Graphical Methods for Identification in Structural Equation Models

Carlos Eduardo Fisch de Brito
June 2004

The dissertation of Carlos Eduardo Fisch de Brito is approved.

_____
Peter Bentler

_____
Adnan Darwiche

_____
Richard Korf

_____
Judea Pearl, Committee Chair

University of California, Los Angeles

2004

To my parents and wife

TABLE OF CONTENTS

LIST OF FIGURES

Brito, C. and Pearl, J. A Graphical Criterion for the Identification of Causal Effects in Linear Models. In *AAAI'02 - The Eighteenth National Conference on Artificial Intelligence*, Edmonton, Alberta, Canada.

Brito, C. and Pearl, J. A New Identification Condition for Recursive Models with Correlated Errors. In *Structural Equation Models*, 2002.

Brito, C. and Silva, E. S. and Diniz, M. C. and Leao, R. M. M. Analise Transiente de Modelos de Fonte Multimidia. In *SBRC'00 - 18o Simposio Brasileiro de Redes de Computadores*, Belo Horizonte, MG, Brazil.

Brito, C. and Moraes, R. and Oliveira, D. and Silva, E. S. Comunicacao Multicast Confiavel na Implementacao de uma Ferramenta Whiteboard. In *SBRC'99 - 17o Simposio Brasileiro de Redes de Computadores*, Salvador, BA, Brazil.

ABSTRACT OF THE DISSERTATION

Graphical Methods for Identification in Structural Equation Models

by

Carlos Eduardo Fisch de Brito

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2004

Professor Judea Pearl, Chair

Structural Equation Models (SEM) is one of the most important tools for causal analysis in the social and behavioral sciences (e.g., Economics, Sociology, etc). A central problem in the application of SEM models is the analysis of Identification. Succintly, a model is identified if it only admits a unique parametrization to be compatible with a given covariance matrix (i.e., observed data). The identification of a model is important because, in general, no reliable quantitative conclusion can be derived from non-identified models.

In this work, we develop a new approach for the analysis of identification in SEM, based on graph theoretic techniques. Our main result is a general sufficient criterion for model identification. The criterion consists of a number of graphical conditions on the causal diagram of the model. We also develop a new method for computing correlation constraints imposed by the structural assumptions, that can be used for model testing. Finally, we also provide a generalization to the traditional method of Instrumental Variables, through the concept of Instrumental Sets.

CHAPTER 1

Introduction

Structural Equation Models (SEM) is one of the most important tools for causal analysis in the social and behavioral sciences [Bol89, Dun75, McD97, BW80, Fis66, KKB98]. Although most developments in SEM have been done by scientists in these areas, the theoretical aspects of the model provide interesting problems that can benefit from techniques developed in computer science.

In a structural equation model, the relationships among a set of observed variables are expressed by linear equations. Each equation describes the dependence of one variable in terms of the others, and contains a stochastic error term accounting for the influence of unobserved factors. Independence assumptions on pairs of error terms are also specified in the model.

An attractive characteristic of SEM models is their simple causal interpretation. Specifically, the linear equation $Y = \beta X + e$ encodes two distinct assumptions: (1) the possible existence of (direct) causal influence of X on Y; and, (2) the absence of (direct) causal influence on Y of any variable that does not appear on the right-hand side of the equation. The parameter β quantifies the (direct) causal effect of X on Y. That is, the equation claims that a unit increase in X would result in β units increase of Y, assuming that everything else remains the same.

Let us consider a simple example taken from [Pea00a]. This model investigates the relations between smoking (X) and lung cancer (Y), taking into consideration the

amount of tar (Z) deposited in a person's lungs, and allowing for unobserved factors to affect both smoking (X) and cancer (Y). This situation is represented by the following equations:

$$X = \varepsilon_1$$
$$Z = aX + \varepsilon_2$$
$$Y = bZ + \varepsilon_3$$
$$cov(\varepsilon_1, \varepsilon_2) = cov(\varepsilon_2, \varepsilon_3) = 0$$
$$cov(\varepsilon_1, \varepsilon_3) = \gamma$$

The first three equations claim, respectively, that the level of smoking of a person depends only on factors not included in the model, the amount of tar deposited in the lungs depends on the level of smoking as well as external factors, and the level of cancer depends on the amount of tar in the lungs and external factors. The remaining equations say that the external factors that cause tar to be accumulated in the lungs are independent of the external factors that affect the other variables, but the external factors that have influence on smoking and cancer may be correlated.

All the information contained in the equations can be expressed by a graphical representation, called *causal diagram*, as illustrated in Figure 1.1. We formally define the model and its graphical representation in section 2.1.

