginning of the ordering. Since q is solved by substituting p into the remaining equation, $\{q\}$ comes next in the order. This covers all the variables in the equations and since p is necessary for solving for q, the formal order among the variables is $\{p\} \rightarrow \{q\}$.

It is important to note that Simon's formal ordering method applies to simultaneous equation models like those used in economics to model equilibrium relations. Consider, for instance,

$$i = \omega$$
$$q = \delta p + \eta i$$
$$q = \phi p + \nu$$

where i denotes income, q the equilibrium quantity transacted of a good and p the equilibrium price of that good. The second equation is the demand equation, while the third is the supply equation. As a whole, it can be read as a simple supply and demand model of a good. This system can be ordered using Simon's method. Here, i can be solved for using just the first equation so it comes at the

beginning of the order, while q and p can only be solved for together using the last two equations once i has been solved for. So Simon's order here is $\{i\} \rightarrow \{p,q\}$. In such a system, p and q are in the same complete subset, they are 'co-determined'. Note also that in this example the 'effect on the left, causes on the right of the equals sign' reading the causal relation does not give the same order. This shows that Simon's ordering method is distinct from this other way of reading a causal order though in some cases, like the simple previous demand example, the two readings coincide.[7]

Before considering whether this formal ordering method helps deal with the earlier conceptual equivalence problem, I first present for purposes of completeness an alternative version of Simon's formal ordering method that is set out in logical rather than mathematical terms.

3.2. An Alternative formalism: Simon's Logic of the Causal Order

Simon (1952) proposes a logic of the causal relation, in the style of Carnap, in which causal order is defined in a way that fits neatly with the way it is defined for equations. In this definition, one takes as given an object language based on a finite number of logically independent and empirically testable atomic sentences, subject to a set of empirically testable laws. Importantly, the atomic sentences are partitioned into two groups, condition and observation sentences, a distinction that corresponds to the distinction made between coefficients and variables in the equations above. Laws are taken to be empirical sentences asserting material conditionals from a condition sentence to a molecular sentence constructed from the observation sentences.[8] In this set up, the causal order is defined as a partition on a complete[9] and consistent[10] set of laws. As in the equation case, the causal

[7] The reason for introducing the 'effect on the left of an equation and causes on the right' method of reading causes earlier was so that the conceptual equivalence problem could be presented. As shown in the simultaneous system here, this simple way of ordering variables in equations is not that of Simon.

[8] Here laws are the analogue to equations in the mathematical version: laws set out implications from condition sentences to observation sentences, just like the set of equations sets out how the values of the coefficients determine the values of variables.

[9] A set of laws is complete if their conjunction determines, given the truth of all the condition sentences, the truth value of every observation sentence. This is the analogue of the solvability constraint on sets of equations in the mathematical version.

[10] In order to get a unique causal order, Simon requires that the set of empirical laws be consistent, that is, if two sets of laws determine the same observation sentence then one of the sets must be a

order can equivalently be taken as defined over sets of observation sentences (*cf.* variables) that are determined (*cf.* solved for) by the correspondingly ordered subsets of laws (*cf.* complete subsets of equations). When the causal order is defined over sets of laws, the key is to give precedence to the smaller subsets of laws that determine observation sentences.[11] In other words, the smallest subset of laws that determine the smallest set of observation sentences comes first in the causal order. Further minimal subsets of laws that determine further observation sentences (and thus properly contain the smallest subset) come subsequently in the order.

To show how this approach gives results consistent with the mathematical approach, I formalise the demand example in terms of Simon's logical analysis. Recall the equations of the demand example:

$$p = \alpha$$
$$q = \beta p + \gamma, (\beta < 0)$$

Define the following as the condition and observation sentences, for some arbitrarily chosen a, b and c, where the condition sentences are constraints on the values of the coefficients, the observation sentences constraints on the values of the variables.[12]

Condition Sentences: C_1 iff $\alpha = a$

C_2 iff $\beta = b$ and $\gamma = c$

Observation Sentences: O_1 iff $p = a$

O_2 iff $q = ba + c$

The next step is to add laws that correspond to the two equations above.

Laws: $C_1 \rightarrow O_1$ (if $\alpha = a$ then $p = a$)

$C_2 \rightarrow (O_1 \leftrightarrow O_2)$ (if $\beta = b$ and $\gamma = c$, then $p = a$ iff $q = ba + c$)

To get the causal order, note that the first law determines the truth value of the observation sentence O_1 if C_1 is true. While both laws together determine both O_1 and O_2 if C_1 and C_2 are true. Since the set of the first law is contained in the set of the two laws together, it follows that the condition sentence determined by the

---

subset of the other. This is the analogue of the linear independence requirement on sets of equations in the mathematical version.

[11] A subset of laws determines an observation sentence if, when the antecedents of all the laws are true, the observation sentence has determinate truth value.

[12] Here C_1, C_2, O_1 and O_2 are assumed to be logically independent. This assumption corresponds to the variation free assumption in the mathematical case, which is discussed later in the chapter.

first law alone causally precedes the condition sentence determined by the two laws. In other words for arbitrary *a, b* and *c*: $p = a$ causally precedes $q=ba + c$, which matches the ordering obtained for the equations in the mathematical causal order.

This logical version of the causal order also shows that Simon's formal definitions fit closely with other analyses of the logic of causal relations. Nancy Cartwright (1989, pp.25-29) shows that Simon's treatment of equations can be interpreted in terms of John Mackie's treatment of causes as *INUS* conditions for their effects. In Mackie's (1974) work, a development of John Stuart Mill's, causes are generally *INUS* conditions for their effects[13] where *A* is an *INUS* condition for *B* if and only if *A* is an insufficient but necessary part of an unnecessary but sufficient set of conditions for *B*. A cause which is not an *INUS* condition because it is sufficient is called a 'complete cause'.

From the presentation above, one can see quite immediately that for the first law, that C_1 is a complete cause for O_1, since C_1 is sufficient for O_1. Similarly, the second law implies that $(C_2 \wedge O_1) \rightarrow O_2$, which shows that C_2 and O_1 are *INUS* conditions for O_2.[14] Returning to the equation form, this is to say the following, for any *a,b* and *c*.

$\alpha = a$ is a complete cause for $p = a$

$\beta = b$, $\gamma = c$ and $p = a$ are each *INUS* conditions for $q=ba + c$

In short, Simon's logical approach implies that causally precedent variables and coefficients are *INUS* conditions or complete causes for the causally antecedent counterparts.[15]

[13] Mackie and Cartwright both make the point that not all *INUS* conditions are causes. I discuss this in more detail in chapter six.

[14] Strictly speaking they are *INS* conditions since they are not unnecessary. This is a result of the simple example used and does not affect the content of the discussion.

[15] More precisely, it is variables constrained to certain values that are the *INUS* conditions.

3.3. Does Simon's Formal Order Help Solve the Conceptual Equivalence Problem?

Having presented Simon's formal methods for ordering variables, I now return to the conceptual equivalence problem. In the version of the problem presented earlier, the set of equations on the left in table 2.1 could be transformed into the mathematically equivalent set of equations on the right. The problem is that the two sets of equations have different causal interpretations.

Table 2.1. Equations, Formal Order and Model Interpretation

Mathematical Equations	$p = \alpha$ $q = \beta p + \gamma$ $(\beta < 0)$	$q = \alpha'$ $p = \beta' q + \gamma', (\beta' < 0)$ *where* $\alpha' = \alpha\beta + \gamma, \beta' = (1/\beta)$ & $\gamma' = -(\gamma/\beta)$.
Intuitive Causal (Model) Interpretation	'price causes demand'	'demand causes price'
Simon's Formal Order	$\{p\} \rightarrow \{q\}$	$\{q\} \rightarrow \{p\}$

How can Simon's formal order help? Well, applying Simon's formal ordering method for variables to the first set of equations gives $\{p\} \rightarrow \{q\}$, while for the second it gives $\{q\} \rightarrow \{p\}$. This is an attractive result because it clearly matches the intuitive causal interpretations for the respective sets.

The formal order also helps with the conceptual equivalence problem because it changes if one transforms the system on the left to that on the right. So, adding Simon' formal order to the mathematical representation would rule out the problematic transformation of the first set into the second that lead to the change in intuitive causal interpretation. This is because the different formal orders, now part of the formal representation, imply that the two sets of equations are no longer formally equivalent. So, by providing an explicit mathematical counterpart for the causal interpretation of the equations Simon's formal order seems to solve the conceptual equivalence problem.

However this is too hasty. Simply adding a formal order to the mathematics cannot be a full solution to the conceptual equivalence problem. This is because to solve the problem requires that an isomorphism hold between the mathematical

and the relevant model interpretation of the equations so that one can move between the mathematical and model interpretation of the equations without jeopardising the way the equations are read under either interpretation. Yet no detailed information has been presented about a model interpretation.[16] Without this one cannot be sure that the required isomorphism holds between the model interpretation of the equations and the mathematical equations to which Simon's formal order is added. The worry is: even if one restricts mathematical transformations on the equations to those which preserve Simon's formal order, what reason is there to believe that this preserves the *model* interpretation of the equations?

Obviously in order to complete the solution to conceptual equivalence problem, more information is required about the model interpretation of the equations. Only in this way can one be sure that the Simon's formal order adequately represents the causal order assumed in the model. If one wants to be sure that Simon's formal order represents causal order in the model interpretation, one needs to set out the *causal order in the model* that corresponds to Simon's formal order. This is what is done in the next section.

4. The Model Reading of the Equations

In this section I set out a model reading for equations by building on some short comments made by Herbert Simon about how his formal order is to be interpreted. In doing this, I *assume* that an isomorphism holds between model terminology that Simon introduces and the mathematical terminology of the set of equations. The resulting model concepts that are isomorphic to the equations I call the *model reading* of the equations.[17] In this way the conceptual equivalence problem discussed earlier is avoided by construction. The key result of this analysis is a model counterpart to Simon's formal order: *the causal order* that is represented by

[16] All that has been provided is an informal, loose way of reading causes on the right of the equation and the effect on the left.

[17] It is important to note that the model reading is *not* the same as the model language, the model language is the language which is used to talk about the model, whereas a model reading is the model interpretation of a set of equations (which by construction is isomorphic to the set of equations). The model reading is expressed using the model language but the model language also contains terminology that is outside the model reading. See appendix 2.1 for a formal definition of the model reading.

the formal order of a set of equations. Finally, to distinguish the position developed here from Simon's, I call my interpretation of Simon's work *the strong reading.*[18]

The section begins with a brief overview of the philosophical influences on Simon at the time he wrote his 1953 paper. This provides some context as to how he goes about providing an interpretation to his formal order.[19] The subsequent part works from Simon's relevant discussion to set out an appropriate model interpretation for sets of equations with formal order.

4.1. Simon's Empiricism

In order to make sense of how equations are to be interpreted *à la* Simon, it helps first to review briefly the philosophical views that were most influential on Simon's 1950's work on causal order. The first influence was operationalism and its influence is evident in repeated parts of his writings. To give just one example, he notes in the introduction to his 1957 volume, following Bridgman, that an 'operational definition of a variable is a specification of the way in which the variable is to be measured' and is necessary to 'relate the model to empirical observations' (1957, p.6). Simon sees his work as extending this process of operationalising concepts to relate them to empirical data. In particular, he views his work on causal order as an attempt to provide an operationalisation of the intuitive asymmetry between cause and effect. Or, as he puts it in his paper, 'the aim of this chapter is ... to provide a clear and rigorous basis for determining when a causal ordering can be said to hold' (1953, p.12). The second obvious influence was logical positivism.[20] This influence is very clear in his 1952 paper which sets out a logic of the causal order in the style of Carnap. Consistent with his admiration of operationalism and logical positivism, Simon espouses a strong scepticism for metaphysics. This is expressed when he states, following Hume,

[18] I call it 'strong' because it adopts a stronger reading of mechanisms than Simon's operationalist position does, this is reflected in the fact that I, unlike Simon, do not require identifiability of sets of equations. For more on this and on Simon's position, see chapter 5.

[19] It also provides some context for later discussion of Simon in chapter five, where I discuss the role identification plays in his view of causal order.

[20] It is debatable just how consistent Simon is being in mixing operationalist and logical positivist approaches (for a brief discussion of the need to distinguish between operationalism and logical positivism, see Suppe (1998)). This said, I do not elaborate on this since the main aim of this chapter is to make explicit the structural view of causality that Simon uses to interpret equations, and this structural view can be held in tandem with a whole range of metaphysical positions.

that '[o]bservation reveals only recurring associations' (1953, p.10) and that '[t]he only "necessary" relationships among variables are the relationships of logical necessity that hold in the scientist's model of the world' (*ibid.*, p.11). So, just like Hume, for Simon it seems that the world provides us merely with regularities, which one then models/interprets as being necessarily connected as cause and effect. Or as Simon puts it 'causal orderings are simply properties of the scientist's model' (*ibid.*,p.11).

This brief overview of Simon's philosophy also helps to explain the way he sets out an interpretation of his formal order. He does this by introducing a 'metalanguage' which is used when speaking about the set of equations. This approach fits with his philosophical position because it attempts to avoid metaphysical assumptions he views as problematic. So, whereas a realist might explain the causal interpretation of mathematical models by setting out what real metaphysical elements the terms in the mathematical language of the equations refer to, Simon's approach takes a reverse approach. He introduces a metalanguage containing terms that have intuitive causal meaning, and associates this terminology with the object language, that is the mathematical (or logical) language of the equations. Of course, this metalinguistic move is not in itself sufficient for avoiding metaphysical assumptions, since one still requires some explanation of what the truth of metalinguistic sentences consists in. Presumably, to maintain his empiricist approach Simon would hold the view that the truth of a metalinguistic sentence can be determined by a some measurement procedure.

In any event, here the focus is on the metalanguage, not what determines the truth of its sentences.[21] In the interpretative analysis of Simon's formal order that follows, I take the metalanguage to correspond to what I call the model language. The analysis builds on Simon's discussion of the metalanguage to make explicit a model reading for sets of equations.

[21] Though this is obviously a very important discussion to have.

4.2. Interpreting Equations Causally

In this section I rationally construct connections between terms in the model language (Simon's metalanguage) and terms in the formal language[22] (mathematical or logical). The rationality assumed in this construction is that the model terms should be isomorphic to the formal terms so that movements between the formal language and the model language are not contentious in respect to the set of equations. This is rational because it ensures that a conceptual equivalence problem is avoided. The model counterpart that results for a set of equations is the model reading.

The reconstruction begins by drawing on various comments of Simon to establish basic connections between terms in the model language and the formal language. These basic connections are set out in the table 2.2.[23] Once these first connections are set out, it then uses Simon's treatment of formal order to stipulate, given the isomorphism assumption, the model terminology that consistently corresponds to the formal terminology. Table 2.3 is the result of this part of the reconstruction.

Table 2.2. Basic Model – Formal Language Correspondences

Model Language	*Formal Language (logical version)*
A mechanism. A mechanism is a relation between a set of directly and indirectly controllable factors,[24] it constrains the possible values that the factors in the set can take as a group.	A linear equation[25] (A law)
Experimenter/Nature	

[22] Ideally a formalisation of both the terms in the model language and the formal language should be presented in order to make fully explicit the isomorphism that holds between the two. An attempt to do this is presented in appendix 2.1.

[23] The rationality in the reconstruction is visible in the fact that corresponding elements in this table and others are chosen so as to be structurally similar i.e. terms correspond as do the relations among those terms.

[24] Here 'factor' is introduced as a model term that corresponds to either coefficients or variables. I assume that factors are quantitative i.e. they can take on a numerical values.

[25] For the more general nonlinear version, simply replace 'linear equation' by 'functional relation' in the right hand column.

A factor in a mechanism that is directly controllable by experimenter/nature	A non-zero coefficient in a linear equation (Condition sentences[26])
A factor in a mechanism that is indirectly controllable by experimenter/nature	A variable in a linear equation (Observation sentences)
A directly controllable factor is intervened into by experimenter/nature to take a particular value	A coefficient has a particular value (A condition sentence has determinate truth value)
Directly controllable factors can be independently controlled.	The coefficients are *variation free*, that is, the set of possible values of a group of coefficients is the Cartesian product of the set of possible values for each.[27] (Logical independence of condition sentences)

Table 2.2 is constructed from a series of comments by Simon. The first correspondence draws on the following quotes:

> 'the phrase ... will perhaps become clearer if we substitute "mechanism" for "equation"'(1957, p.7)

> 'To provide an operational definition for a mechanism is to specify a method for determining whether the mechanism is operative or inoperative (other than by measuring *the variables that the mechanism is supposed to connect*).' (my emphasis, 1957, p.7)

The first quote clearly suggests a correspondence between 'mechanism' and the formal term 'equation' while the emphasised part of the second quote suggests the role in mechanisms in constraining factors.[28] The next four correspondences in the table follow from Simon's comment that:

> 'We suppose a group of persons whom we shall call "experimenters." If we like, we may consider "nature" to be a member of the group...[they] are able to choose the nonzero elements of the coefficient matrix of a linear structure, but they may not replace zero elements by nonzero elements or vice versa (i.e. they are restricted to a specified linear model). We may say that they

[26] Strictly speaking, a condition sentence is a constraint on a coefficient e.g. '$\alpha = 2$' is a condition sentence. Similarly, observation sentences are constraints on variables (that may include coefficients).

[27] In other words, each coefficient can take on any value regardless of the values of others.

[28] To keep model and mathematical terminology separate, I use 'factor' as a model term that corresponds to either a variable or a coefficient.

> *control directly* the values of the nonzero coefficients...[and] *control indirectly* the values of these variables' (original emphases, 1953, p.26)

Thus, directly controllable factors correspond to coefficients, indirectly controllable factors correspond to variables and direct controlling of a factor corresponds with a coefficient being set to a value. The last correspondence in table 2.2 connects 'independence' in the model language with the condition that the coefficients are *variation free*. Formally, this means that the domain of possible values for coefficients as a group is the Cartesian product of the sets of their individually possible values.[29] Informally, it means that the coefficients can take any value as a group as they can individually. This correspondence is slightly difficult since Simon is not very explicit about the need for independence among directly controllable factors/coefficients. It is hinted at in the quote above when mentions that the experimenters are able to choose *presumably freely* the non-zero values of the coefficients. Though this is not sufficient to imply the variation free assumption, I have read it as such[30] since this seems to fit best with the way he changes coefficients freely in the dialogue and with his comments above about the experimenters choosing values of the coefficients.

It is important to note that in the table 2.2 there is no correspondent to 'experimenter/nature' in the formal language, since there is simply nothing in the mathematics that can play that role.[31] It is included nonetheless because the experimenter/nature plays a role in the way Simon discusses what the sets of equations represent. However, the absence of a formal correspondent to 'experimenter/nature' does *not* undermine the isomorphism between the equations and the model reading. This is simply because I *assume* the model reading to be just those model terms which correspond to formal terms.[32] So 'experimenter/nature' is not part of the model reading of a set of equations. Nevertheless, 'experimenter/nature' is a relevant and an important term in the

---

[29] So a set $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$of coefficients is variation free, letting $P(\alpha_i)$ denote the set of possible values for α_i, if and only if $P(\alpha_1, \alpha_2, \ldots, \alpha_n) = P(\alpha_1)xP(\alpha_2)x\ldots xP(\alpha_n)$.

[30] Kevin Hoover also reads it in this way (Hoover, 2001a, p.61-62). James Heckman also assumes that inputs are variation free in his treatment of structural models which is in part based on Simon's work (Heckman, 2000, p.54).

[31] Ultimately it is the person who solves for the variables who seems to be taking the place of the experimenter/nature! But this is clearly not part of the formal language.

[32] See the definition of the model reading in appendix 2.1.

model language because it bears important connections with model terms that *are* in the model reading for a set of equations. For instance, it is relative to the experimenter/nature's 'control' that the distinction between indirect and direct controllable makes sense.[33]

In table 2.3 below, correspondences are constructed by holding fixed the correspondences in the first table above and using the isomorphism assumption to make explicit some model terminology that adequately corresponds to formal terminology used in Simon's definition of the formal causal order. Though this sounds involved, the process is straightforward. The right hand column in table 2.3 is populated with descriptions of important steps/definitions in construction in the Simon's formal order. To infer the corresponding model terminology, one simply 'translates' as much as possible into the model language using the links in table 2.2 with any remaining gaps postulated using the rationality assumption above.

Table 2.3. Implied Model language – Formal Language Correspondences

Model language	*Formal Language (logical version)*
A set of mechanisms whose directly controllable factors are fixed to particular values.	A linear structure, that is, a set of linearly independent equations whose coefficients have particular values. (A set of laws whose condition sentences have determinate truth value)
A set of mechanisms whose directly controllable factors are unfixed, that is, are free to be fixed at any one of a set of possible values.	A linear model, that is, a set of linearly independent equations whose coefficients can have any non-zero values. (A set of laws whose condition sentences can take on different possible truth values)

[33] It could be argued that the terminology used in this chapter should be modified to make this more intuitive since 'model reading' may naturally be conflated with 'model language'. I have some sympathy with this view, but given the clarification presented here, however, I stick with this terminology for now and leave its modification as further work.

A determining set of mechanisms, that is a set of mechanisms, which if it has its directly controllable factors set to particular values, constrains all of its indirectly controllable factors to take a unique value.	A self-contained linear model. A set of linear equations which has a unique solution for its variables in terms of coefficients. (A set of laws for which, if condition sentences have determinate truth value then its observation sentences have determinate truth value.)
A minimal subset of mechanisms of 0^{th} order. This is a subset of mechanisms in which only the directly controllable factors in the mechanisms need to be set to particular values for the indirectly controllable factors in the subset to take particular values.	A complete subset of 0^{th} order. A minimal subset of linear equations in a model for which all variables can be solved in terms of the coefficients. (The smallest set of laws that determines the truth value of an observation sentence.)
A minimal subset of mechanisms of n^{th} order. A minimal subset of mechanisms in which, given that other minimal subsets of lower order have indirectly controllable factors fixed at particular values, when its directly controllable factors are set to particular values all of its unfixed indirectly controllable factors are fixed to particular values.	A complete subset of n^{th} order. A minimal set of linear equations for which all variables can be solved in terms of the coefficients and variables solved in complete subsets of order less than n. (The smallest set of laws that determines the truth value of an observation sentence.)
An indirectly controllable factor is exogenous relative to a minimal subset of mechanisms, if it figures in those mechanisms and it needs to be taken as fixed for that minimal subset of mechanisms to have its unfixed indirectly controllable factors set to particular values.	A variable is exogenous relative to a complete subset, if that variable appears in the equations and its value must be given for the complete subset to be solved for all its variables. (An observation sentence is exogenous to a set of laws which is not the smallest set that determines it.)

An unfixed indirectly controllable factor is endogenous relative to a minimal set of mechanisms, if it has its value is fixed by that minimal set of mechanisms.	A variable is endogenous relative to a complete subset, that is a variable which is solved for when a complete subset is solved for all its variables. (An observation sentence is endogenous to the smallest set of laws that determines it.)
A minimal subset of mechanisms, *C*, is *directly causally dependent* on another minimal subset, *B*, *(B→C)* if at least one fixed indirectly controllable factor in *C* has its value fixed by *B*.	A complete subset of equations, *C*, is *directly causally dependent* on another complete subset, *B*, *(B→C)* if at least one endogenous variable for *B* is an exogenous variable for *C*.[34] (A minimal set of laws is directly causally precedent on another if at least one observation sentence, endogenous for the former, is exogenous for the latter).
A minimal subset of mechanisms, *C*, is *causally dependent* on the minimal subset of mechanisms, *B*, if there exists a sequence of minimal subsets such that that $B \rightarrow B_1 \rightarrow B_2 \rightarrow \ldots \rightarrow B_k \rightarrow C$.	A complete subset (minimal set of laws), *C*, is *causally dependent* on the complete subset of equations, *B*, if there exists a sequence of complete subsets (minimal sets of laws) such that that $B \rightarrow B_1 \rightarrow B_2 \rightarrow \ldots \rightarrow B_k \rightarrow C$.

The second last row of table 2.3 is of particular interest since it provides what I have been aiming for: an explicit definition of *causal order in the model language*:

> A minimal subset of mechanisms, *C*, is directly causally dependent on another minimal subset, *B*, *(B→C)* if at least one fixed indirectly controllable factor in *C* has its value fixed by *B*.

There are other important definitions in the table, namely the definitions of exogeneity and endogeneity for factors and a definition of causal dependence among subsets of mechanisms (the last row).[35]

[34] In his definitions of causal relations (1953, pp.18,22), Simon defines the relations over the sets of endogenous variables for the complete subsets of equations. The versions presented here take the complete subsets as the relata, which Simon also allows. This is done because it fits more neatly into the presentation above.

[35] Note that causal dependence is the transitive closure of direct causal dependence. Therefore the causal relation defined here is transitive.

The correspondences here have been constructed from the formal order defined over equations to give a counterpart causal order over mechanisms. A more intuitive causal order can also be defined over sets of indirectly controllable factors, as a counterpart to the formal order defined over variables. To define this version of the causal order one merely replaces the minimal subsets of mechanisms with the corresponding sets of indirectly controllable factors that are fixed by those mechanisms.[36] This gives us the causal order among factors in the model.

4.3. The Model Interpretation of the Earlier Example

Though the table above provides a full interpretation for the equations, including an interpretation of the causal order from the model perspective, it is all somewhat obscured by the dense presentation style. So to bring out some of the intuitive content of this model interpretation for a set of equations, reconsider the earlier demand example:

$$p = \alpha$$
$$q = \beta p + \gamma, (\beta < 0)$$

Read as a model, the first equation represents a mechanism that determines price. The coefficient α denotes a directly controllable factor that, if set to a value by the experimenter/nature, given the price mechanism causes the price factor to indirectly takes a value (here the price equals to the value of the α-factor). The second equation represents the demand mechanism, if the experimenter/nature directly sets the values of the β and γ factors and also indirectly sets the price factor to a value (by directly setting the α-factor using the price mechanism above) then this, given the demand mechanism, indirectly sets the quantity demanded to a particular value (i.e. $\beta p + \gamma$). The causal order among the indirectly controllable factors is that the price directly precedes quantity demanded.[37]

[36] This is the same trick that was used earlier to get the formal order over variables from that over equations. Though this time it is done in the model reading.

[37] Note that though the experimenter/nature is used in describing the model interpretation of the equations, the experimenter/nature is not itself represented by the equations since there is no formal term that represents it. So, as noted above, 'experimenter/nature' is a term in the model language but *not* the model reading of the equations. Nevertheless the term bears important relations to terms that are in the model reading such as 'directly controllable factor'.

There are two interesting points here. First, the discussion just given exactly mirrors the way p and q were solved for in the demonstration of Simon's formal order earlier in the chapter. Solving for p first using the first equation corresponds to the price being set to a value using the price mechanism, while solving for q using p corresponds to the quantity demanded being set given values for price and the other directly controlled factors in the demand mechanism. This correspondence brings out clearly the isomorphism built into the table above.

Second, this correspondence between the formal order and the way price and quantity are set, also brings out what Simon's formal order represents in the model. In my above example, price causally precedes quantity demanded in the model because in order to set the quantity demanded the price needs to be set to a value. The converse is false, quantity demanded does not need to be set to a value in order for price to be set to a value. Price does not require that any other indirectly controllable factor be set to a value in order for it to be set to a value, which is why it comes at the beginning of the causal order. Only price (among the indirectly controllable factors) needs to be set to a value for quantity demanded to be set, so only price causally precedes quantity demanded. This gives us the causal order *in the model* that price causally precedes quantity demanded. This is how the formal order $\{p\} \rightarrow \{q\}$ for the variables is interpreted using the model language.

To conclude, this section has attempted to make explicit the connections between the formal languages of mathematics and logic, and a model terminology used for interpreting these. This has been done by drawing on comments made by Simon and building model counterparts from these by assuming an isomorphism holds. The result is the model reading.

In addition, note that the absence of identification in the analysis shows that it diverges from Simon's own (1953) treatment. The role identification plays in Simon's treatment is discussed in detail in chapter five. So, unlike Simon, no

assumption is made here that the equations need to be identifiable[38] in order to be attributed a causal order.[39] For this reason, I call the position set out here the strong reading of Simon.

With this model interpretation of the equations and the formal order, it is now possible to explore the properties of the causal order. This is done in the next section.

5. Important Properties of Causal Order in the Model

With this interpretative machinery in place, it is worthwhile to look in some more detail at what a causal order in the model entails. In particular, it is worthwhile to investigate the properties that a system of factors with causal order, represented by a set of equations with formal order, is assumed to have. This is important because ultimately one would like to know when and under what conditions this concept of causal order can be applied in modelling real-world situations.

The section maps out some important properties of the system of factors represented by a set of equations with formal order. It first clarifies the concept of mechanisms before considering the factor properties. These properties are (i) the close relationship between changes in factors and causal order, (ii) the invariance of mechanisms to factor changes, (iii) the independence of directly controllable factors and (iv) the possibility of factors cancelling each other out. Going through these helps to clarify the nature of causal concepts that are assumed in model reading presented above.

Before doing this, however, I first define a few more relations to add to Simon's analysis. This helps with the subsequent analysis and also extends the causal order to apply to intuitive situations not covered by Simon's formal definitions.

[38] Recall that an equation is identifiable in a set of equations if its coefficients can be deduced from knowledge of the form of all of the equations and from observations of the variables in the system of equations.

[39] So the approach here is more general than Simon's since it can be applied to non-identifiable sets of equations.

5.1. Aside: Some Supplementary Causal Relations

In this short aside, I propose some extensions to Simon's definitions. This is to remedy some counterintuitive omissions in Simon's treatment.[40] The first counterintuitive omission is that Simon does not include coefficients in his formal order.[41] From the model perspective, this restriction of the causal ordering to indirectly controllable factors seems unnecessary since, intuitively, directly controllable factors can be causes. So the first extension is to generalise the formal order to cover coefficients. The second counterintuitive omission is that Simon's definition of formal order is defined over sets of entities (equations or variables) rather than the entities themselves; this is odd since causal relations intuitively hold between entities themselves: one is inclined to say that a factor is a cause of another rather than some set of factors is causally ordered prior to another set. So here I define relations that hold between individual variables. I then present model definitions that are counterparts for these formal definitions.

First, I extend Simon's formal order to apply to coefficients in addition to variables. This is done by treating coefficients that appear in a complete subset of equations as if they were exogenous variables to that complete subset of equations. The reason for treating coefficients like exogenous variables in Simon's formal ordering method is that the values of coefficients are taken as given when solving for the endogenous variables, just like exogenous variables.

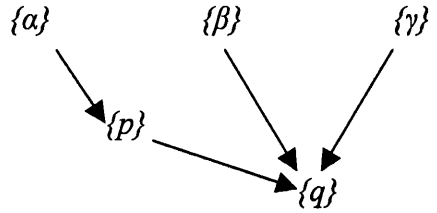
One can add coefficients to Simon's formal order over variables by putting the coefficient in a singleton set and placing that set as directly precedent in the formal order to any complete subset of variables for which the coefficient appears in the equations in which those variables are endogenous (i.e. are solved). So in the demand example

[40] The extensions proposed in this section do not add any content since they are defined using Simon's formal relations and do not make any additional assumptions.

[41] Recall that coefficients denote factors that can change, not constants as in some other treatments (e.g. LeRoy who is discussed in chapter four). This is why it is intuitive to include them explicitly in the causal order.

$$p = \alpha$$
$$q = \beta p + \gamma, (\beta < 0)$$

the formal order among the variables and coefficients, expressed here as a causal graph, where the arrows denote the relationship of direct causal precedence of formal order among sets of variables,[42] would be:



The coefficient *{α}* directly precedes *{p}* because α appears in the equation which is used to solve for p, and likewise for *{β}* and *{γ}* in relation to *{q}*. To keep this new ordering separate from Simon's formal order, I call this order which includes coefficients the *extended formal order.*

The second extension is to define relations that apply to individual variables (or coefficients) rather than sets of variables (or coefficients). The first relation is:

> For x and y any two variables in a self-contained linear model with formal order: x is *causally equivalent* to y if and only if x and y are endogenous for the same complete subset of equations.

In simpler terms, two variables are causally equivalent if and only if they are in the same complete subset in the formal order of the variables. With this new relation, I then define

> For x any variable or coefficient and y any variable in a self-contained linear model with formal order: x is a *direct cause* of y, if and only if x is a coefficient which appears in some equation in the complete subset of equations for which y is endogenous, or x is a variable, or is causally equivalent to a variable, which is exogenous to the complete subset of equations for which y is endogenous.

In simpler terms, a variable or coefficient is a direct cause of another variable if and only if it appears in the complete subset of equations for which y is solved, but is not endogenous to that set. In the extended formal order above, a direct cause appears in a set which directly precedes the set containing the other variable.

---

[42] See, for example, Simon (1953, pp.21, 23).

With this, recursively define a causal relation between variables (or coefficients) as follows.

> For x any variable or coefficient and y any variable in a self-contained linear model with formal order: x is a *cause* of y if and only if
>
> (i) x is a direct cause of y
>
> OR
>
> (ii) x is a direct cause of, or is causally equivalent to, a cause of y.

By construction *cause* is a transitive relation. Also as defined, two causally equivalent variables are not causes of each other because in order to get the recursive definition of cause to work at least one 'link' from a cause to its effect needs to be directly causal. This is done intentionally to avoid the possibility of two variables causing each other which, by transitivity, would imply the counterintuitive result that each of the two variables is a cause of itself. In addition, it is easily checked that the direct cause relation is anti-symmetric and anti-reflexive, from which it follows that cause as defined here is also anti-symmetric and anti-reflexive. Since it is transitive, anti-symmetric and anti-reflexive, the causal relation defined here satisfies some of the basic *a priori* intuitions one has for the causal relations.

Finally, it is straightforward to 'translate' the relations defined above into the model language to obtain counterpart model relations. To do this, simply define direct cause and causal equivalence in the model (using the links in the tables above) as:

> Two factors c_1 and c_2 in a determining set of mechanisms with causal order are *causally equivalent* if and only if they are both fixed by the same minimal set of mechanisms.

> For two factors c and e in a determining set of mechanisms with causal order: a factor, c, is a *direct cause* of a factor, e, if and only if c is a directly controllable factor that figures in the complete subset of equations for which e is endogenous, or c is, or is causally

equivalent to, a factor which is exogenous to the complete subset of equations for which e is endogenous.[43]

The causal relation among factors is defined recursively in exactly the same way, but using the model relations rather than the formal relations.[44]

5.2. Properties of the Causal Relations in the Model

With this framework in place, I now investigate some of the important properties of the causal order.

5.2.1. Clarifying Mechanisms

Before beginning the discussion on causal properties, it helps to elaborate a little on what mechanisms are and how they relate to directly and indirectly controllable factors. To do this, consider the following two systems of equations and their formal orders.

$$\text{(A)}\quad \begin{array}{l} p = \alpha \\ q = \beta p + \gamma \\ \{p\} \rightarrow \{q\} \end{array} \qquad \text{(B)}\quad \begin{array}{l} q = \delta p + \lambda \\ q = \beta p + \gamma \\ \{p, q\} \end{array}$$

Suppose that system (A) denotes the simple demand model presented throughout the chapter, that is, the first equation denotes the mechanism that determines price of a good (say the government sets the price by law) while the second equation denotes the demand mechanism for that good. Suppose that in system (B) the second equation denotes the same demand mechanism as in the first model, but instead of government controlling price as in the model denoted by (A), in this model the market is free, that is, the first equation denotes a supply mechanism relating price and quantity of the good.

These two systems illustrate some important features of mechanisms. First, is that in the first system price is exogenous with respect to the demand mechanism[45] while in the second system, though the demand mechanism is the *same*, the price factor is now endogenous with respect to that mechanism. This shows that *which indirectly controllable factors are exogenous or endogenous for a mechanism*

[43] For definitions of exogenous and endogenous for factors see Table 2.3.
[44] I omit the definition because it would be identical to that above, given the caveat here.
[45] Since p is taken as given in solving for q using the second equation when applying Simon's formal order.

depends on the other mechanisms in the system. More generally, the causal order among the indirectly controllable factors depends on the whole set of mechanisms acting together, so to speak.

Another way of understanding how this arises is to see mechanisms as constraints on the possible values of indirectly controllable factors. So in system (A) the first equation denotes a mechanism that constrains price to be equal to the value of the directly controllable factor denoted by α. Likewise, the second equation denotes a demand mechanism where the possible values of price and quantity are constrained so that their values satisfy the equation $q = \beta p + \gamma$. It is important to note that though indirectly controllable factors' values are constrained by mechanisms, the directly controllable factors values are not. *Mechanisms constrain the values of indirectly controllable factors but not directly controllable factors.* So in system (A), since price is completely constrained by the first (government price control) mechanism, the demand mechanism then fixes the quantity sold given the (already) fixed price. Whereas in system (B) the supply mechanism does not fully constrain price, instead it constrains both the quantity sold and price together. It is only when the supply and demand mechanisms act together that price and quantity take a fixed value (given the values of the directly controllable factors). This shows how the causal order among the indirectly controllable factors is determined by subsets of mechanisms given values of directly controllable factors. Directly controllable factors are not constrained by mechanisms, instead they are determined *outside* the system by the experimenter/nature.

5.2.2. Change and Causal Order

The relationship between causal order and change is easiest to appreciate by looking at the demand example again.

$$p = \alpha$$
$$q = \beta p + \gamma, (\beta < 0)$$

Here the formal order among the variables is $\{p\} \rightarrow \{q\}$. Formally, this is the order of substitution for p and q using Simon's method. But this isn't very enlightening. It is more instructive to consider what happens to the solutions of p

and q if one changes the values of only one coefficient. The different possibilities are set out in table 2.4.

Table 2.4 Changes in Variables Given Changes in only one Coefficient

Only Change Coefficient	Variable Changes
α	*p and q change*
β	*q changes*
γ	*q changes*

Given the equations, if the value for α changes then the solved value for p changes, and since the other coefficients do not change, q also has a different solved value. However, if either β or γ changes alone then only the solved value for q changes. This captures the core idea of Simon's formal order, that changes in values of variables are necessarily accompanied by changes in the values of the variables that follow them in the causal order. Or as Simon puts it, the '[formal] causal ordering specifies which variables will be affected by intervention at a particular point (a particular complete subset) of the structure' (*ibid.*, p.26). He presents a precise version of this claim in a theorem:

> *'Theorem 6.1*: Let A be a self-contained linear structure, let A_1 be a complete subset of order k in A, and let A' be a self-contained linear structure that is identical with A [and of the same linear model] except for a single equation belonging to A_1 ...Then (a) the values of all variables in A that are neither endogenous variables of A_1 nor causally dependent, directly or indirectly, on the endogenous variables in A_1 are identical with the values of the corresponding variables in A'; and (b) the values of all variables in A that are endogenous variables of A_1 or are causally dependent on the endogenous variables of A_1 are (in general) different from the values of the corresponding values in A'.'(*ibid.*, p.25)

Put simply, the theorem states that if one has two systems of equations that are identical except for the value of a coefficient appearing in one equation in a complete subset, then the variables exogenous with respect to that complete subset will be identical cross-systems while the endogenous variables and variables causally dependent on those will generally[46] differ.[47]

[46] The 'in general' in the theorem is there to cover the case where, due to other coefficients happening to have a particular set of values, one or more endogenous or causally dependent

With the model correspondences and supplementary causal relations, light can be shed on the causal order in the model by translating this formal theorem into its model counterpart. To do this, I read the difference in the coefficient(s) in the equation which differs across systems in the theorem 6.1 as a directly controllable factor in the mechanism being changed by the experimenter/nature from the value it had in the first system to the value it had in the second. Interpreted in this way, one gets a counterpart theorem in the model language.

> *Model Counterpart to Theorem 6.1*: Let A be a determining set of mechanisms. If a minimal subset of mechanisms, A', has exactly one of its mechanisms changed by the experimenter/nature changing the directly controllable factor(s) in that mechanism then: (a) indirectly controllable factors that are not fixed by that minimal subset of mechanisms nor caused by factors that are will remain unchanged; (b) indirectly controllable factors that are fixed by that minimal subset of mechanisms and factors caused by these will generally change.

This version is interesting because it helps to make explicit important properties of the causal order. To start with, it follows immediately from this counterpart theorem that causes have the property that when changed by experimenter/nature their effects and factors causally equivalent to them generally change. Conversely, factors that are not causally equivalent nor effects of changed factors do not change. In short, *the causal order maps out how changes in factors go together.*

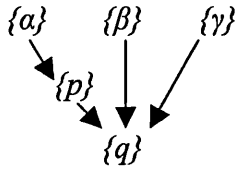
---

variables 'coincidentally' do not have a distinct value across the systems. I analyse this qualification in more detail later in the section.

[47] Though Simon does not discuss this case, theorem 6.1 can be extended to cover cases where more than one equation has distinct coefficients across the two systems. In that case any variable, that is endogenous or causally dependent on an equation which changes, 'in general' changes while variables that are not causally dependent with respect to all equations that change, do not change. Importantly, this extended result shows that it is not a problem to apply Simon's formal ordering methods to sets of equations in which coefficients appear in more than one place in the equations, since the extended result permits changes in more than one equation. Also note that the model reading set out here does not depend in any way on whether coefficients are repeated or not. So sets of equations with repeated coefficients are not a problem for the strong reading set out in this chapter.

To see this, consider the earlier demand example with its extended formal order on the right.

$$p = \alpha$$
$$q = \beta p + \gamma, (\beta < 0)$$



To indirectly change price, the α-factor must be changed directly by the experimenter/nature. If the experimenter/nature does this then the price changes, by the price mechanism. Assuming there is no direct change to either of the β- and γ-factors then the change in the price, directly caused by the change in the α-factor, directly causes a change in the quantity demanded by the demand mechanism. Similar descriptions can be given for changes in the either of the β- and γ-factors. For example, if only the β-factor is directly changed by the experimenter/nature then quantity demanded changes but price does not.

The way that changes in factors go together is particularly easy to appreciate using the extended formal order for the variables. Roughly, the theorem implies that changes 'flow down' the arrows, so in the model reading changes in α-factor are followed by changes in price and quantity. While changes in the β-factor or γ-factor lead to changes in quantity, but do not lead to changes in price.

5.2.3. Invariance of Mechanisms to Change

An important assumption in Simon's theorem 6.1 is that the functional form of the two sets of equations being compared is the same across the two systems in the formal theorem. The only difference between the two systems is that some values of non-zero coefficient(s) in an equation change. When this feature is interpreted from the model perspective, it amounts to an assumption that the mechanisms are *invariant* to factor changes. Invariance is the property of a mechanism that ensures that a mechanism does not change given changes in the directly controllable factors or indirectly controllable factors brought about by experimenter/nature.

Invariance is a strong assumption and it is not difficult to imagine cases where it fails. To give a simple example of invariance failure, imagine an elastic band holding a pack of cards together as a mechanism (it constrains positions of the

cards relative to each other). Then imagine the act of the experimenter slipping in an extra card into the pack as a direct change. Now, there is a point at which the slipping in of an extra card is one too many, the elastic band snaps and no longer constrains the positions of the cards. The elastic band as a mechanism is not invariant to the direct change of introducing the final card. Since the elastic band is invariant to some direct changes of introducing cards, this example of invariance failure in a mechanism suggests that the assumption of the invariance of mechanisms should ideally be made relative to a space of possible factor changes to which it is invariant. [48, 49]

Kevin Hoover (1995, p.69) briefly discusses two hypothetical examples of invariance failure that have been highly influential in econometrics. The first is in Trygve Haavelmo's seminal paper (1944, pp.27-28) on econometric methodology.[50] It is an example of the observed relationship in a car between the depression of the accelerator and the speed at which the car travels. Haavelmo's point is that this relationship can be accurately determined, however it is not invariant to a variety of changes, say to the engine of the car or to the conditions of the track on which the car travels. Changes in any of these would radically change the relation between accelerator depression and speed. So the relationship is not invariant to changes in the engine or in the track.[51]

---

[48] Hausman and Woodward (1999, p.537) use a similar example of a spring to make the same point.

[49] Note that the invariance assumption is even stronger in systems with many mechanisms. This is because the invariance required is that no mechanism changes given changes in any factors in the system, including those that do not appear in the mechanism. Therefore, in complex systems being modelled by this type of formalism, one would ideally be clear about the set of a possible changes for all factors, to which each mechanism is invariant.

[50] Historically, it is not a surprise that Haavelmo's discussion fits well with Simon's since Haavelmo (1944) was influential on Simon.

[51] This is an interesting feature to this example. The accelerator-speed relationship is invariant (*all else being equal*) to changes to the accelerator. So, as set up here whether or not the relationship is invariant or not depends on what is included in the model. For instance, if one just includes the accelerator-speed relation and treats the accelerator position as directly controllable then the relationship is invariant. However, if one also models mechanisms that relate accelerator position to the petrol entering the engine and mechanisms describing the working of the engine and so on, then the relation is not invariant to factors brought in. So, the point here is that the invariance property of mechanisms described here is relative to the factors in the model. As Haavelmo's car example shows, this then implies that the usefulness of a relation will depend on whether the model is large enough to include factors that are relevant to situations to which the model is to be applied. For example, the fact that the accelerator-speed relation is invariant to changes in the accelerator is not particularly helpful if the car is to be used in a wide variety of track conditions and in cases engine may suffer slight changes dues to wear and tear etc..

The other discussion of invariance failure[52] that Hoover mentions is the famous 'Lucas Critique'. In his (1976) Robert Lucas argues that policy interventions based on econometric models that fail to incorporate agents' rational expectations will have results unforeseen by the model. In the context here, the problem is that such models assume 'mechanisms' that are in fact *not* invariant[53] to policy interventions. Lucas constructs a series of hypothetical examples in which this occurs. One of the examples Lucas gives is that of government which aims to stimulate investment by introducing a tax credit. Lucas uses it to show that a model which overlooks the rational expectations of investors about the duration of the tax credit significantly underestimates the impact of a tax credit on investment. For my purposes here, Lucas' example shows that a model that leaves out rational expectations wrongly assumes that observed relations between investor behaviour and government actions are invariant to the government act of introducing a tax credit.

The Haavelmo and Lucas examples show the importance of distinguishing the invariant from non-invariant relations in doing structural modelling, while the elastic band shows the importance of the limits of invariance in mechanisms. All three examples show that invariance is a crucial property of mechanism: it is required to ensure that interventions have results in line with observed functional relations.

5.2.4. Independence of Directly Controllable Factors

Recall that in Simon's theorem 6.1 only one equation is distinct between the two systems. This is possible because the coefficients can take values independent of the values of other coefficients, that is, because the coefficients are variation free. An interesting question is how this variation free assumption is to be interpreted from the model perspective. If one applies the isomorphism used in constructing the model reading earlier, the corresponding model interpretation of the 'independence' of directly controllable factors is that it is possible for the experimenter/nature to set the values of the directly controllable factors independently of the values of the other directly controllable factors.

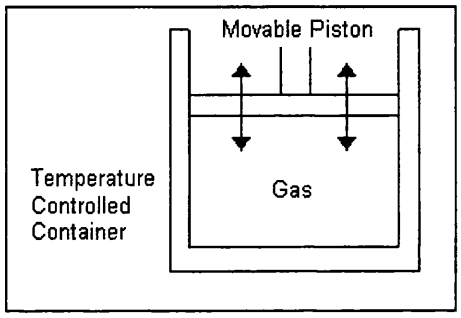
---

[52] Hoover describes it as relating to invariance. Lucas does not describe it in these terms.

[53] The scare quotes are here to highlight that since they are not invariant to factor changes, strictly speaking these are *not* mechanisms, since it is a property of mechanisms that they are invariant.

However, care must be taken not to misread the resulting independence property of the directly controllable factors. In particular, independence does not require that a directly controllable factor has no impact on any other directly controllable factors. To give a simple example from thermodynamics, imagine a container of gas with a movable piston where the container's temperature can be directly controlled, as shown in figure 2.1.

Figure 2.1 – Gas Container



Assume the piston's position can be directly controlled by placing weights onto it, so that it falls to the point at which the pressure exerted by the weights and that of the gas equalises.[54] In this example, one can directly control both the temperature of the gas and the position of the piston. Assuming the ideal gas law holds[55] then increasing the temperature of the gas, assuming constant weight on the piston, expands the gas so that the piston moves upwards. In other words, the directly controllable position of the piston changes as a result of a change in directly controllable temperature. Nevertheless, the values of temperature and position of the piston are variation free. Since regardless of the temperature at which the gas is set one can always adjust weights on the piston to get any desired position of the piston.[56]

The way in which the independence of controllable factors relates with intervention is discussed in more detail in the next chapter. The important point

[54] Simon and Rescher (1966, pp.331-332) present a similar example in a paper giving a concise, tidied up reading of Simon's causal order concepts. However, their example is not used to make the point made here.

[55] The ideal gas law is $PV=nRT$, where P denotes pressure, V denotes volume, T denotes temperature, n is the number of moles of gas and R is the universal gas constant, see Halliday and Resnick (1978, pp.497-509).

[56] One might quibble over whether the position of the piston is really directly controllable here, arguing that it is not really directly controllable because it depends on temperature. However, this would be to strengthen the concept of direct control. Here the aim is to show just how weak the concept of 'direct control' constrained only by the variation free requirement is.

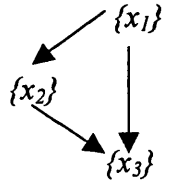
here is simply that an independence assumption holds among the directly controllable factors in the model reading, which requires that the values of directly controllable be variation free. Moreover, it does not imply that directly controllable factors are independent in the stronger sense that changing one directly controllable factor must have no impact on any other directly controllable factor.

5.2.5. Simon's 'In general' Caveat: The Possibility of Cancelling Out

I finish this section on the properties of causal order by considering just what is meant by the 'in general' caveat in the theorem 6.1. It can be illustrated by the following abstract example.

$$x_1 = \alpha_{10}$$
$$x_2 = \alpha_{21} x_1$$
$$x_3 = \alpha_{30} + \alpha_{31} x_1 + \alpha_{32} x_2$$

For these equations the formal order can be represented by the causal graph.



So according to the model version of theorem 6.1 changes in first mechanism, in α_{10}, should lead to changes in all of the factors denoted by x_1, x_2 and x_3. However, suppose that $\alpha_{31} = - \alpha_{32}\alpha_{21}$ happens to hold,[57] then the value of x_3 will not change since the third equation in this case reduces to

$$x_3 = \alpha_{30}.$$

Since this equation holds independent of changes to α_{10}, x_3 will not change value. Under the model reading, in this case the indirect influence of x_1-factor on x_3-factor via the x_2-factor and its direct influence on x_3-factor cancel each other out. In either formal or model readings, this would be a counterexample to the theorem if the 'in general' caveat were not inserted by Simon.

Therefore, the insertion of 'in general' is qualifying the theorem by implicitly bringing in a statement that for *most values of coefficients*, that is, those for which such 'cancelling out' features do not occur (in this case where $\alpha_{31} \neq -\alpha_{32}\ \alpha_{21}$) the

[57] It cannot hold systematically otherwise the coefficients would not be variation free.

conclusion of the theorem holds. This 'in general' caveat says the consequent only holds in certain cases (where no cancelling out occurs). It is similar to what Spirtes *et al.*(1993) call 'the faithfulness condition' and Pearl's calls 'stability' (2000, p.48),[58] which assume that such cancelling out values of coefficients do not occur for a system.[59]

One can question whether or not this restriction is likely to hold in systems modelled in practice. Some, for example Spirtes *et al.*(1993, p.95), argue that this kind of assumption is innocuous since the chance that coefficients will happen to fall on these particular values where changes 'cancel each other out' is remote. However, this view assumes that the inconvenient values for the coefficient are highly unlikely, which is debatable depending on the system. For instance, Kevin Hoover (2001a, pp.168-170) argues that in systems where agents exercise optimising behaviour, coefficients may be chosen by agents for their own purposes to have just such cancelling out values, and thus the likelihood that coefficients in the system happen to meet cancelling out conditions need not be unlikely. Indeed, as Hoover notes, it may even be likely in economic systems in which rational agents exercise optimal control.

6. Conceptual Equivalence Revisited and The Strong Reading

Having presented the key properties of causal relations, I now complete the chapter by returning to the original conceptual equivalence problem, in order to show the importance of specifying what is directly controlled and what are the mechanisms. The section finishes by presenting a formal way of representing the causal systems developed in the chapter.

[58] These conditions apply to indeterministic systems and require that any zero partial correlations be indicative of causal structure, and not arise due to different causal influences 'cancelling each other out' as in the example above.

[59] Though it is not identical to faithfulness since Simon does not assume these cancelling-out values of coefficients do not occur. Nevertheless the 'in general' caveat is similar to the faithfulness and stability assumptions since it limits the usefulness of his theorem 6.1. to cases where cancelling out does not occur.