Figure 1.2 shows a more elaborate model used to study correlations between relatives for systolic and diastolic blood pressures [TEM93]. The squares represent blood pressures for each type of individual, and the circles represent genetic and environmental causes of variation: A the additive genetic contribution of the polygenes; D the dominance genetic contributions; E all the environmental factors; and S those environmental components only shared by siblings of the same sex.
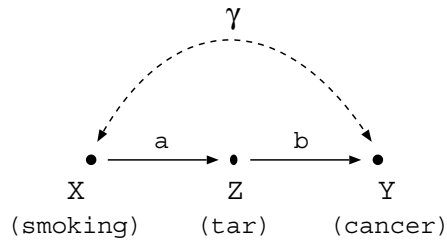
Figure 1.1: Smoking and lung cancer example

1.1 Data Analysis with SEM and the Identification Problem

The process of data analysis using Structural Equation Models consists of four steps [KKB98]:

1. *Specification*: Description of the structure of the model. That is, the qualitative relations among the variables are specified by linear equations. Quantitative information is generally not specified and is represented by parameters.

2. *Identification*: Analysis to decide if there is a unique valuation for the parameters that make the model compatible with the observed data. The identification of a SEM model is formally defined in Section 2.1.

3. *Estimation*: Actual estimation of the parameters from statistical information on the observed variables.

4. *Evaluation of fit*: Assessment of the quality of the model as a description of the data.

In this work, we will concentrate on the problem of Identification. That is, we leave the task of model specification to other investigators, and develop conditions to decide if these models are identified or not. The identification of a model is important because, in general, no reliable quantitative conclusion can be derived from a

3
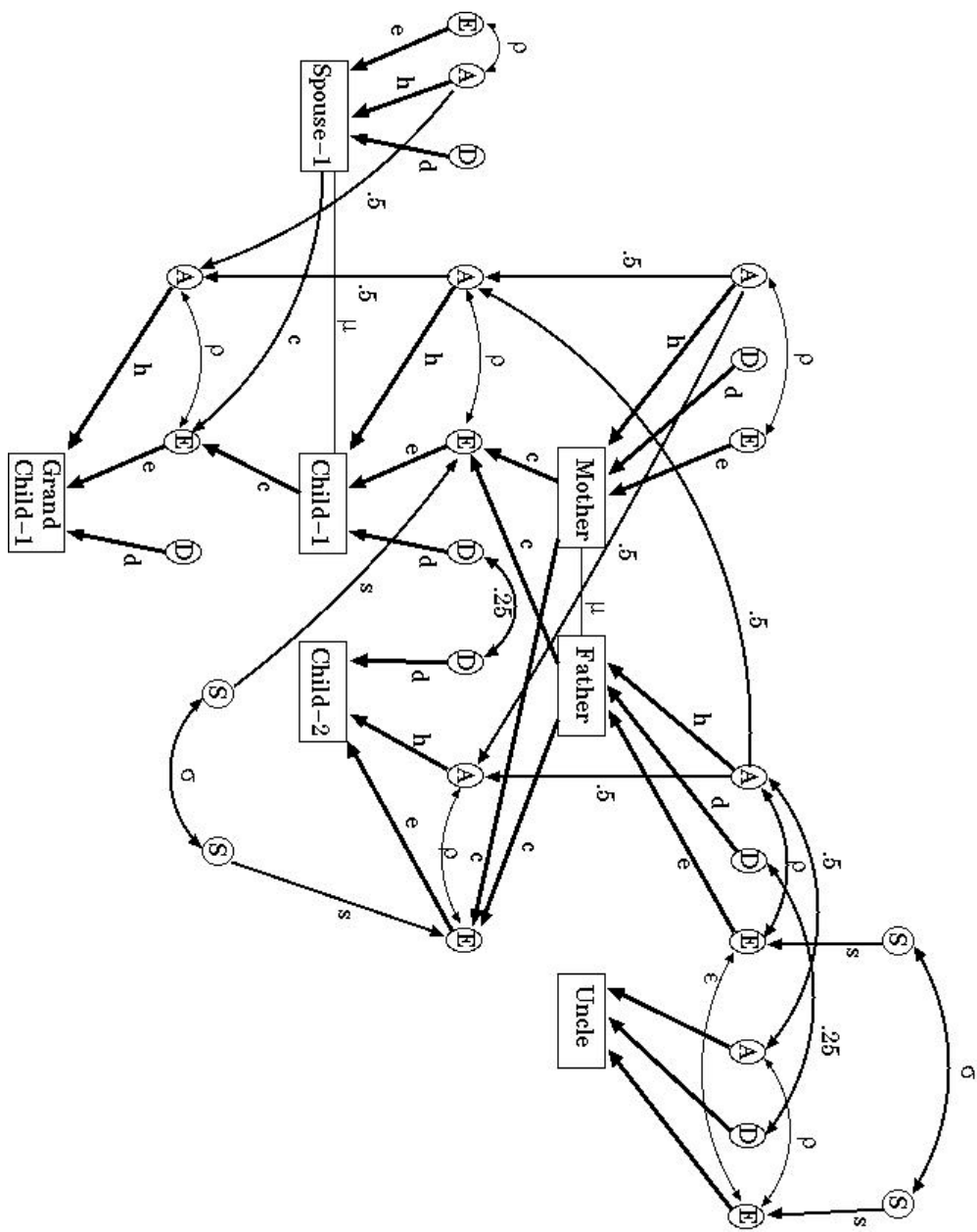
Figure 1.2: Model for correlations between blood pressures of relatives