6.1. Solving the Conceptual Equivalence Problem in the Earlier Example

This chapter began with two mathematically equivalent sets of equations. Using the strong reading of Simon, set out in this chapter, one can see clearly how the conceptual equivalence problem with which the chapter started is avoided. The two sets, now treated as two models are:

Model 1 ($p \rightarrow q$)	Model 2 ($q \rightarrow p$)
$p = \alpha$	$q = \alpha'$
$q = \beta p + \gamma$	$p = \beta' q + \gamma'$

Since the systems are mathematically equivalent the following must hold.

$$\alpha' = \alpha\beta + \gamma, \beta' = (1/\beta) \,\&\, \gamma' = -(\gamma/\beta) \qquad \ldots \qquad (*)$$

Consider a general change in the coefficients. For the first set this is a change from (α, β, γ) to $(\alpha+\Delta\alpha, \beta+\Delta\beta, \gamma+\Delta\gamma)$ while for the second it is a change from $(\alpha', \beta', \gamma')$ to $(\alpha'+\Delta\alpha', \beta'+\Delta\beta', \gamma'+\Delta\gamma')$. It can be shown from (*) that the following relations hold between the shifts in the coefficients in the two sets, given their mathematical equivalence.[60]

$$\Delta\alpha' = \alpha\Delta\beta + \beta\Delta\alpha + \Delta\gamma$$

$$\Delta\beta' = -\Delta\beta \Big/ \beta(\beta + \Delta\beta) \qquad \ldots \qquad (**)$$

$$\Delta\gamma' = \frac{\gamma\Delta\beta - \beta\Delta\gamma}{\beta(\beta + \Delta\beta)}$$

With this background, one can consider how shifts in the values of coefficients are read from the perspectives of model 1 and model 2. This is set out in table 2.5.

[60] For both sets of equations (*) and (**) the reverse equations (for getting model 1 coefficients from model 2) can be got by swapping primed variables for their non primed counterparts and *vice versa*.

Table 2.5. Two Models And Their Two Causal Orders

Intervention	*Model 1 ({p}→{q})*			*Model 2 ({q}→{p})*			*What happens*
	Formal change	Experimenter /Nature	Causal Story	Formal Change	Experimenter /Nature	Causal Story	
1.	$\Delta\alpha \neq 0$ $\Delta\beta = 0$ $\Delta\gamma = 0$	Only changes *p* mechanism	*p* is caused to change and *p* causes *q* to change	$\Delta\alpha' \neq 0$ $\Delta\beta' = 0$ $\Delta\gamma' = 0$	Only changes *q* mechanism	*q* is caused to change and *q* causes *p* to change	*p* and *q* change
2.	$\Delta\alpha = 0$ ($\Delta\beta \neq 0$ *or* $\Delta\gamma \neq 0$)	Only changes *q* mechanism	*q* is caused to change, *p* stays same	$\Delta\alpha' \neq 0$ $\Delta\beta' \neq 0$, $\Delta\gamma' \neq 0$	Changes both mechanisms	*q* is caused to change, *p* stays same: change in other mechanism 'cancels out' effect of *q* on *p*	*q* changes *p* doesn't
3.	$\Delta\alpha \neq 0$ $\Delta\beta \neq 0$, $\Delta\gamma \neq 0$	Changes both mechanisms	*p* is caused to change, *q* stays same: change in other mechanism 'cancels out' effect of *p* on *q*	$\Delta\alpha' = 0$ ($\Delta\beta' \neq 0$ *or* $\Delta\gamma' \neq 0$)	Only changes *p* mechanism	*p* is caused to change, *q* stays same	*p* changes *q* doesn't

The first row considers a change to α in system 1, that is, a shift from (α, β, γ) to $(\alpha+\Delta\alpha, \beta, \gamma)$. Using (**) this is equivalent to a shift from $(\alpha', \beta', \gamma')$ to $(\alpha'+\Delta\alpha', \beta', \gamma')$ for model 2. The respective model interpretations are distinct: in model 1 the intervention is a change in the mechanism that determines *p*, and since *p* causes *q* in model 1, this leads to a change in *q*. In model 2, on the other hand, the intervention is a change in the mechanism that determines *q*, and since *q* causes *p* in model 2, this leads in turn to a change in *p*. For both models *p* and *q* both change. The second and third rows are to be read in a similar way. In these rows, however, the coefficient changes are interpreted by one of the models as changes to two mechanisms; these cells are shaded in the table. These shaded cells are

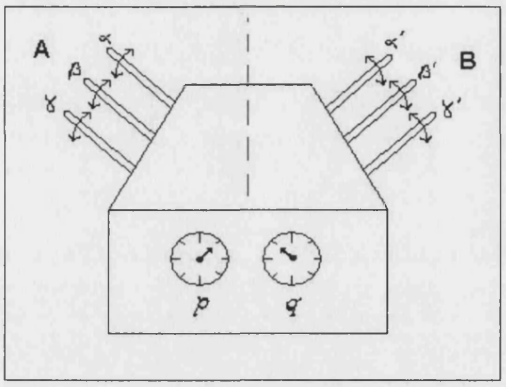
interesting because in these only one indirectly controllable factor changes though both mechanisms are directly changed. As such, they are cases of cancelling out which the 'in general' caveat in Simon's theorem 6.1 served to exclude. The shift in these shaded cells is interpreted as a situation in which the changing of the second mechanism offsets the change in the first, so that a cause changes though its effect does not. In these cases the alternative model interpretation is not one of cancelling out, in the alternative model interpretation just one mechanism is changed and only one indirectly controllable factor changes. So in all rows, the model interpretations are distinct.

This table also shows how to avoid the conceptual equivalence problem noted at the beginning of the chapter. By stipulating what the coefficient and variables are in the equations and the form of the equations, Simon's formal order follows. This stipulation of coefficients, variables and equation form allows a causal story to be given for the set of equations. As the table shows, it can be generated for changes in the values of variables in two mathematically equivalent systems. There is a clear distinct model reading of each set of equations given the distinct stipulation of coefficients, variables and equation form. Thus the conceptual equivalence problem is avoided.

6.2. The Importance of Stipulating the Coefficients and the Form of the Equations

As is clear from the table above, what set of coefficients are chosen in a set of equations plays a key role in solving the conceptual equivalence problem. This can be seen by considering an analogy to the example of the two models above. Imagine a box which has six levers, three levers each on two opposite sides, and two dials on a third side perpendicular to the sides with the levers. Suppose that the levers on one side are labelled α, β and γ, levers on the other side α', β' and γ', and that the two dials are labelled p and q respectively. Suppose also that the angle of a lever corresponds to the value of the coefficient it matches and the dials given the values for p and q. Suppose also that the levers' positions always satisfy (*). Finally, suppose that an experimenter can pull levers on one side at a time and that levers on the side she is moving only change if she pulls them. However, if the experimenter moves the levers on one side, the levers on the other side move to ensure (*) is maintained. This 'lever box' is illustrated in figure 2.2.

Figure 2.2 The 'Lever Box'



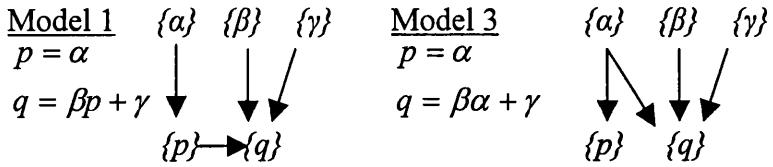
Here a choice of coefficient set is analogous to the experimenter moving levers on a particular side of the box. If the experimenter pulls levers on side A then model 1's set of coefficients is appropriate; if the experimenter pulls levers on side B then model 2's set of coefficients is appropriate. The inconsistency of the two sets, which rules out the conceptual equivalence here, is analogous to the impossibility that the experimenter be on both sides of the box at the same time.[61]

The point here is that which model is appropriate depends on which levers are those that are directly controlled by the experimenter. So, if the α-lever, β-lever and γ-lever are those directly controlled, the experimenter is on side A and model 1 applies. Conversely for the other side. Therefore, the stipulation that certain factors are those that are directly controlled, where I assume that this means that each of these factors can be varied independently of the others like levers, plays an important role in distinguishing the two models. Without it, one cannot be sure which side of the box the experimenter is 'on' and thus which model holds, in that case a conceptual equivalence problem remains.

This shows that it is necessary to stipulate what the coefficients are in order to establish distinct model interpretations for mathematically equivalent sets of equations. However, the stipulation of the set of coefficients to be read as directly controllable is not in itself sufficient to rule out all conceptual equivalence problems. The insufficiency of coefficient stipulation for solving conceptual equivalence can be seen from a simple example. Consider the first of the earlier

[61] Formally, if one set of coefficients $\{\alpha, \beta, \gamma\}$ is stipulated, then these are variation free. This then implies that none of the other coefficients α', β' or γ' can also be variation free in relation to $\{\alpha, \beta, \gamma\}$ given the assumed mathematical relations between the two sets of coefficients.

two systems and another equivalent system of equations, and both of their extended formal orders.[62]



In this case the two sets of equations are mathematically equivalent *and* have the same coefficients and the same variables. Therefore they both have the same directly and indirectly controllable factors in their respective model readings. And *yet* their causal orders are distinct. The model on the right, instead of reading price as a direct cause of quantity like the first model, takes price and quantity to have the α-factor as a common cause.

Clearly, the reason these two systems differ is that the second equation in the system on the right contains α rather than p. This leads to a distinct causal order in the model interpretation because this equation represents a correspondingly different mechanism from that assumed in the model reading of the first set of equations. Therefore, the difference in model readings is due to a difference in equations. It follows that merely stipulating the set of coefficients and variables in a set of equations is not sufficient to remove all conceptual equivalence problems between the two systems of equations. One must also stipulate the form of the equations because different equation forms imply different mechanisms in the model readings. For instance, the second equation in model 1 denotes a mechanism that relates two indirectly controllable factors and two directly controllable factors. Whereas in model 3, the second equation denotes a mechanism that relates three directly controllable factors and one indirectly controllable factor. The two equations respectively denote two different mechanisms, and this leads to different causal relations in their model interpretations.

[62] It could be objected that model 3 below is not acceptable since its second equation isn't linear in the coefficients. Though this is correct, one could easily reformulate it so that it is, for example, by introducing a new variable x and replacing the system above by $x=\alpha$, $p=x$ and $q=\beta x + \gamma$. Alternatively, one could keep the above set up but drop β from the equations in both systems. However, I stick to the example above because it is intuitively clearer. Also, given that Simon's methods can be extended to such non-linear cases it is not a serious problem.

Finally, it is important to note that here I diverge from Kevin Hoover's (2001a) reading of sets of equations, which is also developed from Simon's analysis. As is discussed in chapter four, Hoover's view takes the attribution of direct control to coefficients (or 'parameters' in his terminology) as *sufficient* for determining the causal order of a set of equations.[63] The problem with this is, as Nancy Cartwright (2002) points out, is that Hoover's position cannot then causally distinguish between systems like model 1 and model 3 above. In contrast, the approach adopted here avoids this problem.

6.3. Making the Strong Reading Explicit in Sets of Equations

Ultimately the goal of this chapter has been to set out an explicit causal interpretation of sets of equations that are simpler versions of those used in econometrics. To complete the process I now propose a modification to the mathematical formalism by which causal orders in the model are to be represented. This is necessary because if one simply writes down a set of equations then one is restricted to mathematical symbols that have conventional meanings in the representation. As discussed repeatedly in this chapter, these conventional meanings allow the equations to be transformed into alternative equivalent forms that have different causal orders. So, to explicitly rule this out I propose a change in the mathematical syntax.

The basic idea is to replace '=' with '$=_M$' in the equations to indicate it does not merely indicate an equality but that it denotes a mechanism.[64] Also, I assume that in such an 'M-equation' Greek letters are coefficients and Latin letters are variables. Now '$=_M$' cannot have exactly the same properties as '='. In particular, it is only preserved under transformations that do not change the variables that *appear* in the corresponding equation. This is crucial because Simon's method relies entirely on the form of the equations, that is, which variables appear in which equations. If these change then the formal order changes, so does the model interpretation developed here.

[63] Given the reduced form of the equations, see chapter four.

[64] This proposed symbol '$=_M$' plays a similar role as '$c=$' does in Cartwright's (2003a) analysis and as the causal graphs assumed in Spirtes, Glymour and Scheines (1993) and in Pearl (2000). It provides an extra formal symbol by which causal content can be explicitly represented by the formalism.

Appendix 2.1 presents formal analysis which attempts to precisely define '$=_M$'. However, the detail is cumbersome and rather opaque. So instead of discussing it here, consider the two original equations with which the chapter started, with the coefficients and variables moved to the left hand side of equations.

$$p - \alpha = 0$$
$$q - \beta p - \gamma = 0$$

One can define two functions as equal to the respective left hand sides of these equations: $f_1(\alpha, \beta, \gamma, p, q) = p - \alpha$ and $f_2(\alpha, \beta, \gamma, p, q) = q - \beta p - \gamma$. Then the two equations above can be written as

$$f_1(\alpha, \beta, \gamma, p, q) = 0$$
$$f_2(\alpha, \beta, \gamma, p, q) = 0$$

Now these equations correspond to mechanisms in the model, the coefficients correspond to directly controllable factors, and variables to indirectly controllable factors. To make this explicit, replace the '=' by '$=_M$' to get.

$$f_1(\alpha, \beta, \gamma, p, q) =_M 0$$
$$f_2(\alpha, \beta, \gamma, p, q) =_M 0$$

This addition of '$_M$' to the equals sign signifies that the equation denotes a mechanism. Each of these *M*-equations also is assumed to imply the corresponding equation, so for example, the first implies $f_1(\alpha, \beta, \gamma, p, q) = 0$.

In appendix 2.1, it is shown that one can multiply these *M*-equations by non-zero constants[65] and reorder them to obtain new *M*-equations that represent the same mechanisms so that their model reading is unchanged. Finally, to increase the flexibility of the notation, it is also assumed terms can be moved across between the left and right hand sides of the '$=_M$' just as one does for '='. However, as it is defined in the appendix 2.1, a linear combination of *M*-equations does not yield an *M*-equation because this changes the corresponding model interpretation.

With this background, the two original equations of the demand model that are to be read using the model reading, can be represented using two *M*-equations:

$$p =_M \alpha$$
$$q =_M \beta p + \gamma$$

[65] A constant is fixed relative to coefficients and variables.

These *M*-equations can have terms moved across between left and right hand side and they can be reordered and rescaled. So for instance, the above is equivalent to:

$$q =_M \beta p + \gamma$$
$$2p - 2\alpha =_M 0$$

However, one cannot linearly combine the equations, so the above is *not* equivalent to

$$p + q =_M \alpha + \beta p + \gamma$$
$$q =_M \beta p + \gamma$$

To see the motivation for introducing '$=_M$' note that if one applies Simon's formal order to the sets of equations that correspond to the equivalent sets of *M*-equations above then one gets the same formal order. Whereas one gets a different order for the two *M*-equations that are not equivalent, that is, applying Simon's formal method to either

$$\begin{matrix} p = \alpha \\ q = \beta p + \gamma \end{matrix} \quad \text{or} \quad \begin{matrix} q = \beta p + \gamma \\ 2p - 2\alpha = 0 \end{matrix}$$

gives the same formal order i.e. $\{p\} \rightarrow \{q\}$. While applying it to

$$p + q = \alpha + \beta p + \gamma p$$
$$q = \beta p + \gamma$$

yields formal order $\{p, q\}$ which is distinct from that of the previous two sets.[66] This shows that the reason to introduce '$=_M$' is to modify the equality relation so that mathematical transformations can only be applied to it that respect the formal order and model reading. It formalises the solution to the conceptual equivalence problem presented earlier.

Finally, the introduction of this '$=_M$' makes explicit that the strong reading of Simon's analysis is intended. The strong reading is one which attributes a model interpretation for a set of equations as set out in detail in this chapter. It assumes that the each equation as written denotes a mechanism, in which each coefficient

[66] Strictly speaking they could have the same formal order over the variables. However, formal order is originally defined over equations and such transformations radically change the equations and thus the formal order over these. This does not occur in rescalings however, because I do not take two equations that differ by a scaling constant as different for the purposes of the model reading. This is why I implicitly treated $p = \alpha$ and $2p\text{-}2\alpha = 0$ as identical in the previous example.

denotes a directly controllable factor and so on. It reads Simon's formal order as the causal order in the model, assuming that the mechanisms and causal order relations meet the invariance, independence properties discussed above. In simplest terms, the strong reading takes a set of equations to which Simon's formal order is to be used, such as:

$$p = \alpha$$
$$q = \beta p + \gamma$$

And assumes that these should be modified to read

$$p =_M \alpha$$
$$q =_M \beta p + \gamma$$

so that the equations explicitly are to be interpreted in the way set out in this chapter. This completes the solution to the conceptual equivalence problem by introducing a formal modification that augments the mathematical syntax and semantics in a way that corresponds with the way equations are to be causally interpreted. It makes formally explicit the causal semantics that are to be attributed to the equations.

7. Conclusion

This chapter began with a simple problem. This problem was that equations in a mathematical model of a causal system could be mathematically manipulated in ways that changed the causal meaning of the model. As a beginning of a response to this problem, it presented Herbert Simon's method for deriving formal orders from sets of equations. It then developed Simon's work to make explicit the connections between Simon's formal orders and corresponding model concepts so that mathematical equations can be causally interpreted in an explicit way. This development of Simon's method for causally interpreting equations is called the 'strong reading'. As with any formalisation of concepts, it committed the resulting concepts of causes and causal order to having certain features. Among these features, the chapter discussed how factor changes are related to causal orders, the invariance of mechanisms to factor changes, the independence of directly controllable factors and the possibility that causal impacts cancel out. These properties highlight important constraints that actual systems need to satisfy if they are to be accurately represented using causal models of the sort analysed

here. Lastly, the chapter finished with a discussion of just how the original conceptual equivalence problem was solved and on the importance of stipulating the coefficients and the form of the equations in this solution.

In short, the chapter has presented an explicit interpretation of Simon's formal order and the deterministic sets of equations for which it is defined. In contrast to Simon's approach the method here has not assumed identifiability in the definition of causal order. Instead, the strong reading assumes that it is in virtue of equations denoting mechanisms, that will be muddled if equations are manipulated (in any way but rescaling and reordering) that a unique causal order is attributed to a set of equations. Chapter five picks up the discussion later when it analyses the role identifiability in Simon's approach and how identifiability of a set of equations can be causally interpreted using the strong reading proposed here.

Finally, as is clear from the limited nature of the sets of equations treated, there remains work to extend the interpretation to sets of equations like those actually used in econometrics. This chapter is limited to deterministic, simultaneous equation models. As such it is open to the criticism that it is not immediately relevant to probabilistic models of econometrics. This is why in the next chapter I attempt to extend the strong reading of equations to cover the simplest type of models actually used by econometricians: linear models with errors-in-the-equations.

Appendix 2.1 –Beginnings of a Formalisation of Mechanisms and Causal Order

To develop a formal treatment of $=_M$, first define:

> A *Linear M-System (LMS)* relative to the set of coefficients $C= \{\alpha_1,\alpha_2,...,\alpha_m\}$ and the set of variables $V = \{x_1,x_2,...,x_n\}$ is the ordered triplet $(C, V, E(F))$ where $E(F)$ is a set of equations, $f=0$ for all $f \epsilon F$. F is a set of linearly independent functions f: $P_C x P_V \rightarrow \boldsymbol{R}$ that are linear in variables, V, and linear in the coefficients C, such that the set $E(F)$ of equations is solvable for all of the variables in V in terms of the coefficients in C. For α_i in C, P_{ai} is the set of possible values (the domain) for α_i and similarly for x_i in V, P_{xi} is the set of possible values of x_i in V. To simplify the notation, let $P_C = P_{a1}xP_{a2}x...xP_{am}$ and $P_V = P_{x1}x...xP_{xn}$. The set of possible values for any variable or coefficient is a subset of $\boldsymbol{R}$.

And also define:

> Two linear M-systems relative to (C,V), $(C, V, E(F_1))$ and $(C, V, E(F_2))$, are *M-equivalent* if and only if there exists a bijective operator G: $F_1 \rightarrow F_2$, such that for each function f in F_1 there is a non-zero constant[67] c such that $G(f) = cf$.

To justify it being termed 'equivalent', I prove that it is an equivalence relation.

> *Theorem 2.1*: M-equivalence is an equivalence relation.

> *Proof*: Given $(C_1, V_1, E(F_1))$ an *LMS* relative to (C, V). Define an operator $G(f) = f$ for all f in F_1, since G is the identity operator it is a bijection. Therefore, G meets the conditions of the definition, so $(C_1, V_1, E(F_1))$ is M-equivalent to itself i.e. M-equivalence is reflexive.

[67] A constant is fixed relative to changes in coefficients and variables.

Given $(C_1, V_1, E(F_1))$ M-equivalent to $(C_2, V_2, E(F_2))$ then there is a G such that for each f in F there is a non zero c such that $G(f) = cf$. For any function h in F_2, since G is a bijection, there is a unique f in F_1 such that $h = G(f) = cf$, for some non-zero c. For each such h, define $G_2(h) = (1/c)h = f$. This defines a bijective operator G_2: $F_2 \rightarrow F_1$ which meets the conditions in the definition of M-equivalence. So $(C_2, V_2, E(F_2))$ is m-equivalent to $(C_1, V_1, E(F_1))$ and M-equivalence is a symmetric.

Given $((C_1, V_1, E(F_1))$ M-equivalent to $(C_2, V_2, E(F_2))$ which is itself M-equivalent to $(C_3, V_3, E(F_3))$ then there is a G_1: $F_1 \rightarrow F_2$ that multiplies each function in F_1 by a non-zero constant to get a function in F_2. Likewise there is a G_2:$F_2 \rightarrow F_3$ does the same for each function in F_2 mapping it onto a function in F_3. Then the composition $G_3 = G_2 \, o \, G_1$ multiples each function in F_1 by a non-zero constant to get a function in F_3. Moreover, since G_1 and G_2 are bijective so is G_3, the composition of the two. Therefore $(C_1, V_1,E(F_1))$ is M-equivalent to $(C_3, V_3,E(F_3))$ and M-equivalence is transitive.

Since M-equivalence is reflexive, symmetric and transitive it is an equivalence relation.□

To simplify the notation in an intuitive way now introduce '$=_M$'.

Definition of '$=_M$'. Given $(C, V,E(F))$ a LMS relative to (C, V) where $C = \{\alpha_1, \ldots, \alpha_m\}$ and $V= \{x_1, \ldots, x_n\}$. Denote the LMS by listing the equations for each function f in F where = is replaced by $=_M$:

$$f_1(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) =_M 0$$
$$f_2(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) =_M 0$$
$$\vdots$$
$$f_n(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) =_M 0$$

Assume (to extend the usefulness of this notation) that for f_1 which satisfies $f_1 = h_{11} - h_{12}$ for h_{11} and h_{12} linear in the coefficients and

variables, where these are functions only of coefficients and variables that appear in f_1, that

$$h_{11} =_M h_{12} <=> f_1 =_M 0$$

Given this, one can move terms across the $=_M$ sign in the same way as one does for =.

It is worth noting several important properties that are assumed in the definition:

(i) *A set of M-equations implies the corresponding set of equations* i.e.

If $f_1(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) =_M 0$

$$\vdots$$

$f_n(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) =_M 0$

Then

$f_1(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) = 0$

$$\vdots$$

$f_n(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) = 0$

(ii) *One can reorder and rescale M-equations*:

$f_1(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) =_M 0$

$$\vdots$$

$f_n(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) =_M 0$

Then for any non-zero constants $\{c_1, c_2, \ldots, c_n\}$ and bijection $k:\{1, \ldots, n\} \rightarrow \{1, \ldots, n\}$ which reorders indices,

$c_1 f_{k(1)}(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) =_M 0$

$$\vdots$$

$c_n f_{k(n)}(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) =_M 0$

As discussed in the main body of the chapter, the symbol '$=_M$' is introduced to limit mathematical transformations on sets of equations (to which Simon's formal order is applied) to those that preserve the formal order. To show that it accomplishes this, I need to prove the following.

Theorem 2.2 Two LMS relative to *(C, V)* have the same formal order if and only if they are M-equivalent.

Doing this first requires defining Simon's formal order for an LMS. Loosely this can be done by

> A *formal order* for the LMS *(C, V, E(F))* is the relation over the partition of *E(F)* that results from applying Simon's method for generating a formal order to the linear equations corresponding to the M-equations i.e.
>
> $f_1(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) = 0$
>
> $\vdots$
>
> $f_n(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) = 0$

Since theorem 2.2 makes reference to 'same formal order', it is important to be clear about the identity conditions for formal orders over equations. However, since the formal order is defined over the set of equations, to make sense of when the formal order of two sets of equations is identical, it is first necessary to be clear as to what is required for two sets of equations to be identical.

Aside: Relevant Identity Conditions

In line with the standard identity condition for sets, first assume that two sets of equations are identical if and only if they contain identical equations.

But what is meant by 'identical equations'? Again following convention, assume that two equations, $f = 0$, $g = 0$ are identical only if $f=0$ and $g=0$ are mathematically equivalent, that is both have the identical solution sets:

$$\{(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) \in P_C x P_V \mid f(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) = 0\} =$$
$$\{(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) \in P_C x P_V \mid g(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) = 0\}$$

In addition assume that for $f=0$ and $g=0$ to be identical it is necessary that f and g are functions of the same set of coefficients and variables[68]. By the set of coefficients and variables of which f is a function I mean the subsets of $C \cup V$, on which the value of f depends. For example, if $C=\{\alpha_1,\alpha_2,\alpha_3\}$ and $V = \{x_1,x_2\}$ and $f(\alpha_1, \alpha_2, \alpha_3, x_1, x_2) = \alpha_2 x_1 + \alpha_3$, then the set of coefficients and variables of which f

[68] Here I take it that any coefficient or variable that does not appear in the equation is not a part of that set.

is a function is $\{\alpha_2, \alpha_3, x_1\}$. So, although f can be expressed as a function of all the coefficients and variables, here I mean by the 'set of coefficients and variables of which f is a function' just those that actually figure in the functional expression for f. This condition which will be justified in the discussion of mechanisms that follows the proof of the theorem below.

Summarising the identity conditions for equations:

> Two equations that are linear in the coefficients and variables, $f=0$ and $g=0$ are identical if and only if
>
> (i) $f=0$ holds if and only if $g=0$ i.e. they have the same solution set.
>
> (ii) The set of variables and coefficients of which f and g are functions are the same.

Finally, assume two formal orders over two sets of equations are identical if and only if the two sets of equations are identical and when Simon's formal ordering method is applied the two sets of equations, the same causal precedence, exogeneity and endogeneity relations hold over equations and variables.
End of Aside.

With this background in place I can set out a proof of theorem 2.2.

> *Theorem 2.2* Two *LMS* relative to *(C, V)* have the same formal order if and only if they are M-equivalent.

Most of the proof follows from the following lemma

> *Lemma*: For two *LMS* relative to *(C,V)*, *(C,V, E(F))* and *(C,V, E(G))*, the two sets of equations, *E(F)* and *E(G)*, are identical if and only if the LMS are M-equivalent.

Proof of Lemma:

(i) '*Only If*': *E(F)* and *E(G)* are identical => the LMS are M-equivalent.

For $f=0$ an equation in *E(F)* assume that $g=0$ is the corresponding identical equation in *E(G)*. Since these equations are identical they are mathematically equivalent, that is

$$\{(\alpha_1,\ldots,\alpha_m,x_1,\ldots,x_n) \in P_C x P_V \mid f(\alpha_1,\ldots,\alpha_m,x_1,\ldots,x_n)=0\} =$$
$$\{(\alpha_1,\ldots,\alpha_m,x_1,\ldots,x_n) \in P_C x P_V \mid g(\alpha_1,\ldots,\alpha_m,x_1,\ldots,x_n)=0\}$$

Now define

$$f_\alpha(x) = f(\alpha,x) \text{ and } g_\alpha(x) = g(\alpha,x)$$

$f_\alpha(x)$ is a linear function on the variables defined by $f(\alpha,v)$ where the vector of coefficients, α, takes a particular value. Similarly, $g_\alpha(x)$ is defined in the same way but for g.

By the definition of $f_\alpha(x)$ and $g_\alpha(x)$ and the mathematical equivalence of f and g, it follows that $f_\alpha(x)$ and $g_\alpha(x)$ are mathematically equivalent i.e.

$$\{(x_1,\ldots,x_n) \in P_V \mid f_\alpha(x_1,\ldots,x_n)=0\} = \{(x_1,\ldots,x_n) \in P_V \mid g_\alpha(x_1,\ldots,x_n)=0\}$$

Since $f_\alpha(x)$ and $g_\alpha(x)$ are linear functions in the variables, they can be expressed as

$$f_\alpha(x) = \sum_{i=1}^{n} b_i x_i$$
$$g_\alpha(x) = \sum_{i=1}^{n} c_i x_i$$

where the b's and c's are constants.

So the equivalence between $f_\alpha(x)$ and $g_\alpha(x)$ can be expressed as:

$$\{(x_1,\ldots,x_n) \mid \sum_{i=1}^{n} b_i x_i = 0\} = \{(x_1,\ldots,x_n) \mid \sum_{i=1}^{n} c_i x_i = 0\} \ldots \quad (1)$$

Note that at least one constant, b_i, in the expression for f_α must be non-zero, otherwise $f_\alpha(x) = 0$ and thus $f(\alpha, x) = 0$ at some non-zero coefficient value α. However, this is not possible given that coefficients must be non-zero and $f(\alpha, x)$ is linear in the coefficients and variables,[69] since for functions of this form one

[69] These are conditions specified by Simon, see the main body of the chapter.

can always find some non-zero x, regardless of the (non-zero) value of α, such that $f(\alpha, x) \neq 0$. So,

At least one b_i in the expression for $f_\alpha(x)$ is non-zero ... (2)

Now for $(z_1,z_2,...,z_n)$ any element in $\{(x_1,...,x_n) | \sum_{i=1}^{n} b_i x_i = 0\}$, so from (1) it follows that

$$\sum_{i=1}^{n} c_i z_i = 0 \quad \ldots \quad (3)$$

By (2) one can assume without loss of generality that b_n is a non-zero. Given $\sum_{i=1}^{n} b_i z_i = 0$ then

$$z_n = -(\frac{1}{b_n})\sum_{i=1}^{n-1} b_i z_i \quad \ldots \quad (4)$$

Substituting (4) into (3) yields.

$$\sum_{i=1}^{n-1} c_i z_i - (\frac{c_n}{b_n})\sum_{i=1}^{n-1} b_i z_i = 0$$

$$\sum_{i=1}^{n-1}\left[c_i - (\frac{c_n b_i}{b_n}) \right] z_i = 0 \quad \ldots \quad (5)$$

Now, consider any $(z_1,z_2,...,z_{n-1})$ in $\boldsymbol{R}^{n-1}$, letting

$$z_n = -(\frac{1}{b_n})\sum_{i=1}^{n-1} b_i z_i$$

implies that $(z_1,z_2,...,z_n)$ satisfies $\sum_{i=1}^{n} b_i z_i = 0$, so (5) holds of this $(z_1,z_2,...,z_{n-1})$.

But this implies for *all* $(z_1,z_2,...,z_{n-1})$ in $\boldsymbol{R}^{n-1}$

$$\sum_{i=1}^{n-1}\left[c_i - (\frac{c_n b_i}{b_n}) \right] z_i = 0$$

which implies that

$$c_i - (\frac{c_n b_i}{b_n}) = 0 \quad \forall i$$

that is

$$c_i = \left(\frac{c_n}{b_n}\right) b_i \quad \forall i \qquad \ldots \qquad (6)$$

But this implies

$$f_\alpha(x) = \left(\frac{c_n}{b_n}\right) g_\alpha(x)$$

Or letting $\lambda_\alpha = \left(\frac{c_n}{b_n}\right)$, that $f_\alpha(x) = \lambda_\alpha\, g_\alpha(x)$

Since c_n and b_n are non-zero[70] for $f_\alpha(x)$ and $g_\alpha(x)$, λ_α is a non-zero constant. It follows that g_α is a rescaling of f_α. So,

$$f(\alpha,x) = \lambda(\alpha) g(\alpha,x) \qquad \ldots (7)$$

Note that $\lambda(\alpha)$ may not necessarily be constant, it is only has been shown so far that it is a non-zero constant for fixed α i.e. for f and g when the coefficients are at a particular value.

However, (7) can be strengthened if one considers the second necessary condition for $f=0$ and $g=0$ to be identical. It requires that $f(\alpha,x)$ and $g(\alpha,x)$ be functions of the same set of coefficients and variables. This implies that $\lambda(\alpha)$ can only be a function of coefficients that appear in $g(\alpha,x)$ otherwise f would be a function of more coefficients than g.[71] Moreover, since $f(\alpha,x)$ and $g(\alpha,x)$ are linear in the coefficients, if $\lambda(\alpha)$ is a function of some coefficients (i.e. not a constant) then either $\lambda(\alpha)g(\alpha,x)$ is not suitably linear (since it contains products of coefficients) or, if these products of coefficients all 'cancel out then $\lambda(\alpha)g(\alpha,x)$ is a function of fewer coefficients than $g(\alpha,x)$ which is ruled out by the second identity condition that must hold between f and g. So it follows that $\lambda(\alpha)$ must be a constant, that is

$$f(\alpha,x) = \lambda g(\alpha,x) \qquad \ldots \qquad (8)$$

---

[70] c_n is non-zero because otherwise $g_\alpha(x) = 0$ from (6) which is false.

[71] Recall that the coefficients are variation free, so it is not possible that coefficients in $\lambda(\alpha)$ that do not appear in $g(\alpha,x)$cancel out with coefficients in $g(\alpha,x)$.

So for each f in $E(F)$ there is unique g in $E(G)$ such that (8) holds for some non-zero constant λ. Being identical, it follows that $E(F)$ and $E(G)$ also contain the same number of equations, so there exists a bijection between $E(F)$ and $E(G)$ such that each element in $E(G)$ is an equation in $E(F)$ multiplied by some non-zero constant. In other words, the two LMS are M-equivalent. □

(ii) '*If*': the LMS are M-equivalent => $E(F)$ and $E(G)$ are identical.

Since the two LMS are M-equivalent, for each $f=0$ in $E(F)$ there is a unique $g = 0$ in $E(G)$ such that

$f = \lambda g$ for some non-zero constant λ

It follows trivially that f and g are functions of the same coefficients and variables and that $f=0$ and $g=0$ have the same solution set. Therefore for each $f=0$ in $E(F)$ there is a unique $g=0$ in $E(G)$ which is identical to it. Moreover, since the LMS are M-equivalent there is bijection between $E(F)$ and $E(G)$ so they have the same number of equations. Therefore, $E(F)$ is identical to $E(G)$. □

With the lemma, the proof of the theorem is straightforward.

Proof of Theorem:

(i) *'If'* (M-equivalent LMS => same formal order)

Assume the M-equivalent LMS are $(C,V,E(F))$ and $(C,V,E(G))$ respectively. From the lemma it follows that $E(F)$ and $E(G)$ are identical.

By M-equivalence, the second LMS can be represented as a rescaling and reordering of the first LMS. This does not influence the solution properties of the equations. In particular, it does not change which equations and which variables are solved for in which order when using Simon's formal method for determining formal order. This implies that applying Simon's formal ordering method will yield the same results for a rescaled and reordered set of equations. Therefore, $E(F)$ and $E(G)$ are identical and the relations among the equations and variables

got by applying Simon's formal order will be the same. Therefore, the two formal orders of the two LMS are identical. □

[*Aside*: This is similar to analysis by Simon (1953, pp.29-30) where he shows that multiplying equations by non-zero constants (i.e. rescaling) does not change their formal order.[72]]

(ii) '*Only if*': (Two LMS have the same formal order => The LMS are M-equivalent)

Since the two LMS have the same formal order, *E(F)* and *E(G)* are identical. Therefore we have two LMS *(C,V,E(F))* and *(C,V,E(G))* for which *E(F)* and *E(G)* are identical. By the lemma they are M-equivalent. □

Comment on Theorem 2.2, Identity Conditions for Equations and Mechanisms

Theorem 2.2 formalises the solution to the conceptual equivalence problem presented in the main body of the chapter since it shows that the new symbol '$=_M$', as defined here, limits mathematical manipulations of the set of equations to those that preserve Simon's formal order over equations. The fact that it is the formal order *over equations* and not variables which is preserved is also important. Especially since the identical formal order *over variables* for two LMS does *not* imply M-equivalence (this is essentially noted by Simon in his footnote (1953, [11], p.30) when he observes that some linear combinations of equations may preserve the formal order over the variables). This highlights an important difference between formal orders over variables and over equations which is implicit in Simon's treatment.[73] In the analysis here, it is the fact that identical

[72] Note, however, that in contrast to the strong reading Simon claims reordering equations does *not* preserve the formal order because, as he interprets it, this muddles the interpretation of the equations. While the strong reading is 'reordering blind' i.e. it does not matter what order the equations are written in since a reordering of the equations is assumed to be accompanied by a suitably reordered interpretation for those equations. In any event, the difference between Simon and the position here is *not* significant since for *both* positions the appropriate interpretation of an equation is preserved, either because the interpretation is also reordered (the strong reading) or because the order of the equations is fixed (Simon).

[73] It can be shown the formal order over equations is the more fundamental concept, for instance, from the fact that Simon uses which variables appear *in which equations* to define relations among variables such as direct causal precedence. However, since it is not directly relevant to the analysis of the thesis, I leave discussion of this difference as further work.

formal order over equations requires identical equations that allows one to derive M-equivalence.

Given this, a natural question arises as to what the formalism above has to do with the causal interpretation. To understand this, the identity conditions for equations need to be considered in light of the fact that these equations are taken to denote mechanisms in the strong reading. Specifically, the identity conditions for equations require two things: (i) that the solution set for the equations be the same and (ii) that the functions in the equations (e.g. f in $f=0$ and g in $g=0$) must be over the same variables and coefficients. Both of these conditions have natural interpretations when the equation is read as a mechanism.

To see this, consider the first identity condition which requires that the equations $f=0$ and $g=0$ be identical as constraints on the set of possible values of the variables and coefficients. In the model reading this implies that two mechanisms are identical only if they constrain the possible values of directly and indirectly controllable factors in the same way. This is an intuitive necessary condition for mechanisms to be identical since if one mechanism allowed a value of a factor which was not allowed in another mechanism, one would naturally consider the mechanisms to be distinct.

A similar remark applies to the second identity condition which requires that the respective functions in the identical equations be defined over the same variables and coefficients. In the model reading, this requires the identical mechanisms to relate the same indirectly and directly controllable factors. This too is a natural necessary identity condition for mechanisms since mechanisms that constrain different factors are intuitively distinct.

So, the identity conditions assumed here for equations are motivated by intuitive considerations as to what is necessary for two mechanisms to be identical. Assuming the causal order over mechanisms of two systems of mechanisms can only be identical provided the mechanisms are identical, then provides the basis for the identity conditions on the *formal order over equations*.

To complete this appendix, I now attempt to clarify the concept of a mechanism formally.

> Let D be the set of directly controllable factors and I the set of indirectly controllable factors. For d in D, let $P_d \subseteq \boldsymbol{R}$ be the set of possible values for the factor d, and similarly define P_i for a factor i in I.
>
> For $D=\{d_1,...,d_m\}$ and $I = \{i_1,...,i_n\}$ define a mechanism, m^*, as the ordered triplet $(S_D, S_I, \varphi_{m^*}=0)$ where S_D is a non-empty subset of D, and S_I a non-empty subset of I, and φ_{m^*} is a linear function $\varphi_{m^*}: P_{SD}xP_{SI} \rightarrow \boldsymbol{R}$ (where P_{SD} is the Cartesian product of the individual sets of possible values for d in S_D, and similarly, P_{SI} is the Cartesian product of the sets of possible values for i in S_I). For a mechanism, m^*, let μ_{m^*} a function defined as follows, $\mu_{m^*}: P_DxP_I \rightarrow \boldsymbol{R}$ (where $P_D = P_{d1}x...xP_{dm}$ and $P_I = P_{i1}x...xP_{in}$) such that $\mu_{m^*} = \varphi_{m^*}$ (this is introduced for convenience, and is just the function φ_{m^*} but whose domain is expanded to include all other the sets of possible values for other factors i.e. those factors *not* in the mechanism, that is, not in S_D not S_I). Let M denote a set of mechanisms.

To clarify this, a mechanism is defined using three components (i) S_D, the directly controllable factors that are in the mechanism (ii) S_I, the set of indirectly controllable factors in the mechanism and (iii) $\varphi_{m^*}=0$ the constraint on the values of the factors in S_D and S_I that can co-occur. So a mechanism relates directly and indirectly controllable factors and constrains the possible values of factors.[74]

With this, define the model reading:

[74] Ideally this should be refined so that it is *only* possible values of indirectly controllable factors that are constrained by the mechanisms, as discussed in the main body of the chapter. I leave this refinement as further work, however.

The ordered triplet *(D, I, M)* is the *model reading* for *(C, V, E(F))* under (h_1,h_2,h_3) where h_1: $C \rightarrow D$, h_2: $V \rightarrow I$ and h_3: $E(F) \rightarrow M$ are bijections, if and only if

(i) $\forall \alpha \text{ in } C,\ P_\alpha = P_{h1(\alpha)}.$

(ii) $\forall x \text{ in } V,\ P_x = P_{h2(x)}.$

(iii) $\forall f=0 \text{ in } E(F)$ such that $m^* = h_3(f=0)$:

$$\{(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) \in P_C x P_V \mid f(\alpha_1, \ldots, \alpha_m, x_1, \ldots, x_n) = 0\}$$

$$= \{(w_{h1(\alpha 1)}, \ldots, w_{h1(\alpha m)}, w_{h2(x1)}, \ldots, w_{h2(xn)}) \in P_D x P_I \mid \mu_{m^*}(w_{d1}, \ldots, w_{dm}, w_{i1}, \ldots, w_{in}) = 0)\}$$

This formalises the isomorphism assumed in the chapter in moving from Simon's formal concepts to the model concepts. Though it looks formally involved, it simply performs a relabeling, replacing coefficients by corresponding directly controllable factors, variables by corresponding indirectly controllable factors and functional equations by mechanisms. Importantly, it imposes that the possible values of coefficients must equal the possible values of directly controllable factors, the possible values of variables must equal the possible values of indirectly controllable factors and the constraints on possible values on factors imposed by mechanisms must be identical with those of the equations that denote them.

Given this machinery it is not difficult to define a *causal order* for the model reading of a LMS (one does this analogously to the definition of the formal order) and other model concepts from suitably defined formal concepts (e.g. complete subsets, exogeneity, endogeneity). I don't do this here because it would take too much space and would not add much beyond what is already set out in the chapter. I leave it as further work.

Finally, given the model reading, it follows that by applying the model reading to theorem 2.2 that *two LMS have the same causal interpretation if and only if they are M-equivalent.* This follows simply by stipulation of the unique model reading for an LMS over *(C, V, E(F))* i.e. *(D, I, M)* and by theorem 2.2 which states that M-equivalent and only M-equivalent LMS have the same formal order. This is the formally completed solution to the conceptual equivalence problem.

Chapter 3

Causally Interpreting Simple Models Used In Econometrics and Exploring Intervention

1. Introduction

The last chapter set out a causal interpretation for sets of solvable, linearly independent, deterministic equations by building on Herbert Simon's work on causal order. Using it one can causally interpret a set of equations such as:

$$x_1 = \gamma_{10} + \gamma_{12} x_2$$
$$x_2 = \gamma_{20} + \gamma_{23} x_3 \qquad (x\text{'s variables, } \gamma\text{'s coefficients})$$
$$x_3 = \gamma_{30}$$

This is a first step for interpreting the models actually used by econometricians. However it is only a first step since as it stands it is not sufficient for interpreting any models econometricians actually use.

Models used by econometricians are more complex than the simple, deterministic sets of equations interpreted in the last chapter. They are stochastic, have error terms, and typically treat variables and coefficients differently. To see just how different the systems of equations actually used by econometricians are, consider the following simple simultaneous equation model which might be used for econometric analysis: a supply and demand model for the wheat market.

$$q = \alpha p + \beta r + u_1 \ldots supply$$
$$q = \gamma p + \delta i + u_2 \ \ldots demand$$

Suppose that q denotes equilibrium quantity of wheat transacted, p denotes equilibrium wheat price, r denotes rainfall, i denotes income, and u_1 and u_2 are unobserved error terms denoting omitted factors. The coefficients α, β, γ and δ are unknown but assumed constant. Rainfall and income are assumed to be independent of the supply and demand mechanisms, so r and i are treated as external (exogenous) variables. Quantity and price are determined by the supply and demand mechanisms, so q and p are treated as internal (endogenous) variables. Assume rainfall and income

and the error terms are random so that all variables in the equations are random variables. Also, assume that the error terms are normally and independently distributed of each other and of the *r*, *i* variables. This is the kind of simple structural model one finds in econometrics textbooks.[1]

As is clear from this example, the simplest sets of equations that econometricians actually use to model causal relations differ from the extremely simple sets of equations interpreted in the last chapter in several key ways:

(i) Variables are partitioned into two groups: those that are internal, or endogenous, and those that are external, or exogenous.[2]

(ii) The equations contain error terms.

(iii) Error terms are stochastic, and external variables may be stochastic. As a result internal variables are stochastic.

(iv) Coefficients in the equations are constant.[3]

The work of the last chapter did not cover sets of equations which have these features. The main aim of this chapter is to extend the causal interpretative method, developed in the last chapter, to models that have the four characteristics above. It does this by working from the simple sets of equations, like those interpreted in the last chapter, assuming the strong reading holds for these and then adding further assumptions to allow the more general systems (with the differences above) to be causally interpreted.[4] In this way, it extends the strong reading of the last chapter to causally interpret the simplest models that econometricians use.

---

[1] Using such a model, the econometrician would use sample observations for *q*, *p*, *r* and *i* to estimate moments of the joint distribution for these variables to estimate values for the unknown coefficients.

[2] Here I use 'external' rather than 'exogenous' and 'internal' rather than 'endogenous'. This is because I have already used endogenous and exogenous in the previous chapter to mean something slightly different. In addition, as Stephen LeRoy (2004, p.3) notes, there are different interpretations of 'exogenous' and 'endogenous'. For these reasons I stick to the less ambiguous internal-external terminology.

[3] Though this generally holds in simple analyses, it is not always assumed. In more sophisticated models coefficients may change values; they may be random variables, functions of time etc.. For an example of this kind of econometric modeling, see Cooley and Prescott (1976).

[4] This means that the characteristics of the strong reading apply here. For instance, equations denote mechanisms that are invariant to factor changes, directly controllable factors are independent and so on.

A second aim of this chapter is to use the analysis developed in extending the causal interpretation to present a short exploration of interventions. It is possible to do this because the causal interpretation developed for sets of equations with internal and external variables assumes that these sets of equations are incomplete versions of sets of equations which contain only variables and coefficients (like those of the last chapter). In other words, the sets of equations looked at in the last chapter are taken to be 'complete' with their coefficients taken to denote the ultimate, relevant causes of the factors that are of interest to the modeller.[5] In contrast, sets of equations with internal and external variables are taken to be 'incomplete', that is, taken to represent a subset of the causal relations denoted by some complete set. In addition to providing an intuitive way of introducing external and internal variables, this method has the advantage of allowing one to discuss the properties of the causal inputs modelled by a incomplete set. This allows a discussion of different kinds of intervention.

The chapter is structured as follows. It begins the extension of the causal interpretation to sets of equations with external and internal variables by differentiating between complete and incomplete sets of equations. In doing this, it presents a few formal results that make explicit how incomplete and complete sets relate to each other, when the former represents a subset of the causal relations represented by the latter. Once this is done, the chapter uses this to present a brief discussion of different kinds of interventions. The chapter then returns to the problem of extending the causal interpretation by introducing error terms and stochastic features into sets of equations and presenting a causal reading for these.

2. Introducing External and Internal Variables: Complete vs. Incomplete Sets

In the sets of equations of the last chapter, coefficients could vary. This enabled coefficients to denote the 'causal inputs', the directly controllable factors, which

[5] Though this sounds like a very strong claim, in fact the motivation is pragmatic. It is simply a way of delimiting the causes that are relevant to the purposes of the model. I do not discuss here the conditions under which a model can be taken to be complete, particularly for purposes of causal inference, though clearly this is important. Instead, I assume that one has good reason for treating a model as complete, whatever those reasons are, and leave the analysis of those reasons as further work.

influenced other factors. In the simplest structural models econometricians use, coefficients are typically fixed and it is the external variables that vary and denote causal inputs. This suggests an obvious and straightforward way to extend the interpretation of the last chapter to equations with fixed coefficients and internal/external variables: treat the external variables in the same way that coefficients were treated in the last chapter. In other words, read these as the directly controllable factors. This is simple and would do the job. However, I take the analysis a step further and instead develop the idea that a set of equations which has internal and external variables is an incomplete version of a complete set which has variables and coefficients. This approach is logically stronger[6] but has the advantage of allowing an analysis of interventions later in the chapter.

2.1. Incomplete Sets of Equations and their Causal Interpretation

To begin, I present a definition of a complete set. These are the sets of equations which were interpreted in the last chapter.

> A *complete set* of linear equations is a set of equations that are linear in the (non-zero) coefficients and linear in the variables, where the equations are linearly independent and solvable for the variables in terms of the coefficients. The coefficients are variation free.

This definition essentially matches Simon's definition of linear model (1953, p.14). The only substantive difference is that a variation free requirement on the coefficients is made explicit. Recall from the last chapter that coefficients being variation free means that the coefficients can take any value as a group as they can individually.[7] This was necessary for interpreting the coefficients as directly controllable in the model reading.

An incomplete set is defined in a similar way to a complete set, but some of its variables are termed 'external' and are treated as if they are coefficients.

---

[6] Since it assumes, in addition to the set of equations with internal and external variable that a larger more comprehensive complete set holds. In addition, when causally interpreted it assumes that model readings hold for both sets of equations.

[7] Formally a set $\{z_1,z_2,\ldots,z_n\}$ of coefficients or variables is variation free, letting $P(z_i)$ denote the set of possible values for z_i, if and only if $P(z_1,z_2,\ldots,z_n) = P(z_1)xP(z_2)x\ldots xP(z_n)$.

> An *incomplete set* of equations is a linearly independent set of equations that are linear in the variables and the (non-zero) coefficients and in which variables are partitioned into two sets: external and internal. The equations are such that internal variables can be solved for in terms of the coefficients and external variables. The set of coefficients and external variables together are variation free.

As is clear from the two definitions, if one reclassified the external variables in the incomplete set as coefficients one would obtain a complete set. This implies that one can apply Simon's formal ordering methods to incomplete sets by treating external variables as if they were coefficients.[8]

For example, consider the incomplete set of equations:

$$y_1 = \alpha_{10} + \alpha_{12}y_2 + \beta_{11}x_1 + \beta_{12}x_2$$
$$y_2 = \alpha_{20} + \alpha_{21}y_1 + \beta_{21}x_1 + \beta_{23}x_3$$
$$y_3 = \alpha_{30} + \alpha_{32}y_2 + \beta_{34}x_4$$

(x's external, y's internal, α and β's coefficients)

Suppose one applies Simon's formal ordering method treating the external variables as coefficients and internal variables as variables. First, one solves for y_1 and y_2 from the two first equations to get that the first two equations are a complete subset of 0^{th} order of equations. This also gives $\{y_1, y_2\}$ as the corresponding complete subset of 0^{th} order of the internal variables. The third equation can then be solved for y_3, so it forms the only complete subset of equations of 1^{st} order. Likewise, this gives $\{y_3\}$ as the corresponding complete subset of 1^{st} order in the internal variable ordering. So the formal order over the internal variables is $\{y_1, y_2\} \rightarrow \{y_3\}$.

That one can apply Simon's formal ordering method to these incomplete sets suggests that they can also be interpreted using the model reading of the last chapter. The key to this, as when applying Simon's ordering method, is to treat external variables as if they are coefficients. That said, since external variables and coefficients are

[8] Strictly speaking this requires the non-linear version of Simon's methods.

nevertheless formally distinct, applying the model reading to an incomplete set, one should read coefficients as directly controllable factors, external variables as indirectly controllable factors *that can be treated as if they are directly controllable* and internal variables as indirectly controllable factors.

One could leave it at this, after all this gives a way to causally interpret sets of equations with internal and external variables. Doing it in this way, one simply re-labels the external variables in the equations as 'as if coefficients' and reads the resulting equations using the method of the last chapter. However, there is something unsatisfactory about this re-labelling approach. In particular, it reveals nothing as to why variables can be treated as external. In the model reading, if one reads variables as indirectly controllable factors, why is it that some of these (the external variables) can be read as directly controllable factors?

To answer this question requires a framework within which an explanation can be provided as to how indirectly controllable factors can be treated as directly controllable. To give such a framework, I make an additional assumption that *in the model reading an incomplete set represents just some of the causal relations represented by a complete set.* With this additional assumption, incomplete sets 'abbreviate' complete sets. This assumption provides the necessary extra content for analysing how an incomplete set can have some its variables, its external variables, read as if they denoted directly controllable factors. The next section begins this work of setting out more precisely the relationship between an incomplete set and the complete set which it abbreviates.