non-identified model. The question of identification has been the object of extensive research [Fis66], [Dun75], [Pea00a], [McD97], [Rig95]. Despite all this effort, the problem still remains open. That is, we do not have a necessary and sufficient condition for model identification in SEM. Some results are available for special classes of models, and will be reviewed in Section 1.3.

In our approach to the problem, we state Identification as an intrinsic property of the model, depending only on its structural assumptions. Since all such assumptions are captured in the graphical representation of the model, we can apply graph theoretic techniques to study the problem of Identification in SEM. Thus, our main results consist of graphical conditions for identification, to be applied on the causal diagram of the model.

As a byproduct of our analysis, estimation methods will also be provided, as well as methods to obtain constraints imposed by the model on the distribution over the observed variables, thus addressing questions in steps 3 and 4 above.

1.2   Overview of Results

The central question studied in this work is the problem of identification in recursive SEM models, that is, models that do not contain feedbacks (see Section 2.1 for more details). The basic tool used in the analysis is Wright's decomposition, which allows us to express correlation coefficients as polynomials on the parameters of the model. The important fact about this decomposition is that each term in the polynomial corresponds to a path in the causal diagram.

Based on the observation that these polynomials are linear on specific subsets of parameters, we reduce the problem of Identification to the analysis of simple systems of linear equations. As one should expect, conditions for linear independence of those

systems (which imply a unique solution and thus identification of the parameters), translate into graphical conditions on the paths of the causal diagram.

Hence, the fundamental step in our method of Auxiliary Sets for model identification (see Chapter 3) consists in finding, for each variable Y, a set of variables $\mathcal{A}_Y$ with specific restrictions on the paths between Y and each variable in $\mathcal{A}_Y$.

As it turns out, the restrictions that allow us to obtain the maximum generality from the method are not so easy to verify by visual inspection of the causal diagram. To overcome this problem, we developed an algorithm that searches for an Auxiliary Set for a given variable Y (see Chapter 4). We also provide the Bow-free condition and the Instrumental condition (Section 3.5), which are special cases of the general method, but have straightforward application.

The machinery developed for the method of Auxiliary Sets can also be used to compute correlation constraints. These constraints are implied by the structural assumptions, and allow us to test the model [McD97]. The basic idea is very simple. While in the case of Identification we take advantage of linearly independent equations, it follows that correlation constraints are immediately obtained from linearly dependent equations (see Chapter 5). Despite its simplicity, this is a very powerfull method for computing correlation constraints.

The main goal of this research is to solve the problem of Identification for recursive models. Namely, to obtain a necessary and sufficient condition for model identification. The sufficient condition provided by the method of Auxiliary Sets is very general, and in Chapter 6 we present our initial efforts on our attempt to prove that it is also necessary for identification.

Finally, in Chapter 7, we consider the problem of parameter identification. This problem is motivated by the observation that even on non-identified models there may exist some parameters whose value is uniquely determined by the structural assump-

tions and data. We provide a solution based on the concept of Instrumental Sets, which generalizes the traditional method of Instrumental Variables [BT84]. The criterion for parameter identification involves d-separation conditions, and the proofs required the development of new techniques of independent interest.

1.3   Related Work

The use of graphical models to represent and reason about probability distributions has been extensively studied [WL83, CCK83, Pea88]. In many areas such models have become the standard representation, e.g., Bayesian networks for dealing with uncertainty in Artificial Intelligence [Pea88], and Markov random fields for speech recognition and coding [KS80]. Some reasons for the success of the language of graphs in many domains are: it provides a compact representation for a large class of probability distributions; it is convenient to describe dependencies among variables; and it consists of a natural language for causal modeling. Besides these advantages, many methods and techniques were developed to reason about probability distributions directly at the level of the graphical representation [LS88, HD96]. An example of such a technique is the d-separation criterion [Pea00a], which allows us to read off conditional independencies among variables by inspecting the graphical representation of the model.