2.2. Constructing Incomplete Sets from Complete Sets

To analyse the relationship between complete and incomplete sets, I first consider how an incomplete set of equations can be mathematically derived from a complete set. This is then used to show how an incomplete set can have an intuitively inconsistent causal interpretation from that of the complete set from which it is derived. This helps later in setting out a definition of how an incomplete set can be

causally consistent with a complete set, which provides formal conditions for when an incomplete set represents just some of the causal relations represented by a complete set.

One way to derive an incomplete set from a complete set is straightforward: one linearly transforms the complete set of equations into any other set, drops any equations one likes and then stipulates that sufficiently many variables in the resulting equations be external so that the resulting equations are solvable for the remaining (internal) variables in terms of coefficients and external variables.[9]

To see how this works, consider the complete set:

$$z_1 = \alpha$$
$$z_2 = z_1 + \gamma$$
$$z_3 = z_1 + z_2 + \lambda$$

One can drop the first equation to get.

$$z_2 = z_1 + \gamma$$
$$z_3 = z_1 + z_2 + \lambda$$

Suppose one classifies z_1 and z_2 as internal and z_3 as external, then one gets.

$$z_2 = z_1 + \gamma$$
$$z_3 = z_1 + z_2 + \lambda \qquad (z_1 \text{ and } z_2 \text{ internal, } z_3 \text{ external})$$

For this to be an incomplete set, it is necessary that the set $\{z_3, \gamma, \lambda\}$ be variation free. If one solves z_3 in terms of coefficients in the complete set, one can see that it depends on α. Since $\{\alpha, \gamma, \lambda\}$ are variation free in the complete set,[10] it follows that $\{z_3, \gamma, \lambda\}$ are variation free. Since the set of derived equations is solvable for the internal variables and has variation free external variables and coefficients, it is an incomplete set of equations. Also, since the incomplete set has been derived from the complete set, it is mathematically consistent with it.

[9] Note that the many different ways of doing this shows that a complete set of equations will generally have many different incomplete sets that can be mathematically derived from it.
[10] Recall in the definition of a complete set, the coefficients are variation free.

What is interesting here is the causal order for both the complete set and the incomplete set.[11] These are:

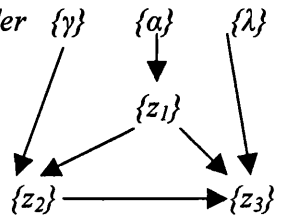
Complete Set 1

$z_1 = \alpha$

$z_2 = z_1 + \gamma$

$z_3 = z_1 + z_2 + \lambda$

Causal Order



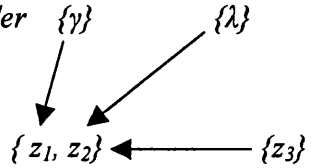
Incomplete Set 1

$z_2 = z_1 + \gamma$

$z_3 = z_1 + z_2 + \lambda$

(z_1 and z_2 internal, z_3 external)

Causal Order



In this example the derived incomplete set has intuitively[12] an inconsistent causal interpretation from the complete set from which it has been derived. This is because in the incomplete set z_3 and λ are direct causes of z_1 and z_2 while none of these relations hold in the causal order of the complete set. So spurious causal relations, absent from the formal order of complete set, have been introduced in constructing the incomplete set. In addition, since α is a direct cause of z_1 for the complete set but not the incomplete set, it also has left out a cause of one its internal variables (i.e. z_1).

Clearly in such a case, the derived incomplete set does not represent just some of the causal relations represented by the complete set. It adds causal relations that are not present in the original causal order (for example, λ directly causing z_1 and z_2) and it leaves relations out (for example, from α to z_1). So not all incomplete sets of equations that can be mathematically derived from a complete set intuitively abbreviate it. So, which derived incomplete sets do abbreviate the complete set?

2.3. Causal Consistency of Incomplete and Complete Sets of Equations

The question that needs to be answered is: under what conditions does an incomplete set have a causal interpretation that is consistent with a complete set from which it is

[11] Here the coefficients and the external variables are included in the causal order, using the concept of extended formal order defined in chapter two.

[12] I say 'intuitively' because a definition for consistency has yet to be given.

derived. As the example above suggests, to be consistent with a complete set, an incomplete set should have a causal interpretation involving only mechanisms, factors and order relations that hold among these that are represented by the complete set and it should not leave out causal relations among the factors that it models. Together, these lead to a natural way of stipulating how an incomplete set represents just some of the causal relations represented by the complete set without leaving important causal information out. It suggests the following definition of causal consistency.

> An incomplete set is *causally consistent* with a complete set if and only if
>
> (a) Each of its equations is an equation in the complete set.
>
> (b) All of the formal order relations, obtained using Simon's formal ordering methods, between equations in the incomplete set, and between its internal variables, also hold for those equations and variables in the formal order of the complete set.
>
> (c) Formal order relations in the complete set that hold among internal variables and equations that appear in the incomplete set also hold in the formal order of the incomplete set.
>
> (d) The external variables and coefficients in the incomplete set of equations are variation free in the complete set.

Requiring (a), that each equation in the incomplete set appear in the complete set, ensures that every mechanism denoted by the incomplete set is also denoted by the complete set.[13] Stipulating (b), that all the formal order relations that hold among equations and variables in the incomplete set hold in the complete set, ensures that all the causal relations in the interpretation of the incomplete set also hold in the interpretation of the complete set. In this way, nothing spurious is introduced by the

[13] At the end of the last chapter it was shown that rescalings of equations also preserve the mechanism denoted by the equation in the model reading. From this it follows that an incomplete set could also consist of rescaled equations of the complete set. I do not include this slightly weaker possibility here because such rescalings are not significant from a causal perspective. Also, it keeps the discussion simpler.

incomplete set. While the (c) condition ensures that any causal information about the factors modelled in the incomplete set *as modelled by the complete set* is modelled by the incomplete set. In this way nothing causally important is left out about the factors and mechanisms to be modelled by the incomplete set. Finally, (d) is added to ensure that the set of external variables and coefficients that appear in the incomplete set, which must be variation free by the definition of an incomplete set, are also variation free in the complete set.

This definition gives a precise answer as to when an incomplete set is causally consistent with a complete set from which it can be derived. In other words, it makes clear what it means for an incomplete set to represent just some of the causal relations represented by a complete set.

2.4. Why some Indirectly Controllable Factors can be Treated as Directly controllable

One reason for developing the concept of an incomplete set representing just some causal relations of a complete set is to attempt to answer: why is it that some indirectly controllable factors can be treated as if they are directly controllable in an incomplete set of equations?

In order to address this, I present a necessary condition for an incomplete set to be causally consistent with a complete set it abbreviates. This gives a more intuitive, causal condition for what is required for a set of equations and variables from a complete set to form an incomplete set. The following theorem, proved in appendix 3.1, gives a necessary condition.

> *Theorem 3.1*: An incomplete set of equations is causally consistent with a complete set of equations only if it meets (NC).
>
> (NC): (I) The incomplete set is a union of complete subsets of equations in the formal order of the complete set.
>
> (II) Its set of internal variables is the union of those variables which are endogenous for those complete subsets.

(III) For any two internal variables y and z such that y causes z in the formal ordering of the incomplete set, then in the formal order of the complete set either y is a direct cause of z or there exists a chain of direct causes such that $y \rightarrow w_1 \rightarrow \ldots \rightarrow w_j \rightarrow z$ where for all j, w_j is an internal variable.

In (NC) 'exogenous' and 'endogenous' are meant in the technical sense used in the last chapter and by Simon (1953). Though the condition may sound opaque, it is an interesting result. For example, (II) and (III) put limits on what variables can be external. (II) requires that all and only internal variables are endogenous (in the ordering of the complete set) for the equations that are included in the incomplete set. This implies that variables that are to be treated as external must be exogenous (in the ordering of the complete set) or unordered with respect to the equations in the incomplete set. Whereas (III) requires that for variables to be treated as external, they must not block all of the 'causal paths' from one internal variable to another (in the ordering of the complete set). If this were not met then treating these variables as external would 'break' the causal path from one internal variable to another, implying that first internal variable would not cause the second internal variable in the incomplete set, while it did in the complete set. So the incomplete set would not then be causally consistent with the complete set.

Note that in the example above, the intuitively inconsistent incomplete set 1 did not meet (NC) for complete set 1, because its external variables were not all exogenous in the ordering of the complete set, that is, it failed condition (II) in (NC) because z_3 was treated as external despite it being endogenous for the complete subset of equations (in the ordering of the complete set) which was included in the incomplete set.

To contrast with this earlier example, consider the following incomplete set which satisfies (NC). First recall the original complete set, and an incomplete set.

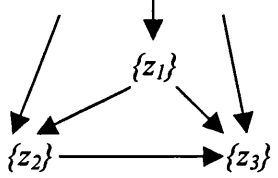
Complete Set 1

$z_1 = \alpha$

$z_2 = z_1 + \gamma$

$z_3 = z_1 + z_2 + \lambda$

Causal Order {γ} {α} {λ}



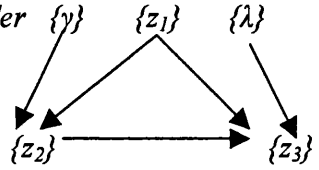
Incomplete Set 2

$z_2 = z_1 + \gamma$

$z_3 = z_1 + z_2 + \lambda$

(z_1 external, z_2 and z_3 internal)

Causal Order {γ} {z1} {λ}



Incomplete set 2 is causally consistent with the original complete set.[14] The first incomplete set simply drops the information about the causes of z_1. Since z_1 is not caused by any other variable in the complete set, this leads to a straightforwardly causally consistent incomplete set.

The following incomplete set is also causally consistent with complete set 1 but it is more interesting since in it the variable, z_2, is treated as external even though it is caused (in the ordering of the complete set) by another variable, z_1, which is internal.

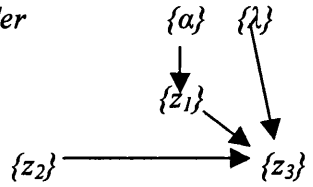
Incomplete Set 3

$z_1 = \alpha$

$z_3 = z_1 + z_2 + \lambda$

(z_1 and z_3 internal, z_2 external)

Causal Order



It may seem surprising that a variable can be treated as external even though it is caused by an internal variable. After all, if external variables denote 'causal inputs' then one would expect them to be determined outside the system and not be dependent on a variable internal to the system. This counterintuitive result suggests that the causal consistency definition provided here may be usefully supplemented by a second stronger version which rules out such cases. So, I also define a stronger form of causal consistency as follows.

> An incomplete set is *strongly causally consistent* with a complete set if and only if it is causally consistent with it and none of its internal

[14] To check this note that z_1 is variation free with respect to γ and λ, also the incomplete set only contains formal order relations that are in that of the complete set, and it does not omit any relation between internal variables and included equations.

variables is a cause of any of its external variables in the order of the complete set.

This stronger form of causal consistency rules out these counterintuitive cases where an internal variable is caused by an external variable. Obviously, since strong causal consistency implies causal consistency, (NC) is also a necessary condition for strong causal consistency.

Returning to (weak) causal consistency, it is important to note that though (NC) is necessary for causal consistency it is not sufficient. Consider the following new complete set, which is the same as the previous complete set except that γ has been dropped from the second equation. Then, reconsider incomplete set *3* above.

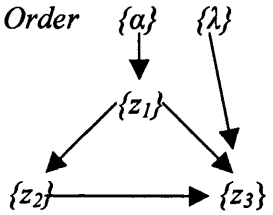
Complete Set 2

$z_1 = \alpha$

$z_2 = z_1$

$z_3 = z_1 + z_2 + \lambda$

Causal Order



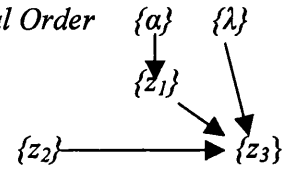
Incomplete Set 3

$z_1 = \alpha$

$z_3 = z_1 + z_2 + \lambda$

(z_1 and z_2 external, z_3 internal)

Causal Order



Incomplete set 3 meets (NC) for complete set 2. In addition, since it is an incomplete set its external variables, z_1 and z_2, must be variation free.[15] However, if complete set 2 holds then z_1 and z_2 are not variation free because they always have identical values (both are equal to α). Therefore, though it meets (NC) incomplete set 3 fails condition (d) of the definition of causal consistency and so is not causally consistent with complete set 2. This shows that (NC) is not sufficient for causal consistency.

In addition, it is interesting to note the that incomplete set 3 is causally consistent with complete set 1 but not complete set 2. The reason is that in complete set 1, z_2 has its own particular cause, γ, which is absent from complete set 2. This extra cause 'gives' z_2 the freedom to vary relative to the coefficients in incomplete set 3 so that it

[15] Recall that this was part of the definition of the incomplete set.

is variation free in complete set 1. This extra cause is absent in complete set 2, so z_2 cannot be variation free as required for causal consistency with that complete set. Therefore, this extra causal input in complete set 1 that is particular to z_2 is important for allowing the variation free condition, (d), of the causal consistency to be met. This suggests a possible connection between the variation freedom requirement and external variables in an incomplete set having separate 'causal inputs' from the other external variables in the complete set. This suggests in turn that it may be fruitful to investigate what causal conditions are necessary and sufficient for the variation free condition to be met. I leave this as further work, however.

To conclude, this section has shown several things. It has provided a method to extend the model reading of the last chapter to sets of equations with internal and external variables. The trick is to take the external variables to denote indirectly controllable factors that can be treated 'as if' they are directly controllable. More interestingly, it has provided a partial explanation of why external variables can be treated in this way in the interpretation of incomplete sets. Given the intuitive definition of causal consistency adopted here, the key requirements are, in the formal order of the complete set from which an incomplete set is constructed, that the endogenous variables for those equations be treated as internal, while the exogenous variables be treated as external. In addition, external variables should not block causal paths between internal variables in the ordering of the complete set. Interestingly, it was then shown that this definition of causal consistency was rather weak in that it permitted internal variables to cause external variables in the complete set. Therefore, a definition of strong causal consistency was offered as a way of ruling this out. Finally, it was noted that the variation free condition of causal consistency, in certain situations, required external variables to have their own 'causal inputs' separate from other external variables. This suggested an interesting area for further work.

3. Using Incomplete Sets to Explore Intervention

One of the reasons for setting out explicit conditions for an incomplete set of equations to be causally consistent with a complete set, is to provide a framework within which to discuss interventions into the factors denoted by an incomplete set. Interventions are important in any discussion of causal order since they are the 'point of entry' into the explicitly modelled causal relationships. They form the bridge between the implicit: what is required to bring about an intervention, and the explicit: what an intervention does to the modelled causal order.

In practice, being clear about intervention is also important for describing and performing experiments. For example, if one wants to test a putative causal relation from a factor, c, to a factor, e, then following John Stuart Mill's method of concomitant variations (1851, pp.398-401), one changes c to see if e changes. However, this method is useless if in varying c one unknowingly varies a common cause of c and e rather than just a cause of c. In that case, any observed co-variation of c and e does not give conclusive grounds for believing c is a cause of e, since it may be due to the common cause varying both c and e rather than any influence from c to e. The point is that to set out when Mill's method of concomitant variations is an effective way to identify causes, it is important to distinguish an intervention of c that does not activate a common cause of c and e (which is desirable) and one which does (which is undesirable). To make distinctions such as these, requires clarity about what is meant by 'intervention'.

This section shows how the analysis of the previous section can be used to explore intervention. By looking at a simple example, the discussion sets out some different ways an incomplete set can be intervened into. The analysis is then loosely generalised to make some general comments about interventions.

To keep the discussion simple, strong causal consistency between incomplete and complete sets is assumed in the example, that is, internal variables do not cause external variables in the ordering of the complete set. This simplifies the discussion

because otherwise the possibility that internal variables cause external variables in the complete set would need to be dealt with. Though it would be interesting to investigate what results hold for interventions in this case, it would be more complex and involved. Since the aim here is to give a brief simple analysis of interventions, I leave this more complex case as further work.

To ease the discussion, I introduce some new terminology. Let an external factor be a factor that is denoted by an external variable in an incomplete set, an internal factor a factor denoted by an internal variable. Let an incomplete model be the model reading of an incomplete set and a complete model that of a complete set. Also, I define an intervention and a basic intervention as follows.

> An *intervention* in a complete model is a change to one or more directly controllable factors in the complete model. An intervention in an incomplete model is a change to one or more directly controllable or external factors in the incomplete model.
>
> A *basic intervention* on a factor, x, in a complete model is an intervention that changes one and only one directly controllable factor, x, in the complete model. A basic intervention on x in an incomplete model is an intervention that changes one and only one external or directly controllable factor x in the incomplete model.

If one uses 'causal input' to denote any factor in a model that is directly controllable or external,[16] then an intervention is a change in one or more causal inputs in a model, whereas a basic intervention in a model is a change in just one causal input in the model. Basic interventions are intuitive – they describe the surgical, clean interventions of changing just one causal input which are particularly desirable for carrying out experiments.

[16] Strictly speaking, causal inputs in a complete model are the directly controllable factors. While in an incomplete model, the causal inputs are the directly controllable or external factors.

Note also that this concept of basic intervention fits neatly with the Simon-based analysis developed in this thesis. This is because Simon's theorem 6.1, discussed in the last chapter, which sets out which variables change value as a result of a change in just one equation, can be recast in terms of basic interventions. The theorem, under the model reading, implies that a basic intervention on a causal input 'in general' changes all factors that are causally dependent on the causal input while all other factors do not change.[17]

3.1. Different Ways to Vary Just One External Factor

With this background, imagine an incomplete model, denoted by the following incomplete set, holds.[18]

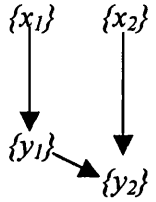
Incomplete Set

$y_1 = x_1$

$y_2 = y_1 + x_2$

(x's external, y's internal)

Causal Order



Suppose that one wants to investigate the strength of the causal relationship between y_1 and y_2. Following Mill's methods, the simplest way to do this is to perform a basic intervention for the x_1-factor[19] and to observe the change in the y_1-factor and the change in y_2-factor. But how can the required basic intervention be brought about?[20] To answer this requires more background information, so in the following four cases, different possible background causal orders for the x's are considered. More precisely, four different complete models each strongly consistent[21] with the incomplete model are looked at. This allows the exploration of different ways a basic

[17] This result is used in the cases below to deduce changes given a basic intervention.

[18] There are no coefficients in this incomplete set to simplify the causal graphs below, though the analysis can be extended for cases where there are coefficients in the incomplete set.

[19] From now on I use 'a basic intervention for x' rather than 'a basic intervention for the x-factor' to make the wording less awkward.

[20] One might wonder if a basic intervention on x_1 is possible at all. In fact, it is because of the variation free requirement in the definition of a incomplete set, which ensures that basic interventions are possible for all directly controllable factors and external factors. Likewise, the variation free assumption ensures that basic interventions are possible for all the directly controllable factors in a complete model.

[21] A complete model is strongly consistent with an incomplete model if a complete set that denotes the complete model is strongly causally consistent with an incomplete set that denotes the incomplete model.

intervention in an incomplete model can come about, depending on the background causal order.

The first case is where the following strongly causally consistent complete set applies.

Case (a)

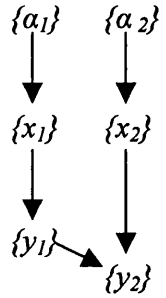
$x_1 = \alpha_1$

$x_2 = \alpha_2$

$y_1 = x_1$

$y_2 = y_1 + x_2$

Causal Order



In this case, the only way to perform a basic intervention[22] on x_1 in the incomplete model is to perform a basic intervention on α_1 in the complete model. The causal order of the complete model makes this straightforward because α_1 is a direct cause of x_1 that is exclusive to x_1. Moreover, since x_1 doesn't cause any other external factor, a change only in α_1 leads to a change only in x_1, giving the desired basic intervention.

A slightly more complex case is where x_1 and x_2 share a common cause in the complete set.

Case (b)

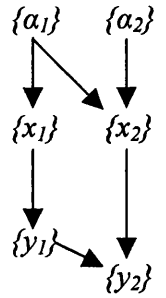
$x_1 = \alpha_1$

$x_2 = \alpha_1 + \alpha_2$

$y_1 = x_1$

$y_2 = y_1 + x_2$

Causal Order



In this case a basic intervention on x_1 can only be brought about by a non-basic intervention on α_1 and α_2 together such that $\Delta\alpha_1 = -\Delta\alpha_2$. This is because the only way to vary x_1 is by varying its cause α_1 which *ceteris paribus* changes x_2, so α_2 also needs to be changed in order to keep x_2 fixed. Here x_1 does not have a direct cause that is

[22] In subsequent discussion, to simplify the discussion I leave out 'relative to ... model' when discussing basic interventions. In the discussion, basic interventions on α's are relative to the complete model, while basic interventions on *x's* are relative to the incomplete model.

exclusive to it among the external factors. Given this, additional direct causes of the other external factors must be varied to keep those external factors from changing.

Another interesting scenario is where x_1 is a cause of x_2.

Case (c)

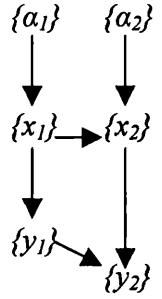
$x_1 = \alpha_1$

$x_2 = x_1 + \alpha_2$

$y_1 = x_1$

$y_2 = y_1 + x_2$

Causal Order



Here a basic intervention on x_1 can only be brought about by a non-basic intervention that changes α_1 and α_2 together so that $\Delta\alpha_1 = -\Delta\alpha_2$. The change in α_2 is necessary to prevent x_2 changing as a result of x_1 changing. Unlike the last case, x_1 has a direct cause that is exclusive to it (α_1), yet x_1 causes x_2 in the complete model, so changing x_1's exclusive direct cause does not bring about a change in x_1 alone. So a direct cause of x_2 must also be changed to keep it fixed.

In the final case x_1 and x_2 are codetermined in the complete set.[23, 24]

Case (d)

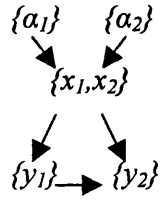
$x_1 = x_2 + \alpha_1$

$x_2 = -x_1 + \alpha_2$

$y_1 = x_1$

$y_2 = y_1 + x_2$

Causal Order



In this case a basic intervention on x_1 can only be brought about by a non-basic intervention that changes α_1, α_2 together so that $\Delta\alpha_1 = \Delta\alpha_2$. The case is similar to the previous case in that x_1 is causally ordered relative to x_2. However, unlike the

[23] By co-determined I mean that both appear in the same complete subset of the causal order.

[24] One might think that since they are codetermined the two x's cannot be variation free. After all, how can two variables be free to vary relative to each other if they are co-determined? However, this intuition is mistaken. The first two equations of the complete set imply that $x_1 = \frac{1}{2}(\alpha_1 + \alpha_2)$ and $x_2 = \frac{1}{2}(\alpha_2 - \alpha_1)$. It is easy to check that $\alpha_1 + \alpha_2 = a$ and $\alpha_2 - \alpha_1 = b$ can be solved for any a and b. Given α_1 and α_2 are variation free, it follows that x_1 and x_2 are variation free (as required by the definition of an incomplete set).

previous case, both α_1 and α_2 both directly cause x_1 and x_2 so there is no direct cause exclusive to x_1.

These four cases, though clearly not exhaustive, show how basic interventions on x_1 in the incomplete model can arise in different ways, depending on the causal order that determines the external factors. It shows the many ways in which a statement such as 'x_1 is varied but not x_2' can hold for an incomplete model. Since these different background causal orders are all strongly consistent with the incomplete model, the external factors of the incomplete model have values that are variation free in all cases. The cases also show how weak the variation free assumption is, for example, that the variation free requirement on the external factors does not require that the external factors have exclusive causes nor that they do not cause each other.[25]

The four cases can also be used to develop some tentative general observations about interventions in incomplete models that are strongly consistent with an underlying complete model. For instance, the only case where a basic intervention in the complete model leads to a basic intervention in the incomplete model is the first case. There x_1 has an direct cause exclusive to it, and x_1 does not cause and isn't causally equivalent to any other external factor. In all the other cases, at least one of these conditions fails and a non-basic intervention is required.

In fact, a general form of this result holds. For a basic intervention in a directly controllable factor, α, to lead to a basic intervention in an external factor, x, then the directly controllable factor must only cause the external factor.[26] This then implies that the directly controllable factor can (i) *only cause an external factor via x* and (ii) *x must not cause nor be causally equivalent to any other external factor*. Conversely, if (i) and (ii) hold then the external factor, x, has a directly controllable factor, α, that

[25] Recall the gas container example in the last chapter which made a similar point: the independence implied in the model reading by a variation free condition on the coefficients does not imply that these cannot denote factors that are causally related.

[26] Purely to keep the discussion simple, here I am setting aside the possibility that the directly controllable factor can cause other factors but that its impact cancels out. In a more rigorous treatment this problem would not be put aside and would instead be dealt with explicitly.

causes it and no other external factor, so a basic intervention on α in the complete model, leads to a basic intervention on x.[27] In this way (i) and (ii) are necessary and sufficient[28] conditions for there to be a basic intervention in the complete model that leads to a basic intervention in the incomplete model.

The importance of conditions (i) and (ii) can be highlighted by returning to the epistemic problem of investigating the causal relation between y_1 and y_2 in the assumed incomplete model

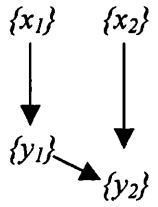
Incomplete Set

$y_1 = x_1$

$y_2 = y_1 + x_2$

(x's external, y's internal)

Causal Order



Here, if one knows that x_1 only causes y_1 among the internal factors, and if x_1 meets condition (i) and (ii) then one can vary x_1 using a directly controllable factor, α, so that only y_1 changes. In this way a suitable experiment for analysing the causal relation between y_1 and y_2 can be carried out. However, if (i) and (ii) are not met, then the directly controllable parameters need to be varied in a 'just so' way, that is a non-basic intervention needs to be performed so that only x_1 changes. As seen in the above cases, this requires that certain causal impacts cancel out, which radically changes what is required of the experimenter. In particular, it requires detailed knowledge of the background causal order, omitted from the incomplete model, in order to carry out the desired experiment.

This relationship between interventions and experiments has been analysed by others, and conditions (i) and (ii) seem to fit well with some of these analyses. For instance, Julian Reiss (2003) gives a definition of an experimental handle:

> 'Z is an "experimental handle" [for X relative to Y] if it satisfies the following assumptions:

[27] Again I am overlooking the possibility that the influence from α to x cancels out. In a more rigorous treatment this would be dealt with explicitly.

[28] Of course, these are only necessary and sufficient if the canceling out possibility is ruled out. See footnotes above.

EH-1 Z Causes X
EH-2 Z Causes Y if at all only through X
EH-3 Z and Y do not have causes in common (except those that might cause Y via Z and X).' (Reiss, 2003, p.12).

As Reiss points out, this is very close to James Woodward's (2003, p.98) definition of an 'intervention variable'. Woodward's definition of an intervention variable also has EH-1 and EH-2 as conditions and has a statistical independence version of EH-3. In addition, Woodward goes further than Reiss by characterizing an intervention on X with respect to Y as the use of some intervention variable Z for X with respect to Y to influence X.

There is a connection between this Reiss and Woodward type analysis and that developed here. For x an external factor such that (1) x causes y_1 (EH-1) and (2) x only causes another internal factor y_2 if at all via y_1 (EH-2) then, if conditions (i) and (ii) are met for x, then there is a directly controllable factor, α, that is an experimental handle for y_1 relative to y_2. To see how, first note that conditions (i) and (ii) ensure that there is a directly controllable factor, α, which is an exclusive cause of x among the external factors. Therefore, by transitivity of the causal relation α causes y_1 and so meets EH-1. Since α exclusively causes x among the external factors and internal factors are caused by the external factors, α can only cause y_2 if at all via x. Since x can only cause y_2 if at all via y_1, α too can only cause y_2 if at all via y_1. In this way α meets EH-2. Finally, EH-3 is met given an important additional assumption that the complete model is suitably 'complete' in the sense that that there are no common causes between a directly controllable factor and an indirectly controllable factor. So provided some important additional assumptions are met (i.e. x meets EH-1 and EH-2 and the complete model is 'complete') then conditions (i) and (ii) can be sufficient for there being a directly controllable factor which can be used for investigating, by experiment, the causal relationship between two internal factors.

In addition, conditions (i) and (ii) also connect with the concept of an 'Open Back Path' which originates from Nancy Cartwright (1989) and which has been adopted by

others such as Daniel Hausman (1998). The connection can be seen from the obvious similarity of Hausman's open back path condition to Reiss's experimental handle conditions and Woodward's definition of an intervention variable.[29] Hausman's open back path condition, for instance, is essentially the conjunction of EH-2 and EH-3 in Reiss' formulation of an experimental handle.

> '(Open back path condition). Every cause *a* of *b* that has any causes has at least one cause *d* such the only path from *d* to *b* is via *a*' (Hausman, 1998, p.83)

Such connections between (i) and (ii) and other characterisations of interventions show that the analysis of interventions begun here, using the incomplete and complete sets, yields some results similar to those already in the literature.

Nevertheless, it is important to mention some differences between the concept of intervention adopted here and that developed by others. Here an intervention is simply a change to one or more causal inputs and a basic intervention is a change to a unique causal input. While additional conditions, such as (i) and (ii) or those of the experimental handle, are desirable for epistemic reasons (since they ensure that it is possible to experiment to investigating causal relations among internal factors) it is *not* assumed here that such epistemically convenient conditions must be met for a change to be an intervention. Crucially, this is in contrast to certain other important analyses of interventions, such as those by James Woodward (2000) and Judea Pearl (2000).

As mentioned above, Woodward's characterises an intervention as influencing a variable using an intervention variable for it, which by definition meets conditions EH-1, EH-2 and a version of EH-3. Moreover, Woodward assumes that causal structures are such that there exists an intervention variable for every variable denoting an effect. His key assumption is that sets of equations representing true causal structures are *modular*, which requires that 'for each equation there is a possible intervention on the dependent variable that changes only that equation while

[29] Indeed, this shouldn't be surprising since Cartwright's (1989) work which developed the open back path was highly influential on the work of Hausman, Reiss and Woodward.

the other equations in the system remain unchanged' (Woodward, 2000, p.329). Since in Woodward's causal interpretation of structural equations, the dependent variable denotes the effect and the independent variables the direct causes of that effect, this ensures that each effect factor in a causal structure has a cause that meets conditions EH-1, EH-2 and EH-3. This means that in Woodward's approach, for any causal structure, one can investigate a putative direct causal connection from x to y from simply by varying the associated intervention variable for x. So, Woodward limits interventions to changes to the system that meet certain experimentally desirable conditions and assumes, rather strongly, that all causal structures are such that there one can intervene cleanly on any variable in that system. Both of these assumptions distinguish it from the approach taken here.

A similar approach to Woodward's is taken by Pearl (2000). Simplifying somewhat,[30] Pearl treats interventions into an effect as follows: replace the structural equation for that effect (where, like Woodward, the dependent variable in the equation denotes the effect, and the independent variables the direct causes of that effect) by an equation assigning the effect-variable to a particular value; in other equations where that effect appears, replace the effect-variable by the value to which it is set.[31] By assuming that structural equations can be individually intervened to in this way, Pearl assumes that for each effect there is a possible intervention that only impacts on that effect (and through it, its effects in turn). So, like Woodward, Pearl assumes that causal structures are modular.

A related divergence between Pearl's treatment of interventions and that taken here can be seen from an interesting criticism of Pearl presented by Stephen LeRoy (2004, pp.25-26). In his discussion, LeRoy considers the following structural equations, taken from Pearl (2000, p.217),

[30] See Pearl (2000, p.85) for further details.
[31] This is a simplification of Pearl since in fact, one *conditions* on the other variables given the effect variable is set to the particular value. However, this more complex treatment reduces to this simpler account above in the deterministic case which is what I am focusing on here.

$$q = \alpha_1 p + i$$
$$p = \alpha_2 q + w$$

where q and p denote internal factors, i and w denote external factors. LeRoy considers the response to the following question raised by Pearl: *what would the value of q be if one were to intervene to set the p to* p_1*?*[32] Following Pearl's calculus of intervention, one answers this question by replacing the second equation by $p=p_1$, and replacing p in the first equation by p_1. From this, one would answer the question à la Pearl that the value of q would be $\alpha_1 p_1 + i$ as a result of this intervention.

LeRoy's criticism of Pearl is that this answer diverges from the standard response given by economists.[33] In contrast, in my set up (and LeRoy's) the question is unanswerable without further information. In my analysis, given that p and q are codetermined by w and i, one needs to know how these external factors change to bring about the intervention on p in order to understand what would happen to q. Another way of seeing the difference is to note that for Pearl the intervention to p 'only goes through' the second equation. This means that some direct cause of p is used to change p without changing any other effect in the system (except via p). In this case, this means that the intervention on p is assumed to be affected purely via w. To use Woodward's terminology, w is the intervention variable for p, and the intervention on p is carried out using it. The relevant contrast between Pearl's approach (or Woodward's) and mine is that I do not assume *a priori* all interventions come via equation-specific intervention variables. And, as LeRoy comments, this has the advantage of being more in keeping with conventional economic interpretations of structural equations.

---

[32] Again, I am simplifying this question a little because I have made the question deterministic since I have yet to introduce indeterminism to models. However, incorporating indeterminism would not alter the substance of the discussion presented here.

[33] Pearl notes that when he presented this example to economists, most could not answer the question as posed (Pearl, 2000, p.216, [10]). Pearl attributes this to a lack of clarity over the causal interpretation of structural equation models for which he sees his formalism as a useful tool for addressing. In contrast, LeRoy considers the economists correct to reject Pearl's reading of the intervention. As noted above, my approach matches the LeRoy's and the economists' view rather than Pearl's.

So in summary, both Woodward and Pearl assume that interventions directly affect only one effect factor (and through it, its effects in turn) in the causal structure. In addition, they assume that all causal structures are modular, that is, that such interventions are possible for any effect in the causal structure. In contrast to Woodward and Pearl, I *do not* assume that causal structures are modular nor that interventions are equation-specific. More generally, I do not assume a concept of intervention that requires epistemically convenient conditions such as (i), (ii), EH-2 or EH-3 to be met.[34] Nevertheless, the analysis here converges with work by others on intervention, in that the results are similar in the sense that when additional epistemically convenient conditions (e.g. (i), (ii), EH-2, EH-3 etc.) are met, then the causal structures are similar to other approaches, such as Woodward and Pearl's approaches.

To conclude, this section used the relationship between incomplete sets and strongly causally consistent complete sets to briefly explore interventions in an incomplete model. It did this by looking at some different ways interventions could be brought about in an incomplete model, depending on the complete model that holds. The analysis shows how, by introducing the complete model strongly consistent with an incomplete model, one can make partially explicit what is involved in intervening to change just one external factor in an incomplete model.[35] Finally, it presented some similarities and differences between this approach to interventions and that of other authors.

4. Introducing Error Terms

This section returns to the problem of extending the causal interpretation to models actually used by econometricians. A way has already been presented for interpreting internal and external variables. This section proposes a way to interpret error terms.

---

[34] In chapters four and five, I discuss further how my approach attempts to avoid building in epistemically convenient assumptions into its concept of causal order.

[35] It is only partially explicit because in the analysis a basic intervention in an incomplete model is causally unpacked relative to a complete model in which an unanalysed intervention (the basic intervention or non-basic intervention on the directly controllable factor) is assumed.

The standard reading of error terms is that these denote causal factors omitted out of ignorance. For example, according to Kevin Hoover 'error terms might be thought to represent those *INUS* conditions that, though they help to determine the effects and are not constant, are not explicitly measured or modelled' (2001a, p.50). While Herbert Simon (1954) states that '"error terms" ... measure the net effects of all other variables (not introduced explicitly) upon the system' (1954, p.40). Finally Nancy Cartwright (1989, p.29) states that the error terms are 'supposed to represent the unknown or unobservable factors that may have an effect'. In short, it is conventional to take error terms to represent the impact of omitted causal inputs from the set of equations.

In his (1954) Herbert Simon analyses sets of linear equations with error terms. Unfortunately, the brief quote above on error terms is as explicit as Simon gets in explaining how to apply his causal ordering method to sets of equations with errors. Nevertheless, his actions speak louder than words since in the paper he reads these sets of equations by implicitly treating the error terms as if they are coefficients.[36] This provides a way to apply Simon's formal method to systems of equations with error terms: treat them as coefficients. It also suggests the following definition for incomplete sets of equations with error terms.[37]

> An *incomplete set of equations with errors* is a linearly independent set of equations that is linear in the variables and the (non-zero) coefficients and in which variables are partitioned into two sets: external and internal. Each equation contains one error term. The equations are such that internal variables can be solved for in terms of the coefficients, external variables and error terms. The set of coefficients, external variables and error terms together are variation free. [38]

[36] See, for example, Simon's discussion of system II (1954, p.40).
[37] I work from incomplete sets not complete sets because an incomplete set can also be a complete set if it has no external variables, in this way the treatment is more general.
[38] The variation free requirement on the external variables implies there can be no equation in an incomplete set that contains only external variables.

This definition follows by intuitively extending the definition of an incomplete set and complete set given earlier. It is defined so that Simon's formal ordering method can be applied to such systems by treating error terms as coefficients. Likewise, a model reading can be applied to these sets of equations by treating the error terms 'as if' they denote directly controllable factors.

With this definition for an incomplete set with error terms, one could perform an investigation analogous to that carried out earlier in setting out the relationship between complete and incomplete sets using causal consistency. This would investigate the relationship between incomplete sets of equations with error terms and those without. It would proceed by extending the definition of causal consistency and then trying to find necessary and sufficient conditions for an incomplete set of equations with error terms to be causally consistent to an incomplete set without error terms. In a similar way that intuitive conditions, i.e. (NC), must hold for indirectly controllable factors in a complete model to be treated as directly controllable in an incomplete model, such an analysis should give conditions under which factors in an incomplete model can be 'omitted', that is, denoted as part of an error term. This would give an explicit interpretation of what the error term in an incomplete set with errors can denote, from the perspective of an incomplete model where no factors are omitted, without jeopardising the causal content of that incomplete model.

However, instead of carrying out this involved analysis, I leave it as further work and take a small shortcut. Here I propose an interpretation of error terms and show that it meets intuitive causal consistency requirements. In line with conventional readings, I propose the following definition of an error term, relative to an incomplete set without error terms.

> Given an incomplete set of equations without error terms, I, define an *error term* for the j^{th} equation, as any non-zero sum of terms in that equation that do not contain any internal variable in I. Define the j^{th} equation with error term to be j^{th} equation in I in which the terms in the sum are omitted and replaced by the error term (which is

their sum). Denote the set of all such equations, where each equation is given an error term, I_E.

This gives a formal way to construct an incomplete set with error terms from an incomplete set without errors. One simply chooses a set of terms from each equation and provided none contains any internal variables, replaces these by an error term that is by definition the sum of these omitted terms. This ensures the resulting equation is mathematically consistent with the original equation. The result is a mathematically consistent set of equations in the internal variables with error terms, I_E, which are solvable in terms of non-omitted coefficients, external variables and error terms.[39] In addition, since no internal variables are omitted, each equation with error term contains all the internal variable terms of the original equation from which it was derived. This implies that applying Simon's formal order to the equations with errors gives the same formal order relations among the internal variables as the formal order for the original incomplete set of equations without error terms.

The mathematical consistency and matching formal order relations between the variables in I and those in I_E suggests that I_E is causally consistent with I. But, as with the earlier insufficiency of (NC) for causal consistency, a variation free condition must also be met. Here it is required that the constructed error terms be variation free relative to any external variables and coefficients that explicitly appear in the incomplete set of equations with error terms. Provided this is also met, then a causally consistent set of incomplete set of equations with error terms, constructed as above, is causally consistent with the incomplete set without error terms from which it was constructed.

All of this can be clarified by way of an example. Consider the following incomplete set without error terms.

[39] Since the error terms are taken as given and no equation or internal variable has been removed, it straightforwardly follows that one can solve for the internal variables in these equations with errors.

$$y_1 = \alpha_{10} + \alpha_{12}y_2 + \beta_{11}x_1 + \beta_{12}x_2$$
$$y_2 = \alpha_{20} + \alpha_{21}y_1 + \beta_{21}x_1 + \beta_{23}x_3$$
$$y_3 = \alpha_{30} + \alpha_{32}y_2 + \beta_{34}x_4$$

(x's external, y's internal, α and β's coefficients)

This system has causal order among the internal variables: $\{y_1, y_2\} \rightarrow \{y_3\}$. Now suppose that some coefficients and external variables are omitted using the method set out above, that is, one picks some terms in equations that do not contain internal variables, drops these and adds an error term to each equation where terms are omitted.

Here suppose one omits the α_{10}, x_1 and x_2 terms from the second equation, the x_3 term from the second equation, and the x_4 term from the third equation, adding an error term, a u, to each equation where one or more terms are omitted. This gives.

$$y_1 = \alpha_{12}y_2 + u_1 \qquad u_1 = \alpha_{10} + \beta_{11}x_1 + \beta_{12}x_2$$
$$y_2 = \alpha_{20} + \alpha_{21}y_1 + \beta_{21}x_1 + u_2 \quad \textit{where} \quad u_2 = \beta_{23}x_3$$
$$y_3 = \alpha_{30} + \alpha_{32}y_2 + u_3 \qquad u_3 = \beta_{34}x_4$$

(x's external, y's internal, α and β's coefficients, u's error terms)

It easy to check that the u's are variation free in relation to each other, the external variable, x_1, and the coefficients. Given this, the derived set of equations is an incomplete set with errors to which the formal ordering methods and model reading can be applied. Moreover, since no internal variables are dropped, Simon's formal order is unchanged, the order among the internal variables is still $\{y_1, y_2\} \rightarrow \{y_3\}$. In this way, the resulting system is causally consistent with the original since its formal relations among its internal variables are the same as the original.[40]

In addition, this way of introducing error terms fits well with the conventional reading of error terms. In the approach set out here, the error term denotes the joint role of some causal inputs (directly controllable factors and/or external factors) in a

[40] Another important part of causal consistency is that no equations are linearly combined in deriving the set of equations with errors, therefore the derived equations can be assumed to denote the same mechanisms as those of the original incomplete set of equations without error terms.

mechanism. In this way, the error terms denote the net 'effect' of omitted exogenous factors ('causes') in a mechanism.

It is important to note that this way of introducing error terms is rather permissive. For instance, it allows external variables and coefficients to be omitted from some equations but not others, like x_1 in the example above. This case is epistemically troubling because if one wanted to intervene in the system using x_1 to find out about the causal order, then x_1 would impact the system in ways that are both explicit and implicit (since it also acts in an error term). This is analogous to the situation where Mill's method of concomitant variation fails because one unknowingly activates a common cause, and like in that case, methods of causal inference can fail to give reliable information.

This possibility might lead some to conclude that the method here for introducing error terms is too permissive and that external variables should either be omitted from all equations or from none when defining error terms. I think this would be a mistake since it rules out using the incomplete sets with errors to describe situations where causal inference would go wrong because an error term hides a factor included elsewhere in the model. The reason for being more general in the conceptual analysis here is to allow the representation and analysis of both epistemically convenient and inconvenient systems.[41]

5. Constant Coefficients and Adding Stochasticity

In this final section, I present a way to deal with the two outstanding differences of the models actually used by econometricians from those interpreted in the last chapter. The first difference is that of coefficients being constant. This is

[41] It might be retorted that keeping the variation free assumption throughout the analysis, as I have done, also restricts the analysis unnecessarily, since it rules out talking about systems where factors are not variation free. I think this is a fair point. For this reason, it would be worthwhile to generalise the analysis of this chapter and the last, without assuming the variation free condition. I leave this as further work however.

straightforward to bring in. The second difference is the stochastic variables and error terms in the sets of equations.

5.1. Constant Coefficients

There is nothing in Simon's treatment of formal order or in the version of it set out here that requires that every coefficient must vary. Simon's formal order and the model reading tell us what factors change given changes in one or more directly controllable factors. If some directly controllable factors do not change, then this is also covered in the analysis by Simon and in the strong reading set out here. This implies that imposing that coefficients in a model are constant is not problematic. Nevertheless, to distinguish the coefficients that can vary from those that are constant, I call a 'constant' any coefficient that does not vary. In the model reading these denote factors that are fixed relative to changes in all other factors.

5.2. Introducing Stochasticity

In order to introduce stochasticity, I begin with complete sets of equations. The extension to the other kinds of sets of equations (incomplete sets and incomplete sets with error terms) then follows by assuming that these are causally consistent with a stochastic complete set of equations.

To begin, consider the following rather unwieldy complete set of equations.[42]

$$\text{(A)}\quad \begin{array}{ll} x_1 = \delta_1 & x_4 = \gamma_2 \\ x_2 = \delta_2 & y_1 = \alpha_1 x_1 + \sigma_1 x_3 + \mu_1 \\ x_3 = \gamma_1 & y_2 = \beta_1 y_1 + \alpha_2 x_2 + \sigma_2 x_4 + \mu_2 \end{array}$$

(Greek letters coefficients, Latin letters variables)

The way to introduce stochasticity is simple, one attributes a joint distribution to a subset of coefficients appearing in the sets of equations. So, for example, assume here that the two γ's are independently normally distributed with mean zero and variance one.

[42] The reason for the unwieldy choice is that it is a complete set that is causally consistent with an incomplete set with error terms which has a very simple and recognizable form, as will be clear later in the discussion.

$$\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$$

In the model reading this implies that the directly controllable factors they denote are directly controlled by nature to take values according to distribution. As for the other coefficients, assume that the α's and β's denote constant factors, assume that σ and μ are non-random coefficients that denote factors directly controllable by nature, while the δ's denote non-random factors that are directly controllable by an experimenter.

If one treats the x's as external and y's as internal to construct an incomplete set,[43] one then gets an incomplete set with stochastic variables.

$$\text{(B)}\quad \begin{aligned} y_1 &= \alpha_1 x_1 + \sigma_1 x_3 + \mu_1 \\ y_2 &= \beta_1 y_1 + \alpha_2 x_2 + \sigma_2 x_4 + \mu_2 \end{aligned} \quad \textit{where} \begin{pmatrix}x_3\\ x_4\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$$

(y's internal, x's external)

The distribution of variables x_3 and x_4 follows from the equations relating to them to the y's in the complete set. This gives an incomplete set of equations with two external random variables, x_3 and x_4 that denote stochastic factors that are indirectly controlled by nature. While x_1 and x_2 are deterministic factors controlled by the experimenter.[44]

Finally, if one assumes that x_3, x_4 , σ and μ's are unobserved and omitted from the equations, one can define error terms for these omitted factors by.

$$u_1 = \sigma_1 x_3 + \mu_1$$
$$u_2 = \sigma_2 x_4 + \mu_2$$

Then, substituting in the error terms, one gets a (causally consistent) incomplete set with stochastic error terms.

$$\text{(C)}\quad \begin{aligned} y_1 &= \alpha_1 x_1 + u_1 \\ y_2 &= \beta_1 y_1 + \alpha_2 x_2 + u_2 \end{aligned} \quad \textit{where} \begin{pmatrix}u_1\\ u_2\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right)$$

(y's internal, x's external and u's error terms)

---

[43] Note that this respects the conditions for causal consistency with (A).

[44] There is no assumption here that the stochastic factors must be controlled by nature, an experimenter may be able to control a stochastic factor by controlling features that determine its distribution. For example, by cutting down the number of cigarettes one smokes, one can lower ones chance of having a heart attack.

At this stage, the original unwieldy complete set has been transformed into an innocuous looking incomplete set of equations with error terms, just like the simplest models used by econometricians. All of the coefficients explicit in the equations are constant and the error terms are stochastic. Under the model reading, the external variables denote variation free exogenous factors for the mechanisms in which they appear,[45] while the error terms denote the joint impact of other external factors and directly controllable factors, some of which take values stochastically.

This example presents the method for introducing stochastic coefficients, variables and error terms to the sets of equations. But just how is the attribution of a joint distribution to coefficients to be understood under the model reading? I read the attribution of a distribution to one or more directly controllable factors in the complete model as the experimenter and/or nature using a randomizing device, or 'rolling a dice', to determine the values of those directly controllable factors. Crucially, if the model relations are still to hold, then the attributed random variation of directly controllable factors should not violate the important assumptions discussed in the last chapter: the invariance of mechanisms to factor changes and the independence of directly controllable factors. To preserve these, I assume that the fact that the values taken by factors are random does not undermine the invariance of mechanisms. Moreover, I assume that in the attributed distribution no coefficients are perfectly correlated, since this would violate the requirement that coefficients be variation free.

Of course, the analogy of the experimenter/nature rolling a dice in the model reading is not very enlightening for explaining stochasticity. However, it could be unpacked according to different views on indeterminism. For example, it could be consistent with a Laplacean view that assumes that the source of the randomness is epistemic uncertainty about an ultimately deterministic system. Conversely, the randomness could denote actual indeterminism of modelled systems. Likewise, a Bayesian interpretation viewing the distribution as rational expectations about uncertain events

[45] I am using (NC) here.

is also feasible. In short, the approach taken here sidesteps the difficult issues in the interpretation of probability by simply attributing a joint distribution to coefficients in a complete set of equations.[46] Given the constraints on invariance and independence mentioned above, it leaves open how this attributed randomness is to be interpreted.

5.3. Aside: The Extended Model Reading, Weak and Super Exogeneity

To finish the chapter, it is interesting to note that the extension of the strong reading to simple stochastic linear models can be used to make an explicit connection to important concepts of exogeneity proposed by Engle, Hendry and Richard (1983).[47] This shows that the extension of the strong reading developed in this chapter is relevant for, and may ultimately be useful for interpreting important concepts in the econometric literature.

To see this, assume system (C) with its model reading holds, that is

$$\text{(C)}\quad \begin{aligned} y_1 &= \alpha_1 x_2 + u_1 \\ y_2 &= \beta_1 y_1 + \alpha_2 x_2 + u_2 \end{aligned} \qquad \textit{where} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$$

(y's internal, x's external and u's error terms)

In this case the marginal distribution for y_1 is given by $D(y_1; x_1, \alpha_1, \mu_1, \sigma_1) = N(\alpha_1 x_1 + \mu_1, \sigma_1^2)$. While the conditional distribution for y_2 on y_1 is given by $D(y_2 | y_1; x_2, \beta_1, \alpha_2, \mu_2, \sigma_2) = N(\beta_1 y_1 + \alpha_2 x_2 + \mu_2, \sigma_2^2)$.

For Engle *et al.* if the coefficients (and here external variable) in the marginal density function are variation free with respect to the coefficients (and here external variable) in the conditional density function, then y_1 is weakly exogenous for y_2 (Engle *et al.*, 1983, p.337).[48] In system (C) y_1 is weakly exogenous for y_2. This is because (C) is assumed to be causally consistent with sets (A) above, so the coefficients in the

[46] Obviously, if one were to tie the analysis of this chapter and the last to a particular metaphysics of causation then this would put constraints on which interpretations of randomness were appropriate.

[47] For a relatively clear exposition of the relevant exogeneity concepts discussed here, see David Hendry (1995, pp.162-164, pp.172-177).

[48] This is a useful condition for estimation purposes since it means one need not estimate the marginal distribution of y_1 in order to estimate the conditional distribution of y_2 on y_1.

respective conditional and marginal distributions, that are coefficients in (A), are variation free.

The second more important connection with Engle *et al.*'s analysis is with super exogeneity. As defined by Engle *et al.*, y_1 is super exogenous for y_2 if it is weakly exogenous for y_2 and intervening to change any of the coefficients (or external variable) in the marginal distribution for y_1 has no impact on the form and the values of the coefficients on the conditional distribution (*ibid.*, p.339). Like weak exogeneity, super exogeneity of y_1 for y_2 also holds in (C). This is because the coefficients in the conditional and marginal distributions denote directly controllable factors in mechanisms in the complete set (A) that underlies (C). Therefore, by the invariance of mechanisms denoted by (A), changes in one directly controllable factor, denoted by a change in just coefficient in the marginal distribution of y_1 do not lead to a change in any directly controllable factor denoted by a coefficient the conditional distribution of y_2 on y_1.

The fact that y_1 is weakly exogenous and super exogenous for y_2 in system (C) and that explanations can be given as to why these hold, suggests that the extended model reading of this chapter may provide a way to interpret Engle, Hendry and Richard's definitions of weak and super exogeneity. This is because in system (C) the relevant features that imply weak and strong exogeneity, the variation freedom of coefficients and external variables and the invariance property of mechanisms, derive from of properties of systems that are assumed in the system (A) that underlies system (C). This suggests that it may be possible to develop a general analysis for when incomplete sets of equations with error terms, that are assumed to be abbreviations of complete set of equations, have internal variables that are weakly exogenous and super exogenous. This could provide a fruitful way of making more explicit conditions under which an internal variable in a set of equations is weakly and super exogenous. More generally, it shows that the extended reading, set out in this chapter, not only allows simple econometric models to be interpreted, but it also

shows promise that it can be used to analyse relevant concepts, such as weak and super exogeneity, that are important in econometrics.

6. Conclusion

This chapter began with simple sets of equations interpreted as in the last chapter and has set out step-by-step how to extend the causal interpretation of the last chapter to sets of equations that differ from those of the last chapter in four ways. Those sets of equations, like the simplest actually used in econometrics, contain external and internal variables, have error terms in the equations, have constant coefficients, stochastic error terms and (sometimes) stochastic external variables.

To interpret sets of equations with external and internal variables, a distinction was introduced between incomplete and complete sets of equations. Complete sets of equations are the sets of equations interpreted in the last chapter, while incomplete sets containing internal and external variables represent just some of the causal relations modelled by a complete set of equations. With this introduction of incomplete sets of equations, the chapter presented a short exploration of how this could be used to analyse interventions. In particular, by considering different possible strongly causally consistent complete sets of equations that could underlie an incomplete set, it was possible to describe using a simple example, the diverse ways in which a simple intervention, changing just one external factor in an incomplete model, could be brought about.

In the second part of the chapter, an interpretation for incomplete sets of equations with error terms was introduced. It presented a way by which error terms could be introduced into equations without jeopardising the causal interpretation for the set of equations. The resulting interpretation for error terms was that these denote the joint impact of omitted, external factors from a mechanism. This reading fit well with conventional views of error terms in structural equations.

The chapter finished by introducing constant coefficients and stochasticity. Constant coefficients were introduced simply by allowing coefficients in a system of equations not to vary. While stochasticity was introduced by attributing a joint probability distribution to a subset of coefficients in the complete set of equations. By assuming that an incomplete set, and an incomplete set of equations with errors were causally consistent with a complete set with some stochastic coefficients, stochasticity was introduced to error terms and variables in these systems of equations.

To conclude, the chapter has extended the causal reading to the very simplest types of sets of equations that econometricians actually use in structural modelling. This is clearly an important step for any attempt to analysis causality in econometrics. In addition, the chapter has attempted to do this in a rigorous way.[49] The key assumption for doing this was to assume that underlying all the sets of equations analysed here was a complete set of equations, read using the strong reading of the last chapter. Though this is highly restrictive, some assumption of this sort is required if one is to make explicit the interpretation of structural equation models used in econometrics. It is an important further question to consider how such assumptions restrict the applicability of this formalisation of causal relations.[50]

[49] Ultimately, a full formalisation of the concepts introduced here would need to be provided.
[50] Recall that I do not claim that the formalisation of causal relations presented here can be used to generally represent causal systems. Instead, the aim of the last two chapter has been to set out an explicit causal content for the simplest models used by econometricians.

Appendix 3.1 - A Necessary Condition for An Incomplete Set to be Causally Consistent with a Complete Set

An incomplete set is *causally consistent* with a complete set if and only if

(a) Each of its equations is an equation in the complete set.

(b) All of the formal order relations, obtained using Simon's formal ordering methods, between equations in the incomplete set, and between its internal variables, also hold for those equations and variables in the formal order of the complete set.

(c) Formal order relations in the complete set that hold among internal variables and equations that appear in the incomplete set also hold in the formal order of the incomplete set.

(d) The external variables and coefficients in the incomplete set of equations are variation free in the complete set.

Theorem 3.1: An incomplete set of equations is causally consistent with a complete set of equations only if it meets (NC).

(NC) (I) The incomplete set is a union of complete subsets of equations in the formal order of the complete set.

(II) Its set of internal variables is the union of those variables which are endogenous for those complete subsets.

(III) For any two internal variables y and z such that y causes z in the formal ordering of the incomplete set, then in the formal order of the complete set either y is a direct cause of z or there exists a chain of direct causes such that $y \rightarrow w_1 \rightarrow \ldots \rightarrow w_j \rightarrow z$ where for all j, w_j is an internal variable.

Proof

Let C denote the complete set of equations, I denote the incomplete set of equations. By (a) $I \subseteq C$. (**NB** - recall that the complete subsets of equations are the subsets of equations obtained by applying Simon's formal ordering method *not* to be confused

with complete set of equations, which is one of the systems of equation to which the formal order is applied.)

First note that by (b) any complete subset of equations for I must be a complete subset of equations in C, since (b) requires that all the properties of formal ordering of I also hold for C. Given this, since the complete subsets of I partition the equations of I, it follows that I is a union of some complete subsets of equations for C. This gives the part (I) of (NC).

Now in determining the formal ordering for I, external variables are treated as coefficients. Therefore, in calculating the formal order each internal variable in I will be solved for (by solvability of equations in I) and only these will be solved for. Therefore, every internal variable is endogenous relative to some complete subset of equations in I. By (b) then it follows that each internal variable is endogenous for the same complete subset of equations in C. Conversely, by (c) any variable which is endogenous for a complete subset of equations in C of which I is the union must be a variable that is endogenous for that complete subset of equations in I. Therefore, any such variable must be an internal variable in I. This shows that the union of endogenous variables for the complete subsets of equations of C of which I is a union is equal to the set of internal variables. This gives the part (II) of (NC).

Consider any two internal variables in I, y and z, such that y causes z in the ordering for I. Since y causes z in the formal ordering for I, then either y is a direct cause for z or there must be some chain of direct causes $y \rightarrow v_1 \rightarrow \ldots \rightarrow v_j \rightarrow z$. If y is a direct cause of z in the ordering for I, then by (b) the same must hold in the ordering for C. If y is not a direct cause of z in the ordering for I, then there must be some chain of direct causes $y_1 \rightarrow v_1 \rightarrow \ldots \rightarrow v_j \rightarrow y_2$ in the ordering for I by which y_1 causes y_2, and since only internal variables are endogenous in the formal ordering for I, it follows that all v's are internal variables. But by (b) this must hold in the ordering for C, so there is a chain of direct causes $y_1 \rightarrow v_1 \rightarrow \ldots \rightarrow v_j \rightarrow y_2$ in the ordering for C by which y_1 causes y_2, in which all v's are internal. So (III) holds and (NC) follows.□

Chapter 4

Alternative Views on Causality based on Simon: Stephen LeRoy and Kevin Hoover

1. Introduction

This chapter presents and critically analyses Stephen LeRoy and Kevin Hoover's respective positions on causal order. Both of these positions are developed from Herbert Simon's (1953) paper, which was the basis for the strong reading of equations in chapter two. Here the aim is to contrast LeRoy and Hoover's views on causal order with that of the strong reading. [1]

The chapter begins with an overview of Stephen LeRoy's treatment of causal order as applied to linear systems of equations. LeRoy's position is highly influenced by, though significantly stronger than Herbert Simon's. His definition of causal order is based on two conditions, the subset condition and the sufficiency condition. The subset condition is particularly interesting because LeRoy uses it to provide an alternative way[2] to characterise Simon's causal order. This part of the chapter sets out LeRoy's definition of causal order, his characterisation of Simon's causal order and makes some relevant criticisms.

The second part of the chapter looks at the treatment of causality by Kevin Hoover who, like LeRoy, is strongly influenced by Simon. It fleshes out Hoover's view and shows ultimately that it is very similar to LeRoy's characterisation of Simon. The chapter concludes that both Hoover and LeRoy's positions build in unnecessary conditions in their definitions of causal order which prevent these from being applied

[1] Ideally, the work of this chapter should be extended to include other important works that have been influenced by Simon. In particular, work by Judea Pearl (2000, chap 7) is important and bears similarities with the strong reading proposed in chapter two. Nevertheless, in this chapter I focus on LeRoy and Hoover because they use simultaneous equation systems and are focused on causal models in economics. Pearl's (2000) does not apply to simultaneous systems nor is economics-specific. For this reason, and also the keep the discussion manageable, I do not analyse Pearl here.

[2] It provides an alternative way (from Simon's) for defining Simon's formal order over variables.

to intuitively causal situations. This is in contrast to the strong reading developed in chapter two.

2. Stephen LeRoy's Treatment of Causal Order

Stephen LeRoy provides his most explicit and general discussion of causality in his paper 'Causal Orderings' (1995). In this paper, he sets out what it means for one variable to cause another in a general non-linear system of functional equations. Since I am concerned here only with linear systems of equations, I focus mainly on LeRoy (2004) where the general approach of (1995) is made specific to the linear sets of equations like those discussed in chapter three. In LeRoy (2004) the focus is on deterministic sets of equations that are linear in the coefficients and variables, and where coefficients are constant. The variables are partitioned into external and internal, the equations are linearly independent and solvable for the internal variables in terms of the external variables and the non-zero coefficients.[3]

2.1. LeRoy Causality – Simple and Conditional Causes

For LeRoy, a variable in a linear set of equations is either structural or non-structural.[4] The structural variables are those which a modeller specifies as either external or internal, while the non-structural are those left unspecified. In his (2004) all variables in the set of equations are assumed to be structural. A structural variable is external if it is 'determined outside the model and subject to direct and independent intervention', internal if it is 'determined by the model and therefore not subject to direct intervention'(1995, p.212). This shows that LeRoy's concepts of 'direct', 'intervention' and 'independent', important primitives in his analysis, fit closely with

[3] The sets of equations he analyses are examples of what I called 'incomplete sets of equations without errors' in the last chapter, but with constant coefficients.

[4] In his earlier paper LeRoy (1995) defines causal order that applies to variables and parameters (i.e. coefficients) for a set of non-linear functions. There LeRoy allows parameters to satisfy the same relations as a those for a variable. So parameters, like variables, can be structural or non-structural and for structural parameters, internal or external. His reason for treating parameters and variables in the same way is that LeRoy does not think (like Kevin Hoover does) that the distinction between parameters and variables is fundamental. In addition, LeRoy claims that his approach allows one to model the distinction in economics between 'shallow' and 'deep' parameters, by treating the former as an internal parameter and the latter as an external parameter. This is important for LeRoy's discussion of the Lucas Critique in his (1995) and (1999).

Simon's corresponding concepts. For example, like Simon's treatment of coefficients, for LeRoy an external variable is directly controllable, its value can be directly changed independently of the equations of the model. Similarly, like Simon's treatment of variables, an internal variable is indirectly controllable, it can be controlled using external variables (*cf.* coefficients for Simon) to take values in line with the equations in the model.

With this background, LeRoy defines two types of causal relation that can occur between external variables in a linear set of equations: *simple* and *conditional*.[5] An important concept for the definitions is that each internal variable, y, is assumed to have an *external set*, $e(y)$, which is the smallest set of external variables which determines that variable. To determine the external set for an internal variable in a set of structural equations, one simply solves for the internal variable using the equations. The external variables that appear in the solution (the reduced form function) for the internal variable are those in its external set.

LeRoy then defines simple causation as follows:

> y_1 is a *simple cause* of y_2 denoted, $y_1 => y_2$, if and only if:[6]
>
> (1) *The subset condition*: $e(y_1)$ is a proper subset of $e(y_2)$.
>
> (2) *The sufficiency condition:* there is a non-zero constant β such that
>
> $y_2 = \beta y_1 + \sum_i \alpha_j z_{2\text{-}1,j}$ where $z_{2\text{-}1}$ is the vector of elements in
>
> $e(y_2) - e(y_1)$

In addition, LeRoy defines conditional causation

> y_1 is a *conditional cause* of y_2 given the set of internal or external variables $Z=\{z_i\}$, denoted $y_k=>y_j|\{z_i\}$, if and only if

---

[5] LeRoy (1995) also defines joint causation as an extension of simple causation that applies to sets of variables. I do not discuss it here because it is absent from LeRoy (2004) and because situations of joint causation can be modelled using conditional causation which is discussed here.

[6] LeRoy (1995, p.214) also includes the 'non-constancy condition' in his definition of simple causation. This requires that $f\big|_{z_{2\text{-}1}=const}(y_1)$ is not constant, so that no matter what value $z_{2\text{-}1}$ takes the resulting function still varies with y_1. LeRoy doesn't mention the non-constancy condition in his 2004 paper, but given that he restricts his analysis there to linear functions with non-zero coefficients, this is unsurprising since the condition is trivially met.

(1) The *subset condition* is met: $e(y_1)$ is a proper subset of $e(y_2)$.

(2) The *sufficiency condition* is met: given the variables in Z are constant, there is some non-zero constant β such that

$$y_2\big|_{\{z_k=constant\}} = \beta y_1\big|_{\{z_k=constant\}} + \sum_i \alpha_j z_{2-1,j}\big|_{\{z_k=constant\}}$$

where z_{2-1} is the vector of elements in $e(y_2) - e(y_1)$

As is obvious from the definitions, conditional causation is simple causation where one or more internal or external variables are fixed. In order to see beyond these formal definitions, it is helpful to flesh out the causal relations using some simple examples.

First, consider the following abstract system of linear equations.

$$y_1 = \alpha_1 x_1$$
$$y_2 = \alpha_2 x_2 + \beta_2 y_1 + \beta_3 y_3$$
$$y_3 = \alpha_3 x_3$$

(x's external, y's internal, α and β's constant coefficients)

The first step in determining what causal relations hold among the internal variables is to find the external sets for each internal variable. If one solves for the y's in terms of the x's one gets the following reduced form equations[7] for the y's:

$$y_1 = \alpha_1 x_1$$
$$y_2 = \alpha_2 x_2 + \beta_2 \alpha_1 x_1 + \beta_3 \alpha_3 x_3$$
$$y_3 = \alpha_3 x_3$$

From these, the external sets are: $e(y_1) = \{x_1\}$, $e(y_2) = \{x_1, x_2, x_3\}$ and $e(y_3) = \{x_3\}$. Since $e(y_1)$ and $e(y_3)$ are both properly contained in $e(y_2)$, y_2 meets the subset condition relative to both y_1 and y_3.

In addition, the following two equations can be derived for y_2.

$$y_2 = \alpha_2 x_2 + \beta_2 y_1 + \beta_3 \alpha_3 x_3$$
$$y_2 = \alpha_2 x_2 + \beta_2 \alpha_1 x_1 + \beta_3 y_3$$

[7] Structural equations are those to which causal order is attributed, while the reduced form is the set of equations obtained by solving the structural equations for the internal variables in terms of the external variables.

The first equation gives y_2 as a linear function of y_1 and the external variables that are in $e(y_2)$ but not $e(y_1)$, so it implies that y_2 meets the sufficiency condition for y_1. The second equation gives y_2 as a linear function of y_3 and external variables that are in $e(y_2)$ but not $e(y_3)$, so y_2 meets the sufficiency condition for y_3. Since y_2 meets the sufficiency and the subset condition relative to y_1 and y_3, it follows that $y_1 => y_2$ and $y_3 => y_2$, that is, y_3 and y_1 are both simple causes of y_2. Finally, note that if x_3 is constant then the sufficiency condition still holds for y_2 and y_1, so y_1 conditionally causes y_2 given x_3. By an analogous analysis, it can be shown that y_3 conditionally causes y_2 given x_1.

This example shows how to calculate LeRoy's causal relations for a system of equations, but what is the intuitive content of those causal relations? To see this, it helps to consider direct changes in the external variables in the example. Given the equations, if x_1 is directly changed alone then both y_1 and y_2 indirectly change. If x_2 is directly changed alone then only y_2 changes. Because the external set of y_2 contains the external set of y_1, changes to y_1 lead to changes in y_2. However, since the external sets of y_1 is *properly* contained in that of y_2, one can change y_2 without changing y_1. Both of these features follow from the subset condition: the external set for y_2 properly contains the external set of y_1.

This last point shows that the subset condition ensures that if an internal variable is a simple cause of another internal variable, then a change in the first variable changes the latter, but it is possible to change the latter internal variable without changing the first. In short, the subset condition is a way of ensuring the following holds for simple (or conditional) causes and their effects: *a change in just one cause is accompanied by changes in each of its effects, but each of its effects can also be changed without changing that cause* (since each of its effects has other causes). Behind this is the standard manipulability intuition for explaining causal asymmetry: causes can be used to bring about their effects but not *vice versa*.

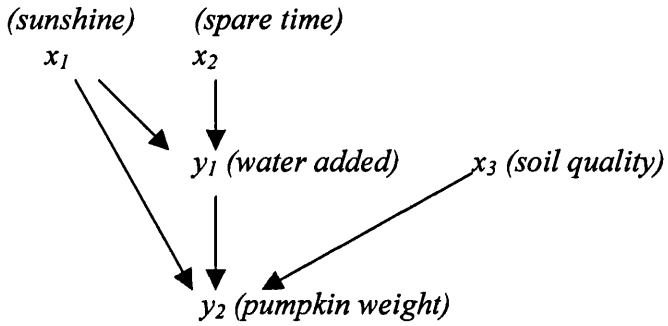
LeRoy's simple causal relation also assumes the sufficiency condition. If y_1 is a simple cause of y_2, the sufficiency condition requires that information about the value of y_1 be sufficient, given the values about external variables on which y_2 depends but not y_1, for determining y_2. This can be seen in the abstract example above where the value of y_1, with that of the two external variables, x_2 and x_3, determines the value of y_2. The intuition behind this condition can be appreciated if one again considers changes to variables. For instance, in the example above if x_2 and x_3 are fixed, but y_1 is directly changed, then the change in y_2 is fully determined by the change in y_1. This means no matter how the change in y_1 came about, it is *ceteris paribus* sufficient for determining the resulting change in y_2. LeRoy himself gives a neat characterisation of his sufficiency condition when he states 'causal statements involving internal variables as causes are *ambiguous except when all interventions consistent with a given change in a cause variable map onto the same change in the effect variable*.' (2004, p.9, original emphasis).

The quote also shows LeRoy's motivation for the sufficiency condition. While the subset condition ensures that changes in causes are accompanied by changes in their effects but not *vice versa*, the sufficiency condition strengthens this by requiring that how much a cause changes be *ceteris paribus* sufficient for how much the effect changes. In this way, an undesirable ambiguity in how much an effect changes is avoided. Putting the two conditions together, LeRoy wants a change in a cause to be *ceteris paribus* sufficient for a change in an effect (the subset condition) *and* he wants how much a cause changes to be (*ceteris paribus*) sufficient for how much an effect changes (the sufficiency condition). As the title of the relevant section in the paper puts it, this is '[c]ausality as sufficiency' (2004, p.9).[8]

[8] As LeRoy points out, systems that meet these criterion provide an unambiguous answer to how much an effect will change if one of it causes changes by a certain amount. This shows the operationalist flavour of LeRoy's position, he is tying his concept of causality to questions that can be unambiguously answered. Indeed, LeRoy introduces his concept of causality in response to a question he poses: 'What is the content of "operationally meaningful" in this context...?' (2004, p.9).

2.2. The Farmer Example and the Restrictiveness of the Sufficiency Condition

To clarify LeRoy's causal relations a little further and to show how restrictive his sufficiency condition is, I consider the following hypothetical example.[9] A farmer chooses to water his prize pumpkin plant according to how much sunshine the plant receives, though he can only water the plant provided he has enough spare time from his other jobs. Assume the pumpkin weight also depends on soil quality, which is assumed to be independent of sunshine levels. The pumpkin's weight is determined by the sunshine levels, soil quality and the water it receives from the farmer. These causal relations are represented in the following intuitive causal graph.



To model this using equations let the external variables be: the sunshine level (x_1), the spare time (x_2) and the soil quality (x_3).[10] Let the internal variables be the amount of water farmer adds to the pumpkin plant (y_1) and the pumpkin weight (y_2). Finally, suppose the following linear structural equations hold, where the coefficients represent the constant contributions along the causal arrows in the graph above.

$$y_1 = \alpha_1 x_1 + \alpha_2 x_2$$

$$y_2 = \alpha_3 x_1 + \alpha_4 x_3 + \alpha_5 y_1$$

(y's internal, x's external)

From these two equations, it is easily checked that the subset condition holds for y_1 and y_2, that is, all the external variables that influence the amount of water added to the pumpkin also influence the pumpkin weight, and it is possible to change pumpkin weight without changing water added (i.e. by changing soil quality, say by adding fertiliser).

---

[9] Kevin Hoover (2001a, p.173) gives a similar example in his discussion of LeRoy, also to show how restrictive the sufficiency condition is.

[10] These are external because they are all determined by unmodelled causal relations.

However, water added to the pumpkin is not a simple cause of pumpkin weight since the sufficiency condition fails.[11] The sufficiency condition requires that the amount of water added, with the soil quality (this is the only external variable in the external set of pumpkin weight but not in that of water added) be sufficient to determine the pumpkin weight. This fails in this example because, given different levels of spare time, the same amount of water may be added under diverse sunshine scenarios, and these different sunshine level scenarios each lead to different pumpkin weights. In other words, the same amount of water added may be added given fixed soil quality in different sunshine scenarios, and since sunshine also directly determines pumpkin weight, the pumpkin weight will vary in these different scenarios, even though the water added and soil quality do not. So the values of water added and the soil quality do not alone determine pumpkin weight, and the sufficiency condition fails. LeRoy would consider this case ambiguous and to be ruled out. So in LeRoy's definition of causal order, water added is not a simple cause of pumpkin weight.

Yet intuitively the amount of water added to the pumpkin plant *is* a contributing cause to the pumpkin weight. The fact that LeRoy's sufficiency condition rules this out shows that it is too restrictive a condition. More generally, the farmer example shows where a cause shares a common cause with its effect then the sufficiency condition rules out using LeRoy's simple cause relations to model the relationship between the cause and its effect.[12] This is a strong restriction on what causal relations can be modelled since it is intuitive to have a cause and effect which share a common cause. The sufficiency condition, though it provides an attractive property of

[11] Formally this follows from the fact that though $y_1 = f(x_1,x_2)$ and $y_2 = g(x_1,x_2,x_3)$, $y_2 \neq h(y_1,x_3)$ is not met for any linear function h as required by the sufficiency condition.

[12] There is an apparent exception to this claim, where the cause's only cause is the common cause with its effect. In that case, the value of the cause will be consistent with only one value of the common cause and thus sufficient to determine the effect. This suggests that in that case LeRoy's simple cause relation could be used to model the relation between cause and effect. However, this is incorrect because in the case where the cause only has one cause (the common cause with the effect) then the external set of the cause and its cause are identical, so the subset condition fails. So, this situation cannot be modelled by LeRoy's simple cause relation.

connecting a particular change in a cause to a particular change in the effect (i.e. it avoids ambiguity) does not apply to a wide range of causally intuitive systems.

The sufficiency condition makes simple cause a restrictive causal relation. However, LeRoy's conditional causal relation is weaker and more flexible.[13] By holding variables fixed, one can obtain conditional causal relations where simple causality fails. In general conditional causality can hold wherever the subset condition holds, by holding fixed sufficiently many external (or internal) variables that cause the conditional cause variable. For instance, in the farmer example holding fixed the level of spare time, one obtains conditional causality from water added to pumpkin weight. This holds because if spare time is fixed, then how much the farmer waters the pumpkin is consistent only with one sunshine level. In that case the farmer's watering amount and soil quality, since they are consistent with just one sunshine level, are sufficient to determine pumpkin weight.

Despite this, using conditional causation to model the intuitive causal relation between the farmer's watering and the pumpkin weight is still somewhat unsatisfactory. This is because intuitively the farmer watering is a straightforward cause of the pumpkin weight and not one which is conditional on spare time levels being constant.[14] So, despite the ability of LeRoy's conditional causal relation to formally model the farmer example, it is debatable whether it really does capture the intuitive causal connection from water added to pumpkin weight.

2.3. Summary

To summarise LeRoy's concept of causal order, the idea is that for two internal variables x and y in a system of equations: x is a (simple) cause of y if any thing which changes x also changes y but not *vice versa* (the subset condition) and if the

[13] Note that in spite of developing a concept of conditional causation, LeRoy is concerned (2004, p.11) that it is not consistent with his treatment of external variables. In particular, he worries that the introduction of fixed variables in conditional causal relation violates a requirement that external variables are suitably independent of each other. As a result, LeRoy prefers his strong concept of simple causality.

[14] Why, after all, should the causal relation from the water added to the pumpkin weight be contingent on how much spare time the farmer has?

value of x, along with those of other causes of y that do not cause x, carries sufficient information for determining the value of y (the sufficiency condition). However, as shown in this farmer example, the sufficiency condition in particular imposes a rather strong restriction on the concept of causal relation, since it rules out the possibility of common causes between a simple cause and its effect. This is restrictive and rules out certain causal systems from being modelled using LeRoy's simple causal relation. It shows the price to be paid in ruling out 'ambiguous' systems like the farmer example above.

3. LeRoy's Characterisation of Simon

In addition to presenting his own view of causality for linear systems of equations, Stephen LeRoy (2004) also presents an interpretation of Herbert Simon's (1953) work on causal order, using his subset condition. LeRoy sets out, much as described in the previous chapter, how to determine Simon's formal order among the internal variables by solving for the internal variables in terms of external variables using the smallest subsets of equations for which these can be solved.[15] In addition, LeRoy interprets Simon as solving the conceptual equivalence problem[16] using what he calls the 'exclusion condition'.

This section begins by exploring LeRoy's claim that Simon solves the conceptual equivalence problem using the exclusion condition. Here the aim is not to determine whether or not LeRoy correctly interprets Simon[17] but rather to show that there is a problem in using the exclusion condition to solve the conceptual equivalence problem. To show this, it presents a counterexample against the claim that the exclusion condition is sufficient for solving the conceptual equivalence problem, and also criticises LeRoy's response to the counterexample. Having done this, the section

[15] See chapters two and three. Recall that Simon also defines a formal order over the equations, LeRoy does not discuss this alternative.

[16] Recall the conceptual equivalence problem is that the causal content attributed to equations can be changed by mathematically acceptable transformations.

[17] The discussion of what Simon assumes in his treatment of causal order is discussed in the next chapter. The analysis there agrees with LeRoy that Simon makes an exclusion condition assumption, but differs slightly from LeRoy in the interpretation of that condition.

then sets out how LeRoy's subset condition and the exclusion condition, when suitably qualified taking the counterexample into account, provide a characterisation of Simon's causal order.

3.1. How Simon solves the Conceptual Equivalence Problem according to LeRoy

In his discussion of Simon's causal order for sets of equations, LeRoy notes the conceptual equivalence problem that 'innocuous mathematical operations alter causal orderings' (2004, p.5). Recall that in the strong reading adopted in chapter two, the conceptual equivalence problem was solved by assuming that each equation in the set denoted a separate mechanism. This ruled out all but the most trivial mathematical manipulations of equations (reorderings and rescalings) because transformations that linearly combined equations were taken to mix up the separate mechanisms denoted by the equations. In this way, formal-order-changing mathematical manipulations were ruled out and the conceptual equivalence problem avoided.[18]

LeRoy, on the other hand, outlines a different approach for solving the conceptual equivalence problem. He reads Simon as imposing an 'exclusion condition' on the sets of equations to which his formal order is applied.[19]

> (*LeRoy's Exclusion Condition*) 'each equation contain[s] at least one external variable not found in any other equation' (2004, p.5, original emphasis removed).

To see why this should solve the conceptual equivalence problem, consider two mathematically equivalent sets of equations with different formal orders,[20] where one meets the exclusion condition (system A) while the other (system B) does not (y's external, x's internal).

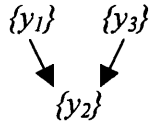
[18] Specifically, the manipulations were ruled out by introducing a new equality operator '$=_M$' that made explicit that an equation denoted a mechanism.

[19] As LeRoy notes, Simon (1953) is not very clear about exactly what is assumed in his definition of causal order. This is why I have called it 'LeRoy's exclusion condition'.

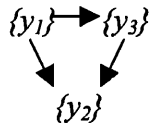
[20] The formal orders are obtained using Simon's method.

Set of Equations *Formal Order Among Internal Variables*

(A)
$$y_1 = \alpha_1 x_1$$
$$y_2 = \alpha_2 x_2 + \beta_2 y_1 + \beta_3 y_3$$
$$y_3 = \alpha_3 x_3$$



(B)
$$y_1 = \alpha_1 x_1 - \delta y_3 + \delta \alpha_3 x_3$$
$$y_2 = \alpha_2 x_2 + \beta_2 y_1 + \beta_3 y_3$$
$$y_3 = \alpha_3 x_3$$



To derive system B from system A one adds δ times the third equation to the first in A, to derive the first equation in B. System B is mathematically equivalent to A but has different formal order, so this is an example of the conceptual equivalence problem. Yet system A meets the exclusion condition since every one of its equations contains an external variable unique to it, while system B violates the exclusion condition since its third equation does not contain an external variable not contained in any other equation. Therefore, in constructing system B the exclusion condition has been violated. So imposing that the set of equations meet the exclusion condition rules out the problematic transformation from system A to system B and avoids the conceptual equivalence problem.

Intuitively one would expect the exclusion condition to be sufficient for solving the conceptual equivalence problem, because whenever one linearly combines two equations from a system meeting the exclusion condition and introduces the resulting equation in place of one of the original equations, then the other original equation contains only external variables that appear in the new equation so the resulting set of equations does not meet the exclusion condition. This is the intuition behind LeRoy's claim that '[t]he exclusion condition rules out algebraic operations that involve more than one equation (because if the original model satisfies the exclusion conditions, the modified model will not).' (2004, p.5). However, this intuition overlooks a problem, since a counterexample can be constructed to the claim that the exclusion condition is sufficient to solve the conceptual equivalence problem.

3.2. A Counterexample to LeRoy's Exclusion Condition

Here I show that the exclusion condition is not sufficient for solving the conceptual equivalence problem. To see this consider the following set of equations (y's internal, x's external).

$$\text{(C)}\begin{matrix} y_1 = \alpha x_1 + \beta x_2 ...(1) \\ y_2 = \gamma x_2 + \delta x_3 ...(2) \end{matrix} \qquad \textit{Formal Order: } \{y_1\}, \{y_2\} \text{ (i.e. unordered)}$$

In this set of equations, x_1 appears only in (1) while x_3 appears only in (2), so the exclusion condition is met for this system. Moreover, applying Simon's ordering method to it yields that y_1 and y_2 are not causally ordered relative to each other.

To construct the mathematically equivalent system with different causal order, first solve (2) for x_2.

$$x_2 = \frac{1}{\gamma}(y_2 - \delta x_3) ...(3)$$

Substituting (3) for x_2 in (1) and rearranging one gets (4) below. If one combines (4) with (2) one gets a new set of equations, D, which is mathematically equivalent to system C.

$$\text{(D)}\begin{matrix} y_1 = \alpha x_1 + \beta/\gamma (y_2 - \delta x_3) ...(4) \\ y_2 = \gamma x_2 + \delta x_3 ...(2) \end{matrix} \qquad \textit{Formal Order: } \{y_2\} \rightarrow \{y_1\}$$

Crucially, system D also meets the exclusion condition since x_1 only appears in (4) and x_2 only appears in (2). However, applying Simon's ordering method to it gives that y_2 causally precedes y_1. Therefore, these are two mathematically equivalent systems of equations both meeting the exclusion condition, but both yielding a different causal order when Simon's ordering approach is applied. Thus a set of linear equations meeting the exclusion condition is *not* sufficient to solve the conceptual equivalence problem.

More generally, counterexamples of this type can be constructed in any set of equations with more external variables than internal variables and in which there is an

external variable that appears in more than one equation.[21] But what does the counterexample imply for the exclusion condition? Can this solution to the conceptual equivalence problem be salvaged?

An answer to this can be found in observing that in system D if the compound coefficients are relabelled as simple coefficients then these satisfy a functional relationship amongst themselves. To see this, reconsider system D.

$$\text{(D)}\quad \begin{aligned} y_1 &= \alpha x_1 + \beta/\gamma\,(y_2 - \delta x_3)\ldots(4) \\ y_2 &= \gamma x_2 + \delta x_3 \ldots(2) \end{aligned}$$

Re-label the compound coefficients in equation (4) as follows to get system D'.

$$\text{(D')}\quad \begin{aligned} y_1 &= \alpha x_1 + \beta' y_2 + \delta' x_3 \ldots(4') \\ y_2 &= \gamma x_2 + \delta x_3 \ldots(2) \end{aligned} \qquad \textit{where} \qquad \beta' = \beta/\gamma \;\&\; \delta' = -\beta\delta/\gamma$$

In system D', a functional relationship holds between three of its coefficients namely $\delta' = -\beta'\delta$. So system D, though it satisfies the exclusion condition and is mathematically equivalent to C, has a functional dependency among its coefficients.[22]

In an earlier version of the paper[23] LeRoy was not aware of this counterexample, however he responds to it in his (2004) by introducing a caveat. Instead of concluding the exclusion condition is sufficient for unique causal order,[24] he concludes that it is sufficient for the causal order to be generically unique. In other

---

[21] This can be done by first deriving an expression for the multiply occurring external variable in terms of other variables by picking one of the equations in which it occurs and rearranging. Then one substitutes this expression for the multiply occurring external variable in all the places in which that variable appears, with the exception of the equation which was used to derive the expression for that variable. At this stage, one should be left with a different, yet mathematically equivalent system that still meets the exclusion condition, because the variable that was multiply occurring now occurs only in the substituting equation, while the external variable that occurred only once in that substituting equation now appears in all the equations in which the original multiply occurring variable appeared. So by 'swapping' one external variable that appeared in only one equation for another multiply occurring variable in this way, one constructs a visibly different system of equations that has different formal order from the original.

[22] One might be tempted to conclude that since the coefficients are variation free this problem is ruled out. However, this would be to forget that in the systems analysed here by LeRoy, the coefficients are assumed constant. It is meaningless to assume that the constants are variation free, since constants cannot vary.

[23] See LeRoy (2003), the counterexample was shown to LeRoy in correspondence.

[24] Note that the conceptual equivalence problem is equivalent to the problem of defining a unique causal order for sets of equations.

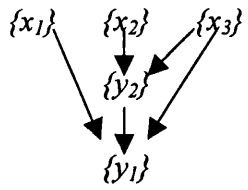
words, he introduces an assumption that for most systems, that is, those with coefficients whose values do not have functional dependencies among the coefficients (like those in D' above), the exclusion condition rules out problematic mathematical manipulations and solves the conceptual equivalence problem.

LeRoy then claims that such problematic cases (where these functional dependencies occur) have already been ruled out. He states '[s]ince we have already ruled out non-generic special cases (see note 2), it is seen that Fennell's observation about nonuniqueness of causal orderings ... does not involve anything new' (2004, p.5). This is surprising because the note 2 he refers to rules out non-zero values for coefficients, not the functional dependencies above. So, *pace* LeRoy, ruling out of coefficients that meet the functional dependencies, like those in system D', does involve adding something new. It is different from assuming coefficients have non-zero values and in fact amounts to ruling out causal relations that cancel out.

To see this, consider System D' once again, with its extended formal order.

$$\text{(D')}\quad \begin{aligned} y_1 &= \alpha x_1 + \beta' y_2 + \delta' x_3 \\ y_2 &= \gamma x_2 + \delta x_3 \end{aligned}$$

where $\beta' = \beta/\gamma$ & $\delta' = -\beta\delta/\gamma$



According to the formal order, y_1 is causally dependent on x_3. So following Simon's theorem 6.1 [25] (1953, p.25) one expects changes in x_3 to imply changes in y_1. However, given that system D' implies equation (1) of system (C) (recall D' and C are mathematically equivalent) y_1 satisfies $y_1 = \alpha x_1 + \beta x_2$. It follows from this equation that changes in x_3 are *not* accompanied by any change in y_1. In other words, this is an example of 'cancelling out'. System D, read causally, assumes that x_3 is a cause of y_1 that has no net impact on y_1.[26] This illustrates that ruling out the inconvenient functional dependencies among the coefficients, as LeRoy does in his

[25] See chapter two.
[26] This was discussed in chapter two in relation to Simon's 'in general' caveat in his theorem 6.1.

caveat, amounts to an assumption of ruling out certain systems that include causal relations in which the influence of a cause cancels out.[27, 28]

In conclusion, LeRoy needs to rule out these problematic functional dependencies among coefficients if he is going to solve the conceptual equivalence problem to equations using the exclusion condition.

3.3. Relating LeRoy's Causality and Simon's Formal Order

Returning to LeRoy's interpretation of Simon, there is a neat relationship that holds between LeRoy's treatment of causality and Simon's formal order for systems of equations that meet the exclusion condition. LeRoy claims (2004, p.11) that in a system that meets the exclusion condition an internal variable causally precedes another in Simon's formal order if and only if LeRoy's subset condition is met. Surprisingly, LeRoy does not prove this important claim. So first I present an outline of a proof.

3.3.1. Aside: Proof for LeRoy's Equivalence claim

A linear system of equations that meets the exclusion condition is such that whenever an internal variable, y, is solved for using Simon's formal ordering method, its solution will contain (*provided there is no cancelling out*) every external variable that appears in the equations for which it is endogenous or on which it is causally dependent. Since the exclusion condition ensures every equation has an exclusive external variable, an internal variable z that is causally dependent on y in Simon's formal order must depend on extra external variables than y (for instance those that appear in equations for which z is endogenous but y is not). Moreover, since y is used to solve for z, all of the external variables on which y depends, z must depend on also.

---

[27] This, as noted in chapter two, is similar to the faithfulness assumption made by Spirtes, Glymour and Scheines (1993).

[28] This needs a more careful formulation, because it may be possible to have a system that meets the exclusion condition which has some cancelling out causal relationships, but when the system is mathematically transformed into another system which does not have any cancelling-out relations, the resulting system does not meet the exclusion condition. So, a more precise explication of exactly which systems are ruled out by LeRoy's caveat is required. I leave this as further work.

These two features imply that the external set for y is properly contained in that z, that is, LeRoy's subset condition is met.

Conversely, assume the subset condition is met between two internal variables y and z, in a linear system of equations that meets the exclusion condition. Then the solution (reduced form) for y contains all the external variables in the solution for the other variable, z. Since all of the external variables that are exclusive to the equations used to solve for y appear in its reduced form (*provided there is no cancelling out*) they must also appear in the solution for z. This then implies that all of the equations used to solve for y are necessary for solving z, since there are no other equations that contain those variables. However, by the subset condition, z also depends on some external variables on which y does not depend. This implies that at least one equation was used to solve for z that was not required for solving for y. This implies that y must be causally precedent to z in the formal order. This completes the proof.

3.3.2. Characterising Simon's Causal Relation Using the Subset Condition

As expected from the counterexample discussed above, LeRoy's equivalence claim relies on the functional dependencies above being ruled out so that no cancelling out of external variables occurs when solving for internal variables. This can be seen in the emphasised statements in the proof. The no-cancelling-out condition is important because it ensures that the external set for an internal variable is equal to the set of external variables that appear in the equations necessary to solve for it. This underpins LeRoy's equivalence claim.

So correctly stated LeRoy's equivalence claim is that: *for solvable sets of linear equations in external and internal variables where the exclusion condition holds and there are no problematic functional dependencies among the coefficients, then an internal variable y is causally precedent to another z in Simon's formal order if and only if the external set for y is contained in that of z.* This shows how LeRoy's subset

condition given the exclusion condition and no functional dependencies among coefficients provides an alternative characterisation Simon's causal relation.[29]

Finally, note that this equivalence also shows that the difference between LeRoy's treatment of causal order and Simon's rests with the sufficiency condition. LeRoy adds the sufficiency condition to the subset condition to obtain a stronger concept of causal relation (simple cause) than Simon's causal relation. This can be illustrated by looking at the equations of the farmer example above.[30]

$y_1 = \alpha_1 x_1 + \alpha_2 x_2$

$y_2 = \alpha_3 x_1 + \alpha_4 x_3 + \alpha_5 y_1$

(y's internal, x's external)

From the earlier discussion, y_1 is not a simple cause of y_2. However, if one applies Simon's formal order one obtains the intuitive causal order $\{y_1\} \rightarrow \{y_2\}$. So, unlike LeRoy's simple cause relation, Simon's formal order captures the intuitive causal relation for this system (from water added to pumpkin weight). In short, the reason's LeRoy's simple cause relation fails to hold while Simon's does, is that LeRoy further strengthens his causal position by imposing the sufficiency condition. This is done to rule out systems like the farmer case, which he views as ambiguous.

Having set out and critically discussed LeRoy, I now look at another economist's work on causal order: Kevin Hoover's. The analysis there shows that his position closely matches LeRoy's characterisation of Simon.

4. Kevin Hoover on Causality

Kevin Hoover's views on causality in macroeconomics are presented in his two recent books (2001a, 2001b). Like Stephen LeRoy's position on causal order, Hoover's position on causality is a development of Simon's (1953) work on causal

[29] Importantly, it characterises the causal *not* the direct causal relation.

[30] Note that this system of equations meets the exclusion condition. I assume also that the coefficients have values that rule out the problematic functional dependencies. By LeRoy's equivalence claim, this implies that by the subset condition holding for y_1 and y_2, y_1 causally precedes y_2 in Simon's formal order.

order.[31] Indeed, in many respects Hoover's reading of Simon's formal order appears very close to the strong reading of chapter two. In particular, Hoover stresses the distinction between direct and indirect control, the independence of directly controllable factors and the invariance of causal structure to intervention in factors. Despite this, however, there is an important divergence between Hoover's position and the strong reading proposed in chapter two. The difference is in the attitude to the equations and what they represent. This difference is important and Hoover's position is significantly different from the strong reading as a result.

This section gives a brief presentation of Hoover's treatment of causal order fleshing out how his approach is similar to and differs from the strong reading. The result of the analysis is that Hoover's position is seen to fit closely the way LeRoy characterised Simon. Hoover's definition of causal order can also be interpreted by a subset condition holding in systems where an exclusion condition is met.

4.1. Hoover's Simon-based Reading of Sets of Equations

To understand how Hoover reads equations and Simon's formal order, it is first necessary to be clear about the systems of equations which he attributes causal order to. These are sets of equations that are solvable for variables in terms of coefficients.[32] In this respect, he stays faithful to the systems Simon analyses in his (1953).[33] However, unlike Simon and the discussion of chapter two, Hoover does not assume equations to be linear.[34]

---

[31] Unlike Simon (1952, 1953), Hoover rejects the metaphysical scepticism on causality inspired by David Hume. Instead Hoover points out the causal richness of Hume's writings on political economy to support his own causal realist position (Hoover, 2001a, pp.2-11; 2001b, pp.98-99). Hoover believes that sets of equations are attributed causal content in virtue of their claim to denote causal structures in the world.

[32] Hoover uses the term 'parameter' instead of Simon's 'coefficient', I stick to coefficient here to keep the discussion clear, and to make the connections with earlier analyses explicit.

[33] In the terminology of the last chapter, Hoover looks at complete sets of equations.

[34] In addition, Hoover includes sets of equations with error terms, where error terms represent omitted causal factors, as in the interpretation of error terms in the last chapter (2001a, pp.49-51). To keep the presentation simple, I do not discuss sets of equations with error terms here.

There are many similarities between Hoover's reading of systems of equations with Simon's formal order and the reading in chapter two. In particular, Hoover's interpretation of the distinction between coefficients and variables, of the independence assumption for coefficients and of the invariance assumptions appear to match closely the reading in chapter two.

The similarities are clear from the way Hoover describes and builds from Simon's reading of the equations. Describing Simon, he writes

> '[a]ssume there exist experimenters...who can alter the parameters [coefficients] of a causal system. The class of interventions defines a higher-order relation called *direct control.* If by altering a parameter [coefficient]...the experimenter can change the value of a variable...he has direct control over [that variable]'(2001a, pp.38-39, original emphasis).

While in his generalisation of Simon's formal order (2001a, chap 3), Hoover interprets the independence assumption as a requirement that the coefficients are variation free.

> 'The idea that true parameters [coefficients] may be chosen independently is embodied in the definition of P as a Cartesian product (every possible option is open).' (2001a, p.62)[35]

In addition, Hoover stresses the importance of invariance of causal structure to intervention.

> 'Models of causal structure trace out the claims of modal invariance. Given the structures, a change in one part of the structure – i.e, a change in parameterization – is transmitted according to the causal order in a reliable way.'(2001a, p.56)

So in summary, the way that Hoover builds on Simon appears to fit very closely with the strong reading presented earlier. Both readings draw on Simon's comments about direct control, both read the independence assumption of coefficients as the variation free condition and both set out the importance of the invariance of causal structure to

[35] P is the domain of the vector of coefficients. Given this, the quote here amounts to a claim that the domain of the vector of coefficients is the Cartesian Product of the domains of the individual coefficients, that is, the set of coefficients is variation free.

changes in factors, that is, to interventions.[36] This similarity is perhaps not surprising given that both readings, Hoover's and that of chapter two, take Simon's comments on how equations are to be interpreted as their starting point.

Despite these similarities, there is a crucial difference between Hoover's reading and the strong reading. In particular, recall that in the strong reading it was assumed the attribution of direct control to factors and the interpretation of equations as mechanisms was necessary for solving the conceptual equivalence problem.[37] In short, the strong reading takes equations as fundamental entities to which causal content is to be attributed.

Hoover, however, takes a different view. For him it is the choice of coefficients and variables *not the equations* that is fundamental. He makes this clear when he writes.

> 'Simon himself may lead readers astray by writing as if the equations were the fundamental building blocks of his system. A sympathetic reading, I believe, would take the choice of parameterization [coefficients] to be fundamental as I do here.'(2001a, p.39, [6])

Equations, for Hoover, can be mathematically manipulated provided the variables and coefficients in the equations stay the same.

> '[I]t is the choice of parameterization [coefficients] that assigns the arrowheads to the causal linkages represented in a graph. If we respect that distinction, any mathematically equivalent syntax will equally represent the same causal structure.' (2001a, p.40)

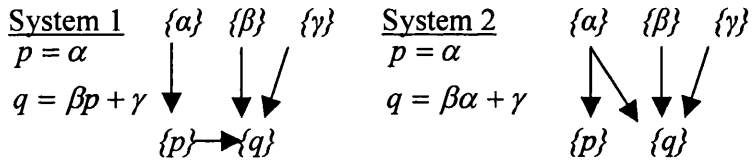
In other words, Hoover takes the attribution of direct and indirect control to factors denoted by variables and coefficients to be fundamental and in this way avoids the conceptual equivalence problem.

---

[36] See chapter two for the strong reading of these features.
[37] See Section 6.2 chapter two.

That said, equations still do play an important role for Hoover since they determine the solution for variables in terms of coefficients (the reduced form equations[38] for the variables). It is this mapping from variables to coefficients, the reduced form equations implied by the structural equations, that denotes the causal structure for Hoover. Nevertheless, Hoover's treatment of the structural equations is weaker than that assumed in the strong reading, since it only takes the reduced form equations implied by the structural equations to be causally significant.[39] In contrast, the strong reading takes the structural equations themselves, not just the solutions that follow from them, to be causally significant. It takes each structural equation to denote a mechanism, and as such it cannot be linearly combined with any other equation. In simple terms, Hoover takes the reduced form equations as a group to denote the causal structure, whereas the strong reading takes each structural equation to denote a separate mechanism in the causal structure.

To see the difference between Hoover's position and the strong reading more clearly it helps to reconsider an example of two mathematically equivalent systems given in chapter two.



As pointed out in chapter two, according to the strong reading of equations, these two mathematically equivalent systems have different causal interpretations which is visible from their distinct causal orders. This stems from the different second equations in the two systems: they represent different mechanisms. However for Hoover these two systems are not just mathematically equivalent, they are causally

[38] The structural equations are those to which Simon's formal order is to be applied, while the reduced form equations are the equations derived from these that give the variables solved in terms of coefficients.

[39] This explains why in the quote above Hoover states that one can mathematically manipulate the structural equations. This is because the reduced form functions do not change if one manipulates the set of structural equations without changing the variables and coefficients that appear in them. This contrasts with the strong reading where one could only reorder and rescale equations, because each particular equation denoted a particular mechanism that would change if the equations were manipulated in other ways.

equivalent. This is because what matters is the choice of coefficients and variables. Since both systems contain the same variables, coefficients and are mathematically equivalent, the systems imply the same reduced form equations. So their causal order *à la* Hoover is the same.

This shows the important difference in the treatment of causal order of both Hoover and the strong reading. But what does Hoover's causal order mean? How is it to be read intuitively? I now turn to this.

4.2. Hoover's Causal Order

For Kevin Hoover the causal order attributed is fully determined by the reduced form equations and the sets of coefficients and variables. Nevertheless, Hoover still uses Simon's formal order for sets of structural coefficients, which seems inconsistent given that this method is sensitive to the form of the equations, as seen in systems 1 and 2 above.

I think the way to make sense of the apparent inconsistency here is to assume that Hoover reads Simon's formal order in the same way as LeRoy. That is, system 2 has the same formal order as system 1, $\{p\} \rightarrow \{q\}$ because any changes to the value of p are accompanied by changes in q, since both depend on α. Whereas, there are changes in q that are not accompanied by changes in p, because it depends on β and γ whereas p does not. This reads Hoover as assuming a subset condition like LeRoy.[40] This solves the inconsistency provided one applies Simon's formal ordering method to systems where the exclusion condition is met, since then LeRoy's equivalence claim holds and Simon's formal order always matches the orderings obtained by applying the subset condition.

Importantly, this reading fits well with Hoover's assumption that it is the reduced form equations rather than the structural equations that denote the causal structure,

[40] Though for Hoover, the external set would be for variables, rather than internal variables, and would contain coefficients rather than external variables.

because the sets of coefficients on which any variable depends (the external sets in Hoover's systems) are determined by the reduced form equations. Therefore, the reduced form equations carry all the information (the external sets) for determining whether the subset condition holds between two variables in a system of equations, and thus, for determining the causal order between two variables. So this reading makes sense of Hoover's reliance on reduced forms.

Of course, if Hoover's causal order matches LeRoy's characterization of Simon and his view of causal order is characterized by the subset condition applied to systems that meet the exclusion condition, then one should find evidence in Hoover's writings that he restricts his analysis to systems of equations that meet the exclusion condition. In fact, there is clear evidence At one point, Hoover states that Daniel Hausman's independence assumption (1998, p.64) applies to the systems he analyses. Paraphrasing Hausman's independence assumption, Hoover writes:

> 'If A [causes] B (or if A and B are causally connected only as effects of a common cause) then B has a cause that is distinct from A and is not causally connected to A. The implication of independence is that all effects have multiple causes and not all causes are directly or indirectly causally connected...*The point to notice is that independence arises naturally in the structural [Hoover's] account with its emphasis on the causal field (error terms in econometric applications) and parameters [coefficients]'* (emphasis added, Hoover, 2001a, pp.103-104).

So it seems that Hoover thinks that the structural view implies Hausman's independence assumption. This implies that a set of structural equations, read using Hoover's approach, must be such that if a variable x causally precedes a variable y, then there must be a coefficient on which y depends but x does not. Since if this were not the case then the factor denoted by y would only have directly controllable factors that are causally connected[41] to directly controllable factors causing the x-factor, which would violate Hausman's independence assumption. From this, it is necessary if one variable is to cause another in Hoover's causal order that the effect-variable be

[41] According to Hausman (1998, p.59) two factors are causally connected if and only if one causes the other or they share a common cause.

dependent on a coefficient on which the cause-variable does not depend. But, for this to work in systems of equations where Simon's formal order is applied, requires that each equation contain an exclusive coefficient.[42] This follows from LeRoy's equivalence claim.[43]

These considerations support a reading of Hoover's causal order as one that matches LeRoy's characterisation of Simon's formal order. So I read Hoover's casual order as applying Simon's formal ordering method to systems which meet the exclusion condition, that is, have an exclusive coefficient in each equation.[44] This means that, for Hoover, given a system satisfying the exclusion condition one variable causally precedes another if and only if the subset condition is met.[45] Since the subset condition relations are determined entirely by the reduced form equations, this is consistent with Hoover's taking the reduced form equations as denoting causal structure.

All of this shows that Hoover's treatment of causal order closely matches LeRoy's reading of Simon. In both, 'x causes y' is equivalent to all the direct controllable causes of x being directly controllable causes of y, and y having some directly controllable cause that is not a directly controllable cause of x. In short, for both LeRoy and Hoover, *x causes y if whatever changes x also changes y, but not vice versa.* Note that unlike LeRoy, however, Hoover does not introduce an additional sufficiency condition on causal relations.

---

[42] Strictly speaking it requires only that each complete subset of equations contain an exclusive coefficient. But since this is essentially the same as the exclusion condition, and it would only add unnecessary complexity to the discussion to introduce this caveat everywhere, I assume the exclusion condition follows instead.

[43] See the earlier proof outline of the equivalence claim. Also, strictly speaking it holds for LeRoy's equivalence claim, properly re-labelled to apply to the sets of equations Hoover analyses that have variables and coefficients rather than internal variables and external variables.

[44] Obviously, since the causal inputs in Hoover's equations are denoted by coefficients, the exclusion condition for Hoover assumes that a unique coefficient, rather than a unique external variable appears in each equation.

[45] I leave out the earlier caveat about cancelling-out relations to keep the discussion simple.

5. The Advantage of The Strong Reading over Hoover and LeRoy

The discussion of Kevin Hoover has shown that his view of causal order closely matches Stephen LeRoy's characterisation of Simon's formal order. The problem with both positions is that they limit the application of Simon's formal ordering method to equations that meet the exclusion condition. In contrast, the strong reading permits a causal interpretation of systems of equations where the exclusion condition is not met.

The motivation for imposing an exclusion condition is operationalistic. When a system of equations meets the exclusion condition, then the mechanism denoted can be intervened into separately from any of the others, which makes causal inference easier. This can be seen from the fact that Daniel Hausman's independence condition, seen in the discussion of Hoover above to imply the exclusion condition, is very similar to his Open Back Path condition mentioned in chapter three.[46] As seen there, a causal order that meets the Open Back Path condition allows certain causal inferences to be made. Therefore, LeRoy and Hoover, like Hausman, build into their concepts of causal order a condition which makes it convenient for these things to be known about.[47] I think this is a mistake since it rules out using causal order as a concept to describe situations where causal inference is difficult or impossible. These are important situations to be able to describe if one is to make sense of the possibility of mistaken causal inference. After all, the world may well present us with causal systems that are difficult to find out about, why rule out conceptualising such systems *a priori*? It leaves us without no formal language for discussing situations where causal relations are difficult to discover. More generally, it leaves no formal language for discussing some of the limits of causal inferential methods.

In addition, the earlier discussion of LeRoy's characterisation of Simon showed that it relied on ruling out causal systems which have causal relations that cancel out.

---

[46] Recall Hausman's Open Back Path condition: 'every cause a of b that has any causes has at least one cause d such the only path from d to b is via a' (Hausman, 1998, p.83)

[47] It is a surprising position for Kevin Hoover who criticises those who 'conflate the concept of cause with the method of inferring cause' (2001a, p.22).

This is problematic in a similar way to the exclusion condition, since it too rules out a large range of natural and social systems from being modelled. The strong reading, in contrast, has no problem analysing such systems.

Finally, LeRoy's treatment of causal order also builds in the sufficiency condition. This is a particularly strong requirement which, as shown in the earlier farmer example rules out analysing causal relations in which a cause shares a common cause with its effect. This is particularly restrictive, and though LeRoy's conditional causal relation seems to offer an alternative approach for analysing these situations, it does so at the price of introducing an artificial conditionality on the other causal relation being analysed.

In summary, the strong reading of causal order has an advantage of being applicable more generally than both LeRoy and Hoover's definitions of causal order. It avoids building in conditions that permit causal inference into the very concept of causal order. This is important if one is to make formally explicit the pitfalls and limits of causal inference using one's concept of causal order.

6. Conclusion

This chapter has set out both Stephen LeRoy and Kevin Hoover's views of causal order. It has critically presented LeRoy's characterisation of Simon's formal order and shown it to match closely with Hoover's treatment of causal order. The essence of both LeRoy and Hoover's views on causal order *is that changes in causes lead to changes in their effects but not vice versa since it is possible for effects to change without changes in causes.* LeRoy also makes an additional assumption that information about a cause must be sufficient for determining its effect. This sufficiency condition was seen to be particularly restrictive, which lead it to fail in describing some intuitive causal systems (the farmer example).

In addition, it was shown that Hoover and LeRoy's positions, when applied to linear systems, require that an exclusion condition holds. This requires that each equation

have its own external variable (or coefficient in Hoover's case) that does not appear in any other equation in the system. This epistemically motivated assumption unnecessarily restricts the scope of both Hoover and LeRoy's concepts of causal order. The strong reading, however, has the advantage of not making this assumption and can be used to conceptualise a wider range of causal systems.

Chapter 5

Identification and Causal Order

1. Introduction

This chapter looks at one important part of econometric methodology: *identification.* Put generally, relations that hold among observable and unobservable entities are identifiable if given some *a priori* knowledge about the relations, previously unknown characteristics of those relations can be deduced from observations. Typically in the econometrics literature, the identification problem is presented as the problem of inferring unknown coefficient values in systems of equations from observations. Yet a natural question comes to mind when these equations are structural (taken to denote a causal structure): *what does requiring identifiability for a set of structural equations require of the causal structure represented?* What features of causal structures ensure that they can be denoted by identifiable equations?

The main aim of this chapter is to answer this question by clarifying in an intuitively causal way what identifiability of a set of structural equations requires of the causal structure it represents. The ultimate aim is to 'translate' the classic, mathematical conditions for identification of structural equations in econometrics into intuitive conditions on the causal order denoted by a set of structural equations.

In order to provide some context, the chapter begins with a discussion of what is typically known as 'the identification problem'. The next section critically reviews Simon's (1953) discussion of the relationship between identifiability and causal order, and concludes that the way Simon requires identifiability of systems of equations to operationalise causal order precludes his analysis from being used to causally interpret the identification conditions. It argues instead that the strong reading, developed in chapter two, should be used since it does not make identifiability necessary for attributing causal order to sets of equations. The chapter then presents a useful theorem showing how identifiability of a structural equation is equivalent to it being possible that any two variables in the equation

can vary relative to each other while all other variables in it remain unchanged (a 'two-variable experiment'). Though this result applies to the functional equations and not the causal order, the strong reading of equations is then applied to develop an analysis of what identifiability requires of a *causal* order. The final section ties up the chapter with a brief discussion of the role that identifiability plays in causal inference.

2. The Identification Problem

Mary Morgan's (1990) book on the historical development of ideas in econometrics dedicates a chapter to identification. According to Morgan, the development of ideas on 'the identification problem' was tied to practical difficulties faced in measuring supply and demand elasticities from observation. As Morgan discusses, in early applied empirical work on supply and demand measurement concerns were sometimes raised that measurements did not actually measure what was claimed (pp.163-168). The difficulty was (and remains) that observed prices and quantities of goods transacted result from supply and demand mechanisms acting together. This raised a practical challenge of how to identify properties of one mechanism without mistakenly mixing in properties of the other mechanism. The subsequent development of identification concepts in econometrics arose to clarify the conditions under which one could claim to have measured properties of the individual demand and supply mechanisms.

In this section, I present a brief account of the identification problem which is similar to accounts one finds in introductory textbooks in econometrics.[1] However, it is also slightly different because I present a deterministic example. This is done because the analysis of the rest of the chapter focuses on deterministic systems of equations.[2]

[1] See, for example, Maddala (2001), Gujarati (1995).
[2] Specifically, I focus on linear systems of equations with internal variables, external variables and constant coefficients in which the internal variables are solvable in terms of the external variables. These were called 'incomplete sets of equations' in chapter three.

2.1. A Deterministic Example

Suppose an economist is in the following situation, she knows that observations of quantity and price are generated by a causal system represented by the following pair of structural equations. In the equations q and p are two internal variables denoting equilibrium quantity and price respectively, while t is an external variable denoting a tax that the government imposes on all transactions, paid in part by the consumer and in part by the supplier.

$$q = \alpha_1 p + \alpha_2 t \quad \ldots \quad demand$$
$$q = \alpha_3 p + \alpha_4 t \quad \ldots \quad supply$$

Suppose also that she knows that $\alpha_1 < 0$, $\alpha_3 > 0$, $\alpha_2 < 0$ and $\alpha_4 < 0$ in line with what is known about demand and supply mechanisms respectively.

The economist's aim is to measure the values of the coefficients in the equations from the observations she has for p, q and t. Obviously, if the tax level remains fixed then equilibrium quantity and price will be constant. But suppose that the government cannot make up its mind about the level at which to set the tax, so it changes the tax level. Then price and quantity will change in response to the tax shift. The graph below shows how two observations of price and quantity would be generated by the supply and demand mechanisms if the government increased tax.

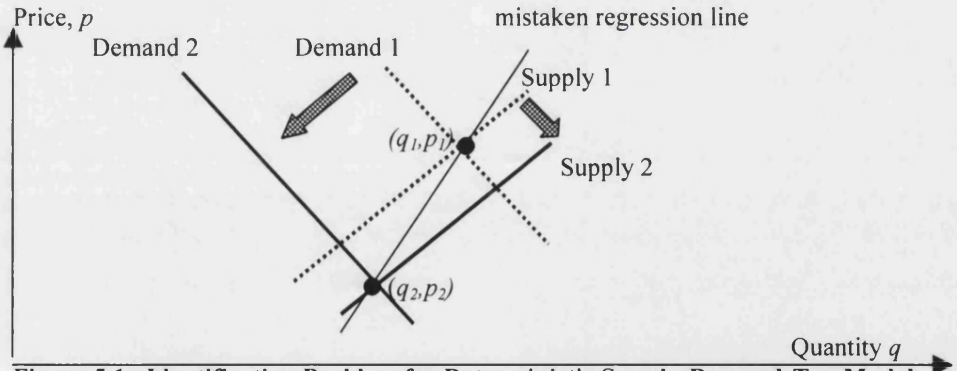


Figure 5.1 –Identification Problem for Deterministic Supply-Demand-Tax Model

Now the aim of the economist is to measure the structural coefficients from observations. Unfortunately, these coefficients cannot be *identified*. As one can see from the graph, if the economist were to straightforwardly regress a line through the two observations for price and quantity, then she would obtain the 'mistaken regression line' shown in the graph. Such a regression would not be

measuring the slope coefficient of either the supply or demand curves. Instead, since both curves have shifted (tax shifts both intercepts in the price-quantity plane as shown in figure 5.1) she would be measuring some unknown linear combination of the slope coefficients in the demand and supply equations.[3]

2.2. Koopmans' Supply-Demand Example

A more standard example of the identification problem, presented by Tjalling Koopmans in his influential paper on identification (1949, p.127) and often given as an example in textbook discussions of identification, is that of a supply and demand model in which there are two error terms u_1 and u_2 covering the factors not explicitly represented in the demand and supply equations. In this case, the structural equations are:

$$q = \alpha_1 p + u_1 \quad \ldots \textit{ demand}$$
$$q = \alpha_3 p + u_2 \quad \ldots \textit{ supply}$$

Since the u's are error terms they are unobservable.[4] The identification problem in this case arises in part because one does not know, given two (or more) observations for quantity and price, whether and by how much u_1 and u_2 differ between observations.

To see this, suppose one observes two distinct price and quantity observations which one knows were generated by mechanisms correctly represented by the structural equations above. First note that there is no way to know whether the change was due to a shift in u_1, a shift in u_2 or both. For instance, suppose (q_2, p_2) lies 'northeast' of (q_1,p_1), then there are two possibilities to explain the shift:[5] (1) a shift in both mechanisms (u_1 and u_2) and (2) a shift in just the demand mechanism (in u_1). These are represented in the two graphs below

[3] Note that this is not a problem of having too few observations, even if the government changes tax repeatedly, all the observations for price and quantity will lie on the same spurious line above. All the observations lie on the same spurious line in this system because both price and quantity shift in proportion to the tax shift (easily checked if one calculates the reduced form for p and q). Since this ratio of the observed shifts of price and quantity is constant and independent of the tax shift, the observations must lie on the same (spurious) line.

[4] I am assuming in line with convention that the error terms denote factors omitted out of ignorance.

[5] I haven't included a case where only the supply mechanism shift occurs as possible, because that would require that the demand mechanism were upward sloping, which I assume is known not to be the case.

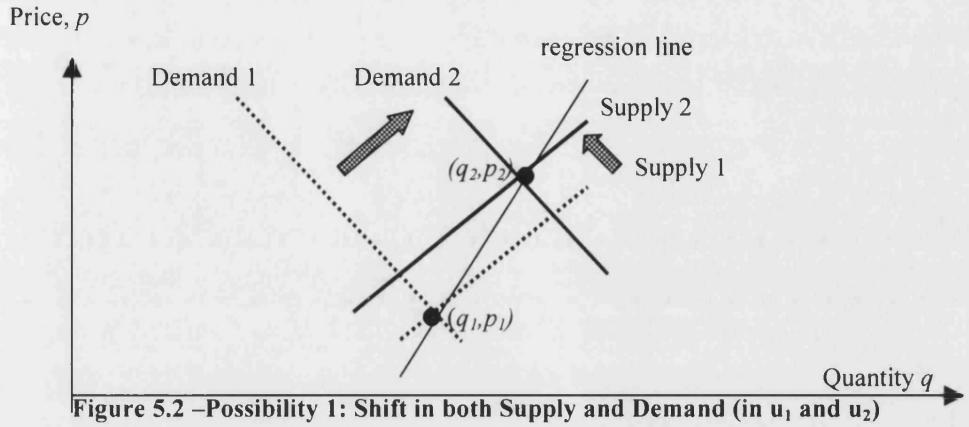


Figure 5.2 –Possibility 1: Shift in both Supply and Demand (in u_1 and u_2)

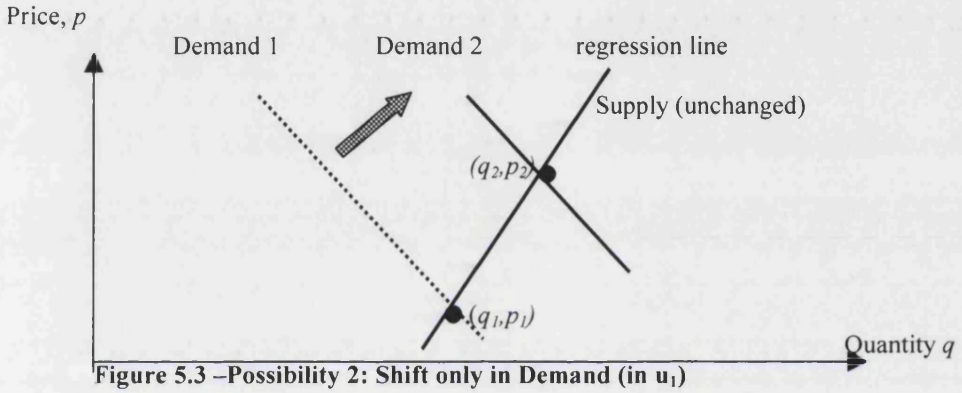


Figure 5.3 –Possibility 2: Shift only in Demand (in u_1)

The first graph shows how the observations would have been generated if both the supply and demand mechanisms were shifted, while the second shows how they would have been generated if the change was only due to a shift in the demand mechanism. Since the changes in the *u*'s are not observed, there is no way of knowing which of the two possibilities above it is. Moreover, even if one somehow knew that both error terms had shifted, one would not know by how much and so that knowledge would not help to identify the equations. However, if one knew that only the demand mechanism shifted, as shown in the second graph, then one *could* regress on the observations to measure the supply equation because, as the figure 5.3 illustrates, in this case the regression line fits on the supply equation.

2.3. Solving the Identification Problem

This second possibility above suggests a way out: attributing shifts to particular mechanisms. If one has a way of observing shifts that can be attributed to one mechanism but not another then identification may be become possible. To see

this, suppose that both examples above are modified so that there is an extra observable external factor, income, which figures in the demand mechanism but not the supply mechanism (and so income appears in the demand equation but not the supply equation). Then one can identify the supply equation in both systems.[6]

In the first (deterministic) example, income could allow the supply mechanism to be measured in the following way. In this case the correct structural equations would become (where i is the external variable denoting income).

$$q = \alpha_1 p + \alpha_2 t + \alpha_5 i \quad \ldots \quad \textit{demand}$$
$$q = \alpha_3 p + \alpha_4 t \quad \ldots \quad \textit{supply}$$

To identify the supply equation, one would ask the government to keep the tax level fixed and then wait to observe a change in income.[7] Once income changes, one gets a situation like that represented in Figure 5.3 above. Income leads the demand mechanism to shift while the supply mechanism does not shift (since income is 'excluded' from it). Here, since one observes that only income changes (among the external variables) and since one knows the form of the structural equations, one can simply fit the supply equation to the observed two points. This is possible because knowing that only the demand mechanism has shifted it is known that the two points must lie on the unchanged supply equation.

As an aside, note that here we have an 'experiment' to observe the coefficients of the supply equation in the following way: tax is held fixed while income changes lead to a systematic change in the equilibrium price and quantity. Since the observed changes in price and quantity must satisfy the known form of the supply equation, this can then be used to infer the strength of connection between price and quantity in the supply mechanism. Later in the chapter, I discuss in more detail how such experiments relate to a necessary and sufficient condition for identification, the rank condition.

---

[6] In the second system one also needs that the error terms are uncorrelated with income and each other, otherwise this correlation in the error terms will lead to a bias in the estimates for coefficients.

[7] It is not necessary that the government hold tax fixed, identification is also possible if tax varies sufficiently. Later in the chapter, I discuss this in more detail.

In the second example (with errors), things are somewhat complicated by the fact that one cannot hold the u's fixed, as was done with tax in the first example. In this case, adding in income, one obtains the set of structural equations.[8]

$$q = \alpha_1 p + \alpha_5 i + u_1 \quad \ldots \quad demand$$
$$q = \alpha_3 p + u_2 \quad \ldots \quad supply$$

Provided that the u's are not correlated with one another nor with income,[9] then identification of the supply equation is possible. To see this, imagine that income changes to six different levels while the error terms change in an uncorrelated way with each other and income. The graph below shows how six such observations might be generated given the shifts in income and error terms.

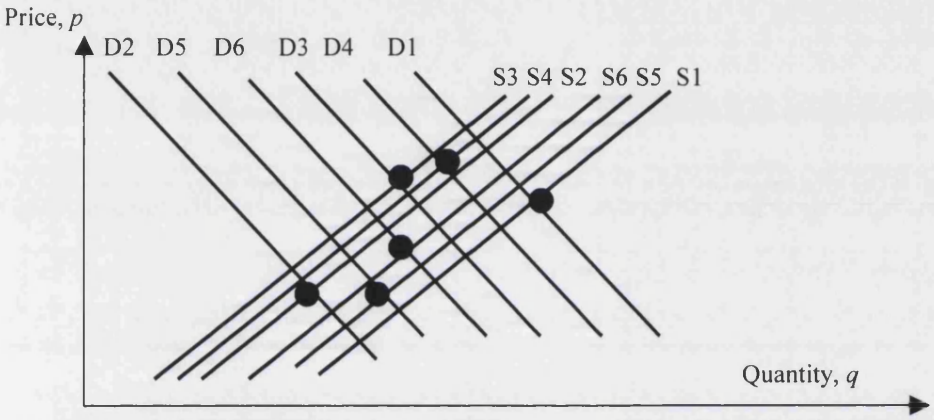


Figure 5.4 Multiple Observations for Income changes with Error term changes.

Under these conditions,[10] the regression line fitted through the observations[11] (which is shown in the graph below) should have a slope which is a good fit with the inverse coefficient for price in the supply equation.[12]

---

[8] This case follows closely Koopmans' own discussion (1949, p.129) and countless examples in introductory econometric discussions of identification.

[9] Stephen LeRoy calls this requirement 'the uncorrelatedness assumption' see LeRoy (2004, pp.16-17).

[10] As an aside, note that if the impact on the mechanisms of the variation in income is small relative to the impact due to the variation in the error terms, then estimates of the coefficients will be highly inaccurate. In the graph above we have implicitly assumed that this is *not* the case, by assuming that the 'spread' of shifted demand equations is greater (since income changes it not supply) than the spread of the supply equations.

[11] Using a suitable estimation procedure.

[12] In the case where the error terms were correlated with each other or with income then these points would trend away from the underlying supply equation, introducing a bias in the measurement of the slope coefficient. This is a problem of identification not merely of statistical inference because, not knowing the correlation, one cannot infer back to the correct coefficient, no matter how well the sample of observations represents the population. In cases where these correlations are known, identification can become possible again. Indeed, specifying constraints

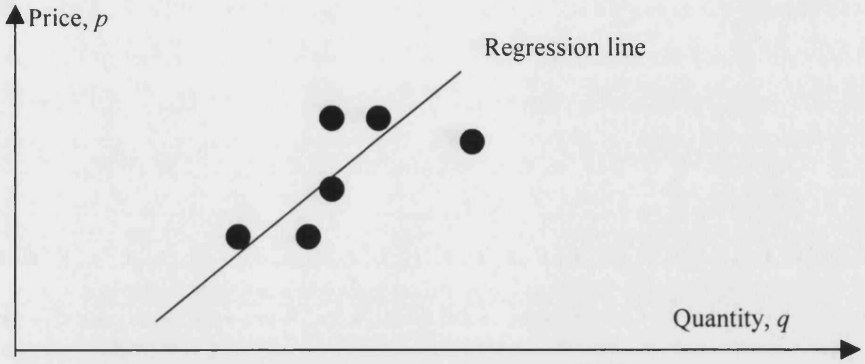


Figure 5.5 Regression Line for Income changes with Error term changes.

In both examples, identification of the supply mechanism is possible given that the demand mechanism contained a factor which was excluded from the supply mechanism.[13] This is an example of identification being secured due to variables being excluded from equations, or at the causal level, due to factors being absent from mechanisms.

This 'exclusions' approach can be generalised to give conditions for identifiability of equations in sets of linear equations. The conditions for identification of an equation in such systems of equations, where all variables are observable, are:[14, 15]

> *The Order Condition (Necessary for Identification)*
> Given a system of *n* linear equations in *n* internal variables and *m* external variables in which all variables are observable, a necessary condition to identify all the coefficients in an equation is that the equation exclude at least *n-1* variables (internal or external).

---

on the covariance matrix for error terms is another way identification can be secured, for a detailed discussion of this see Fisher (1966, chapter 4).

[13] Note that the method used in the two examples above cannot be used to identify the demand equations. This is because there are no observable variables which the demand equation excludes. Therefore, there is no way of attributing an observable shift to the supply mechanism alone, which is what is required to make inferences about the demand equation.

[14] For more on these conditions, see Fisher (1966, pp.39-41), Gujarati (1995, pp 657-669) and Maddala (2001, pp.348-352).

[15] In line with the standard econometric treatment, I assume that all coefficients in the system of equations are unknown and that there are no constraints on the external variables (i.e. they are variation free). Given this, the only way to secure identifiability is by exclusions of variables from equations.

The Rank Condition (Necessary and Sufficient for Identification)
Given a system of n linear equations in n internal variables and m external variables, in which all variables are observable, a necessary and sufficient condition to identify all the coefficients in an equation is that the submatrix of coefficients formed from the columns of the coefficients of the variables (internal and external) excluded from that equation has rank $n\text{-}1$.[16]

These conditions are not discussed in detail here. Instead, later in the chapter a theorem is presented that allows one to interpret the rank condition in a way that makes explicit a general connection between it and the possibility of experiments like that described above in the tax example. The aim in doing this is to make clear just what requiring a structure to be identifiable entails, from an explicitly causal perspective.

Before that however, I return to Herbert Simon's important work on causal order. This is because Simon ties his causal order intimately with identification conditions. This is done for the purposes of clarifying what Simon said and also for critically evaluating the claims that he makes in relating identification and causal order. This is important since this chapter aims to flesh out the relationship between identifiability and causal order.

3. Simon on Identification and Causal Order

In chapter two, a strong reading for causally interpreting linear systems of equations was developed by building on Herbert Simon's (1953) paper on causal order. Unlike Simon however, my reading was developed without assuming that identifiability of those systems was necessary for attributing them a meaningful causal order. There the conceptual equivalence problem was dealt with by assuming the equations denoted mechanisms, which ruled out the equations being mathematically manipulated to change the formal order. Of course, Simon also wants to avoid the conceptual equivalence problem, or, as he puts it 'we sought an

[16] The rank of a matrix is the dimension of the space spanned by its rows or columns.

operational basis for the concept of causal ordering, a basis that would make of the ordering something more than an arbitrary property of a particular (arbitrary) way of writing the equations governing certain empirical variables' (1953, p.27). In contrast to the strong reading and in line with his operationalism, however, Simon uses an identifiability condition for sets of equations to avoid the conceptual equivalence problem. This section sets out the role identification plays in Simon's (1953) treatment of causal order.

3.1. Interpreting Simon on Identifiability and Causal Order

Clarifying Simon's position on identification and causal order is not easy because Simon's paper is sometimes ambiguous. To give just one example, consider the following quote in which Simon apparently makes a claim for an equivalence between causal order and identifiability.

> 'the conditions under which the causal ordering of a structure is operationally meaningful are generally the same as the conditions under which the structural equations can be distinguished from nonstructural equations, and the same as the conditions under which the question of identifiability of the equations is meaningful'. (Simon, 1953, p.27)

Unfortunately, this quote can be read in (at least) two ways. One reading is that Simon is claiming that operationally meaningful causal order and identifiability are equivalent. This would read 'the conditions under which the structural equations can be distinguished from nonstructural equations' as conditions for the identifiability of structural equations. But a problem with this reading comes immediately from the third part of the quote: what does Simon mean by 'conditions under which the question of identifiability is meaningful'? This last part suggests a second reading: that operational causal order is equivalent to a *precondition* for considering identifiability *not* to the identifiability conditions themselves.

As the quote shows, it is not always very clear what Simon means. So in order to avoid such interpretative difficulties, I look at the more explicit, formal analysis that Simon presents to develop a concrete interpretation of the logical relationship he claims holds between causal order and identifiability.

In his formal discussion Simon sets out which mathematical transformations preserve his formal order for sets of equations. He shows that for any set of equations, rescalings[17] of the equations (transformations that multiply each equation by a non-zero constant) preserve the formal order among the variables.[18] Other transformations are problematic, for instance, linear combinations of equations since these typically change the variables which appear in the equations and thus change the formal order over variables.[19,20] So, in short, Simon shows:

(1) The formal order of a system of equations is preserved by rescalings of the equations.

A connection between this result and identifiability then follows from Simon's claim that '[t]he definition of identifiability implies that a linear structure is completely identifiable if and only if the *a priori* restrictions on the model...are such as to permit only [rescalings of equations]'(*ibid.*, p.30). Here by *a priori* restrictions Simon means exclusions of variables from equations, as in the above discussion of the identification problem. By exclusions permitting only rescalings, he means that there must be sufficiently many zeros in the coefficient matrix so that any linear combination of equations leads to a system with fewer

---

[17] Simon's uses 'S-transformations' to denote mathematical transformations that rescale equations, I think this terminology is unnecessarily cumbersome, so I use 'rescalings' instead.

[18] Note this result fits with the properties of M-equations in appendix 2.1, since there it was shown there that linear combinations of equations preserve the formal order over equations and therefore over variables.

[19] Simon rules out transformations that reorder equations because he sees reordering as rearranging the interventions that are associated with each equation. I think this is unnecessary since I doubt anyone would consider a model to have a different meaning merely because its structural equations were written in a different order. Arguably, interventions associated with equations would be reordered along with the equations. This is why in the strong reading M-equations can be reordered. That said, these reordering transformations are not important since either they are ruled out (like Simon) or easily brought into the set of acceptable transformations (like the strong reading). I do not discuss them here as they merely complicate the discussion without contributing to its substance.

[20] Importantly, there may be some non-rescaling transformations that do *not* change the formal order over variables. In a footnote Simon points out that linear combinations of equations in the *same* complete subset of equations does not change the formal order among the variables (1953, p.30, [11]). Though this may sound like it contradicts theorem 2.2 in Appendix 2.1, where it was shown that the formal order is preserved *only* by rescalings and reorderings, this is not the case. This is because, as set out there, non-rescalings (nor reordering) transformations always change the formal order *over equations*. This was motivated by the particular interpretation of equations (as denoting specific mechanisms) adopted in the strong reading, which is muddled given any linear combination of equations. So, though there are some non–rescaling and non-reordering transformations that preserve the formal order *over variables*, these change the identity of equations (in the strong reading) and thus change the formal order over equations. So the contradiction is only apparent. See Appendix 2.1 for details.

exclusions or exclusions of variables in different parts of the equations.[21] So Simon claims:

> (2) A system is identifiable by *a priori* exclusions if and only if the exclusions permit only rescalings of the equations.

To understand the connection between this and Simon's operationalism, consider Simon's comment:

> 'An important guiding principle in the relationship between mathematical models and empirical data is that a property of a mathematical model cannot be regarded as reflecting a property of the empirical world the model purports to describe unless this property is invariant under permissible *(operationally nonsignficant)* transformations of the equations specified by the model.'(my emphasis, ibid., p.28)

But what transformations on sets of equations are operationally nonsignificant? To answer this, consider a set of equations whose coefficient values are unknown. If that set of equations fits a set of observations, then any set of equations that is an invertible linear transformation of that set of equations will fit the observations equally well. Therefore, if one is restricted to fitting a set of equations with unknown coefficients to the observations as *the* operation for measuring coefficients, then all invertible linear transformations of that set of equations are operationally nonsignificant. As noted by Simon, the problem with this is that it makes the formal order of the set of equations an arbitrary feature of the way equations are written because then the formal order can be changed by an operationally nonsignificant transformation. Therefore, without further conditions, formal order is not 'operationally unique' and is operationally meaningless!

In contrast, if one has operations for specifying sufficiently many exclusions of variables from a set of equations (i.e. the exclusions are operationally meaningful) then one can avoid the operational meaningless of formal order. This is because if one has a set of equations which has sufficiently many exclusions of variables so that the set of equations is identifiable, then the formal order of the set of equations is operationally unique. This follows because the set of transformations

[21] It is crucial that one reads the exclusions as having a *particular* location. What is important is that the locations of the exclusions are preserved by transformation, not just the number of exclusions in an equation.

that preserves the exclusions in an identifiable set of equations, by (2), are rescalings of the equations. Since, by (1), rescalings preserve formal order the formal order is then the same under all operationally nonsignificant transformations. So the formal order is operationally unique and is operationally meaningful. In this way, identifiability for a set of equations ensures an operationally unique formal order for a set of equations. In summary, (1) and (2), assuming operationally meaningful exclusions of variables, imply

(3) A set of equations has operationally unique formal order if it is identifiable.

Given this, it seems Simon could simply require that sets of equations be identifiable so that they have operationally unique formal order. Interestingly however, Simon's operationalism[22] drives him even further. Simon states that '[operationalism] requires us to associate with each equation a procedure (set of operations) for altering its constant term or coefficients' (ibid., p.27, original emphasis removed). Simon connects this with his analysis of formal order and identification when he notes:

> '[i]f with each equation of a structure we associate a specific power of intervention, then, under S-transformations [rescalings] this one-to-one correspondence between equations and interventions between equations will retain its identity. But under [other transformations], the equations will be scrambled and combined' (1953, p.30).

To interpret this, recall that 'experimenters' intervene into equations using external variables,[23] so requiring that each equation have a specific power of intervention is best interpreted as a requirement that each equation have a unique external variable that only appears in that equation. In other words, Simon requires that the following hold.

(4) *Simon's Exclusion Condition*: Each equation has a specific external variable unique to that equation.

---

[22] See chapter two for a short discussion of Simon's operationalism.

[23] Strictly speaking, for Simon it is coefficients not external variables that denote the factors by which experimenters intervene. However, since in this chapter I focus on linear systems of equations in internal and external variables, I reformulate Simon's exclusion condition for these systems. I focus on these systems of equations because these are more like those analysed in econometrics (see chapter three), which makes it more straightforward to relate Simon's discussion to the standard identification discussions.

This exclusion condition is related to rescalings by the subsequent part of the quote, which can be restated as a claim that.

(5) If each equation has an external variable unique to that equation then *only* rescalings of equations preserve each equation having its unique external variable.

This last claim ensures that the exclusion condition by permitting only rescalings of equations, ensures that the formal order attributed to the set of equations is operationally unique.

Given this, Herbert Simon's position on the relationship between formal order and identifiability can be summarized by two important claims. The first is

Sufficiency of Identifiability for Operationally Unique Causal Order A set of equations has operationally unique formal order if it is identifiable.

This is complemented by Simon's operationalist requirement that the set of equations satisfy his exclusion condition. So his second key requirement is that the systems of equations to which his formal order is applied, satisfy.

Simon's Operationalist Requirement In a system of equations, the exclusion condition must hold, that is, each equation must have an external variable unique to that equation.

Simon's operationalist requirement ensures that the system of equations is identifiable, which in turn ensures that the set of equations has a operationally unique formal order.

Interestingly, Simon's approach can also be seen as a solution to the conceptual equivalence problem. Recall that the conceptual equivalence problem is that sets of equations to which causal order is attributed can have their causal order changed by mathematically manipulating equations. Simon solves the conceptual equivalence problem by requiring that each equation have its own unique external variable that is unique to that equation. This implies that the system of equations is identifiable,[24] which fixes a operationally meaningful unique formal order for the set of equations. So the exclusion condition ensures that the formal order

[24] See appendix 5.1 for a proof.

attributed to a set of equations is unique, which solves the conceptual equivalence problem.

Finally, note that this interpretation of Simon is very similar to Stephen LeRoy's reading discussed in chapter four. Like LeRoy, I read Simon as assuming an exclusion condition. However, I read Simon's exclusion condition as stronger than LeRoy's version. To see the difference, contrast LeRoy's version with Simon's.

> (*LeRoy's Exclusion Condition*) '[E]ach equation contain[s] at least one external variable not found in any other equation.' (LeRoy, 2004, p.5).

> (*Simon's Exclusion Condition*) Each equation has a specific external variable unique to that equation.

Despite the obvious similarity, the two conditions are not equivalent. The Simon version requires that a particular external variable uniquely appear in a particular equation. So, for instance, that variable x_1 appears only in the first equation. In contrast, a system that meets LeRoy's exclusion condition need not require that a particular variable be specific to a specific equation. All it requires is that in each equation contain *some* external variable that does not appear in any of the other equations. In certain circumstances, systems that meet LeRoy's conditions can be mathematically manipulated to get a mathematically equivalent set where the equations are scrambled to have *different* unique variables in the equations. An example of such a system was presented in the discussion of LeRoy in chapter four.[25] In contrast, the version of the exclusion condition I attribute to Simon rules out this possibility. Any system created by a linear combination of equations in a system that meets Simon's exclusion condition, changes which exclusive variable is specific to which equation, and so violates his exclusion condition.

There are two reasons why I read Simon as making this stronger exclusion condition. The first is that Simon says 'with each equation of a structure we

[25] See system (D) in chapter four.

associate a *specific* power of intervention' (emphasis added, 1953, p.20) and I take the stronger exclusion condition to be suggested by his use of 'specific'. The second reason is that, unlike LeRoy's exclusion condition, this stronger condition implies identifiability. This is necessary if Simon's claims, set out above, are to be valid. So, being charitable, I read Simon as adopting the stronger exclusion condition above rather than LeRoy's.

3.2. How Simon Contrasts with the Strong Reading and Related Criticism

Recall that my strong reading assumes that interpreting equations *as mechanisms* imposes the constraint that equations cannot be linearly combined. As an example, consider the earlier unidentifiable system of equations, the supply and demand example with tax as an external variable common to both equations.

$$q = \alpha_1 p + \alpha_2 t \quad \ldots \quad demand$$
$$q = \alpha_3 p + \alpha_4 t \quad \ldots \quad supply$$

In my reading, I assume that there are principled reasons for taking the first equation to represent a demand mechanism and the second a supply mechanism. Though the two equations relate the same variables (representing equilibrium price, quantity of a good and tax), they cannot be linearly combined without jeopardizing the mechanistic interpretation of the equations. In this case, linearly combining the demand equation and the supply equation may give a new equation (here with the same variables) but it would not give an equation that represents a mechanism.

In contrast, Simon would not consider the set of equations here to have an operationally meaningful causal order because neither equation has an exclusive external variable. In Simon's view, for sets of equations like those above, though they can be attributed a unique formal order (using his method) this formal order is not operationally meaningful because there is no independent way of intervening into the respective equations, that is, Simon's exclusion condition is not met. This condition operationalises the causal order because it allows each equation to be varied independently of the others and thus allows each complete subset in the causal order to be varied independently. This naturally fits with the possibility of experimenting to investigate each part of the causal order, or, in

others words allows an operation for discovering how different variables (or equations) vary under intervention.[26] In this way, Simon's operationalism leads him to tie his concept of causal order to a condition for finding out about causal order.

In contrast, under my reading the supply-demand example has a unique causal order simply because the equations are taken to denote mechanisms. In the strong reading, it is not the exclusions of variables that ensures unique causal order, but rather an appeal to content which is not explicit in the equations, or at least, not explicit in the exclusions of variables from equations.[27]

Simon's reading achieves uniqueness of causal order by relying on excluding sufficiently many variables. This is a rather strong requirement, and problematic because it denies a causal interpretation to unidentifiable (underidentified) models. Of course, the motivation behind an approach like Simon's is to avoid metaphysics, or more prosaically, to avoid talking about things about which we cannot know. So, a defender of Simon might argue: what sense is there in talking about a model whose relations are unknowable? And he would claim that this makes talking about underidentified models (like that above) meaningless, since in these cases one cannot deduce the values of coefficients from observations.

As with Stephen LeRoy and Kevin Hoover's position at the end of chapter four, I think this is mistaken. To see why, remember that in the discussion of the identification problem in section two above, measuring the values of coefficients relied on knowing the correct structural form of the equations. Now the obvious question is: where does this *a priori* knowledge come from? There are clearly a lot of possibilities. For instance, in our demand-supply example, one might appeal to everyday 'folk' knowledge that consumers buy less when prices go up, or one might appeal to more sophisticated rationality claims about the utility of consumers. However, in almost all of these cases, the *a priori* causal knowledge has not been gained by setting up some more general set of equations, excluding

[26] And, as is shown in detail below, identifiability is equivalent to the possibility of experiments.
[27] The content is partially explicit in the equations because the coefficients in the demand and supply equations are different with different interpretations. Also, recall at the end of chapter two that to make the strong reading explicit I suggested that a new equality symbol $=_M$ be used.

variables to secure identification and fitting data to them to see what values of coefficients fit best.[28] My criticism isn't that *a priori* knowledge is required, but is instead that to restrict causally ordered systems to those that are identifiable reveals an inconsistent attitude. On one the one hand, the attitude makes free use of *a priori* knowledge that does not rely on identification to support the functional form which is used to interpret the observations. While on the other, it requires that the functional form be identifiable to be meaningfully causally ordered. It seems simply arbitrary that identifiability (or even stronger, an exclusion condition) should be necessary for reading equations causally in the second case but not the causal claims underlying the functional form in the first.

On a closely related issue, Nancy Cartwright (2001) presents criticisms against those who claim modularity is necessary for causal relations. Modularity is a similar, though stronger requirement than identifiability, that requires that each factor have its own causal factor that influences it alone.[29] It is attractive, according to Cartwright, because it implies 'epistemic convenience'. Epistemic convenience is essentially another name for identifiability, it ensures that coefficients in the systems of equations to be measured from observed values for the variables (given other conditions are met). One criticism she gives of modularity, which is particularly relevant here, is her argument that operationalism does not give a good reason for adopting modularity (pp.73-74). Earlier in the paper, she shows that modular systems allow coefficients to be identified using a simple method of concomitant variation, that is, one can vary one particular cause of an effect to observe the strength of its influence.[30] However, she argues that this method of concomitant variation is just *one* of the methods that are open to operationalists. For example, more complicated versions of concomitant variation and other methods for finding out about causal

[28] Model selections methods are an example of this. However, even if it is done this way, that is, by testing a more general identifiable set of equations one is still left with the same problem 'one level up'. Where does the *a priori* knowledge for this more general set of equations come from? Eventually, we must rely on some method that does not involve inferring coefficients in identifiable systems.

[29] The similarity of modularity to the exclusion condition should be obvious.

[30] The next section of the paper presents a similar result, but generalized to cover simultaneous equation systems.

connections are possible. This implies that operationalism is not restricted to modular systems in what it can operationalise.

My argument above makes a similar point. The claim is that in order to make use of an identifiable system, one uses methods other than those that require identifiability to obtain the background '*a priori*' knowledge which is needed to identify the coefficients. Given this is the case, it seems odd to tie having a causal order to identifiability because that seems to arbitrarily privilege one method for knowing over others. It ties the causal interpretation to just one method *among many* for obtaining causal knowledge.

3.3. Concluding Comments on Simon

This section has presented an interpretation of Simon's discussion on the relationship between causal order and identifiability. It has shown that Simon (1953) restricts his definition of formal order, for operationalistic reasons, to systems that meet his exclusion condition of having a unique external variable in each equation. Systems of equations that meet his exclusion condition are identifiable, which Simon shows to be sufficient for a unique formal order equations. In this way he avoids what I call the conceptual equivalence problem.

In contrast, my strong reading, developed in chapter two, does not build in identifiability as a condition for solving the conceptual equivalence problem. Instead, it takes equations to denote particular mechanisms which alone is sufficient for ruling out the problematic transformations that lead to conceptual equivalence problems. This approach has the advantage over Simon's that it does not limit the causal interpretation of systems of equations to those that are identifiable.

This advantage of the strong reading is exercised in the remainder of the chapter, where I attempt to use it to discuss what identifiability requires of a causal order. This question is meaningful in the strong reading, where both identifiable and unidentifiable systems of equations can be causally interpreted. This is ruled out in Simon's approach since, in his reading, in order for a set of equation to have a meaningful causal order it must be identifiable. This leaves no scope for

investigating what interesting features, if any, causal orders must have in order to be identifiable.

4. What Identifiability Requires of Causal Order

The aim of this section is to set out an explicit causal interpretation of identifiability using the strong reading of chapter two. The underlying goal is to facilitate the understanding of identifiability of causal structures in a way that makes causally intuitive how the identification conditions allow one to measure the strength of causal connections.

In standard discussions of identification in econometrics, two conditions are presented for linear systems of equations to be identifiable. These are the *order condition* and the *rank condition* presented at the end of section two. The rank condition is the more powerful of the two conditions since it is a necessary and sufficient condition for identifiability of an equation, whereas the order condition is only necessary. The rank condition requires that a submatrix (formed using exclusions) of the coefficient matrix of a linear system of equations have a certain rank, that is, as a transformation it preserves sufficiently many dimensions. As this description makes clear, the rank condition is a purely mathematical condition on a matrix, it does not make explicit what special features, if any, a causal order denoted by an identifiable system of equations has. Though the rank condition ensures identifiability for a system of equations, it does not give any clue as to what is special about a causal order denoted by an identifiable system of operations.

Yet, it is intuitive that causal orders that are denoted by identifiable systems of equations should have interesting properties. After all, if a system of equations is identifiable, then one can measure its coefficients from observations. If the system of equations is structural, and thus denotes some causal order, then these coefficients are structural, they measure the strength of causal connections. In this case identifiability allows strengths of causal connections to be measured. Intuitively, one expects this to require something of the causal order which it represents since not all causal orders 'will permit' the strengths of their causal

connections to be deduced from observations. This intuition suggests that identifiability of systems of equations that denote causal orders, should imply that their causal orders satisfy certain conditions that make them epistemically convenient. It is the aim of this section to flesh out what these properties are.

The section begins by presenting and discussing a theorem that shows that the rank condition is equivalent to an alternative condition, which is easier to interpret causally. It then introduces the strong reading in order to interpret just what identifiability requires of causal order.

4.1. An Equivalence between Identifiability and Possible Experiments

In appendix 5.2 I prove a theorem which shows that identifiability of an equation is equivalent to any two variables in that equation being able to vary, while all other variables in the equation remain constant. This situation where at least one of two variables in an equation change but all other variables in the equation do not change, I call a 'two-variable experiment' since it has features one associates with ideal experiments. Namely, it is a situation where two factors can vary while some other relevant factors are fixed.

This relationship between identifiability and possible changes in variables has been suggested by others. For example, Stephen LeRoy states what is essentially the same result in his discussion of identification in his recent paper[31] (2004, p.19) attributing the result to James Heckman (2004). Though I have not been able to trace a clean statement of the result that LeRoy gives,[32] a similar claim is made for a specific supply-demand model in recent works by Heckman (2000, pp.57-59) and (2001, p.34). Also, Nancy Cartwright in various works (e.g. 2003a) has

[31] Stephen LeRoy writes 'the coefficient a_{ij} represents the effects of internal variable j on internal variable i condition on the other variables in equation i being held constant, if and only if a_{ij} is identified.' (2004, p.19, original emphasis removed).

[32] I have discussed the reference with LeRoy. I now think LeRoy may be referring to Heckman's statement that 'the causal effects are defined if the parameters are identified in the Cowles definition of identification' (Heckman, 2004, p.39). Since Heckman defines causal effects using *ceteris paribus* manipulation, like a two-variable experiment, this may be the source of LeRoy's claim.

stated and proved similar results but these do not cover the simultaneous equation systems covered by the theorem here.[33]

The theorem and the relevant concept of experiment are:

> *Theorem 5.2*: Given an incomplete set of equations,[34] the rank condition holds for an equation if and only if a two-variable experiment is 'possible' between any two variables in that equation.
>
> Let z_1 and z_2 be two variables that appear in an equation, a *two-variable experiment* occurs for z_1 and z_2 in that equation if and only if
>
> (i) All variables in the equation except z_1 and z_2 do not 'change'.
>
> (ii) At least one of z_1 and z_2 changes.

The terms in scare quotes require clarification. First, 'possible' is to be understood as being constrained in the following ways.

- Each value in the domain of an external variable is possible in some primitive sense.
- The variation free assumption on the external variables is to be read as each individually possible value of an external variable being possible independent of the values taken by other external variables.[35]
- A value is possible for an internal variable if and only if there are some possible values for the external variables which, given the equations, imply that value for the internal variable.[36]

---

[33] Also, Cartwright's approach is slightly different since she derives further causal knowledge from limited causal knowledge and knowledge about functional relations. In the work here, the analysis essentially stays at the level of functional relations, since as we will see in the later part of this section, causal order needs to be assumed separately in order for identifiability (or the experiments here) to yield causal knowledge.

[34] Recall from chapter three that incomplete sets of equations are linear systems of equations in internal and external variables, just like those being discussed in this chapter. I use the term again here because it is a convenient way of specifying the linear systems of equations in internal and external variables to which the theorem applies. Note that these systems do not contain error terms. The extension of this analysis to systems with error terms is left as further work.

[35] So the set of jointly possible values for the external variables is the Cartesian product of the sets of the individually possible values for each external variable.

[36] Formally, the set of possible values for an internal variable is the range of the reduced form function for that variable. Here I also assume that the domains of the external variables have 'nice' properties, that is the domains are open intervals in the set of real numbers. I also assume that these domains are such as to allow the joint changes in one or more external variables to

The second term 'change' refers to a difference in two values for a variable. A possible change is a difference between two possible values of a variable. Finally, a two-variable experiment between z_1 and z_2 in an equation is possible if there is a set of possible changes in the external variables, which given the equations, implies that at least one of z_1 and z_2 changes while all other variables in the equation do not change.[37]

To clarify this by example, consider the system of equations (p and q are internal, the x's external) where the external variables are variation free.[38]

$$p = \alpha x_1$$
$$q = \beta p + \gamma x_2 \qquad \textit{Formal Order } \{p\} \rightarrow \{q\}$$

It is easily checked that the second equation is identifiable using the rank condition. The theorem also allows us to see why, in terms of two-variable experiments. By definition, a two-variable experiment is possible between p and q in the second equation if it is possible that the external variables, x_1 and x_2, change so that at least one of p and q varies without any other variable in the second equation varying. Since the external variables are variation free, it is possible that x_1 changes but not x_2. Such a change (in x_1 but not x_2) leads to a change in p (by the first equation) and this change in p leads to a change in q (by the second equation). Since x_2 does not change, then only p and q change in the second equation and a two-variable experiment occurs. It follows then that a two-variable experiment is possible between p and q in the second equation. By analogous reasoning, one can also show that two-variable experiments are possible in the second equation for q and x_2, and p and x_2.[39] By theorem 5.2 above, this is sufficient for the second equation to be identifiable.

---

cancel out in the systems of equations analysed. These assumptions would be made formally explicit in a fuller, more rigorous treatment. However, they are not central to the analysis of the chapter so I do not discuss them in depth here.

[37] These interpretations are chosen so as to be as weak as possible, while consistent with the treatment using functional relations. Any stronger view of 'possible' that is consistent with these requirements could also be assumed.

[38] Recall the requirement that the external variables be variation free in order to causally interpret systems of equations, that is, for the external variables to denote suitably independent factors in the strong reading. See chapter two.

[39] Though for the two variable experiment between p and x_2, q must be held fixed by the joint impact of p and x_2 cancelling out.

This clarifies the theorem but what is the connection with identifying values of coefficients? After all, the value of identifiability is that it allows one to deduce values of unknown coefficients given known functional forms and observations of variables. So it is necessary to assume that the form of the equations above is known, that the values of the coefficients are not known and that the variables are all observable.[40] In this case, identifiability of an equation should allow one to deduce the value of the unknown coefficients from the known functional form and observations of the variables. One advantage of the theorem above, is that it makes particularly intuitive how coefficients can be measured.

To see this, reconsider the system above with its identifiable second equation. By the theorem, this means that a two-variable experiment is possible for p and q. Such a two-variable experiment occurs when only x_1 changes among the external variables. Suppose such a two-variable experiment occurred, then one would (since the variables are observable) observe changes in x_1, p and q but not x_2. In addition, since the form of the equations are known, it is known that this is a two-variable experiment for the second equation.[41] Therefore one would know that,

$$\Delta q = \beta \Delta p + \gamma \Delta x_2 = \beta \Delta p + \gamma 0 = \beta \Delta p$$

from which it follows that it is known that

$$\Delta q = \beta \Delta p$$

$$\beta = \left. \Delta q / \Delta p \right|_{\Delta x_1 \neq 0, \Delta x_2 = 0}$$

This last equation is known to hold and its right hand side can be calculated from the observed changes in p and q. The left hand side, the originally unknown coefficient β, can now be deduced from the ratio of shifts in q and p. In this way, a two-variable experiment that occurs given observable variables and known equation forms, allows coefficient values to be deduced, that is, it allows coefficients to be identified in the equations.

In summary, the theorem above gives an alternative necessary and sufficient condition for identifiability of an equation. It has the advantage over the rank

[40] These are the conditions under which the rank condition above permits the measurement of coefficients.

[41] Because it is known that only p, q and x_2 appear in the second equation.

condition of showing in intuitive terms just how identification of coefficients from observations can take place. Moreover, it does this using a concept which appears to fit closely with an intuitive concept of experiments.[42, 43]

Having given a flavour of the theorem by example, there are a couple of important points to be made. First, it is important not to mix up two-variable experiments being possible and such two-variable experiments actually occurring. The second important point to note is that the discussion here holds independent of causal order. I consider each of these in turn.

4.2. Possible vs. Actual Experiments

The theorem shows identifiability requires that it be *possible* that functional relations generate an experiment *not* that they in fact do. The latter condition is much stronger and likely to occur only in systems which can be suitably controlled or happen to be naturally shielded. In practice, if one knows the form of the equations and can observe variable values, then if these equations are identifiable then one can infer values of coefficients even if two-variable experiments do not occur.

To see this consider the example again, where as before, external variables are variation free, the form of the equations is known, but the values of the coefficients are not.

$$p = \alpha x_1$$
$$q = \beta p + \gamma x_2$$

Suppose that one observes two separate shifts in the variables. Suppose that one observes a first shift $(\Delta x_1^1, \Delta x_2^1, \Delta p^1, \Delta q^1)$ and later a second shift $(\Delta x_1^2, \Delta x_2^2,$

[42] The theorem only shows how to identify slope coefficients. However, identifying an intercept coefficient is also possible once one has measured all the slope coefficients. In that case, it can be done by substituting all the observed values of the variables in the equation. The intercept coefficient is equal to the sum of these values multiplied by their corresponding slope coefficients (assuming the equation is written with the intercept on one side of the equation and all other variables and coefficients on the other).

[43] The theorem suggests an obvious generalisation for non-linear systems. The generalised theorem would be an equivalence claim between the existence of partial derivatives for the reduced form at a point and local identifiability at that point. For an example of non-linear analysis, see Heckman (2000, 2001).

Δp^2, Δq^2). Then substituting into the associated difference equations (which are known, given the equations are known) gives four known equations.

$$\Delta p^1 = \alpha \Delta x_1{}^1$$
$$\Delta p^2 = \alpha \Delta x_1{}^2$$
$$\Delta q^1 = \beta \Delta p^1 + \gamma \Delta x_2{}^1$$
$$\Delta q^2 = \beta \Delta p^2 + \gamma \Delta x_2{}^2$$

These are four known equations in three unknown coefficients, so provided the shifts are independent[44] one can solve for the unknown coefficients. Importantly, this case can happen without any two-variable experiment occurring, that is, where all the variables change in both shifts. Nevertheless, the coefficients can be measured. So clearly an actual two-variable experiment is not required. The connection with possible experiments is simply that only a set of equations for which a two-variable experiment is possible, will be such that it uniquely fits the observed variable shifts.

Obviously, requiring that some condition be possible is much a weaker requirement that requiring that the condition applies. Therefore, the theorem might tempt some to conclude that identifiability is a rather weak requirement. Strictly speaking, this is correct in the sense that it is weaker than requiring that an experiment actually occur. However, as seen above, identifiability is only useful for finding out the values of coefficients if other strong conditions are met. In particular, it is required that (i) *the form of the equations is known* and (ii) *that variables are observable.*[45] The point is that identification may be a 'weak' condition relative to certain others, but it is only useful when other 'strong' conditions are met.

4.3. The Missing Causal Order

The second important point is that this experimental interpretation of the identification in the theorem is independent of the causal order for a system of

[44] That is, the shifts in x_1 and x_2 must not be of the same ratio in both cases. Since the variables are variation free this cannot be systematically the case, so here I assume that they are not related in this way.

[45] In the examples here I have assumed all variables are observable. However, one can derive further identification conditions for cases where some variables are unobservable.

equations. This shows up an important limitation of the concept of experiment presented here. To see this, consider the following two mathematically equivalent systems with different formal orders.

$$p = \alpha x_1 \qquad\qquad q = \beta\alpha x_1 + \gamma x_2$$
$$q = \beta p + \gamma x_2 \qquad\qquad p = {}^{1}\!/\!_{\beta} q - {}^{\gamma}\!/\!_{\beta} x_2$$
$$\{p\} \rightarrow \{q\} \qquad\qquad \{q\} \rightarrow \{p\}$$

It is easily checked that the both systems of equations are identifiable, using the rank condition. By the theorem, two-variable experiments are possible for both systems and coefficients can be measured in either system.

If one assumes that functional relations have the form on the left, then those coefficients can be measured and conversely, if the functional form on the right is chosen. Since both systems are correct *qua* functional relations, identifiability allows the measurement of coefficients in either set of equations. So which one gives us causal connections? It depends on which, if any, has the 'true' causal order. If one knew that the first system had the correct causal order, then the coefficients inferred from observation would measure causal strengths. If the second, then its coefficients measure the causal strengths. *So, in order for identifiability to measure causal strengths one must know, not only a correct functional form, but that it has the correct causal order.*

This implies that the two-variable experiments are in the absence of knowledge about causal order, not 'experiments' at all, at least not experiments that measure *causal* connections. In fact, this is not very surprising if one looks back at how two-variable experiments are defined, since the concepts used are entirely functional concepts. All they require is that two variables in an equation have different values while all other variables in the equation have the same values. As is clear from it being put in these terms, this does not use any causally substantive concepts. The causal content must come from elsewhere, as the above example of the two systems shows. The next section adds this causal content, using the strong reading of chapter two, to obtain a picture of what identifiability requires of causal order.

4.4. Identifiability and Constraints on Causal Order

This section asks for a system of equations read using the strong reading what identifiability requires of the *causal order* denoted by that set of equations. Another way of putting this is: what makes one causal order identifiable, and another one not?

The first step in understanding what identifiability requires of causal order is to introduce causal content using the strong reading. So suppose that one has a linear system of equations in external and internal variables, that is causally interpreted using the strong reading. Suppose that an equation in this system is identifiable. This implies, by the theorem above, that in that equation it is possible that any two variables change relative to each other while all other variables remain fixed. In the strong reading this implies that the mechanism, corresponding to that equation, in the causal system is such that for any two factors in the mechanism it is possible for those two factors change relative to each other while no other factor changes. Call a situation in which just two factors change relative to each other in a mechanism a *two-factor experiment.*

This moves us from two-variable experiments to two-factor experiments. It moves from the functional domain to the causal domain by using the strong reading. However, this merely 'translates' the earlier condition for identifiability into a condition on a mechanism denoted by an identifiable equation in a causally interpreted systems of equations. What is lacking is some causal story as to what is required of a causal system in order for a two-factor experiment to be possible for a mechanism.

To help with this, recall the model version of Simon's theorem 6.1 discussed in chapter two. It stated that an indirectly controllable factor in a causal order changes 'in general' if one of its directly controllable factors causing it changes, but does not change if none of the directly controllable factors causing it change. The 'in general' caveat covers the case where one or more directly controllable factors change but the joint impact of these cancel out, which implies that it is possible for causal directly controllable factors to change without causing a

change in an indirectly controllable factor. These features can be summarised in a useful 'Change condition'.

> (*Change Condition*) If an indirectly controllable factor, y, in a system changes then at least one directly controllable factor causing it changes. Conversely, if an indirectly controllable factor, y, does not change then either no directly controllable factor causing it changes OR one or more directly controllable factors change and these changes cancel out to have no impact on y.

This is useful, because it allows one to infer back from factors that do or do not change in a mechanism to conditions on the directly controllable factors. This allows one to infer from a two-factor experiment to conditions on directly controllable factors.

So, suppose that a two-factor experiment between z_1 and z_2 is possible for a mechanism, m; what kind of causal order must be in place to allow this?

To simplify the discussion, let S be the set of directly controllable factors that either appear in the mechanism, m, or are causes of the indirectly controllable factors that do. This is the set of directly controllable factors that can be used for varying factors in the mechanism m. Let S_{free} be the set of directly controllable factors that are either identical to, or causes of z_1 and z_2 (the two factors that are free to vary in the two-factor experiment). Likewise, define S_{fixed} to be the same set, but for those factors in m that are to remain fixed in the two-factor experiment.[46]

The two-factor experiment requires that at least one of z_1 and z_2 vary, while all other factors remain fixed. Given this, the change condition implies that some directly controllable factor in S_{free} must be changed (to vary the one or more of the two factors z_1 and z_2) to perform a two-factor experiment. Similarly, the change condition implies that the factors in S_{fixed} must either be left unchanged, or varied in a particular way, with cancelling out, so that the other factors in the mechanism remain fixed as required by the two-factor experiment.

---

[46] By definition, both S_{free} and S_{fixed} are contained in S, and their union is S. They need not be disjoint however.

Now there are two possibilities either S_{free} and S_{fixed} overlap or they do not. If they do not overlap, then the two-factor experiment is straightforward. Simply vary some directly controllable factors in S_{free} and leave those in S_{fixed} unchanged. Then provided there is no cancelling out,[47] then at least one of the two factors will change, while the other factors will remain unchanged since none of their causes have been changed. So, in a case where S_{free} and S_{fixed} do not overlap there is no problem carrying out the two-factor experiment.

However, this nice result is not relevant because it is *not* possible for S_{free} and S_{fixed} not to overlap.[48] To see why, recall that every factor in a mechanism is either exogenous or endogenous for the mechanism, and that every mechanism contains at least one endogenous factor.[49] Also recall that any exogenous factor in a mechanism is a direct cause of every endogenous factor in the mechanism.[50] Given this, if z_1 and z_2 are both exogenous factors, then there must be some other factor, z_3, in the mechanism which is endogenous. In that case, both z_1 and z_2 are direct causes of z_3 so then, either z_1 and z_2 are in S_{fixed} or some of their causes are; in either case S_{free} and S_{fixed} overlap. On the other hand, if either z_1 and z_2 is endogenous, then if there is another exogenous factor in the mechanism, z_3, then it causes the factor of z_1 and z_2 which is endogenous. In this case, z_3 or some cause of it is in S_{free}, so S_{free} and S_{fixed} overlap. Whereas if z_3 is endogenous, then it is causally equivalent with either z_1 or z_2, whichever is endogenous. Since z_3 is causally equivalent it then has the same causes as either z_1 or z_2.[51] In that case S_{fixed} and S_{free} must overlap since the causes of z_3 must all be in S_{free}. In

---

[47] Since identifiability is assumed two-factor experiments are possible for any two factors in the mechanism (recall that the question is what this implies for the mechanism) then it must be possible to vary elements in S_{free} so that one or more of z_1 and z_2 changes.

[48] Bar the trivial exception where z_1 and z_2 are the only two factors in the mechanism. In that case, S_{free} and S_{fixed} trivially do not overlap since S_{fixed} is the empty set as there are no factors to be held fixed.

[49] Recall from chapter two, that an exogenous factor is one that is not determined by the mechanism, while an endogenous factor is one which is. Each mechanism must contain an endogenous factor, otherwise it would be denoted by an equation which is not used to solve for any internal variable in the system. That equation would be redundant to solving for the internal variables which violates the conditions of the systems of equations that are attributed causal content. See chapters two and three.

[50] See chapter two for the definition of direct cause.

[51] Recall from chapter two that two factors are causally equivalent if they are both endogenous for the same mechanisms.

conclusion, however one chooses the two factors, z_1 and z_2, in the mechanism S_{free} and S_{fixed} will overlap.[52]

Since S_{free} and S_{fixed} must overlap, this leaves two possibilities. The first is where there is a part of S_{free} which is not in S_{fixed}. In that case one can vary those factors in S_{free} that are not in S_{fixed} and leave factors in S_{fixed} unchanged. This implies that one can vary the two factors using their directly controllable causes without varying any cause of the factors that must remain fixed. In this case a two-factor experiment is possible because there exists a cause of one or more of the two factors that does not cause any of the other factors in the mechanism.[53] This is a similar to what Nancy Cartwright (1989, p.33) calls an 'Open Back Path'.[54, 55]

The second possibility is where S_{free} is contained in S_{fixed}. In that case one cannot keep the other factors unchanged by not changing the factors in S_{fixed}, since if all the directly controllable factors in S_{fixed} were fixed then no factors in S_{free} would change, and neither of the two factors (of which one must vary for a two-factor experiment) would change. This implies that to vary these factors some change must be made to some factor(s) in S_{fixed}. In that case, a two-factor experiment will be possible if and only if the changes made to vary the two factors, can be accompanied by changes to other factors in S_{fixed} that cancel out any impact on the factors to be fixed. Only in this way can the other factors remain fixed while the two factors can vary.

So to summarise, the two ways a two-factor experiment is possible for z_1 and z_2 in a mechanism, m, are:[56]

---

[52] Assuming there are more than two factors in the mechanism. See footnote 45 above.
[53] Or, if the factor is directly controllable, then it does not cause any the other factors (those that must be fixed for the two-factor experiment) in the mechanism.
[54] There are important differences too. I discuss Cartwright's work and open back paths in more detail in the next chapter.
[55] The name is suggestive because at least one of the two factors has a 'back path' of causes such that at the 'top' of that back path there is a directly controllable factor that does not cause any of the other factors (those fixed for the two-factor experiment) in the mechanism. In this way, the back path is 'open' with respect to those other factors.
[56] The wording of the condition may sound awkward, but it is put this way to cover the case where either one or both of z_1 and z_2 are directly controllable factors.

(a) Either of z_1 and z_2 has as a cause, or is equal to, some directly controllable factor that does not cause, and is not, one of the other factors in the mechanism.

OR

(b) Either of z_1 and z_2 has as a cause, or is equal to, some directly controllable factor that does cause some of the other factors in the mechanism. However, it is possible to change this factor to vary one or more of z_1 and z_2, while not changing the other factors, though this may require changing other directly controllable factors to cancel out any influence of the first directly controllable factor on the other factors.

In case (a) some directly controllable factor(s) are varied that change z_1 and z_2, but these do not cause any of the other factors, so by not changing any other directly controllable factors, a two-factor experiment for z_1 and z_2 results. In case (b), some directly controllable factor(s) are varied that change z_1 and z_2, but since these cause some of the other factors, that must remain fixed, then other directly controllable factors may need to be varied to cancel out this unwanted influence.[57]

To complete the connection with identifiability, recall that an equation being identifiable is equivalent to a two-variable experiment being possible for any two variables in that equation. In the corresponding mechanism, m, this requires that the causal order be such that for any two variables z_1 and z_2 in m either (or both) (a) or (b) hold. That is, a *mechanism in a causal order is 'identifiable'*[58] *if and only if the causal order is such that for any two factors in that mechanism either (a) or (b) holds.*

This gives an answer to what identifiability requires of a causal order. In short, identifiability of systems of equations that are read causally (using the strong reading) requires, roughly, that the causal factors in a causal order not be too 'connected' with one another. To see how, note that (a) puts limits on common

[57] It is also possible that varying other directly controllable factors is also not necessary in this case, if the changing directly controllable factor only causes other factors on which it has no net impact, that is, for which it 'cancels itself out'.

[58] I put scare quotes because I haven't defined identifiability for mechanisms.

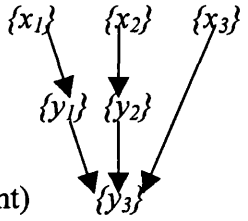
causes between factors in mechanisms.[59] Whereas (b) requires in cases where there are common causes, that there be sufficiently many other causal inputs that can be used to neutralize any unwanted influence (on the factors that are to remain fixed in a two-factor experiment). Generally, the conditions require that the casual order be 'open enough' so that any two factors in a mechanism can vary alone, without any other factors in the mechanism varying. In summary, requiring identifiability for systems of equations that denote causal orders puts strong limits on the causal orders that can be modelled.

4.5. An Example

To simplify all of this, consider the following identifiable system of equations with its causal graph on the right.

$$y_1 = \alpha_1 x_1$$
$$y_2 = \alpha_2 x_2$$
$$y_3 = \beta_1 y_1 + \beta_2 y_2 + \alpha_3 x_3$$

(x's external, y's internal, coefficients constant)



Suppose one wishes to identify the coefficients in the third mechanism. To perform a two-factor experiment for y_3 and x_3 is straightforward, it is an example of case (a). Here the directly controllable factors that control the two factors, x_3 and y_3 are given by the set $S_{free} = \{x_1, x_2, x_3\}$. The other factors, y_1 and y_2, are controlled by the factors in $S_{fixed} = \{x_1, x_2\}$. Since S_{free} is not contained in S_{fixed} it is a case of (a). By varying x_3 one can activate an 'open back path' that directly varies x_3 and only causes y_3 to vary. As long as x_1 and x_2 are not varied, y_1 and y_2 stay fixed. So the two-factor experiment is performed by varying x_3 but not x_1 and x_2.

In this case, one can see how, when complemented by the causal interpretation, the concept of a two-variable experiment fits neatly with an idealised concept of an experiment. The idealised picture is that of an experimenter changing one

[59] Unsurprisingly, (a) is a similar condition to that seen to be necessary for Mill's method of concomitant variations in chapter three. The discussion here can be seen as a generalisation of the discussion of interventions in chapter three. Though here, the discussion is directly related to an epistemic virtue for simultaneous systems of equations (identifiability).

cause, while holding all other causes of an effect fixed, to see by how much the effect changes due to the change in that cause.[60, 61]

One can also perform a two-factor experiment between y_1 and y_2. In that case $S_{free} = \{x_1, x_2\}$ and $S_{fixed} = \{x_1, x_2, x_3\}$. Since S_{free} is contained in S_{fixed} it is a case of (b). This implies that to perform the experiment one must vary the factors in S_{fixed} so that y_3 and x_3 do not change. The only way not to change x_3 is to directly fix it, so one sets $\Delta x_3 = 0$. To get y_3 not to change requires that

$$\Delta y_3 = 0 = \beta_1 \Delta y_1 + \beta_2 \Delta y_2$$

Substituting this requires

$$0 = \alpha_1 \beta_1 \Delta x_1 + \alpha_2 \beta_2 \Delta x_2 \quad \ldots \quad (*)$$

This constraint (*) can be met since x_1 and x_2 are variation free. Suppose x_1 and x_2 are both varied so that (*) is met but x_3 is not changed, then y_1 and y_2 are both caused to change but y_3 does not change because its direct causes 'cancel out'. This is how this two-factor experiment is possible in this case.

This last two-factor experiment is rather odd since it requires changes in two causes of y_3 so that their impacts cancel out. In this way one can measure the proportional effect each has on y_3. Interestingly, in this case this 'odd' type of experiment isn't in fact necessary to identify the mechanism. In fact, one can perform (a) type experiments for the following pairs of factors $\{y_3, y_1\}$, $\{y_3, y_2\}$ and $\{y_3, x_3\}$, which is sufficient for measuring the coefficients in the third equation. In these cases one is always testing to see the effect of a cause (y_1, y_2 or x_3) on the effect (y_3) so there is no need to hold the effect constant in the experiments.[62, 63] So one does not always need to perform an 'odd' type (b) experiment.

---

[60] Note the similarity between this and the discussion of interventions in chapter three.

[61] Not all two-factor experiments will be readable in this ideal way; only those that are between one factor which is a cause of another. Recall that a two-factor experiment can take place between any two factors in an equation, so the two factors may be causally ordered, in the same complete subset or even causally unordered (if both factors are exogenous to the mechanism).

[62] In this case the fact that the cause in each two-factor experiment has an 'open back path' relative to other causal factors in the mechanism is sufficient for identification. This is essentially an instance of what Cartwright (1989, p.37-38) proves generally for time ordered systems.

[63] This suggests an interesting avenue for further work. I suspect one can show that identifiability of mechanism holds if and only if every two-factor experiment between an endogenous and exogenous factor in a mechanism is possible. If this result held then one would have an arguably

However, there are also identifiable systems in which one can only perform two-factor experiments of type (b). Consider, for example, the classic supply-demand example presented in the earlier discussion of the identification problem.

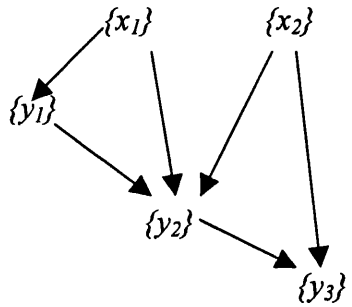
$$q = \beta_1 p + \alpha_1 x_1 ...supply$$
$$q = \beta_2 p + \alpha_2 x_2 ...demand$$

Suppose one wants to identify the supply mechanism and this is to be done by performing a two-factor experiment between p and x_1. In that case one must vary x_1 but this *ceteris paribus* changes q. So to keep q fixed one must vary x_2 to offset the impact of x_1 on q, so that q remains fixed. Similar remarks apply for all the other two-factor experiments in the system: between q and x_1; q and x_2; p and x_2; and p and q. So, in this classic identifiable economic system it is not possible to perform an experiment that meets condition (a), that is an experiment that uses some 'open back path'.

4.6. The Rank Condition vs. The Order Condition from a Causal Perspective

To finish, I show that this causal reading of identifiability can be used to highlight an important difference between the order and rank conditions for identification. Consider the following set of equations and its causal graph in which the third equation meets the order condition but not the rank condition.

$$y_1 = \alpha_1 x_1$$
$$y_2 = \beta_1 y_1 + \alpha_2 x_1 + \alpha_3 x_2$$
$$y_3 = \beta_2 y_2 + \alpha_4 x_2$$
$$where\ \beta_1 = -\left(\alpha_2 / \alpha_1\right)$$



Consider what happens if an experimenter tries to perform the following two-factor experiment: varying y_2 and y_3 while holding x_2 fixed in order to measure β_2. Since he must hold x_2 fixed, the only directly controllable factor that he can vary is x_1. According to the causal order, changing x_1 directly causes y_1 and y_2, with y_1 in turn also directly causing y_2. So its sounds that it should be possible to vary y_2

---

nicer and more intuitive result that relates identifiability with possible experiments to measure relationships from (direct) causes to their effects in a mechanism.

by changing x_1. However, there is a problem, if the first equation is substituted into the second equation, one gets

$$y_2 = \alpha_3 x_2$$

This implies that x_1 has no net impact on y_2 because its indirect impact on y_2 through y_1 and its direct impact on y_2 cancel out (due to the value of β_1). So, when the experimenter attempts to vary y_3 and y_2 holding x_2 fixed, he also keeps y_2 fixed. But then all the direct causes of y_3 are fixed, so it too is fixed. Therefore in this system that two-factor experiment is not possible (as expected given that the third equation is not identifiable).

In contrast, in an alternative system which differs from the system above only in that $\beta_1 \neq -\alpha_2/\alpha_1$, then the rank condition holds for the third equation and it is identifiable. All of the two-factor experiments for the third mechanism are then possible.

More generally, cases in which the order condition holds for an equation but the rank condition fails, are situations like the example presented here. In these situations, the values of the coefficients cancel out some influence from some directly controllable factor(s) which would have otherwise permitted identification. These are also systems like those discussed in chapter two in which changes to some cause of another factor did not lead that factor to change. As discussed in chapter two, these are systems which violate 'faithfulness'. If one rules out such unfaithful systems then the order condition becomes sufficient for identification (for the kinds of systems of equations analysed here).[64, 65]

5. Causal Inference using Identifiable Systems

In this final section, I consider two ways in which identifiable sets of equations are useful for causal inference in econometrics: (i) for measuring strengths of

[64] I do not prove this rigorously. However, the proof is straightforward, if an equation in a system meets the order condition, then since none of the impact of the excluded variables vanishes (I have ruled out such unfaithful systems) then one has sufficient variation in the observed variables to uniquely fit the equation to be identified.

[65] I suspect that an implicit faithfulness assumption lies behind the standard econometric textbook emphasis of the order condition over the rank condition. For example, Andrew Harvey (1990, p.328) states '[f]ortunately, the order condition is usually sufficient to ensure identifiability, and although it is important to be aware of the rank condition, a failure to verify it will rarely result in disaster'.

causal connections and (ii) for doing limited inferences about causal order. Hence, in this section I move away from the earlier question of what identifiability requires of causal structures, to the epistemic question of the role identifiability plays in causal inference.

Identifiability of an equation ensures that its unknown coefficients can be inferred from observations and knowledge of the form of the system of equations. Reconsider the earlier identifiable supply-demand example.

$$q = \beta_1 p + \alpha_1 x_1 ... supply$$
$$q = \beta_2 p + \alpha_2 x_2 ... demand$$

If all variables are observable and the forms of the equations are known but coefficients are unknown, then coefficients can be calculated from observation.[66] All that is required is some sufficiently diverse variable observations from which the values of the coefficients can be solved. In this example, all one needs is two distinct observations, say (q^1, p^1, x_1^1, x_2^1) and (q^2, p^2, x_1^2, x_2^2). Substituting these observations into the known forms of the equations one gets.

$$q^1 = \beta_1 p^1 + \alpha_1 x_1^1 ... supply1$$
$$q^1 = \beta_2 p^1 + \alpha_2 x_2^1 ... demand1$$

$$q^2 = \beta_1 p^2 + \alpha_1 x_1^2 ... supply2$$
$$q^2 = \beta_2 p^2 + \alpha_2 x_2^2 ... demand2$$

Since the equations are identifiable, these four equations can be solved for the four unknown coefficients.

However, as noted above, identifiability of a system of equations is insufficient for determining the causal relationships between quantity, price and the external factors. The method for determining coefficients does not ensure that the equations are rightly interpreted as structural,[67] that is, denote the correct causal order. Of course, if one knows that the equations do have correct causal order, then it is known that the measured coefficients are structural and do measure the strength of causal connections between factors.

---

[66] Recall that these are standard assumptions for the simple identification problems looked at in this chapter.

[67] Here, to simplify the discussion I introduce the term 'structural' to characterise equations read using the strong reading. So structural equations are those interpreted using the strong reading, while structural coefficients are coefficients in these equations.

So there are two important points here. The first is that the method for measuring coefficients above gives structural coefficients, that measure the strength of causal connections, only if background knowledge can be used to infer that the causal order of the set of equations is correct. If this condition is met, and if the set of equations are identifiable then the coefficients can be measured and interpreted as structural. Conversely, if there is insufficient background knowledge for knowing whether the equations denote the correct causal order then, if the equations are identifiable the coefficients can be measured but there is no guarantee that the coefficients are rightly interpreted as structural, as measuring the strength of causal connections.

This suggests that identifiability of systems of equations is tangential to finding out if functional equations can be causally interpreted, since to measure structural coefficients one must already know the causal order of the system *a priori*. If it is necessary to know the causal order *a priori* and the form of the functional equations that represent the causal order before identifiability of these equations can be exploited to measure coefficient values, then identifiability plays a small role in causal inference. Identifiability appears to be simply a condition on functional relations that, when these are known to correctly denote causal order, implies that the strengths of causal connections of that causal order can be inferred from observation.

However, there is a case in which one can make an inference *to* causal order using a system of identifiable equations. To see how, reconsider the two equations above.

$$q = \beta_1 p + \alpha_1 x_1 ... supply$$
$$q = \beta_2 p + \alpha_2 x_2 ... demand$$

As discussed above, where it is known that these equations hold and known that they denote the correct causal order, then structural coefficients can be inferred that measure the strengths of causal connections. There is a however, a small generalising move that can be made. Instead of assuming that the equations denote the correct causal order, one can weaken the assumption very slightly, and

assume that the equations are consistent with the correct structural equations, where one or more coefficients may be zero.[68]

In other words, instead of assuming that the equations above are known to hold and to be structural, assume instead that it is known that *whatever the structural equations are*, they will be of the form above, where some of the coefficients may be zero. In this case, this is equivalent to knowing that one of the sixteen following sets of structural equations in fact holds (in each of these coefficients are all non-zero):

$$(1)\ \begin{matrix} q = 0...supply \\ q = 0...demand \end{matrix} \qquad \text{OR} \qquad (2)\ \begin{matrix} q = 0...supply \\ q = \beta_2 p...demand \end{matrix}$$

$$\text{OR}\quad (3)\ \begin{matrix} q = \beta_1 p...supply \\ q = 0...demand \end{matrix} \qquad \text{OR} \qquad (4)\ \begin{matrix} q = \beta_1 p...supply \\ q = \beta_2 p...demand \end{matrix}$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$\text{OR}\quad (15)\ \begin{matrix} q = \beta_1 p...supply \\ q = \beta_2 p + \alpha_2 x_2...demand \end{matrix} \qquad \text{OR}\ (16)\ \begin{matrix} q = \beta_1 p + \alpha_1 x_1...supply \\ q = \beta_2 p + \alpha_2 x_2..demand \end{matrix}$$

However, one must be careful because some of the possible structural equations are *not* solvable for the internal variables in terms of external variables which is a condition for the systems of equations to be attributed causal order, that is, to be structural. For instance, system (4) is not solvable. Since these unsolvable systems have no causal interpretation, they are assumed not to be possible. A second problem is that some of the possible sets of equations are not identifiable. For instance, in system (15) the first equation is not identifiable. One cannot allow as possible unidentifiable systems because the method used to determine the value of coefficients assumes identifiability.[69]

[68] The idea here stems from the possibility that structural coefficients are inferred to be zero, so that one could revise the original belief that the equations are structural with non-zero coefficients, to assume that instead a more restricted set of structural equations holds. The proposal here revises the interpretation of the identifiable system of equations to permit this.

[69] Strictly speaking, this should be equation specific. One could allow systems with underidentified equations, provided the equation one is investigating is identifiable in all possible systems.

So here the set of possible systems must be restricted to those that are solvable and wholly identifiable.[70] This reduces the number of possibilities to the three.

$$\text{(A)} \begin{matrix} q = \beta_1 p + \alpha_1 x_1 ... supply \\ q = \alpha_2 x_2 ... demand \end{matrix} \quad \text{OR} \quad \text{(B)} \begin{matrix} q = \alpha_1 x_1 ... supply \\ q = \beta_2 p + \alpha_2 x_2 ... demand \end{matrix}$$

$$\text{OR} \quad \text{(C)} \begin{matrix} q = \beta_1 p + \alpha_1 x_1 ... supply \\ q = \beta_2 p + \alpha_2 x_2 ... demand \end{matrix}$$

At this stage, the earlier assumption that the causal order is known has been weakened to an assumption that it is known that true structural equations either have form (A), (B) or (C).[71] Since each of these is identifiable the coefficient values can be measured regardless of which system actually holds. So measuring the coefficients from observations, one can then determine which of the three systems above is the correct set of structural equations.

More generally, the logic of this method for inferring causal order is as follows.

(i) A set of *identifiable* linear equations is known to hold, and it is known that the true structural equations are identifiable and may be obtained by setting one or more, if any, of the coefficients in this general set of linear equations to zero.

(ii) A sufficiently varied set of observations for the variables is obtained, so that the coefficients of the general set of equations can be measured from observation.

THEN By measuring coefficients and finding out which if any are zero, one can deduce which of the possible systems of structural equations holds, and thus deduce the causal order.

Though this sets out a way in which identifiable system can be used to make an inference to causal order, it is important to note just how restrictive the assumptions are for being able to infer causal order in this way. One must know *a priori* that the true structural equations representing the mechanisms are identifiable and must be consistent with some known, identifiable general set of

[70] Again, if one is interested only in identifying one structural equation, then one could weaken this restriction to the set of systems in which that equation is identified.

[71] Where all of these have non-zero coefficients.

equations. This is particularly strong. It also is rather odd in some ways, for instance, how is one to know that the true structural equations are identifiable even if one does not know their form? Clearly this method would require some serious justification of these *a priori* claims in order to justify its use for inferring causal order. It is only a slight weakening of the previous situation where identifiable systems of equations can be used to infer structural coefficients, given the equations were known to have correct causal order.

To conclude this section, note that this method for inferring causal order provides an interpretation of a process for selecting an appropriate causal model from a known set of possibilities. In this way, this approach suitably generalised may provide an interpretative framework with which to interpret model selection methods that are current in econometrics. Though this would require a great deal more work, it is interesting to note the possibility. Model selections methods such as the LSE methodology[72] developed by David Hendry and others might be one particularly suitable candidate. This is because this method is based on working from general models to more specific ones by eliminating irrelevant variables, that is, those whose coefficients are inferred to be zero from observations. The similarity of the LSE methodology and the approach here of using inferred zero coefficients to select a causal model is obvious. In cases where the LSE methodology is used to select models that are to have a causal interpretation, a version of the approach above, developed to a much greater sophistication, might be able to help understand these model selection methods in a causally explicit way.

6. Conclusion

This chapter has attempted to relate one particular part of econometric methodology, identification, with more intuitive concepts of causal order and experiment. Focusing on the simplest case of linear deterministic sets of equations, it has begun by looking at relevant work by Herbert Simon, mapping out the relationships he claims hold between causal order and identifiability. In the process, I have clarified Simon's position as strongly operationalist. This

[72] For more on the LSE methodology, see David Hendry (2000).

position has been criticised as unnecessarily restrictive. It is not necessary, even from an operationalist perspective, to limit oneself to sets of equations whose causal order is uniquely specified, purely in virtue of excluded factors.

The second part of the chapter analysed what identification of a set of equations that is taken to denote a causal order requires of that causal order. Or in simpler terms, what constraints does identifiability place on causal orders? Using a theorem relating the rank condition to conditions for two-variable experiments, it was shown that 'identifiable' causal orders must either have limited common causes among factors, or have sufficiently many causal inputs to ensure that the effect of common causes can be suitably cancelled out to permit causal inference. This provided an alternative, causally explicit reading of identification which differs from the typically a-causal mathematical presentation of identification found in econometrics textbooks.

The last section looked at the role of identification in causal inference. Importantly, it was seen that identifiability, in order to be useful for learning about the strength of causal connections, needed to be supplemented with strong background knowledge for determining the causal order. In an extension of the use of identifiable sets of equations to determine structural coefficients, it was shown that identifiable sets of equations could also be used to make inferences to causal order. However, this method also depended on very strong background assumptions, in particular, an assumption that it is known what the possible sets of structural relations are and that these are all consistent with a known identifiable system of equations.

Appendix 5.1. Simon's Exclusion Condition Implies Identifiability

Theorem 5.1 Given an incomplete set of equations that meets Simon's exclusion condition then the rank condition holds for any equation.

Notation: The incomplete set of equations are given by:

$$\underset{nxn}{A}\ \underset{nx1}{y} = \underset{nxm}{B}\ \underset{mx1}{x} + \underset{nx1}{c}$$

$$\Leftrightarrow (A \mid -B)\begin{pmatrix} y \\ x \end{pmatrix} = c$$

where

y is the $nx1$ vector of internal variables.

x is the $mx1$ vector of external variables.

A is the nxn non singular matrix of (constant) coefficients.

B is a mxn matrix of (constant) coefficients.

c is the $nx1$ vector of (constant) intercept coefficients.

Proof:

Assume without loss of generality that we are interested in the identifiability of the last equation.

Now, since the structural equations meet the exclusion condition, there is a unique external variable that appears in that equation. So, assume without loss of generality that for all i: x_i is the unique equation in the i^{th} equation.[73]

In that case the structural equations have form:

$$Ay = (D \mid F)\begin{pmatrix} x_{unique} \\ x_{other} \end{pmatrix} + c$$

[73] At this point, by assigning particular variables to particular equations, I am assuming Simon's version of the exclusion condition, as discussed in the chapter.

Where x_{unique} is the $nx1$ vector of the exclusive external variables, x_{other} is the $(m-n)$x1 vector of other external variables. D is a diagonal matrix, with non-zero diagonal elements of order n. F the submatrix for the 'remainder' of B i.e. $B = (D|F)$.

Now, to apply the rank condition for the last equation we construct the submatrix of $(A|-B)$ formed by of the columns that correspond to excluded variables for the last equation. We also exclude the row for the last equation. In this case the resulting submatrix must include the first $n-1$ rows and columns of D, since these external variables are excluded from the last equation, by the exclusion condition. This implies that the submatrix contains a diagonal matrix of order $n-1$ (which has non-zero diagonal elements) so it has rank of at least $n-1$. But the submatrix only has $n-1$ rows (since it contains coefficients from the other $n-1$ equations), so it has rank of at most $n-1$. It follows that the submatrix has rank $n-1$. This is the rank condition (see Appendix 5.2), so the result follows. ■

Appendix 5.2. An Alternative Necessary and Sufficient Condition for Identification

Theorem 5.2: Given an incomplete set of equations, the rank condition holds for an equation if and only if a two-variable experiment is possible between any two variables in that equation.

Definition 'Two-variable experiment': Let z_1 and z_2 be two variables that appear in a equation, a two-variable experiment occurs for z_1 and z_2 in that equation if and only if

(i) All variables in the equation except z_1 and z_2 do not change.
i.e. $\Delta z_k = 0$ *for all* z_k *in the equation such that* $k \neq 1$ *and* $k \neq 2$

(ii) At least one of z_1 and z_2 changes i.e.
i.e. $\Delta z_1 \neq 0$ *or* $\Delta z_2 \neq 0$

Finally, a two-variable experiment for z_1 and z_2 is possible for z_1 and z_2 in an equation if and only if there exist a set of changes in external variables $\{\Delta x_1, ..., \Delta x_m\}$ which imply that a two-variable experiments occurs for z_1 and z_2 in their equation.

Notation:

The incomplete set of equations is given by:

$$\underset{nxn}{A}\ \underset{nx1}{y} = \underset{nxm}{B}\ \underset{mx1}{x} + \underset{nx1}{c}$$

$$\Leftrightarrow (A \mid -B)\begin{pmatrix} y \\ x \end{pmatrix} = c$$

where

y is the $nx1$ vector of internal variables.

x is the $mx1$ vector of external variables.

A is the nxn non-singular matrix of (constant) coefficients. [74]

B is a mxn matrix of (constant) coefficients.

[74] A is invertible because the systems of equations analysed are solvable for the internal variables in terms of the external variables. In other words, the structural equations $Ax + By = c$ are solvable uniquely for the y's in terms of the x's. This is one of the conditions for the systems of equations to be attributed causal order, using Simon's method.

c is the $nx1$ vector of (constant) intercept coefficients.

The corresponding reduced form is given by:

$$\underset{nx1}{y} = \underset{nxm}{C}\ \underset{mx1}{x} + d$$

C is the mxn reduced form matrix, and $C = A^{-1}B$.

d is the $nx1$ vector of the reduced form intercept coefficients $d=A^{-1}c$

Assume without loss of generality that the equation we are concerned with is the first equation. Divide up the internal and external variables into those that are included in the equation and those that are excluded. Let

k_x = number of external variables excluded from the equation.

k_y = number of internal variables excluded from the equation.

Assume also without loss of generality that the included variables are labelled with lower indices than the excluded variable. With this set up, we can partition the matrices of the structural form as follows.

$$\begin{pmatrix} \underset{1x(n-k_y)}{A_{11}} & | & \underset{1xk_y}{A_{12}} \\ - & + & - \\ \underset{(n-1)x(n-k_y)}{A_{21}} & | & \underset{(n-1)xk_y}{A_{22}} \end{pmatrix} \begin{pmatrix} \underset{(n-k_y)x1}{y_{incl}} \\ - \\ \underset{k_yx1}{y_{excl}} \end{pmatrix} = \begin{pmatrix} \underset{(n-k_y)x(m-k_x)}{B_{11}} & | & \underset{(n-k_y)xk_x}{B_{12}} \\ - & + & - \\ \underset{(n-1)x(m-k_x)}{B_{21}} & | & \underset{(n-1)xk_x}{B_{22}} \end{pmatrix} \begin{pmatrix} \underset{(m-k_x)x1}{x_{incl}} \\ - \\ \underset{k_xx1}{x_{excl}} \end{pmatrix} + \begin{pmatrix} \underset{(n-k_y)x1}{c_{incl}} \\ - \\ \underset{k_yx1}{c_{excl}} \end{pmatrix}$$

However, since the excluded variables do not appear in the first equation, we must have $A_{11} = 0$ and $B_{12} = 0$. So the structural form is, by construction:

$$\begin{pmatrix} A_{11} & | & 0 \\ - & + & - \\ A_{21} & | & A_{22} \end{pmatrix} \begin{pmatrix} y_{incl} \\ - \\ y_{excl} \end{pmatrix} = \begin{pmatrix} B_{11} & | & 0 \\ - & + & - \\ B_{21} & | & B_{22} \end{pmatrix} \begin{pmatrix} x_{incl} \\ - \\ x_{excl} \end{pmatrix} + \begin{pmatrix} c_{incl} \\ - \\ c_{excl} \end{pmatrix}$$

Also partition the reduced form as follows.

$$\begin{pmatrix} y_{incl} \\ (n-k_y)x1 \\ - \\ y_{excl} \\ k_y x1 \end{pmatrix} = \begin{pmatrix} C_{11} & | & C_{12} \\ (n-k_y)x(m-k_x) & & (n-k_y)xk_x \\ - & + & - \\ C_{21} & | & C_{22} \\ k_y x(m-k_x) & & k_y xk_x \end{pmatrix} \begin{pmatrix} x_{incl} \\ (m-k_x)x1 \\ - \\ x_{excl} \\ k_x x1 \end{pmatrix} + \begin{pmatrix} d_{incl} \\ (n-k_y)x1 \\ - \\ d_{excl} \\ k_y x1 \end{pmatrix}$$

By definition of the reduced form, $C=A^{-1}B$, this is equivalent to $B=AC$, so in our partitioned versions of the matrices we have.

$$\begin{pmatrix} B_{11} & | & 0 \\ - & + & - \\ B_{21} & | & B_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & | & 0 \\ - & + & - \\ A_{21} & | & A_{22} \end{pmatrix} \begin{pmatrix} C_{11} & | & C_{12} \\ - & + & - \\ C_{21} & | & C_{22} \end{pmatrix}$$

$$\begin{pmatrix} B_{11} & | & 0 \\ - & + & - \\ B_{21} & | & B_{22} \end{pmatrix} = \begin{pmatrix} A_{11}C_{11} & | & A_{11}C_{12} \\ - & + & - \\ A_{21}C_{11} + A_{22}C_{21} & | & A_{21}C_{12} + A_{22}C_{22} \end{pmatrix}$$

So we must have for the matrices as partitioned that $A_{11}C_{12} = 0$

Note that with this set up the Rank Condition for identification is

The Rank Condition for Identification.

An equation is identifiable if and only if $Rank\ (A_{22}|-B_{22}) = n-1$, where $(A_{22}|-B_{22})$ is the submatrix of $(A|-B)$ formed by dropping the row (in A and B) corresponding to the equation and dropping any columns (in A and B) containing a variable included in the equation.

Before proving the theorem, we start with two useful preliminary lemmas.

Lemma 1 A matrix, M_0, which has the form (I is the identity matrix of order k, 0 a zero matrix)

$$M_0 = \begin{pmatrix} I & | & 0 \\ - & + & - \\ M_1 & | & M_2 \end{pmatrix}$$

satisfies

$$Rank(M_0) = k + Rank(M_2)$$

Proof Consider the two submatrices of M_0: $\begin{pmatrix} I \\ - \\ M_1 \end{pmatrix}, \begin{pmatrix} 0 \\ - \\ M_2 \end{pmatrix}$

No column in the first can be a linear combination of the columns in the second because the 1's in the identity matrix in the first submatrix cannot be generated by linearly combining the zeros in the upper part of the second submatrix. In addition, the columns of the first matrix are linearly independent (because of the identity matrix). This implies that adding the first submatrix to the second (to form M_0) gives us a columnspace of M_0 whose dimension is the sum of the dimensions of the columnspaces of the two submatrices, that is

$$Dim\ Colspace\ (M_0) = Dim\ Colspace \begin{pmatrix} I \\ - \\ M_1 \end{pmatrix} + Dim\ Colspace \begin{pmatrix} 0 \\ - \\ M_2 \end{pmatrix}$$

But this is simply

$$Rank(M_0) = Rank \begin{pmatrix} I \\ - \\ M_1 \end{pmatrix} + Rank \begin{pmatrix} 0 \\ - \\ M_2 \end{pmatrix}$$

But $Rank \begin{pmatrix} I \\ - \\ M_1 \end{pmatrix} = k$ & $Rank \begin{pmatrix} 0 \\ - \\ M_2 \end{pmatrix} = Rank(M_2)$

The first follows because it has k linearly independent columns from the identity matrix while the second is immediate. Substituting then gives the desired result.

$$Rank(M_0) = k + Rank(M_2)\ \square$$

[*Aside:* It is also straightforward to prove that the same conclusion holds where M_0 has a form where the *0* and *I* matrix are appear in different parts of the matrix

(provided that M_2 is diametrically opposite to the identity submatrix). For example, the result holds for M_0 of the form:

$$\left(\begin{array}{ccc} I & | & M_1 \\ - & + & - \\ 0 & | & M_2 \end{array}\right),\left(\begin{array}{ccc} 0 & | & M_2 \\ - & + & - \\ I & | & M_1 \end{array}\right),\left(\begin{array}{ccc} M_2 & | & 0 \\ - & + & - \\ M_1 & | & I \end{array}\right),\left(\begin{array}{ccc} M_2 & | & M_1 \\ - & + & - \\ 0 & | & I \end{array}\right),\left(\begin{array}{ccc} 0 & | & I \\ - & + & - \\ M_2 & | & M_1 \end{array}\right) etc.]$$

Lemma 2. If z_1, z_2, z_{fixed} (a vector of variables), and x a vector of a set of variation free variables (partitioned into two vectors x_{fixed} and x_{free}) satisfy the following linear difference matrix equation.

$$\begin{pmatrix} \Delta z_1 \\ \Delta z_2 \\ - \\ \Delta z_{fixed} \end{pmatrix} = \left(\begin{array}{ccc} M_{11,1} & | & M_{12,1} \\ - & + & - \\ M_{11,2} & | & M_{12,2} \end{array}\right)\begin{pmatrix} \Delta x_{fixed} \\ - \\ \Delta x_{free} \end{pmatrix} \qquad (+)$$

Then it is possible that $\begin{pmatrix} \Delta z_1 \\ \Delta z_2 \end{pmatrix} \neq 0$ while Δz_{fixed} and $\Delta x_{fixed} = 0$, if and only if

$$Rank(M_{12}) > Rank(M_{12,2}) \quad where \quad M_{12} = \begin{pmatrix} M_{12,1} \\ - \\ M_{12,2} \end{pmatrix}$$

Proof:

Given (+) it is possible that $\begin{pmatrix} \Delta z_1 \\ \Delta z_2 \end{pmatrix} \neq 0$ while Δz_{fixed} and $\Delta x_{fixed} = 0$, if and only if

$$\begin{pmatrix} \Delta z_1 \\ \Delta z_2 \end{pmatrix} = M_{12,1}\Delta x_{free} \neq 0 \text{ and } \quad \Delta z_{fixed} = 0 = M_{12,2}\Delta x_{free}$$

These together are equivalent to

$\Delta x_{free} \notin Nullspace(M_{12,1})$

and

$\Delta x_{free} \in Nullspace(M_{12,2})$

[*Aside:*

By the fundamental theorem of linear algebra[75] for any linear transformation M, we have:

[75] This is a well known result in linear algebra, see Strang (1980, pp.84-88) for a discussion.

$$Nullspace(M) \cup Colspace(M^T) = Domain(M)$$

and

$$Nullspace(M) \cap Colspace(M^T) = \{0\}$$

Where the *colspace*(M^T) is the *rowspace(M)* with all of its vectors transposed[76])
End of Aside]

So using the Fundamental Theorem, the first of the condition above is equivalent to

$$\Delta x_{free} \in Colspace(M_{12,1}^T) \; and \; \Delta x_{free} \neq 0$$

While the second is equivalent to

$$\Delta x_{free} \notin Colspace(M_{12,2}^T) \; or \; \Delta x_{free} = 0$$

But since the set of columns in $M_{12}{}^T$ is the union of the columns in $M_{12,1}{}^T$ and $M_{12,2}{}^T$

$$Colspace(M_{12}^T) = Colspace(M_{12,1}^T) \cup Colspace(M_{12,2}^T)$$

Given this, the two conditions above hold if and only if

$$\Delta x_{free} \in Colspace(M_{12}^T) \setminus Colspace(M_{12,2}^T)$$

Since the *x's* are variation free, Δx_{excl} can take any value so this holds if and only if

$$Colspace(M_{12}^T) \setminus Colspace(M_{12,2}^T) \neq \Phi$$

Which holds if and only if

$$Colspace(M_{12}^T) \neq Colspace(M_{12,2}^T)$$

This holds (transposing vectors) if and only if

$$Rowspace(M_{12}) \neq Rowspace(M_{12,2})$$

Given that $M_{12,2}$ is a proper submatrix consisting of rows of M_{12} this holds if & only if

$$Rank(M_{12}) > Rank(M_{12,2}) \quad \square$$

[76] Transposing is necessary because the vectors in the nullspace are column vectors whereas the vectors of the rowspace are row vectors.

Lemma 2 is useful because if the x's are the external variables, z_1 and z_2 are two variables in an equation and z_{fixed} are the other variables in that equation, and the matrix equation (+) follows from the structural equations, then it gives a necessary and sufficient condition in terms of matrices for a two-variable experiment is possible for z_1 and z_2 in the equation in which they appear.

With these preliminary lemmas, we can now begin the proof of the theorem. First we prove an alternative rank condition for identification.

Lemma 3 (A Reduced Form Rank Condition): The Rank condition for an equation holds if and only if $Rank(C_{12})=n-k_y-1$ for C_{12} defined as above.[77]

Proof:

The rank condition is $Rank(A_{22}|-B_{22}) = n-1$. So first we establish a useful identity involving the matrix $(A_{22}|-B_{22})$.

To do this, note first that it follows from $B=AC$ that

$$B_{12} = A_{11}C_{12}+A_{12}C_{22}$$

$$B_{22} = A_{21}C_{12}+A_{22}C_{22}$$

But from above we know $A_{12} = 0$ and $B_{12}=0$ so we have

$$B_{12} = A_{11}C_{12} = 0$$

So

$$\begin{pmatrix} 0 & | & 0 \\ - & + & - \\ A_{22} & | & -B_{22} \end{pmatrix} = \begin{pmatrix} 0 & | & -A_{11}C_{12} \\ - & + & - \\ A_{22} & | & -(A_{21}C_{12}+A_{22}C_{22}) \end{pmatrix}$$

But the right hand side is equal to

[77] This result was first proved by Koopmans and Hood (1953). Fisher also has a proof (1966, p.54). I include my own proof here for completeness.

$$\begin{pmatrix} 0 & | & -A_{11}C_{12} \\ - & + & - \\ A_{22} & | & -(A_{21}C_{12}+A_{22}C_{22}) \end{pmatrix}$$
$$= \begin{pmatrix} A_{11} & | & 0 \\ - & + & - \\ A_{21} & | & A_{22} \end{pmatrix} \begin{pmatrix} 0 & | & -C_{12} \\ - & + & - \\ I & | & -C_{22} \end{pmatrix} = A \begin{pmatrix} 0 & | & -C_{12} \\ - & + & - \\ I & | & -C_{22} \end{pmatrix}$$

Substituting,

$$\begin{pmatrix} 0 & | & 0 \\ - & + & - \\ A_{22} & | & -B_{22} \end{pmatrix} = A \begin{pmatrix} 0 & | & -C_{12} \\ - & + & - \\ I & | & -C_{22} \end{pmatrix}$$

Since A is invertible it preserves rank, so it follows that

$$Rank \begin{pmatrix} 0 & | & 0 \\ - & + & - \\ A_{22} & | & -B_{22} \end{pmatrix} = Rank \begin{pmatrix} 0 & | & -C_{12} \\ - & + & - \\ I & | & -C_{22} \end{pmatrix}$$

But the rank of the right hand side matrix is simply $rank(A_{22}|\text{-}B_{22})$, so

$$Rank(A_{22} \mid -B_{22}) = Rank \begin{pmatrix} 0 & | & -C_{12} \\ - & + & - \\ I & | & -C_{22} \end{pmatrix}$$

The identity in the right hand side matrix is of order k_y, so applying lemma 1, we have

$$Rank \begin{pmatrix} 0 & | & -C_{12} \\ - & + & - \\ I & | & -C_{22} \end{pmatrix} = Rank(C_{12}) + k_y$$

Substituting, we get

$$Rank(A_{22} \mid -B_{22}) = Rank(C_{12}) + k_y$$

It then follows that

$$Rank(A_{22} \mid -B_{22}) = n-1 \Leftrightarrow Rank(C_{12}) = n - k_y - 1 \quad \square$$

We can now prove the theorem.

Theorem 5.2: Given an incomplete set of equations, the rank condition holds for an equation if and only if a two-variable experiment is possible between any two variables in that equation.

Proof

Part I '*Only if*' (Rank Condition→ Experiment possible between any two variables).

There are three distinct cases, when one picks two variables from an equation to see if an experiment is possible between them.

(i) Both variables are internal.

(ii) One is internal the other external.

(iii) Both variables are external.

Case(i):

For any two internal variables y_i and y_j, assume without loss of generality (simply by reordering the indices on the variables) that they are y_1 and y_2. The reduced form difference equations for the internal variables that appear in the equation is given by[78]

$$\Delta y_{incl} = C_{11}\Delta x_{incl} + C_{12}\Delta x_{excl}$$

To simplify the analysis, partition y_{incl} as follows:

$$y_{incl} = \begin{pmatrix} y_1 \\ y_2 \\ - \\ y_{fixed} \end{pmatrix}$$

Where y_{fixed} denotes the internal variables other than y_1 and y_2 in the equation, which are labelled as 'fixed' since these would not change in the relevant possible experiment.

Correspondingly we can partition the reduced form (here the difference) equations for the variables that appear in the equation of interest as follows:

[78] In what follows the 'excl' subscript applies to variables that do not appear in the structural equation and 'incl' to those that do.

$$\begin{pmatrix} \Delta y_1 \\ \Delta y_2 \\ - \\ \Delta y_{fixed} \end{pmatrix} = \begin{pmatrix} C_{11,1} & | & C_{12,1} \\ - & + & - \\ C_{11,2} & | & C_{12,2} \end{pmatrix} \begin{pmatrix} \Delta x_{incl} \\ - \\ \Delta x_{excl} \end{pmatrix} \qquad \ldots \qquad (/)$$

$$\text{where} \quad \underset{(n-k_y)x(m-k_x)}{C_{11}} = \begin{pmatrix} \underset{2x(m-k_x)}{C_{11,1}} \\ - \\ \underset{(n-k_y-2)x(m-k_x)}{C_{11,2}} \end{pmatrix} \quad \text{and} \quad \underset{(n-k_y)xk_x}{C_{12}} = \begin{pmatrix} \underset{2xk_x}{C_{12,1}} \\ - \\ \underset{(n-k_y-2)xk_x}{C_{12,2}} \end{pmatrix}$$

Now (/) has the form appropriate for lemma 2, so an experiment is possible between y_1 and y_2 if and only if $Rank(C_{12}) > Rank(C_{12,2})$, call this condition (*).

Now, the rank condition holds so, by lemma 3, $Rank(C_{12}) = n\text{-}k_y\text{-}1$. In addition, $C_{12,2}$ has $n\text{-}k_y\text{-}2$ rows so $rank\ (C_{12,2}) \leq n\text{-}k_y\text{-}2$, so it follows that $Rank(C_{12,2}) < Rank\ (C_{12})$. In other words, condition (*) is met, and an experiment is possible between the two chosen internal variables. □

Case (ii):

Assume without loss of generality that we want to show an experiment is possible between y_1 and x_1 (we can always change indices otherwise). Then we can split up the reduced form as.

$$\begin{pmatrix} \Delta y_1 \\ - \\ \Delta y_{fixed} \end{pmatrix} = \begin{pmatrix} \underset{1xm-k_x-1}{C_{11,1}} & | & \underset{1x1}{\gamma_{11,1}} & | & \underset{1xk_x}{C_{12,1}} \\ & + & & + & - \\ \underset{(n-k_y-1)xm-k_x-1}{C_{11,2}} & | & \underset{(n-k_y-1)x1}{\gamma_{11,2}} & | & \underset{(n-k_y-1)xk_x}{C_{12,2}} \end{pmatrix} \begin{pmatrix} \Delta x_{fixed} \\ - \\ \Delta x_1 \\ \Delta x_{excl} \end{pmatrix}$$

where

$$C_{11} = \begin{pmatrix} \underset{1xk_x}{C_{11,3}} & | & \underset{1x1}{\gamma_{11,1}} \\ - & + & - \\ \underset{(n-k_y-1)xk_x}{C_{11,4}} & | & \underset{(n-k_y-1)x1}{\gamma_{11,2}} \end{pmatrix}, \; C_{12} = \begin{pmatrix} \underset{1xk_x}{C_{12,1}} \\ - \\ \underset{(n-k_y-1)xk_x}{C_{12,2}} \end{pmatrix} \text{ and } \Delta x_{incl} = \begin{pmatrix} \underset{1x1}{\Delta x_1} \\ - \\ \underset{m-k_x-1X1}{\Delta x_{fixed}} \end{pmatrix}$$

Adding the equation for $\Delta x_1 = \Delta x_1$ to the matrix equation above we get.

$$\begin{pmatrix} \Delta x_1 \\ \Delta y_1 \\ - \\ \Delta y_{fixed} \end{pmatrix} = \begin{pmatrix} 0 & | & 1 & | & 0 \\ C_{11,1} & | & \gamma_{11,1} & | & C_{12,1} \\ 1xm-k_x-1 & | & 1x1 & | & 1xk_x \\ & + & & + & - \\ C_{11,2} & | & \gamma_{11,2} & | & C_{12,2} \\ (n-k_y-1)xm-k_x-1 & | & (n-k_y-1)x1 & | & (n-k_y-1)xk_x \end{pmatrix} \begin{pmatrix} \Delta x_{fixed} \\ - \\ \Delta x_1 \\ \Delta x_{excl} \end{pmatrix}$$

This has the form appropriate for lemma 2, so an experiment is possible between x_1 and y_1 is possible if and only if.

$$Rank\begin{pmatrix} 1 & | & 0 \\ - & + & - \\ \gamma_{11,1} & | & C_{12,1} \\ - & + & - \\ \gamma_{11,2} & | & C_{12,2} \end{pmatrix} > Rank\begin{pmatrix} \gamma_{11,2} & | & C_{12,2} \end{pmatrix}$$

By lemma 1 the left hand side equals

$$Rank\begin{pmatrix} 1 & | & 0 \\ - & + & - \\ \gamma_{11,1} & | & C_{12,1} \\ - & + & - \\ \gamma_{11,2} & | & C_{12,2} \end{pmatrix} = Rank\begin{pmatrix} C_{12,1} \\ - \\ C_{12,2} \end{pmatrix} + 1 = Rank(C_{12}) + 1$$

Substituting we get

Condition**: An two variable experiment is possible between y_1 and x_1 if and only if $Rank(C_{12}) + 1 > Rank\begin{pmatrix} \gamma_{11,2} & | & C_{12,2} \end{pmatrix}$

Since the rank condition is met $Rank(C_{12}) = n-k_y-1$, so substituting this into condition ** an experiment is possible if and only if

$$n - k_y > Rank\begin{pmatrix} \gamma_{11,2} & | & C_{12,2} \end{pmatrix}$$

Now the matrix $\begin{pmatrix} \gamma_{11,2} & | & C_{12,2} \end{pmatrix}$ has $n-k_y-1$ rows and so

$$Rank\begin{pmatrix} \gamma_{11,2} & | & C_{12,2} \end{pmatrix} \leq n - k_y - 1$$

i.e. $$Rank\begin{pmatrix} \gamma_{11,2} & | & C_{12,2} \end{pmatrix} < n - k_y$$

So condition ** holds and the two variable experiment is possible. We have shown that if the rank condition is met for an equation then the experiment is possible between any internal and external variable in the equation.□

Case (iii)

In this case assume without loss of generality that the two external variables we are considering are x_1 and x_2. Using the same set up as in Case (ii), the situation can be represented by

$$\begin{pmatrix}\Delta x_1\\ \Delta x_2\end{pmatrix}=\left(\begin{array}{cc|c}1 & 0 & 0\\ 0 & 1 & \end{array}\right)\begin{pmatrix}\Delta x_1\\ \Delta x_2\\ -\\ \Delta x_{excl}\end{pmatrix}$$

and $$\Delta y_{incl}=\Delta y_{fixed}=\left(\gamma_{11}\ \middle|\ C_{12}\right)\begin{pmatrix}\Delta x_1\\ \Delta x_2\\ -\\ \Delta x_{excl}\end{pmatrix}$$

Putting these two equations into one we get

$$\begin{pmatrix}\Delta x_1\\ \Delta x_2\\ \Delta y_{fixed}\end{pmatrix}=\begin{pmatrix}1 & 0 & | & 0\\ 0 & 1 & & \\ - & & + & -\\ \gamma_{11} & & | & C_{12}\end{pmatrix}\begin{pmatrix}\Delta x_1\\ \Delta x_2\\ -\\ \Delta x_{excl}\end{pmatrix}$$

This has a form appropriate for lemma 2, so an experiment is possible between x_1 and x_2 if and only if.

$$Rank\begin{pmatrix}1 & 0 & | & 0\\ 0 & 1 & & \\ - & & + & -\\ \gamma_{11} & & | & C_{12}\end{pmatrix}>Rank\left(\gamma_{11}\ \middle|\ C_{12}\right)$$

By lemma 1, the left hand side satisfies

$$Rank\begin{pmatrix}1 & 0 & | & 0\\ 0 & 1 & & \\ - & & + & -\\ \gamma_{11} & & | & C_{12}\end{pmatrix}=Rank\left(C_{12}\right)+2$$

Substituting we get:

Condition***: A two-variable experiment is possible between y_1 and x_1 if and only if $Rank(C_{12})+2>Rank\left(\gamma_{11}\ \middle|\ C_{12}\right)$

Now the matrix $(\gamma_{11} \mid C_{12})$ has $n\text{-}k_y$ rows. So

$$n - ky + 1 > Rank(\gamma_{11} \mid C_{12})$$

Since the rank condition holds, by lemma 3, $n\text{-}k_y\text{-}1 = Rank(C_{12})$. So

$$Rank(C_{12}) + 2 > Rank(\gamma_{11} \mid C_{12})$$

In other words, condition (***) is met. So we have shown that if the rank condition holds then an experiment is possible between any two external variables.

Since the rank condition implies a two variable experiment is possible in cases (i), (ii) and (iii), it covers all possible cases and the rank condition implies an experiment is possible between any two variables in the equation. □

Part II '*If*' (Experiment possible between any two variables → Rank Condition)

If an experiment is possible for any two variables then for all possible $C_{12,2}$ submatrices[79] of C_{12}, we have $Rank(C_{12,2}) < Rank\ (C_{12})$ from lemma 2. Since $C_{12,2}$ is a submatrix of C_{12} consisting of rows of C_{12} this implies

$$Rowspace(C_{12}) \neq Rowspace(C_{12,2})$$

We first show that this is not possible if $Rank(C_{12}) < n\text{-}k_y\text{-}1$.

If $Rank(C_{12}) < n\text{-}k_y\text{-}1$ then there are two variables that have a $C_{12,2}$ such that the rows of $C_{12,2}$ span the rowspace of C_{12}. This must be the case since if the rowspace of C_{12} has dimension less than $n\text{-}k_y\text{-}1$, it follows that one can pick a subset of at most $n\text{-}k_y\text{-}2$ rows of C_{12} to span its rowspace. Then one can simply pick the two variables so that $C_{12,2}$ (which consists of $n\text{-}k_y\text{-}2$ rows) contains those rows that span C_{12}. In that case $Rowspace(C_{12}) = Rowspace(C_{12,2})$ which implies an experiment is not possible, a contradiction. Therefore, $Rank(C_{12}) \geq n\text{-}k_y\text{-}1$.

So if an experiment is possible for any two variables then $Rank\ (C_{12}) \geq n\text{-}k_y\text{-}1$.

Now C_{12} has $n\text{-}k_y$ rows, so $Rank\ (C_{12}) \leq n\text{-}k_y$.

[79] Since the choice of two variable in an equation determines the rows in C_{12} that are rows in $C_{12,2}$.

Therefore, we must have $n\text{-}k_y\text{-}1 \leq Rank\ (C_{12}) \leq\ n\text{-}k_y$

Finally, note that $Rank(C_{12}) \neq n\text{-}k_y$. This immediate because given that $A_{11}C_{12} = 0$ for $A_{11} \neq 0$ (see preliminaries) so the rows of C_{12} are linearly dependent, therefore the dimension of its rowspace, its rank, is less than its number of rows so $Rank\ (C_{12})<n\text{-}k_y$.

We have shown that $n\text{-}k_y\text{-}1 \leq Rank\ (C_{12}) < n\text{-}k_y$ so $Rank(C_{12}) = n\text{-}k_y\text{-}1$.
Applying lemma 3, the rank condition holds for the equation. ■

Chapter 6

Deducing Causal Order from Observation: Herbert Simon and Nancy Cartwright

1. Introduction

The last chapter finished with a brief analysis of a way in which identifiable systems of equations can be used to make inferences about causal orders.[1] This chapter continues this analysis by looking in more detail at the conditions required for making inferences about causal order. Specifically, the chapter focuses on two ways of doing this. The first approach is set out by Simon in his 1954 paper on spurious correlation. The second, alternative approach is set out by Nancy Cartwright in chapter one of her 1989. In both of these, the logic of the inferential method is deductive, that is, new knowledge claims are deduced from observations and existing knowledge claims. This approach to testing and warranting hypotheses is what Clark Glymour and Nancy Cartwright call 'bootstrapping'.[2]

The problem of how to infer causal order is discussed in many forms and in this chapter it is discussed under three slightly different terminologies. The first and most explicit terminology is simply to describe it as the problem of inferring causes from correlations, probabilities or observations. This is the way Nancy Cartwright (1989) discusses the problem. A second way the problem is discussed is as 'the problem of spurious correlation'. In this form the problem is how to tell if a correlation between two factors is due to a direct causal relationship or instead due to something else, such as common causal factors. This is a variant of the first problem where the focus is on two factors. Herbert Simon's 1954 paper, discussed here, aims to provide a solution to this problem. The third and final way the problem is presented in this chapter is as 'the problem of observational equivalence'. In this case the problem is how to distinguish between different

[1] As elsewhere in the thesis, causal order does not simply mean the order of causation between two factors, it is a term that denotes which factors directly cause which for a set of factors. Though 'causal structure' may be a better term, I use causal order in line with Simon's usage. For a definition of causal order, see chapter two.

[2] This is not the same way that 'bootstrapping' is used in statistics, see Cartwright (1989, p.22).

causal orders that are consistent with the same observations. Clearly, this is the problem of inferring causal order put in a different way. Instead of asking what causal order generated observations *tout court*, it asks 'which of the causal orders that are consistent with the observations (observationally equivalent causal orders) generated the observations?' Obviously, any solution to the problem of inferring causal order must be able to solve this problem of observational equivalence, since it must be able to rule out all but one causal order as being responsible for observations.[3]

This chapter begins with a discussion of the solution Simon gives to the problem of spurious correlation in his 1954 paper. In that paper Simon's key claim is that the causal order among internal variables in a linear system of equations can be deduced given knowledge of uncorrelated error terms and of a strict time order for the internal variables. This claim, however, is criticised by Nancy Cartwright as unsuccessful. Given this, I explore the limits of Simon's key claim by attempting to construct counterexamples to it. From this analysis, it is seen that Simon's key claim *seems* to hold provided one takes the set of internal variables in the equations as given. However, I say 'seems to allow a unique causal order' because there is a deeper problem for Simon. The problem is that no matter how many internal variables are introduced into the set of equations, it is always possible that the causal relations asserted by the system of equations are spurious. In this way, Simon's claim is undermined. To avoid the problems Simon's approach faces, an alternative 'S-approach' using the strong reading is suggested. This is very close to Simon's approach and avoids Cartwright's criticism, but does so at the price of making some very strong assumptions about background knowledge. The chapter then presents and criticises Nancy Cartwright's alternative approach for inferring causal order. The final section briefly compares the strong reading approach with Cartwright's.

[3] For the 'bootstrapping' logic of causal inference assumed here, this requires that one be able to *deduce* the correct causal order from observations and background knowledge. Other approaches to inference may treat the problem of inferring a unique causal hypothesis differently, for instance, the hypothetico-deductive (H-D) method. With the H-D method the set of possible hypotheses is always *logically* underdetermined by observation. Therefore, for this method of testing hypotheses some other principle, such as a principle of induction, could be used to choose (here inductively rather than deductively) a unique hypothesis from the different observationally equivalent hypotheses.

2. Simon's Method for Inferring Causal Order from Correlations

In his 1954 paper, Herbert Simon shows how one can solve the problem of spurious correlation, provided certain *a priori* conditions are met. In this section, I briefly describe his analysis in the paper and set out his key claim that causal relations can be deduced from knowledge of correlations and time ordering.

Before discussing Simon, it is important to note some differences in the systems of equations Simon analyses in his 1954 paper from those looked at in his 1953 paper. First, the 1954 paper restricts analysis to cases where there is a strict time order between causally ordered factors. For these systems, one factor can causally precede another only if it occurs earlier in time. This implies that the complete subsets in his causal orders now contain only one factor and that the systems of equations analysed can be reordered to be lower triangular.[4] Second, unlike his earlier paper, Simon's models here have an error term in each equation to cover omitted factors. In his 1954 analysis, the direct control passes through the error terms.[5]

2.1. Simon's Solution to the Problem of Spurious Correlation

Herbert Simon's (1954) focuses on recursive models of the following kind, where the z's are internal variables written with indices that indicate their time order and the u's are the error terms that denote omitted factors.

$$z_1 = u_1$$
$$z_2 = a_{21}z_1 + u_2$$
$$\vdots$$
$$z_n = a_{n1}z_1 + \ldots + a_{nn-1}z_{n-1} + u_n$$

These equations are to be read using his causal ordering method of his 1953 paper, treating the error terms as if they are external variables. In such models, it is easy to show from Simon's definition of causal order that one variable, z_i, is a

[4] In contrast with the simultaneous equation systems, which Simon considers in his (1953) and analysed in the previous chapters of the thesis, whose equations can be reordered to be *block* triangular.

[5] As in the previous chapter, I reformulate Simon's systems so that his variables are treated as internal variables and his coefficients are treated as external variables. With this relabelling, Simon's paper considers systems that contain internal variables but no external variables. In these systems the error terms are source of 'direct control' into these systems; they are treated by Simon as if they are external variables.

direct cause of another, z_j, if and only if in the equation with z_j on the left hand side, z_i appears with a non-zero coefficient on the right hand side of the equation.

The problem Simon tackles is the problem of spurious correlation. Suppose that two factors z_1 and z_2 are correlated, that z_1 precedes z_2 in time, but one is suspicious that the correlation is spurious. In other words, one is not happy to accept a structural equation in which z_1 causes z_2 i.e. $z_2 = z_1 + u$. Simon notes that this situation is typically dealt with by looking for an earlier third factor, z_0, which is correlated with both z_1 and z_2. One can then test whether the original correlation is spurious by conditioning on z_0. If the correlation vanishes the original correlation between z_1 and z_2 is deemed spurious.

In his paper Simon shows that this method for testing spurious correlation can be formalised in a way that shows that one can deduce whether or not the correlation between z_1 and z_2 is spurious, given the time order of the variables and uncorrelated errors in structural form equations relating z_0, z_1 and z_2. Simon analyses the following general form of equations for three internal variables.[6]

$$z_0 = u_0$$
$$z_1 = a_{10}z_0 + u_1$$
$$z_2 = a_{20}z_0 + a_{21}z_1 + u_2$$

Simon shows that if one assumes that the error terms are uncorrelated, then one can deduce the values of the coefficients from observation. He sets this out explicitly by showing how the zero correlation of the error terms implies that a set of equations holds for covariances of the variables and coefficients in the system. These equations can then be solved for the coefficients in terms of the covariances. Since the covariances can be estimated (the variables are observable) this allows one to calculate estimates for the coefficients.[7] For instance, it is easily derived from the fact that u_0 and u_1 are uncorrelated that a_{10} must satisfy.

$$a_{10} = E(z_0z_1)/E(z_0^{\,2})$$

[6] This is the general three variable system in which there are no simultaneous relations and no a's appear in the equations that would imply that a later factor causes an earlier one (so the equations respect the time order of: z_0 then z_1 then z_2).

[7] This is essentially a version of the case, discussed in the previous chapter, in which one calculates the coefficients by fitting the set of equations to observations. Though the process here is slightly complicated by the error terms, which is why it is necessary to take expectations.

Since $E(z_0z_1)$ and $E(z_0{}^2)$ can be estimated from sample data, a_{10} can be estimated. Though Simon does not mention it explicitly in the paper, the time order assumption among the variables ensures the above general equations are identifiable, and this is why one can deduce the values of the coefficients from observation.

In much the same way as in the last section of the previous chapter, Simon explicitly associates certain coefficients being zero with different causal orders relative to the set of error terms u_0, u_1, and u_2. In his analysis Simon considers all the different cases for the three equations above (1954, pp.44-45). For instance, he notes that if only $a_{10} = 0$ then z_0 does not cause z_1 but both cause z_2. Since z_1 causes z_2 in this case, the correlation between z_1 and z_2 is not spurious. Similarly if only $a_{21} = 0$, then z_0 causes z_1 and z_2, but z_1 does not cause z_2, so in this case the correlation between z_1 and z_2 is found to be spurious.[8] In this way, one can deduce whether the correlation is spurious or not.

So, as was done at the end of the last chapter,[9] Simon uses estimated 'zeros' for coefficients in the proposed general equations to make inferences to more restricted sets of equations. And, in the three variable case he considers, this allows him to show that one can deduce by introducing an earlier factor whether or not the correlation between two factors is spurious.

2.2. Simon's key Claim and his General Approach to Inferring Causal Order

In describing the result of his paper, Simon does not hold back. He claims that his paper shows '*correlation is proof of causation in the two-variable case if we are willing to make the assumptions of time precedence and non-correlation of the error terms*' (1954, p.43, original emphasis). As can be seen from its emphasis in the paper, this is the key claim in Simon's paper. This is a strong and contentious claim with which many people would take issue.[10] In this chapter I consider

[8] Simon proceeds to cover other cases in a similar way.

[9] Recall the discussion of identification at the end of the previous chapter, in which a method was set out where by measuring certain coefficients to be zero a causal order could be inferred.

[10] Consider the well-known maxim that Simon quotes in the first sentence of his paper: 'Even in the first course in statistics, the slogan "Correlation is no proof of causation!" is imprinted firmly in the mind of the aspiring statistician or social scientist.' (1954, p.37). Simon's key claim is

Nancy Cartwright's objection to the claim, because I later consider her own proposal for inferring causal order.

It is also important to note that Simon's method can be generalised to systems of equations with more than three variables and three equations. One can extend his method to any lower triangular system of equations in which the variables are time ordered and the error terms are uncorrelated. As discussed later, all such systems are identifiable so one can follow Simon's method above to deduce, from sample estimates for population covariances of the variables, the values of the coefficients.

As Simon notes, this approach assumes that the form of the equations is known, that the variables in those equations are strictly time ordered and that error terms in those equations are uncorrelated. Simon's key claim is that one can solve for unknown coefficients in this case, and thus solve for the causal order among the variables. Having presented this claim, I now consider Cartwright's objection to it.

3. Cartwright vs. Simon: Can Correlations Really be Used to Infer Causal Order?

This rather long section first presents Nancy Cartwright's criticism that Simon's method fails to deduce causal order from observation. In other words, Cartwright claims that it fails to solve the problem of observational equivalence.[11] This is followed by a presentation of some counterexamples Cartwright constructs to show that Simon's claim fails to solve the observational equivalence problem. Unfortunately no counterexample is given that applies to the key claim presented in Simon's 1954 paper. Given this, I present an additional counterexample to Simon's claim and conclude from this that Simon's key claim fails to solve the observational equivalence problem. With this done, I discuss a way in which Simon's claim might be weakened to avoid the counterexample. However, it is

---

controversial because it *does* claim that correlation can be a proof of causation under certain conditions.

[11] Recall that two systems are observationally equivalent if they are consistent with the same observations. The problem of observational equivalence is how to narrow down the possible causal systems to just one from observation. As discussed in the introduction, it is a variant of the problem of inferring causal relations from correlations.

then noted that even if Simon's claim is weakened in this way, it still faces an observational equivalence problem. The section then concludes by presenting an alternative 'S-approach' that assumes a strong reading for equations. In contrast to Simon's approach, this method solves the observational equivalence problem. However, it constitutes limited progress since it does so at the price of 'almost' assuming the problem of observational equivalence away.

3.1. Cartwright's Criticism of Simon

In the first chapter of her 1989 book, Nancy Cartwright is concerned to show how causes can be inferred from probabilities (or correlations). After a brief review of how Simon's analysis in his 1954 paper, Nancy Cartwright makes a bold criticism:

> '...there is a stock objection: the bulk of Simon's paper is devoted to showing that the parameters can be determined from the probabilities. But the problem occurs one stage earlier, in the interpretation of the data and the selection of the variables. The argument given here assumes, roughly, that dependent variables are effects and independent variables are causes. But the facts expressed in a system of simultaneous equations do not fix which variables are dependent and which are independent.' (1989, p.20).

To support her claim, she presents a long quote by Clark Glymour (1983) which criticises the contentious section in Simon's 1953 paper where Simon distinguishes between different mathematically equivalent systems by 'wiggling' coefficients (Simon, 1953, p.24). According to Glymour, Simon is attempting to solve the problem of distinguishing between observationally equivalent systems. Under his reading of Simon, Glymour argues that Simon's 'wiggling' approach does not suffice to resolve the equivalence. Glymour does this by constructing an observationally equivalent case in which the ordering is reversed. The logic of Glymour's point is essentially the same as that of some of the discussions in chapter two, where it was shown that one could change the causal order by mathematically manipulating the equations. However, Glymour uses this logic to claim that Simon fails to resolve the observational equivalence problem, whereas in chapter two I restricted the discussion to the conceptual equivalence problem.

In any event, Nancy Cartwright disagrees with Glymour's view that Simon was attempting to solve the observational equivalence problem. Instead, she uses Glymour's analysis to note that 'whether it was Simon's intent or not, there is a prima facie plausibility to the hope that it will solve the [observational] equivalence problem, and it is important to register clearly that it cannot do so' (1989, p.22). Is this right? Well, as seen in chapter two, one must assume something beyond the mathematical equations if the causal order is to have a well defined meaning. This was the solution to the conceptual equivalence problem. Given this, Glymour is clearly correct in that one cannot use *observed* facts about the equations alone to distinguish which system is correct. Depending on how the system is written and what coefficients are changed, different orders, and different changes in the variables follow. As long as one simply infers mathematical equations, one cannot distinguish between the causal orders of different systems. So Cartwright's point stands.

As shown in the quote above, Cartwright's general criticism is that one cannot use facts about mathematical equations to determine facts about causal relations. Another version of this criticism is that Simon's method can fail to distinguish between observationally equivalent systems. But if this is the case, then it should be possible to construct observationally equivalent systems, that is systems that are mathematically equivalent and thus consistent with the same observations, but which have different causal orders. Constructing such examples would then be explicit counterexamples to Simon's method for determining whether correlation is spurious, and more generally to his method for inferring causal order.

In a recent unpublished work, Cartwright attempts to construct such examples, cases of mathematically equivalent systems which have different causal order, that would falsify Simon's key claim that 'correlation is proof of causation ... if we are willing to make the assumptions of time precedence and non-correlation of the error terms' (1954, emphasis removed). These counterexamples are important because they make explicit why Simon's method cannot merely rely on facts about equations in order to learn about causes. Without such examples, Cartwright's criticism, though persuasive, might be dismissed as a merely

philosophical scepticism, of no consequence to practical causal inference from correlation. For this reason, I now look at various proposed counterexamples.

3.2. Attempted Counterexamples To Simon's Claim

To begin the discussion, consider a well-known example presented in Cartwright (1995, p.51), (2003b, p.9) of observationally equivalent systems with different formal orders.

$$\text{A.}\ \begin{matrix} z_1 = u_1 \\ z_2 = az_1 + u_2 \end{matrix} \qquad \text{B.}\ \begin{matrix} z_2 = v_2 \\ z_1 = cz_2 + v_1 \end{matrix} \quad \textit{where} \quad \begin{matrix} c = \dfrac{a}{a^2+1} \quad v_1 = (1-ac)u_1 - cu_2 \\ v_2 = au_1 + u_2 \end{matrix}$$

$\{z_1\} \rightarrow \{z_2\}$ $\qquad$ $\{z_2\} \rightarrow \{z_1\}$

In this case, A and B are mathematically equivalent. Moreover, the error terms in A are uncorrelated if and only if those in B are. Also, A and B have reverse causal orders. So, this is an example of two mathematically equivalent, and therefore observationally equivalent, systems with different causal orders. This shows more simply than Glymour's discussion, and in a way which is more relevant to Simon's analysis in the 1954 paper (given that the errors are uncorrelated in both systems), the problem of determining causal order from correlations. The problem is that both systems will be consistent with the same observations, so how can they distinguished by observation? In such a case it is clearly impossible to distinguish, *merely by using correlations* of z_1 and z_2, which causal order is correct.

This example clearly rebuts a generalised version of Simon's claim that one is able to infer causes from correlations in the case of *simultaneous* equation systems. However, a problem remains. In his 1954 paper Simon deals only with time ordered variables, which rules out the example of A and B above as a counterexample to his 1954 analysis. Since if z_1 precedes z_2 in time then system B is ruled out, while if z_2 precedes z_1 then A is ruled out. Cartwright herself notes this: '[i]n that case, our counterexample is no counterexample because time ordering will fix the causal order' (2003b, p.9). So, though this counterexample is a counterexample to a version of Simon's claim extended to apply to simultaneous systems, it fails to rebut Simon's 1954 claim. This also shows how Simon's

stipulation of time order among the internal variables helps rule out some observationally equivalent systems with different formal orders.

Since Cartwright is aware that the last counterexample fails in the time ordered case, she proposes another counterexample where the time order among the variables is respected. In this case, she presents the following two systems (2003b, p.9).

$$\begin{array}{lll} & z_1 = u_1 & z_1 = u_1 \\ \text{C} & z_2 = az_1 + u_2 & \text{D} \quad z_2 = az_1 + u_2 \\ & z_3 = bz_1 + u_3 & z_3 = cz_1 + dz_2 + v \end{array} \qquad \textit{where} \quad c = \frac{b}{a} + 1, d = -\frac{1}{a}$$

$$v = -z_1 + \frac{z_2}{a} + u_3 - \frac{u_2}{a}$$

Unfortunately, this counterexample fails. Though it is not immediately obvious, appendix 6.1 shows that though D has uncorrelated error terms and preserves the time order among the internal variables, it is not consistent with system C unless it is identical to it. The reason is that from the definition of v it follows that either the formulae above for c and d are incorrect, or D is not consistent with system C. If the formulae for c and d are incorrect, then D must be identical to system C so in that case the counterexample fails. Whereas if D is not consistent with C, then it cannot be observationally equivalent with it. So the counterexample fails. Unfortunately in her discussion Cartwright overlooks the inconsistency and concludes that this provides a time ordered counterexample to Simon. The result here shows this to be mistaken.

An obvious next question is whether or not a counterexample like the one attempted by Cartwright can be constructed. If it is possible then there exists a system D' below, distinct from system C, which can be derived from C, where D' has uncorrelated error terms.

$$\begin{array}{lll} & z_1 = u_1 & z_1 = u_1 \\ \text{C} & z_2 = az_1 + u_2 & \text{D'} \quad z_2 = az_1 + u_2 \\ & z_3 = bz_1 + u_3 & z_3 = cz_1 + dz_2 + v \end{array}$$

However, appendix 6.1 shows this is not possible. It shows that if one assumes the above form for D', and assume that its error terms are uncorrelated then system D' *must* be identical to system C.

So, as it stands, Simon's key claim in 1954 paper, that causal order can be inferred from time order and uncorrelated errors, remains unrebuted by the examples of observationally equivalent systems considered here. This, then raises a question: does Cartwright's criticism that correlations and time-ordering alone are insufficient for determining a unique order apply to the systems that Simon analyses in his 1954 paper? Or perhaps Simon *is* correct, and imposing a strict time order among variables is sufficient to rule out observationally equivalent systems with different causal orders. I now consider this.

3.3. A Time Ordered Counterexample to Simon's Claim

In fact, there is a good reason why counterexamples are difficult to construct. This is because there is a general impossibility result that shows that no such counterexample can be built. To see why, first recall that the systems that Simon considers in his 1954 paper have, on time ordering the equations, the following lower triangular form.

$$\begin{aligned} z_1 &= u_1 \\ z_2 &= a_{21} z_1 + u_2 \\ &\vdots \\ z_n &= a_{n1} z_1 + \ldots + a_{nn-1} z_{n-1} + u_n \end{aligned}$$

In appendix 6.2 it is shown that such systems are identifiable when the error terms are uncorrelated. This seems to put the matter to rest, since identifiability (by definition) prevents systems being manipulated into mathematically equivalent systems with the same functional form but with different coefficient values. This is exactly what one does in attempting to construct a counterexample to Simon's claim. Identifiability rules this out.

Therefore, it appears hopeless to attempt to construct a time ordered counterexample. This is because any observationally equivalent system, to be mathematically equivalent, must be constructed from the original system. Since the equivalent systems must have the same functional form as the original (lower triangular with uncorrelated errors) they must, since the original system is identifiable, have all the same coefficient values as the original. Therefore, it appears impossible to construct a counterexample like A and B for the time ordered case. As a result, Simon's position appears strengthened.

However, one should not be too quick to opt in favour of Simon here. There is in fact an important implicit assumption in the impossibility result just presented. It is a condition that Cartwright notices in the quote presented earlier: '...the problem occurs one stage earlier, in the interpretation of the data and *the selection of the variables.*' (1989, p.20, my emphasis). In turns out that changing the internal variables that appear in the equations is key to constructing a counterexample.

To show this, I construct a counterexample from system C. Assume as before that system C holds, where the variables are time ordered according to index, where the u's have zero mean, variance *1*, and are uncorrelated with each other.

$$\text{C} \qquad \begin{aligned} z_1 &= u_1 \\ z_2 &= az_1 + u_2 \\ z_3 &= bz_1 + u_3 \end{aligned}$$

Appendix 6.3 shows that one can derive the following system from A.

$$\text{D'} \qquad \begin{aligned} z_2 &= v_2 \\ z_3 &= \frac{ab}{1+a^2} z_2 + v_3 \end{aligned} \quad \textit{where} \quad \begin{aligned} v_2 &= au_1 + u_2 \\ v_3 &= \frac{b}{1+a^2} u_1 - \frac{ab}{1+a^2} u_2 + u_3 \end{aligned}$$

By construction, system D' is mathematically consistent with system C (though it is not equivalent since it omits z_1) and has uncorrelated error terms. The causal order for D' is that z_2 causes z_3, whereas in C, z_2 and z_3 are jointly caused by z_1. So, unlike the earlier examples, the two systems C and D' *do* present a counterexample against Simon's claim in his 1954 paper. This is because if the equations in system C hold, then those of system D' do also. According to Simon's claim, System D' is a system that meets the required time order and uncorrelated error assumptions[12] for deducing causal order and *yet*, if system C is the system with the correct causal order then system D' has *incorrect* causal order. So, Simon's claim that one can deduce the causal order from time ordered variables and uncorrelated error terms fails for D' and Simon's key claim is false.

[12] The error terms in D' will not have variance *1* in this system. This is not important however since which, if any, error terms have variance *1* depends entirely on the choice of scale for the variables being analysed. To rely on this choice of scale to break the observational equivalence between the systems would be tantamount to claiming that which causal order is correct depends on a choice of scale for the variables, which would be absurd.

There is something surprising about this counterexample. If system C and its causal order are true, System D' is the system one infers if one regresses z_3 on z_2. In that case, if one reads system D' causally then one makes the mistake of spuriously reading the correlation between z_2 and z_3 as indicative of a causal connection. This is the problem of spurious correlation all over again. So, the counterexample is particularly damaging to Simon, since it shows that his method fails to address the paradigmatic case it was meant to deal with, that is, the problem of spurious correlation among two variables.

In spite of this counterexample, there is a clearly a way out for Simon, that is, to assume that it was known that the set of internal variables being considered are somehow 'given', that is, no important internal variables have been omitted like in the example above.[13] This is because in that case the earlier impossibility result holds, so it is not possible to construct, keeping the time order among the *same* variables and the error terms uncorrelated, an observationally equivalent system with different causal order for the *same* set of variables. This rules out the possibility of a counterexample like that presented here.

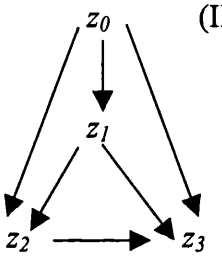
However, it should be immediately obvious that this move won't do for Simon. The method proposed in his 1954 paper is to solve the problem of spurious correlation by introducing a third earlier variable to fit the correlated variables in a larger model to see whether or not the original two variables are spuriously correlated. If Simon solves the problem of spurious correlation by introducing a variable, he can hardly put a limit on the number of variables that can be included in the model, in order to avoid the possibility that the causal connections, in a model that showed another a correlation was spurious, turn out themselves to be spurious in a larger model.

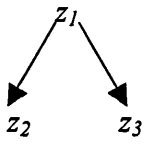
[13] The conditions under which the internal variables modelled can be taken as 'given' or suitably 'complete' is therefore a crucial question. It is a difficult question however, given that there will always be omitted causes not explicitly represented by the system of equations. Bayes-net methods, for instance, assume a modelled set of factors is 'causally sufficient' which requires that any common cause of two modelled factors also be modelled and in this way avoid the problem above, see Spirtes *et al.* (1993, p.45). Also, chapter three, by attempting to set out when equations could model only some causal relations and when factors could be omitted using error terms, provides the beginnings of an attempt to set out the conditions under which variables can be left out, and when they can be covered in error terms, given the strong reading.

3.4. A Way to Salvage Simon's Claim?

This section considers another way of salvaging Simon's claim. To do this, I attempt to construct a more damaging counterexample to his analysis. When it is shown that it is not possible to do so, a way Simon's claim seems to hold in spite of the above counterexample is observed.

The attempt is to construct from a system, with variables z_0, z_1, z_2 and z_3 in which z_2 does cause z_3, a smaller system without z_0 which read causally asserts that z_2 and z_3 are both caused by z_1 but do not cause each other. Let (I) denote the larger system, (II) the smaller. In other words, the attempt is to construct equations for (II) from those of (I) where the causal graphs for each system are given by.

(I): 

(II): 

If successful, this would be particularly devastating to Simon, because system (II) is just like System A, and we have just shown in our counterexample that such a system is in turn consistent with another system, say (III), in which z_2 causes z_3. So, if this new counterexample can be constructed then (I) and (II) would both be consistent with

(III) $z_2 \longrightarrow z_3$.

But then the following situation would be possible. Suppose that (I) is the true but unknown underlying system with correct causal order. It is consistent with the equations of systems (II) and (III). Now suppose that a correlation between z_2 and z_3 is observed, but that it is assumed that the causal relation between z_2 and z_3 is spurious. Suppose also that it is known that z_1 is a common cause of z_2 and z_3. Then, following Herbert Simon's approach and good scientific practice, z_1 is incorporated into the model relating z_2 and z_3. Estimating the coefficients in such equations would then yield system (II),[14] which would suggest that the causal

[14] I am glossing over some details here, but if it is possible to construct a system (II) from (III) then it can be shown that applying Simon's method of inferring coefficients from observed covariances, will in fact yield system (II), if one assumes equations of the form of (II).

connection between z_2 and z_3 is spurious when in fact, given system (I) holds, it is not.

Clearly, if it were possible to construct such an example, it would be a serious problem for Simon since it would imply that one could not be certain that a spurious correlation discovered using Simon's method in fact held, because for all one knew there might in fact be another larger system in which a causal connection between the 'spuriously' correlated variables obtained.[15]

Luckily for Simon, however, it is *not* possible to construct such a case. Appendix 6.4 shows that it is *not* possible to construct a system (II) where z_2 does *not* cause z_3, provided the coefficients in the system (I) are not functionally related. Why this happens can be seen from analysis of the more general case. Suppose, analogously to (I), that a lower triangular system and its causal order holds among some set of time order factors, where the error terms are uncorrelated. Generalising the attempt above, the aim is to derive another system from this system, that relates a proper subset of the original variables but with different causal order.

In this general case, the system to be derived has form with uncorrelated error terms.

$$
\begin{aligned}
z_1 &= u_1 \\
z_2 &= a_{21} z_1 + u_2 \\
&\vdots \\
z_n &= a_{n1} z_1 + \ldots + a_{nn-1} z_{n-1} + u_n
\end{aligned}
$$

To derive this system, one must solve for its coefficients from those of the original system. To do this, one uses the $n(n-1)/2$ pairwise orthogonality constraints to be met by the error terms.[16] In addition, there are, provided none are

[15] The problem of causal connections changing when extra variables are introduced is general problem of causal inference. It is a form of Simpson's paradox. In its general form, Simpson's paradox is the problem that conditional probability relations e.g. $P(A|B) > P(B)$ can always be reversed or removed by conditioning on some other variable. Causal systems where the introduction of another variable leads a correlation to vanish is an example of the paradox, see Malinas and Bigelow (2004).

[16] This follows because there are $C(n,2) = n(n-1)/2$ ways of choosing two distinct error terms from the n equations. So this gives the number of constraints implied by the pairwise independence of errors. Since joint independence of errors is not implied by pairwise independence however, it

assumed to be zero, exactly $n(n-1)/2$ unknown coefficients[17] (a's) to be solved for in order to solve for the derived system. Therefore, functional dependencies in the coefficients of the original system aside,[18] one can solve for the coefficients of the derived system if and only if no coefficient is assumed to be zero in the derived system. Otherwise there would be fewer coefficients than constraints and there would be no solution for the derived system. Therefore, if the system can be derived it has only non-zero coefficients. In this way systems like (II) are ruled out since in these it is assumed that some causal connections are absent, that is, there are equations with some zero coefficient(s).

Taking a step back from the formalism, it follows that the only systems that can be derived from larger systems with greater numbers of variables are those that, when read causally, assume that *all* possible causal connections among the variables hold. This implies that one *cannot* derive a system from a larger system, which read causally asserts that some earlier variable does not cause a later variable.

So, for example, if we observe a correlation between z_2 and z_3, which we suppose to be spurious, and suppose that we introduce an earlier variable z_1 which screens off the correlation between z_2 and z_3. Read causally, the resulting model states that z_2 does not cause z_3. The above analysis shows that any larger model containing z_1, z_2 and z_3, along with other time ordered variables and uncorrelated error terms, will be such that when read causally, z_2 will *not* cause z_3.

This result highlights an important asymmetry in what can be inferred. From the counterexample we presented above, it was shown that a causal connection between z_2 and z_3 in a model which respected time order and had uncorrelated errors, could in fact be due to an omitted, earlier common cause z_1. However, it

---

may be possible to get more constraints by assuming the error terms are jointly independent. This in turn may help ruling out further observationally equivalent systems. This suggests an interesting topic for further exploration, which I leave as further work.

[17] This is because if none of the a's are assumed to be zero, then by counting there are: $1+2+...+n-1 = n(n-1)/2$ a's to be solved in deriving the system.

[18] This is an important and now familiar caveat since it rules out original systems that have cancelling out relationships. In a fuller, more rigorous analysis these systems would not be excluded.

has just been shown that in a model where z_2 does not cause z_3 it is not possible to construct a larger model[19] including all the variables of the smaller model, in which z_2 does cause z_3.

Putting these two results together, it has been shown that:

> (*) If a linear model of time ordered variables with uncorrelated errors is such that, when read causally, it asserts that z_1 directly causes z_2, then there may be a larger model with more variables in which this causal connection vanishes. However, if in the model z_2 does not directly cause z_3 then there is *no* larger model,[20] which read causally, asserts that z_2 directly causes z_3.

This inferential asymmetry is important, it suggests that although one cannot be sure of causal connections inferred using Simon's method, one *can be* sure about the absence of causal connections. In this way part of Simon's claim appears to be salvaged. The analysis here appears to show that one can deduce from time ordered variables and uncorrelated errors that two factors are *not* causally connected.

However this respite for Simon is short lived. Since, in spite of appearances, even this modified claim is undermined. I now show how.

3.5. A Further Problem for Simon's Claim

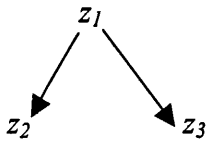
The above appears to suggest a way in which Simon's claim can be weakened, in light of the counterexample, so that it holds. The suggestion is to limit his claim that only absences of causal claims can be deduced from observations given uncorrelated error terms and time order. However, this claim also fails.

Suppose one infers, using Simon's method, that z_1, z_2 and z_3 satisfy a set of three equations that has the following causal graph.

---

[19] Strictly speaking, it is not possible to construct a larger model with no cancelling out relationships, see previous footnote.

[20] Assuming no functional dependencies among its coefficients, see two previous footnotes.



Now, (*) above states that one *can be sure of the lack of causal connection* between z_2 and z_3 but cannot be sure that the connections between z_1 and z_2 (z_3) are spurious. And yet, the inference to a lack of causal connection between z_2 and z_3 is *contingent* on the causal connections holding from z_1 to z_2 and z_3. This is clear because one can always fit a *smaller* model without z_1, which read causally asserts that z_2 causes z_3.[21] So, how can one be sure that there is no causal connection between z_2 and z_3 if the claim relies on an uncertain claim of causal connections from z_1 to z_2 and z_3? It seems then that we can't be sure of the lack of causal connection either. Moreover, it doesn't help to enlarge the model by introducing earlier variables since, at best, this merely introduces further causal connections of which one cannot be sure.

This problem raises a serious problem. It shows that:

> (**) *No matter how many variables one introduces one cannot be sure that the equations, when read causally, are not in fact spurious.* Moreover, since claims as to the lack of direct causal connection between variables also *depend* on other causal connections holding in the model, this means that ultimately inferences to absences of causal connections among variables are also not secure.[22]

To conclude, Simon's weakened claim fails to hold. This undermines confidence in inferences to a lack of direct causal connection in spite of the result (*) shown above: that any time ordered linear model that assumes that two variables are not directly causally connected, cannot be embedded in a more general linear model in which the same two variables are directly causally connected.

---

[21] This is what above time ordered counterexample to Simon (systems C and D') does.

[22] Bayes-nets methods are interesting here, since these analyse all possible causal graphs for a causally sufficient set of factors. These are interesting because causal sufficiency rules out the possibility of any common causes being omitted, on which (**) ultimately depends. Whether or not this is progress, though, depends on how the causal sufficiency assumption is justified.

So what, if anything, can be done to salvage Simon's approach? The answer is to bring in the strong reading developed in chapter two.

3.6. The Right way to Salvage Simon: Introduce the Strong Reading

Nancy Cartwright's criticism of Simon is based on one key point. It is encapsulated in her statement that 'the facts expressed in a system of simultaneous equations do not fix which variables are dependent and which are independent.' (1989, p.20). This problem was also made clear in chapter two: the facts in the equation are insufficient to determine the causal interpretation. The time ordered counterexample (of C and D') shows also that stipulating time order doesn't fix the problem because then the causal interpretation is contingent on the variables that are included in the equations. While (**) shows that adding more variables doesn't help because whatever variables are added, one still has a model that can change its causal interpretation when even more variables are added. In short Cartwright's point still applies. A set of equations relating time ordered variables with uncorrelated error terms, underdetermines the causal interpretation. So why should one believe that such a set of equations describes the causal relations among factors?[23]

Clearly something is missing, one needs to stipulate that the equations represent causal relations, and one must have some way for justifying the truth of these causal claims. A set of equations among variables with error terms does not fix the causal semantics, nor does a set of *true* equations among variables fix the *truth* of causal claims among the factors those variables represent.

So, what is the solution? One needs to explicitly assume that relations represented by the equations are causal, and to assume that to learn about causal relations one must have background knowledge about those causal relations. This is the point of Nancy Cartwright's (1989) maxim *'no causes in, no causes out'*.

In fact, these conditions for dealing with Cartwright's criticism are met in the approach for inferring causal order presented at the end of the last chapter (see

[23] This point is made in Cartwright (forthcoming).

section 5). Under this approach there are two key assumptions that are not made by Simon.[24] The first is that the strong reading of the equations is adopted. This strong reading, as set out in chapter two, ensures that one cannot manipulate equations (in all but non trivial ways) without changing their causal content. It assumes that there are principled reasons for taking equations to represent mechanisms and that causal order among factors follows in virtue of relations among the mechanisms. This strong reading introduces content *outside the equations* which ensures the causal order as represented by a set of equations is unique. In other words, it adds the content required to fix the causal interpretation of the equations.

The second assumption necessary for overcoming Cartwright's criticism is made in the discussion of the role identifiability plays in causal inference at the end of the last chapter. There strong background knowledge was assumed in order to make an inference to causal order using identifiable systems of equations. In particular, it was assumed that a set of possible causal orders was known, that were consistent with a known identifiable system of equations. By measuring coefficients for this system from observation, then one could deduce from observation which of the possible causal orders is correct. The key point is that background causal knowledge is assumed and used to deduce further knowledge.

I call this approach to inferring causal order the 'S-approach'. The S-approach avoids Cartwright's criticism, because it accepts that one cannot have causes out without causes in, both in the interpretation of equations and in making causal inferences. Nevertheless, the method bears very close similarities to the method set out by Simon in his 1954 paper. In fact, the method is essentially Simon's but with corrections made to deal with the criticism of Cartwright and Glymour.

To see this more clearly, recall the logic of the inferential method (the S-approach) presented at the end of the last chapter (see section five).

[24] Perhaps one should say 'not made explicitly by Simon' if one wants to be charitable. Either way, the point made here stands.

(i) A set of *identifiable* of linear equations is known to hold, and it is known that the true structural equations are identifiable and may be obtained by setting one or more, if any, of the coefficients in this general set of linear equations to zero.

(ii) A sufficiently varied set of observations for the variables is obtained, so that the coefficients of the general set of equations can be measured from observation.

THEN By measuring coefficients and finding out which if any are zero, one can deduce which of the possible system of structural equations holds, and thus deduce the causal order.

Unlike Simon's approach, the S-approach for inferring causal order does not rely on mere correlations and time orders. It works by building in a strong knowledge claim about what causal orders there may be among the factors. This is why it avoids Cartwright's criticism. Otherwise the approach is like Simon's, formally it requires fitting identifiable equations to observations and inferring, from coefficients that are found to be zero, which causal order obtains.[25]

In addition, Simon's 1954 analysis can be put in these terms to make his arguments valid. In that case, the assumption that it is known which variables are relevant for the analysis and their time order, is read as an assumption that it is known that only causal orders among those factors with that time order are possible. Since all of these are identifiable by the time order restriction and the uncorrelated errors, then, just as Simon shows, one can deduce the causal order from observation.

Finally, one should note that this S-approach is not circular. Using it, one does not infer to a causal order by assuming that the correct causal order holds *a priori*. Though the assumption that one knows the set of possible causal orders is strong, it does not imply that the actual causal order is known.

[25] There is an important problem which is by-passed in this discussion. The S-approach here needs to be generalised to cover systems with error terms. This is a very important but difficult analysis which I leave as further work. In discussion of the S-approach in the remainder of the chapter I focus on systems of equations without error terms.

That said, as the discussion at the end of the last chapter made clear, the background knowledge assumptions are very strong. One needs to know an awful lot about the causal order before one can use an identifiable system of equations to infer causal order like Simon does in his (1954). So, though the S-approach avoids Cartwright's criticism of Simon, in some ways it does not provide a very satisfying analysis of how to find out about causal order because so much needs to be known before it can be used. Another way of seeing this is that the background knowledge assumed requires that many of the problems discussed above for Simon's claim have been somehow dealt with. For instance, in assuming that it is known that the only causal orders that are possible are those consistent with a known identifiable system of equations, one takes as 'given' the set of internal variables in that system of equations. This was criticised in discussing Simon's claim. Therefore, a more complete analysis would ultimately be desirable to unpack how the background knowledge required to use the S-approach might be obtained.

4. Cartwright's Alternative Approach for Deducing Causal Order

In this section, I consider another way to infer causal order. This approach is developed by Nancy Cartwright as an alternative way to solve the problem of observational equivalence.

In her analysis, Cartwright (1989) focuses on time ordered systems like those considered by Simon (1954). To make clear the causal content of equations, she stipulates that a causal equation be such that there be one variable on the left hand side which denotes the effect, while the variables on the right hand side of the each denote causes of that effect. Given this assumption and the time order assumption, the systems Cartwright analyses also have lower triangular form (where indices of variables denote time order):

$$z_1 = u_1$$
$$z_2 = a_{21}z_1 + u_2$$
$$\vdots$$
$$z_n = a_{n1}z_1 + \ldots + a_{nn-1}z_{n-1} + u_n$$

In her 1989 work, an equation, that has one variable on the left hand side of the equality, is *causally correct* if and only if it is functionally correct and all of the variables on the right hand side denote causes of the effect denoted by the variable on the left hand side. Cartwright's analysis on how to infer causal order then proceeds by analogy with the problem of spurious *INUS*[26] conditions and that of observationally equivalent linear models. She then proposes a solution for *INUS* conditions from which she develops an analogous solution for linear models.

4.1. Using Spurious INUS conditions and Open Back Paths to Infer Causal Order

In her discussion, Cartwright briefly reviews John Mackie's claim that causes are *INUS* conditions for their effects. She then presents Mackie's (1974) famous example of the two factories that shows that being an *INUS* condition is not sufficient for being a cause, which shows how one can construct 'spurious' *INUS* conditions for effects.

In Mackie's example there are two factories, one in London and one in Manchester. In both, if it is five o'clock this causes the hooters to blow in the respective factories. Quoting Mackie, Cartwright formalises the example as follows (1989, p.26).

$X_2 \equiv AX_1 \vee W$

$X_3 \equiv BX_1 \vee V$

where

X_1: It is five o' clock

X_2: Manchester hooters sound

X_3: London hooters sound

A: Conditions under which ensure Manchester hooters blow if it is 5 o' clock

B: Conditions under which ensure London hooters blow if it is 5 o' clock

W: Conditions under which the Manchester hooters will blow when it is not 5.

[26] Recall that an *INUS* condition is an insufficient but necessary part of an unnecessary but sufficient condition.

V: Conditions under which the London hooters will blow when it is not 5.

In the correct causal representation, X_1 is an *INUS* condition for both X_2 and X_3, which is to be expected since it is assumed that X_1 is a cause of both X_2 and X_3.

In this case one can, in a similar way to Cartwright,[27] derive the following proposition from the two propositions above.

$$X_3 \equiv BX_2 \neg W \vee BAX_1W \vee B \neg AX_1 \vee V$$

In this proposition, if the Manchester hooters blow (X_2) and the conditions are met under which the London hooters blow if it is five o' clock (B) and conditions are not met under which the Manchester hooters blow when it is not five o' clock ($\neg W$), then the London hooters blow (X_3). Of course, this works because this conjunction ensures that it is five o' clock *and* conditions for the London hooters to blow at five are met, which implies the London hooters blow. However, the proposition is spurious since, if one was to read its *INUS* conditions as causes, then this would mistakenly imply that that Manchester hooters blowing is a cause of the London hooters blowing.

Cartwright's key point in setting out this example is that if one has sufficient background causal knowledge then one can rule out this spurious proposition. In this case, she notes, if one knows that neither W nor $\neg W$ can be a cause of X_3 except possibly via X_2, then one can rule out the spurious proposition. This is because in that case, any attempt to derive a proposition in which X_2 is an *INUS* condition for X_3 also introduces W or $\neg W$ as an *INUS* condition. Then, if it is known that neither of these can cause X_3 not via X_2, then the derived propositions in which X_2 is an *INUS* condition for X_3 are spurious propositions. So, if all of the propositions in which X_2 is an *INUS* condition are known to be spurious then X_2 is known not to be a cause of X_3. Finally, note that this is plausible in this hooters example, since one could easily rule out the conditions under which the Manchester hooters blow when it is not five as not independently causally

[27] This is slightly simpler than the proposition that Cartwright derives (1989, p.27). However, it makes Cartwright's point just as well since in it X_2 is a spurious *INUS* condition for X_3.

connected with the London hooters blowing and in this way rule out the derived proposition as spurious.

Cartwright also presents an analogous problem for linear models. Instead of following her analysis strictly, one can also see her point by making a direct analogy with the two propositions above. Suppose the following are the true but unknown causal equations.

$$x_2 = ax_1 + w$$

$$x_3 = bx_1 + v$$

Substituting x_1 out from the second equation we get an equation analogous to the derived proposition above.

$$x_3 = b/a(x_2 - w) + v$$

Finally, the analogous conclusion to Cartwright's conclusion for the *INUS* conditions, is that if we know that w doesn't cause x_3 except possibly via x_2 then we know that this equation must be spurious.

Cartwright's general conclusion from this is that if we have a set of possible causes for an effect, and we know that each of these possible causes, c, has a cause, u, which cannot cause the effect or interest except via c, then if an equation holds between the effect and the possible causes then any attempt to derive another equation from this introduces a factor which *cannot* cause the effect in question. In that case any derived formula is spurious and the original formula must be causally accurate, and all the possible causes genuine.

This condition is formulated in Cartwright's requirement that each putative cause have an open back path relative to the effect. Her definition of an open back path is:

> '*OBP*: $x(t)$ has an *open back path* with respect to $x_e(0)$ just in case at any earlier time t', there is some cause, $u(t')$, of $x(t)$, and it is both true, and known to be true, that $u(t')$ can cause $x_e(0)$ only by causing $x(t)$.' (1989, p.33)

In order to use this condition to infer causal order, Cartwright also makes explicit the need for some further assumptions. She assumes what she calls the 'Generalised Reichenbach Principle' which assumes that all true functional

relations can be derived from the true set of causally correct equations, and also assumes transitivity of causality.[28]

Given these assumptions, she then proves that her inferential claim above holds generally (1989, p.37). Specifically, she proves that if an equation $x_e = \sum_i a_i x_i$ is known to hold, and if every factor on the right hand side has an open back path with respect to x_e, then the equation is *causally correct*, that is, every factor on the right hand side is a cause of x_e. In this way, Cartwright generalises her observation of how the problem of spurious *INUS* condition could be resolved, to propose a distinctive approach for inferring causal order information from observations.

Cartwright's distinctive approach to inferring causal order obviously raises many questions. The most interesting one is: what is the relationship between the method here outlined by Cartwright for inferring causal order and that of the S-approach? Before an attempt is made to answer this, however, I highlight a few problems with her OBP definition and her inferential claim. Once this is done, I compare the S-approach with Cartwright's.

4.2. A Few Criticisms of Cartwright

The discussion above shows that Nancy Cartwright's open back path requirement lies at the heart of her approach for deducing the causal order from a correct

[28] Recently, there has been renewed debate as to whether or not causation is transitive. Many counterexamples have been given to show that causation is not transitive (e.g. a dog bites a terrorist's left hand which causes him to push the button of a bomb with this right hand, which causes the bomb to explode, but one would not say that the dog biting the terrorist's hand caused the bomb to explode). There has been much discussion of this and other examples, and various solutions have been proposed. See, for example, Hall (2000). This discussion ultimately has relevance for the theories of causal relations discussed in this thesis (i.e. Simon's approach, the strong reading, Cartwright's etc.) since all of these assume a transitive causal relation. Therefore, it would be important to analyse whether the counterexamples to transitivity of causation imply that there are important situations that could not be treated using the theories of causal relations analysed here. However, since the thesis focuses on making clear, comparing and contrasting the different causal positions discussed, the question is not of immediate relevance for the discussion of the thesis. For this reason, I flag the problem and leave its investigation as further work. In passing, it is interesting to note that some have used structural equations approach to fruitfully analyse the intransitivity examples, for example, Hitchcock (2001). So, in addition to the intransitivity examples being relevant to structural approaches to causality, the converse may also hold, that is, structural approaches may be beneficial for discussing the intransitivity counterexamples. This too suggests further work.

functional relation. However, the definition is challenging in several respects. In this section, I suggest a slight reformulation. Second, I set out an example to emphasise that Cartwright's (1989) analysis allows the inference of causally correct equations, it is insufficient for inferring certain direct causal relationships. Lastly, a counterexample is presented to the proof that shows the need for another condition to be added to safeguard her inferential claim.

The first issue considered here is the following unusual feature in Cartwright's OBP definition: it incorporates both an ontological and an epistemic element. This is explicit in the definition which requires that 'it is both *true*, and *known* to be true' (1989, p.33, emphasis added). This is an unusual philosophical move since it intertwines ontology and epistemology in the same definition. Cartwright's motivation is clear enough from her preceding discussion though. In the Manchester hooters example, one needs not only that W be spurious but one must also *know* that it is not a cause except possibly via X_2. Since the *OBP* requirement follows by generalisation from this example, it is not surprising that it mixes the epistemic requirement with the ontological requirement.

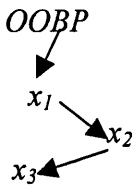
I think this mixing of epistemic and ontological elements in the open back path is unnecessary and liable to lead to confusion. So, I prefer the following reformulation.

> *OOBP*: *x(t)* has an *ontological open back path* with respect to $x_e(0)$ just in case at any earlier time *t'*, there is some cause, *u(t')*, of *x(t)*, and *u(t')* can cause $x_e(0)$ only by causing *x(t)*.
>
> *KOBP*: *x(t)* has an *known open back path* with respect to $x_e(0)$ just in case that *x(t)* has and is known to have an ontological open back path with respect to $x_e(0)$.

Clearly, Cartwright's open back path matches what I call here a known open back path (KOBP).

The second critical point about Cartwright's analysis can be seen from the following interesting case. Consider the following simple causal structure, where

x_1 is a direct cause of x_2, and x_2 is a direct cause of x_3, and where x_1 only causes x_3 by x_2.



Suppose x_1 has an OOBP with respect to x_3 where this back path is denoted by the OOBP in the causal graph. Here the interesting question is: does x_2 have an OOBP with respect to x_3? Since x_1 only causes x_3 by causing x_2, and there is a back path of causes of x_1 which only cause x_3 by causing x_1, then all of the causes on this back path also cause x_3 only by causing x_2 (since x_2 is an intermediate cause on the path from x_1 to x_3).[29] So it follows that x_2 also has an OOBP with respect to x_3.

To see why the case is interesting, suppose that the following are the causal equations associated with the causal graph above (where a and b are non-zero constants).

$$x_3 = ax_2$$

$$x_2 = bx_1$$

Then it is easily derived that the following equation holds for any non-zero constant d.

$$x_3 = dax_2 + (1\text{-}d)abx_1 \qquad (+)$$

Now suppose that this functional equation is known to hold and it is known that both x_1 and x_2 have OOBP's. Then in that case the conditions for Cartwright's proof are met, and it follows by her proof that (+) is causally correct, that is, both x_1 and x_2 are causes of x_3.

The interesting point is that though (+) *is* causally correct, in the sense that every right hand variable is a cause of the left hand variable, it is unlike the two original equations in that every right hand side variable does not denote a *direct* cause of the factor denoted by the left hand side variable. This interesting case makes clear that Cartwright's (1989) inferential result allows one to infer causally correct

[29] Note that the argument uses the transitivity of causality, which is assumed by Cartwright in her analysis.

equations but not direct causal relationships, like those represented in the causal graph above.

That said, it might be possible to strengthen Cartwright's inferential claim to allow inference of direct causal relationships. To see how, note that I have read the OOBP condition in a particular way. In the example I read the OOBP of x_1 with respect to x_3 as implying that any effect, x_2, of x_1 that are causes of x_3 such that the influence from x_1 to x_3 must pass via x_2, also has an OOBP for x_3. One way to strengthen the inferential claim would be to modify the definition of an open back path to rule this out. However, I think this would be a mistake since it would require a very strong stipulation along the lines of: if x_1 has an OOBP for x_3 it cannot have any intermediate cause between it and x_3. This would turn the OOBP condition into a requirement that x_1 is a direct cause for x_3 which is clearly very restrictive.

There are other features which might be used to strengthen Cartwright's inferential claim. For instance, in the example x_1 being a cause of x_2 is crucial in constructing the causally correct equation where the right hand variables do not denote direct causes. If situations like this are ruled out then one cannot infer to the causally correct equation (+) in the example above. Therefore, the following condition might be added to strengthen Cartwright's inferential claim: it is known that no right hand side variable (in the known functional relation) is a cause of any other. However, the problem with this restriction is that it is also very strong since it rules out causal inference in cases in which a right hand factor directly causes another. Perhaps a better, less restrictive alternative is to require that each right hand variable be known to have a distinct OOBP from those of the others.[30] This would be less restrictive while also ruling out the inference to the undesirable causally correct equation, (+), because in the system, x_1 and x_2 share an OOBP.

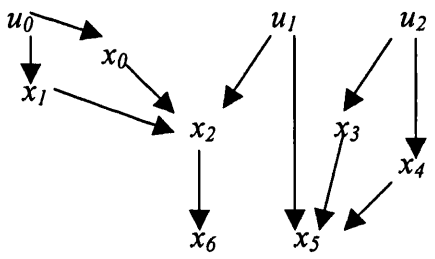
Indeed this last option is suggested by Cartwright in a later work (2003a). Describing her (1989) result, Cartwright (2003a) claims '[t]he [known] equation for x_e is thus a true causal law, *so long as nothing appears on the right-hand side*

[30] How exactly to formulate what makes two OOBP's distinct would need to be fleshed out.

that is from the back path of any other factor that appears there' (emphasis added, 2003a, p.214). By 'causal law' Cartwright has in mind equations like those that correspond to the causal graph above where the right hand variables denote direct causes of that denoted by the variable on the left.[31] So, this claim is a restatement of her (1989) result but with a stronger consequent and with a condition that no two variables share an OOBP. Whether or not this stronger claim is justified by her later paper (she presents some very extensive formal analysis there) I leave as further work. Nevertheless, the statement at least shows the intent of the (later) Cartwright to extend the 1989 result to make stronger causal inferences.

To finish this section, I present a counterexample to Cartwright's inferential claim. This shows a need for a further strengthening of the conditions for inferring causal order. In the example, assume the following causal graph and causally correct equations hold, that the variables are time ordered according to index and that x_2 and x_5 both have KOBP's with respect to x_6. Also, assume that the OOBP for x_2 passes through u_0, while that for x_5 passes through u_2.

$x_0 = \gamma_2 u_0$
$x_1 = -\gamma_1 u_0$
$x_2 = u_1 + \gamma_1 x_0 + \gamma_2 x_1$
$x_3 = \beta_2 u_2$
$x_4 = -\beta_1 u_2$
$x_5 = \alpha_1 u_1 + \beta_1 x_3 + \beta_2 x_4$
$x_6 = \alpha_2 x_2$



It is easy to see that both $x_5 = \alpha_1 u_1$ and $x_2 = u_1$ hold. This is because the bifurcated causal paths from u_0 and u_2 into x_2 and x_5 respectively cancel themselves out. It then follows that $x_5 = \alpha_1 x_2$. From the last equation it follows that $x_6 = \lambda\alpha_2 x_2 + (1-\lambda)\alpha_2 x_2$ for any λ. Substituting x_5 for x_2 in the first term, one gets.

$$x_6 = \lambda\alpha_2\alpha_1 x_5 + (1-\lambda)\alpha_2 x_2 \qquad (++)$$

By construction, (++) holds for any λ. The problem for Cartwright is that if (++) is known to hold for a non-zero λ, then we have a known equation which is causally incorrect but functionally correct, in which every factor on the right hand

[31] Strictly speaking, Cartwright avoids using this terminology of direct causes. A fuller analysis would give a more careful description of her characterisation of a causal law. However, this simpler characterisation is an acceptable simplification of the concept for my purposes here

side has a KOBP with respect to x_6. This is a counterexample to Cartwright's 1989 theorem, since the theorem implies that x_5 is a genuine cause of x_6 when it is not.

So Cartwright's inferential claim can fail in cases where possible causes' OOBP cancel themselves out. Note also that imposing that the OOBP be known, that is, that a KOBP holds is of no help here, since one can know that a factor has an OOBP without knowing that it cancels itself out.

I propose to fix this by adding an additional condition to the definition of an OOBP: *no OOBP for a factor can have coefficient values that imply that the factors along it would have no net influence on the factor whose OOBP it is.* Since for almost all possible values of the coefficients this will hold, it is not a very restrictive additional assumption. In the subsequent analysis I assume that this condition holds.

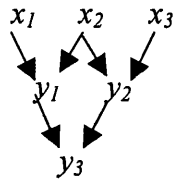
To conclude, the discussions here have shown up an ambiguity in Cartwright's analysis, emphasised that her result allows inference of causally correct equations but need not capture direct causal information and has presented a counterexample to her proof. To avoid the counterexample, a modification has been proposed that the OOBP condition be strengthened so that the influence of an OOBP not cancel out. With these clarifications, I now address the earlier question: how do the Cartwright and the S-approaches to inferring causal order compare?

5. Cartwright's Approach to Inferring Causal Order vs. the S-Approach

Ideally, this section would present a generalised formulation that would give conditions under which Cartwright's approach and the S-approach to inferring causal order would imply one another, and conditions when they did not. Unfortunately, the work carried out has yet to reach this level of analysis. So instead, I work by way of example to compare the two approaches.

5.1. An Example of Inferring Causal order using the Cartwright and the S-approach

The example assumed is the following. Assume that the following causal graph denotes the actual causal relations that are of interest.



Assume that the following is *known* by an experimenter:

- y_1 is caused by at least one of x_1 and/or x_2 and at most by both.
- y_2 is caused by at least one of x_2 and/or x_3 and at most by both.
- y_1 and y_2 have OOBP's with respect to y_3.
- $y_3 = ay_1 + by_2$ for known a and b.

The first two assumptions are made to keep analysis of the example simple.[32] While the second two assumptions match the conditions for Cartwright's inferential claim. Also assume that Cartwright's Generalised Reichenbach Principle holds and that the causal relation is transitive, in line with the other conditions for her inferential claim.

Given these assumptions, it follows from Cartwright's inferential claim that $y_3 = ay_1 + by_2$ is causally correct. Therefore, if the above assumptions are met then it can be deduced that both y_1 and y_2 are causes of y_3.

Now consider the same case using the S-approach. Recall that the S-approach works by assuming that the some general, identifiable equations are known to hold and that it is known that the true structural equations are identifiable and consistent with the general equations where some coefficients in the general equations may be zero. One then infers from observations which, if any, coefficients in this general form are zero to determine which of the possible structural forms holds.

[32] They can be relaxed and since Cartwright's conditions are still met, one can make the inference made here. However, doing this in this example would require a much larger number of possibilities be analysed, so I impose these extra assumptions.

So, assume that the experimenter knows that the general identifiable set of equations holds (for unknown coefficients).

$y_1 = \alpha_1 x_1 + \alpha_2 x_2$
$y_2 = \alpha_3 x_2 + \alpha_4 x_3$
$y_3 = \alpha_5 y_1 + \alpha_6 y_2$
(x's external, y's internal)[33]

Also assume, as above, that it is known that

- y_1 is caused by at least one of x_1 and/or x_2 and at most by both.
- y_2 is caused by at least one of x_2 and/or x_3 and at most by both.

In line with the S-approach, assume that it is known that the true structural equations are identifiable and consistent with the general set above, where one or more of the coefficients can be zero.

First, consider those possible sets of structural equations in which the knowledge about the possible causes of y_1 and y_2 is taken into account. This amounts to the assumption that one of the following sets of equations (where coefficients are non-zero in the first two equations)[34] is the set of true structural equations.

(a) $y_1 = \alpha_1 x_1 + \alpha_2 x_2$; $y_2 = \alpha_3 x_2 + \alpha_4 x_3$; $y_3 = \alpha_5 y_1 + \alpha_6 y_2$

(b) $y_1 = \alpha_1 x_1$; $y_2 = \alpha_3 x_2 + \alpha_4 x_3$; $y_3 = \alpha_5 y_1 + \alpha_6 y_2$

(c) $y_1 = \alpha_2 x_2$; $y_2 = \alpha_3 x_2 + \alpha_4 x_3$; $y_3 = \alpha_5 y_1 + \alpha_6 y_2$

(d) $y_1 = \alpha_1 x_1 + \alpha_2 x_2$; $y_2 = \alpha_3 x_2$; $y_3 = \alpha_5 y_1 + \alpha_6 y_2$

(e) $y_1 = \alpha_1 x_1 + \alpha_2 x_2$; $y_2 = \alpha_4 x_3$; $y_3 = \alpha_5 y_1 + \alpha_6 y_2$

(f) $y_1 = \alpha_1 x_1$; $y_2 = \alpha_3 x_2$; $y_3 = \alpha_5 y_1 + \alpha_6 y_2$

(g) $y_1 = \alpha_1 x_1$; $y_2 = \alpha_4 x_3$; $y_3 = \alpha_5 y_1 + \alpha_6 y_2$

(h) $y_1 = \alpha_2 x_2$; $y_2 = \alpha_3 x_2$; $y_3 = \alpha_5 y_1 + \alpha_6 y_2$

(i) $y_1 = \alpha_2 x_2$; $y_2 = \alpha_4 x_3$; $y_3 = \alpha_5 y_1 + \alpha_6 y_2$

Since the S-approach assumes that the true structural equations are identifiable, this rules out case (h) because in that case the third equation is not identifiable. In all of the other possible systems the three equations are identifiable. So if it is known that one of the above systems bar (h) is the true set of structural equations, then following the S-approach, one can infer the values of α_5 and α_6 are non-zero

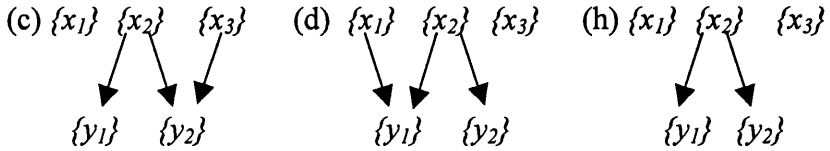
[33] The x's are external because they are determined outside the equations.
[34] I do not explicitly write down all the possibilities where the coefficients in the last equations take different non-zero values, because this is not important for the identifiability of the equations. Also, it helps keep the presentation of the set of possible systems to a manageable number.

in the third equation from observation in order to determine whether x_3 is caused by x_1 and x_2. Therefore, if it is known that $x_3 = ax_1 + bx_2$ for non-zero a and b, as assumed in the Cartwright analysis, this then implies that $a_5 = a$ and $a_6 = b$, so it can be inferred that both y_1 and y_2 are causes of y_3. So in this example, applying the S-approach gives the same result as Cartwright's method.

5.2. Comparing the Two Approaches

The key difference between the two approaches is that the S-approach assumes a known identifiable general form of equations with which the true identifiable structural relations are consistent, while Cartwright assumes that a KOBP holds for the factors in a known functional relation. But just how significant is this difference?

In the example, treated using the S-approach, a key assumption is that the experimenter knows the true structural form is identifiable. This rules out the case (h). To contrast this with Cartwright's OBP assumption, consider the different possible cases above (from (a) to (i)) where either y_1 or y_2 fails to have an OOBP with respect to y_3. These are cases where the conditions for Cartwright's inferential claim are not met. There are only three such cases: (c), (d) and (h). Their failure to meet the OOBP condition can be seen in their causal graphs for y_1 and y_2.



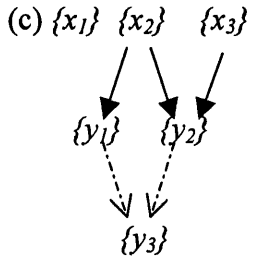
In each of these graphs, at least one of the two factors, y_1 and y_2, fails to have an OOBP with respect to y_3, because in each one factor only has a single cause which also causes the other factor. For instance, in case (c) y_1 is caused only by x_2 which may also cause y_3 via y_2 (if y_2 causes y_3), so y_1 does not have an OOBP with respect to y_3.

This shows that the KOBP assumption plays a somewhat similar role as the assumption in the S-approach that it is known that whatever the true structural equations are, they are identifiable. This is because the KOBP assumption, like the identifiability condition in the S-approach, rules out case (h). However, the

example also shows that it differs from identifiability condition since, unlike the identifiability condition, the KOBP condition also rules out cases (c) and (d).

The analysis of the last chapter, where it was shown that the identifiability required a certain sparseness of causal structure, is relevant here. Specifically, it was shown there that for a two-factor experiment to be possible between two factors in a mechanism (part of what was shown to be required for identifiability) either an 'open back path'[35] had to obtain between these two factors and the other factors[36] in the mechanism, or there had to be sufficiently many causal inputs to allow, by compensating 'cancelling out' changes, to change just those two factors.[37]

These two possible ways in which two-factor experiments can be carried out can be illustrated by case (c). Consider its causal graph (where the dashed arrows represent possible causal connections).



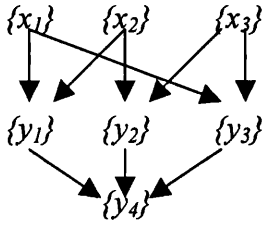
The first way a two-factor experiment is possible can be illustrated by y_2, which has an OOBP (x_3) with respect to y_3. Here it is possible by varying x_3 (y_2's OOBP for y_3) to perform a two-factor experiment between y_2 and y_3. This suggests that a factor having an OOBP implies that a two-factor experiment is possible. The second way a two-factor experiment is possible can be seen from y_1 which has no OOBP with respect to y_3. Here a two-factor experiment between y_1 and y_3 requires varying both x_2 and x_3, because x_2 needs to be varied to vary y_1 while x_3 needs to be varied to cancel out any unwanted influence of x_2 on y_2. The fact that having an OOBP seems to permit a two-factor experiment of the first type, while its absence appears to rule out this type of two-factor experiment, justifies the use of the 'open back path' terminology to describe the first kind of

[35] I use scare quotes because it is not exactly the same as Cartwright's OOBP.
[36] More precisely, there was a directly controllable factor that caused the two factors but not the other factors, see (a) section 4.4, chapter five.
[37] See (b), section 4.4, chapter five.

two factor experiment in chapter five.[38] It suggests a clear connection between Cartwright's OOBP and two-factor experiments.[39]

Finally, it is interesting to note that a system can be identifiable even though it has no open back paths. This can be seen in the following, simple system with its causal order on the right.

$$y_1 = \alpha_1 x_1 + \alpha_2 x_2$$
$$y_2 = \alpha_3 x_2 + \alpha_4 x_3$$
$$y_3 = \alpha_5 x_1 + \alpha_6 x_3$$
$$y_4 = \beta_1 y_3 + \beta_2 y_2 + \beta_3 y_3$$



It is easily checked, using the rank condition, that every equation in this system is identifiable. Importantly, the fourth equation can be read, in Cartwright's terms, as causally correct. However, no cause of y_4 (i.e. y_1, y_2 or y_3) has an OOBP with respect to y_4 in the causal order, because every cause of y_1, y_2 or y_3 also causes y_4 via another y. In this causal order the two-factor experiments that are possible for the pairs of factors in the last mechanism (by identifiability) can only be carried out by varying common causes together (i.e. x_1, x_2 and x_3). Here identifiability holds without open back paths.

This two last examples appear to show that the S-approach can be used in cases where Cartwright's cannot. In particular, there are identifiable systems with insufficient OBP's to use Cartwright's approach for which causal inference can be carried out using the S-approach. This may seem to show that the S-approach is more powerful that Cartwright's. However, this is too hasty a conclusion. I think judgment should be suspended until the strong assumptions made by the S-approach are considered in more detail. Recall that the S-approach assumes that it is known that only a finite set of structural equations are possible where each is identifiable. Although, if these assumptions are met, the S-approach may be able to perform causal inferences not possible using Cartwright's method, the really important question is what is required to limit the set of causal orders known to be possible in the way assumed by the S-approach. The concern is that in cases like

[38] See (a), section 4.4, chapter five.
[39] Though I leave a detailed analysis of this connection for further work.

the example just presented, the S-approach simply builds in more background knowledge than Cartwright's method, which allows it to be used in situations where Cartwright's cannot be used. So, until there is analysis of how this background knowledge is to be obtained, I think it is premature to claim an advantage of the S-approach over Cartwright's from examples like that just presented.

6. Conclusion

This chapter has asked how one can deduce causal order from observations and background knowledge. It has considered Simon's approach in his 1954 paper, proposed an alternative S-approach, and looked at Cartwright's OBP approach presented in her 1989 book. The S-approach and Simon's method are very similar, though the S-approach makes stronger assumptions about how structural equations are to be read and about the background causal knowledge to hand. It was shown that the failure of Simon's 1954 method to make these assumptions left it open to Nancy Cartwright's criticism that his key claim (that one can deduce causal order from knowledge of time order and uncorrelated errors) fails.

In addition, important counterexamples were constructed to Simon's key claim. One counterexample was of two observationally equivalent systems that met the time order and correlation assumptions required by Simon, but which had different causal order. This was then used to show a deeper problem with Simon's 1954 method: it relied on causal connections holding in order to solve the spurious correlation problem, but since the causal connections on which it relied were also possibly spurious, the method fails. In response, the S-approach was proposed as an alternative.

The last part of the chapter presented Nancy Cartwright's alternative approach and some criticisms of it. These criticisms were that there was an important ambiguity in her definition of the open back path and that her conditions for inferring causal order needed to be strengthened to rule out a case where a factor's open back path 'cancelled itself out'. Finally, the two inferential methods, Cartwright's and the S-approach were compared for simple examples. This showed important

similarities and differences between the two methods, and suggested interesting avenues for further work.

Appendix 6.1. Cartwright's Observationally Equivalent Counterexample

Cartwright's Time Ordered Example

In her example Cartwright presents two systems C and D. In system C, the error terms are orthogonal.

$$\begin{array}{lll} & z_1 = u_1 & z_1 = u_1 \\ \text{C} & z_2 = az_1 + u_2 & \text{D} \quad z_2 = az_1 + u_2 \\ & z_3 = bz_1 + u_3 & z_3 = cz_1 + dz_2 + v \end{array} \qquad \text{where} \quad \begin{array}{l} c = \frac{b}{a} + 1, d = -\frac{1}{a} \\ v = -z_1 + \frac{z_2}{a} + u_3 - \frac{u_2}{a} \end{array}$$

First note that for system D

$$v = -z_1 + \frac{z_2}{a} + u_3 - \frac{u_2}{a}$$

$$v = -u_1 + u_1 + \frac{u_2}{a} + u_3 - \frac{u_2}{a}$$

$$v = u_3$$

Since v is identical to u_3, v is orthogonal to u_1 and u_2, so the error terms in D are orthogonal.

However, System D is not consistent with System C. To see this subtract the last equation of system C from the last equation in D.

$$z_3 - z_3 = (c-b)z_1 + dz_2 + (v - u_3)$$

$$0 = (c-b)z_1 + dz_2$$

Substituting in equations for z_1 and z_2 from System C we get.

$$0 = (c-b)u_1 + d(bu_1 + u_2)$$

Multiplying by u_2 and taking expectations implies $d=0$

While multiplying by u_1 and taking expectations implies

$$0 = (c - b + db)$$

$$\Rightarrow \; c - b = 0$$

So, $c = b$ and $d = 0$.

Therefore, for system D with v as defined $c=b$ and $d = 0$, this implies that system D is identical to system C.

This shows that if one assumes like Cartwright that

$$c = \frac{b}{a} + 1, d = -\frac{1}{a}$$

$$v = -z_1 + \frac{z_2}{a} + u_3 - \frac{u_2}{a}$$

Then System D as defined is not consistent with System C because in order for D to be consistent given the definition of *v*, one must have *c=b* and *d=0*. □

Attempted Reformulation: The Impossibility of a suitable Equivalent System

$$\begin{array}{ll} z_1 = u_1 & z_1 = u_1 \\ \text{C } z_2 = az_1 + u_2 & \text{D' } z_2 = az_1 + u_2 \\ z_3 = bz_1 + u_3 & z_3 = cz_1 + dz_2 + v \end{array}$$

The aim is to construct C system D' from C , which has orthogonal errors in which *d* is non-zero. Now solving for *v* in terms of *u*'s one gets.

$$v = bz_1 + u_3 - cz_1 - d(az_1 + u_2)$$

$$v = (b - c - ad)z_1 + u_3 - du_2$$

$$v = (b - c - ad)u_1 - du_2 + u_3$$

For the error terms in D' to be orthogonal one must have $E(vu_1) = E(vu_2) = 0$.

$$E(vu_1) = (b\text{-}c\text{-}ad) = 0$$

$$E(vu_2) = -d = 0$$

Therefore for D' to have orthogonal error terms *d* must be zero. Moreover, if *d=0* zero then *c* = *b*, so D' is then identical to C. Therefore, it is not possible to construct the required counterexample in this case. □

Appendix 6.2. Identifiability of Lower Triangular Systems Simon Analyses

Preliminaries:

Consider the following system of n equations, the aim is to show that it is identifiable.

$$z_1 = u_1$$

$$z_2 = a_{21}z_1 + u_2$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots$$

$$z_n = a_{n1}z_1 + \ldots + a_{nn-1}z_{n-1} + u_n \quad \text{where} \quad E(u_i) = 0 \text{ for all } i,$$

$$Cov(u_i, u_j) = 0, \text{ for } i,j \text{ distinct}$$

Bringing all the z terms to the left hand side, we can represent the set of equations as the following:

$$Az = u$$

where

$$A = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -a_{21} & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ -a_{n1} & -a_{n2} & \cdots & 1 \end{bmatrix}, \; z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \text{ and } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}.$$

[*Aside*: Note that A has full rank since it is impossible to make a non-trivial linear combination of any subset of its rows sum to zero. Thus the inverse of A, A^{-1}, exists. Since A is lower triangular its inverse A^{-1} is lower triangular (the inverse of a lower (upper) triangular matrix is itself lower (upper) triangular). Also it is important to note the following that the product of two lower (upper) triangular matrices has as a diagonal the product of each of the corresponding diagonal terms, that is,

$$\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ k_{21} & d_2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ k_{n1} & k_{n2} & \cdots & d_n \end{bmatrix} \begin{bmatrix} e_1 & 0 & \cdots & 0 \\ l_{21} & e_2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ l_{n1} & l_{n2} & \cdots & e_n \end{bmatrix} = \begin{bmatrix} d_1e_1 & 0 & \cdots & 0 \\ m_{21} & d_2e_2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ m_{n1} & m_{n2} & \cdots & d_ne_n \end{bmatrix} \quad (*)$$

which in turn implies that A^{-1} has the same form as A i.e. has a diagonal of 1's.]

The problem is to show that the following representation is unique, that is, identifiable. So the aim is to prove the following theorem.

Theorem 6.1 Given $Az = u$, where A is lower triangular with a diagonal of ones, $E(u) = 0$ and $E(uu^T) = D$, where D is diagonal with positive diagonal coefficients, then the system is identifiable, that is there does not exist any B and u^* such that $Bz = u^*$, and either $B \neq A$ or $u^* \neq u$, where B is lower triangular with a diagonal of ones, $E(u^*) = 0$ and $E(u^*u^{*T}) = D^*$, where D^* is diagonal with positive diagonal coefficients.

Proof: Assume it is false, that is, there exists, B and u^* such that $Bz = u^*$, and either $B \neq A$ or $u^* \neq u$, where B is lower triangular with a diagonal of ones, $E(u^*) = 0$ and $E(u^*u^{*T}) = D^*$, where D^* is diagonal with positive diagonal coefficients.

Given its form B (like A) is invertible. Therefore we have the following.

$$z = A^{-1}u = B^{-1}u^* \qquad (1)$$

It follows that

$$u^* = BA^{-1}u \qquad (2)$$

and $u = AB^{-1}u^*$

Hence,

$$\begin{aligned} E(u^*u^{*T}) &= E(BA^{-1}u(BA^{-1}u)^T) \\ &= E(BA^{-1}uu^T(A^{-1})^TB^T) \\ &= BA^{-1}E(uu^T)(A^{-1})^TB^T \\ &= BA^{-1}D(A^{-1})^TB^T \end{aligned}$$

But the left hand side $= D^*$ so we have,

$$D^* = BA^{-1}D(A^{-1})^TB^T$$

Thus,

$$AB^{-1}D^* = D(A^{-1})^TB^T$$

Now the left hand side is the product of three lower triangular matrices and is thus lower triangular. Similarly, the right hand side is the product of three upper triangular matrices and is thus upper triangular. Therefore, the identity asserts that both sides are both upper and lower triangular, that is diagonal. So $AB^{-1}D^*$ is diagonal, let us say, D_1.

$$AB^{-1}D^* = D_1$$

$$AB^{-1} = D_1D^{*-1} \qquad (D^{*-1} \text{ exists since } D^*\text{'s diagonal coefficients are all positive, also } D^{*-1} \text{ is diagonal})$$

The right hand side is diagonal, so AB^{-1} is diagonal also. But now consider (*) it implies that B^{-1} has only 1's in its diagonal, which in turn means that the product of A and B^{-1} by (*) must also have only 1's in the diagonal, since both A and B^{-1} do. Thus AB^{-1} is the diagonal matrix with 1's in its diagonal, that is, the identity. We have shown that

$$AB^{-1} = \boldsymbol{I}$$

so $$A = B$$

Substituting into (2) it follows that

$$u^* = AA^{-1}u = u$$

Hence, $B = A$, $u^* = u$. This is a contradiction so the result follows.□

Appendix 6.3. A Time-Ordered Counterexample to Simon's 1954 Claim

Assume as before that system C holds, where the variables are time ordered according to index, where the *u*'s have zero mean, variance *1*, and are orthogonal.

$$\text{C} \qquad \begin{aligned} z_1 &= u_1 \\ z_2 &= az_1 + u_2 \\ z_3 &= bz_1 + u_3 \end{aligned}$$

Now assume that another system D' for some *c*, v_2 and v_3 also holds.

$$\text{D'} \qquad \begin{aligned} z_2 &= v_2 \\ z_3 &= cz_2 + v_3 \end{aligned}$$

By construction the variables in D' have correct time order. If D' is to serve as a counterexample, it must have orthogonal errors. So if we can solve for *c*, v_2 and v_3 so that v_2 and v_3 are orthogonal, we will have constructed the required counterexample.

To solve for *c*, v_2 and v_3, note that if D' holds then by substituting out z_2 and z_3 from D' using the equations in C we get.

$$\begin{aligned} au_1 + u_2 &= v_2 \\ bu_1 + u_3 &= cau_1 + cu_2 + v_3 \end{aligned} \quad \Rightarrow \quad \begin{aligned} v_2 &= au_1 + u_2 \\ v_3 &= (b - ac)u_1 - cu_2 + u_3 \end{aligned}$$

Since the *u*'s have zero mean, so must the *v*'s. Therefore v_2 and v_3 are orthogonal if and only if $E(v_2v_3) = 0$. But given $E(u_1u_2) = 0$, and $E(u_1^2)$ and $E(u_2^2)$ are both *1*, $E(v_2v_3) = 0$ is equivalent to.

$$a(b\text{-}ac) - c = 0$$

Solving for *c* we get

$$c = \frac{ab}{(1+a^2)}$$

Substituting *c* back into the equations gives the required system

$$\text{D'} \qquad \begin{aligned} z_2 &= v_2 \\ z_3 &= \frac{ab}{1+a^2} z_2 + v_3 \end{aligned} \quad \textit{where} \quad \begin{aligned} v_2 &= au_1 + u_2 \\ v_3 &= \frac{b}{1+a^2} u_1 - \frac{ab}{1+a^2} u_2 + u_3 \end{aligned}$$

Since the system was derived from C, and has orthogonal errors and correct time order among the variables it meets the conditions to be a counterexample.□

Appendix 6.4. An Attempted Extension of the Simon Counterexample

Assume that a system (I), with its causal order, time ordered variables and orthonormal error terms (the w's) holds. Assume all c's are non-zero.

$$\text{(I)} \qquad \begin{aligned} z_0 &= w_0 \\ z_1 &= c_{10}z_0 + w_1 \\ z_2 &= c_{20}z_0 + c_{21}z_1 + w_2 \\ z_3 &= c_{30}z_0 + c_{31}z_1 + c_{32}z_2 + w_3 \end{aligned}$$

The aim is to construct a spurious system (II) with orthogonal errors which, read causally, asserts that z_1 causes both z_2 and z_3, but z_2 does not cause z_3. In other words, the equations of (II) will have form.

$$\text{(II)} \qquad \begin{aligned} z_1 &= u_1 \\ z_2 &= a_{21}z_1 + u_2 \\ z_3 &= a_{31}z_1 + u_3 \end{aligned}$$

To construct this system from (I) we must solve for the a's in terms of the c's requiring that the error terms in (II) are orthogonal. This is done, by substituting out the z's to solve for the u's in terms of w's, and then imposing the orthogonality constraints on the u's to solve for the a's. Solving for the u's in terms of the w's yields:

$$\begin{aligned} u_1 &= c_{10}w_0 + w_1 \\ u_2 &= \left[c_{20} + (c_{21} - a_{21})c_{10}\right]w_0 + (c_{21} - a_{21})w_1 + w_2 \\ u_3 &= (c_{30} + c_{20}c_{32})w_0 + (c_{31} - a_{31} + c_{32}c_{21})(c_{10}w_0 + w_1) + c_{32}w_2 + w_3 \end{aligned}$$

Since the w's have zero mean, the u's have zero mean, so the orthogonality requirement on the error terms in (II) is equivalent to $E(u_1u_2) = 0$, $E(u_1u_3) = 0$ and $E(u_2u_3) = 0$. Now, if one calculates these $E(u_iu_j)$'s and imposes the three orthogonality conditions, one obtains three equations in the c's and the a's. If one can solve these for the a's one is done. However, there is a problem. There are only two unknown a's (a_{21} and a_{31}) and there are three equations relating c's to the a's (one for each orthogonality condition). This implies that there will be *no* solution for the two a's unless these three equations are functionally dependent. It is easily checked that the equations are only functionally dependent provided the c's are functionally related in some appropriate way in (I). Since for almost all values of the c's this will not be the case, it follows that for almost all values of c it will not be possible to solve for the a's. In other words, barring the occurrence

of particular functional dependencies among the c's in (I), it will not be possible to construct the spurious system (II) from (I). So, the attempted counterexample fails.

Moving Forward From Here

This thesis developed an explicit causal interpretation, a 'strong reading', for simple sets of simultaneous linear equations used in econometrics. This was done by building on Herbert Simon's definition of causal order. It then explored some important features of the causal interpretation such as the relationship between causal order and changes in factors, the invariance of mechanisms to factor changes and the independence of directly controllable factors. It also investigated different kinds of intervention, how the standard identification conditions could be causally interpreted and how unknown causal orders could be inferred from observation. In the thesis relevant work of important philosophers and economists, Nancy Cartwright, Kevin Hoover, Stephen LeRoy and Herbert Simon was also presented and critically analysed. This brought out important similarities and differences between their definitions of causal order, their methods for finding out about causal order and those of the strong reading.

As discussed in the introduction, the aim of the thesis is to clarify the causal concepts assumed in econometric modelling and to clarify the methods by which causal relationships are discovered in econometrics. Both of these are crucial if one is to understand clearly the claims of econometric studies and to set out the strengths and weaknesses of the methods that these studies use. Ultimately, this is motivated by policy relevance. Econometric studies inform economic policy decisions that influence us all, therefore it is important that is clear just what the econometric 'view' of causes is, and just how it goes about finding out about causes.

This aim suggests one way the work of the thesis might be extended. Ideally, the strong reading would be extended to the point at which it can be used to clarify actual, important econometric studies, for example, ground-breaking studies such as Adams *et al.* (2003).[1] Of course, to do this requires a great deal of further work. For some studies, this would require setting out a clear relationship between

[1] See chapter one.

the reading of causal order set out here and Granger causality.[2] The relationship between Granger causality and structural views of causality, such as that presented in this thesis, is an important topic in econometrics.[3] In addition, the extension of the strong reading to cover more complex sets of equations might also allow an analysis of different exogeneity concepts used in econometrics. This is suggested by the brief analysis of weak and super exogeneity set out at the end of chapter three. Parallel to this work of extending the sets of equations to which the strong reading can be applied, would be work to analyse methods for finding out about causal relations. For instance, how might econometric tests for exogeneity be understood causally? What about model selection methods, such as the LSE methodology? As was done here with the identification conditions in chapter five, there remains valuable work to be done in clarifying what causal interpretations such important methods of econometrics have.

The above shows a rich potential for further work in econometrics. However, in setting out an explicit causal reading of sets of equations, the thesis also presents the beginnings of a theory of causal relations.[4] This gives the work many possible avenues of development in relation to current philosophical analyses of causal relations. For instance, the strong reading sets out that causal relations arise from the joint actions of mechanisms. This shows potential to connect the strong reading in this thesis with current philosophical analyses of mechanisms. What exactly is a mechanism? What concepts of mechanism are appropriate for the reading developed here? Progress on these questions could be made by investigating the recent literature on causality and mechanisms and connecting it with the strong reading proposed here.[5]

Another way the work here could be developed would be to extend the formal analysis. This work, begun in appendix 2.1, would develop a rigorous

[2] This would be the case for Adams *et al.* since Granger causality plays a key role in their analysis.
[3] See, for example, Hoover (2001a, pp.150-155).
[4] Though it does not claim to provide a theory of causal relations that can apply in all situations. A very important outstanding question is when a formal treatment of causal relations, like the strong reading, can be applied successfully and when it can't.
[5] See Steel (2004), Glennan (1996) for some recent work on mechanisms and causality. Cartwright's work (1989; 1999) on capacities and nomological machines is also relevant, as are Woodward (2003) and Elster (1998).

formalisation of causal relations in the strong reading.[6] In addition, a set theoretical treatment could also facilitate a rigorous introduction of probabilities using measure-theoretical probability theory. Moreover, a suitably developed strong reading should yield interesting connections with other formal theories of causal relations. In particular, there is good reason to believe that the algebraic approach of the strong reading here should, under certain conditions, be compatible with the graph theoretical, Bayes-net approaches to causality. Indeed, this hope is expressed by Hoover (2001a, p.191-192) and some work connecting structural theories of causal relations with Bayes-nets approaches has already been carried out.[7]

As this very brief survey shows, there are many ways in which the work of this thesis could be developed. Moreover, the proposed work is ultimately of relevance to the larger goal of clarifying the strengths and limits of structural modelling in econometrics. For example, exploring what concepts of mechanisms can be joined to the strong reading opens up possible ontological discussions about econometrics. In other words, if econometric methods assume that structural equations denote mechanisms with certain features, one can then investigate the extent to which the systems studied by economics *do* have these features. Similarly, work that extends the formalism of the strong reading by making explicit connections with other well-developed analyses, such as Bayes-nets methods, opens a way of bringing existing rich work, such as that on Bayes-nets, to bear on econometrics.[8] So in conclusion, though this thesis takes a first few steps in exploring these interesting issues, a long and exciting road remains to be travelled.

[6] Ideally, the analysis would be extended to cover mechanisms denoted by non-linear functions.
[7] See Pearl (2000, chap 7).
[8] There is existing work on Bayes-nets and econometrics see, for example, Spirtes (2005) which applies Bayes-net semantics to econometric models.

References

Adams, Peter, Micheal D. Hurd, Daniel McFadden, Angela Merrill, and Tiago Ribeiro (2003) 'Healthy, Wealthy and Wise? Tests for direct causal paths between health and socioeconomic status', *Journal of Econometrics*, 112, pp.3-56.

Adda, Jérôme, Tarani Chandola, and Michael Marmott (2003) 'Socio-economic status and health: causality and pathways', *Journal of Econometrics*, 112, pp.57-63.

Carnap, Rudolf (1932) 'The Elimination of Metaphysics Through Logical Analysis of Language' in A.J. Ayer (ed.), *Logical Positivism*. Glencoe, Ill: The Free Press.

Cartwright, Nancy (1983) *How the Laws of Physics Lie*, Oxford: Clarendon Press.

Cartwright, Nancy (1989) *Nature's Capacities and Their Measurement*, Oxford: Clarendon Press.

Cartwright, Nancy (1995) 'Probabilities and Experiment', *Journal of Econometrics*, 67, pp.47-59.

Cartwright, Nancy (1999) *The Dappled World: A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.

Cartwright, Nancy (2001) 'Modularity: It Can - and Generally Does – Fail' in D. Costantini, M.C. Galavotti and P. Suppes (eds.) *Stochastic Dependence and Causality*, Stanford: CSLI Publications.

Cartwright, Nancy (2002) 'How Not to Use Invariance in Causal Explanation', *Abstract for International Congress on Causation and Explanation in Natural and Social Sciences*, Centre for Logic and Philosophy of Science, Ghent University, Belgium (May 2002). http://logica.rug.ac.be/censs2002/abstracts/Cartwright.htm.

Cartwright, Nancy (2003a) 'Two Theorems on Invariance and Causality', *Philosophy of Science*, 70, pp.203-223.

Cartwright, Nancy (2003b) 'Causes from Statistics à la Simon: a Primer', Manuscript, London School of Economics.

Cartwright, Nancy (2003c) 'Causation one Word Many Things', *Causality: Metaphysics and Methods Technical Report,* CTR 07-03, Centre for the Philosophy of the Natural and Social Science, London School of Economics.

Cartwright, Nancy (forthcoming) *Hunting Causes and Using Them*, Cambridge: Cambridge University Press.

Collins, John, Ned Hall and L. A. Paul (2004) *Causation and Counterfactuals*, Cambridge, Mass: M.I.T. Press.

Cooley, Thomas and Edward Prescott (1976) 'Estimation in the Presence of Stochastic Parameter Variation', *Econometrica*, 44, pp.167-184.

Elster, Jon (1998) 'A Plea for Mechanisms' in Peter Hedstrøm and Richard Swedberg (eds.) *Social Mechanisms: An Analytical Approach to Social Theory*, Cambridge: Cambridge University Press.

Engle, Robert, David Hendry and Jean-Francois Richard (1983) 'Exogeneity' reprinted in David Hendry *Econometrics - Alchemy or Science?*, 2nd edition, Oxford: Oxford University Press.

Fisher, Franklyn (1966) *The Identification Problem in Econometrics*, Huntington: Krieger Publishing Company.

Florens, Jean-Pierre (2003) 'Some technical issues in defining causality', *Journal of Econometrics*, 112, pp.127-128.

Frisch, Ragnar (1938) 'Autonomy of Economic Relations' reprinted in David Hendry and Mary Morgan (eds.), *The Foundations of Econometric Analysis*, Cambridge: Cambridge University Press.

Geweke, John (2003) 'Econometric issues in using the AHEAD panel' *Journal of Econometrics*, 39, pp.115-120.

Glennan, Stuart (1996) 'Mechanisms and the Nature of Causation' *Erkenntnis*, 4, pp.49-71.

Glymour, Clark (1983) Seminar Notes, reprinted in Nancy Cartwright, *Nature's Capacities and Their Measurement*, Oxford: Clarendon Press.

Granger, Clive (1980) 'Testing for Causality: A Personal Viewpoint', *Journal of Economic Dynamics and Control*, 2, 4, pp.329-352.

Granger, Clive (1988) 'Some Recent Developments in a Concept of Causality' *Journal of Econometrics*, 39, pp.199-211.

Granger, Clive (2003) 'Some aspects of causal relationships', *Journal of Econometrics*, 112, pp.69-71.

Gujarati, Damodar (1995) *Basic Econometrics*, 3rd edition, New York: McGraw-Hill.

Haavelmo Trygve, (1944) 'The Probability Approach in Econometrics', *Econometrica*, 12, supplement., pp.iii-vi, pp.1-115.

Hall, Ned (2000) 'Causation and the Price of Transitivity', *Journal of Philosophy*, 97, pp.198-222.

Halliday, David and Resnick, Robert (1978) *Physics Parts 1 & 2*, New York: John Wiley and Sons.

Harvey, Andrew (1990) *The Econometric Analysis of Time Series*, 2nd edition, Cambridge, Mass: M.I.T. Press.

Hausman, Daniel M. (1998) *Causal Asymmetries*, Cambridge: Cambridge University Press.

Hausman, Daniel M. and James Woodward (1999) 'Independence, Invariance and the Causal Markov Condition', *British Journal for the Philosophy of Science*, 50, pp.521-583.

Hausman, Jerry A. (2003) 'Triangular structural model specification and estimation with application to causality', *Journal of Econometrics*, 112, pp.107-113.

Heckman James (2000), 'Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective', *Quarterly Journal of Economics*, 115, pp.45-97.

Heckman, James (2001) *Econometrics, Counterfactuals and Causal Models*, Keynote Address, International Statistical Institute, Seoul, South Korea.

Heckman, James (2003) 'Conditioning, causality and policy analysis', *Journal of Econometrics*, 112, pp.73-78.

Heckman, James (2004) *The Scientific Model of Causality*, manuscript, University of Chicago.

Hendry, David (1995) *Dynamic Econometrics*, Oxford: Oxford University Press.

Hendry, David (2000) *Econometrics: Alchemy or Science?*, 2nd edition, Oxford: Oxford University Press.

Hitchcock, Christopher (2001) 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy*, 98, pp.273-299.

Hoover Kevin (1995), 'Facts and Artifacts: Calibration and the Empirical Assessment of Real-Business Cycle Models", *Oxford Economic Papers*, 47, pp.24-44.

Hoover, Kevin (2001a) *Causality in Macroeconomics*, Cambridge: Cambridge University Press.

Hoover, Kevin (2001b) *The Methodology of Empirical Economics*, Cambridge: Cambridge University Press.

Hoover, Kevin (2003) 'Some causal lessons from macroeconomics', *Journal of Econometrics*, 112, pp.121-125.

Hoover, Kevin (2004) 'Lost Causes', *Journal of the History o f Economic Thought*, 26, 1, pp.149-164.

Hume, David (1739) *A Treatise of Human Nature*, L.A. Selby-Bigge (ed.), Oxford: Oxford University Press.

Koopmans, Tjalling (1949) 'Identification Problems in Economic Model Construction', *Econometrica*, 17, 2, pp.125-144.

Koopmans, Tjalling (1950) 'When is an Equations System Complete for Statistical Purposes?' reprinted in David Hendry and Mary Morgan (eds.), *The Foundations of Econometric Analysis*, Cambridge: Cambridge University Press.

Koopmans, Tjalling and William Hood (1953) 'The Estimation of Simultaneous Linear Economic Relationships' in Tjalling Koopmans and William Hood, *Studies in Econometric Method*, New York: John Wiley and Sons.

LeRoy, Stephen (1995) 'Causal Orderings' in *Macroeconometrics: Developments, Tensions and Prospects*, Hoover, Kevin (ed.), Boston: Kluwer Academic Publishers.

LeRoy, Stephen (1999) 'On Policy Regimes' in *The Legacy of Robert Lucas Jr.* Vol. 2, Hoover, Kevin (ed.), Aldershot: Edward Elgar.

LeRoy, Stephen (2003) 'Causality in Economics', Manuscript, University of California, Santa Barbara, California.

LeRoy, Stephen (2004) 'Causality in Economics', *Causality: metaphysics and methods technical reports*, CTR 20/04, Centre for the Philosophy of the Natural and Social Sciences, London School of Economics.

Lucas, Robert (1976) 'Econometric Policy Evaluation: A Critique' in (1981) *Studies in Business Cycle Theory*, Oxford: Basil Blackwell.

Mackie, John L. (1974) *The Cement of the Universe: A Study of Causation*, Oxford: Clarendon Press.

Maddala, G. S. (2001) *Introduction to Econometrics*, 3rd edition, New York: John Wiley and Sons.

Malinas, Gary and John Bigelow (2004) 'Simpson's Paradox' in Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2004 Edition). http://plato.stanford.edu/archives/spr2004/entries/paradox-simpson

Mealli, Fabrizia and Donald B. Rubin (2003) 'Assumptions allowing the estimation of direct causal effects', *Journal of Econometrics*, 112, pp.79-87.

Mill, John Stuart (1851) *A System of Logic Ratiocinative and Inductive – Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*, Books I-III, Robson, J. M. and McRae (eds.), University of Toronto Press, Routledge and Kegan Paul.

Morgan, Mary (1990) *The History of Econometric Ideas*, Cambridge: Cambridge University Press.

Morgan, Mary (1991) 'The Stamping Out of Process Analysis from Econometrics' in Neil de Marchi and Mark Blaug (eds.), *Appraising Economic Theories: Studies in the Methodology of Research Programmes*, Aldershot: Edward Elgar.

Orcutt, Guy (1952) 'Actions, Consequences and Causal Relations' reprinted in David Hendry and Mary Morgan (eds.), *The Foundations of Econometric Analysis*, Cambridge: Cambridge University Press.

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

Poterba, James M. (2003) 'Some observations on health status and economic status', *Journal of Econometrics*, 112, pp.65-67.

Reiss, Julian (2003) 'Practice Ahead of Theory: Instrumental Variables, Natural Experiments and Inductivism in Econometrics.' *Causality: metaphysics and methods technical reports*, CTR 12/03, Centre for the Philosophy of the Natural and Social Sciences, London School of Economics.

Robins, James M. (2003) 'General Methodological Considerations', *Journal of Econometrics*, 112, pp.89-106.

Russell, Bertrand (1913) 'On the Notion of Cause', *Proceedings of the Aristotelian Society*, 13, pp.1-26.

Simon, Herbert (1952) 'On the Definition of the Causal Relation' reprinted in Herbert Simon, *Models of Man*, New York: John Wiley and Sons.

Simon, Herbert (1953) 'Causal Ordering and Identifiability' reprinted in Herbert Simon, *Models of Man*, New York: John Wiley and Sons.

Simon, Herbert (1954), 'Spurious Causation: A Causal Interpretation' reprinted in Herbert Simon *Models of Man*, New York: John Wiley and sons.

Simon, Herbert (1957) *Models of Man*, New York: John Wiley and Sons.

Simon, Herbert and Nicolas Rescher (1966) 'Cause and Counterfactual', *Philosophy of Science*, 33, pp.323-240.

Spirtes, Peter (2005) 'Graphical models, causal inference, and econometric models', *Journal of Economic Methodology*, 12, 1, pp.3-34.

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

Steel, Daniel (2004) 'Social Mechanisms and Causal Inference' Philosophy of Social Science, 34, 1, pp.55-78.

Strang, Gilbert (1980) *Linear Algebra and its Applications*, New York: Academic Press.

Suppe, Frederick (1998) Operationalism. In E. Craig (Ed.), Routledge Encyclopedia of Philosophy. London: Routledge. Retrieved January 19, 2004, from http://www.rep.routledge.com/article/Q077SECT1

Suppes, Patrick (1970) *A Theory of Probabilistic Causality*, Amsterdam: North-Holland Publishing Company.

Tinbergen, Jan (1939) *Statistical Testing of Business Cycle Theories: A Method and its Application to Investment Activity* reprinted in David Hendry and Mary Morgan (eds.) *The Foundations of Econometric Analysis*, Cambridge: Cambridge University Press.

Williamson, Jon (2004) *Bayesian Nets and Causality: Philosophical and Computational Foundations*, Oxford: Clarendon Press.