The Identification problem has been tackled in the past half century, primarily by econometricians and social scientists [Fis66, Dun75]. It is still unsolved. In other words, we are not in possession of a necessary and sufficient criterion for deciding whether the parameters in a structural model can be determined uniquely from the covariance matrix of the observed variables.

Certain restricted classes of models are nevertheless known to be identifiable, and these are often assumed by social scientists as a matter of convenience or convention

(1)   (2)   (3)
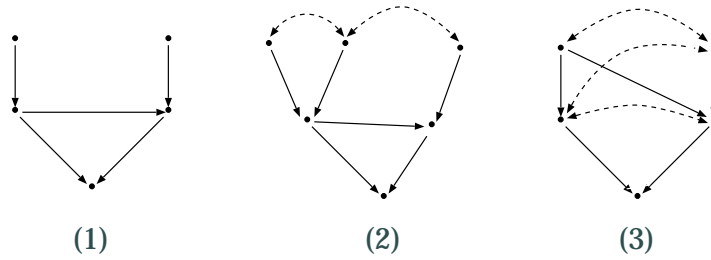
Figure 1.3: McDonald's regressional hierarchy examples

[Dun75]. McDonald [1997] characterizes a hierarchy of three such classes (see Figure 1.3): (1) uncorrelated errors, (2) correlated errors restricted to exogenous variables, and (3) correlated errors restricted to pairs of causally unordered variables (i.e., variables that are not connected by uni-directed paths.). The structural equations in all three classes are regressional (i.e., the error term in each equation is uncorrelated with the explanatory variables of that same equation) hence the parameters can be estimated uniquely using Ordinary Least Squares techniques.

Traditional approaches to the Identification problem are based on algebraic manipulation of the equations defining the model. Powerful algebraic methods have been developed for testing whether a specific parameter, or a specific equation in a model is identifiable . However, such methods are often too complicated for investigators to apply in the pre-analytic phase of model construction. Additionally, those specialized methods are limited in scope. The rank and order criteria [Fis66], for example, do not exploit restrictions on the error covariances (if such are available). The rank criterion further requires precise estimate of the covariance matrix before identifiability can be decided. Identification methods based on block recursive models [Fisher, 1966; Rigdon, 1995], for another example, insist on uncorrelated errors between any pair of ordered blocks.

Recently, some advances have been achieved on graphical conditions for identi-

fication [Pea98, Pea00a, SRM98]. Examples of such conditions are the "back-door" and "single-door" criteria [Pea00a, pp. 150–2]. The backdoor criterion consists of a d-separation test applied to the causal diagram, and provides a sufficient condition for the identification of specific causal effects in the model. A problem with such conditions is that they are applicable only in sparse models, that is, models rich in conditional independence. The same holds for criteria based on instrumental variables (IV) (Bowden and Turkington, 1984), since these require search for variables (called *instruments*) that are uncorrelated with the error terms in specific equations.

CHAPTER 2

Problem Definition and Background

2.1 Structural Equation Models and Identification

A structural equation model M for a vector of observed variables $\mathbf{Y} = [Y_1, \ldots, Y_n]'$ is defined by a set of linear equations of the form

$$Y_j = \sum_i c_{ji} Y_i + e_j \qquad , \text{ for } j = 1, \ldots, n.$$

Or, in matrix form

$$\mathbf{Y} = C \cdot \mathbf{Y} + \varepsilon$$

where $C = [c_{ji}]$ and $\varepsilon = [e_1, \ldots, e_n]'$.

The term e_j in each equation corresponds to an stochastic error, assumed to have normal distribution with zero mean. The model also specifies independence assumptions for those error terms, by the indication of which entries in the matrix $\Psi = [\psi_{ij}] = Cov(e_i, e_j)$ have value zero.

In this work, we consider only recursive models, which are characterized by the fact that the matrix C is lower triangular. This assumption is reasonable in many domains, since it basically forbids feedback causation. That is, a sequence of variables $Y_1, \ldots, Y_k$ where each Y_i appears in the right-hand side of the equation for Y_{i+1}, and variable Z_k appears in the equation for Z_1.

The structural assumptions encoded in a model M consist of:

(1) the set of variables omitted in the right-hand side of each equation (i.e., the zero entries in matrix C); and,

(2) the pairs of independent error terms (i.e., zero entries in Ψ).

The set of parameters of model M, denoted by Θ, is composed by the (possibly) non-zero entries of matrices C and Ψ.

A parametrization π for model M is a function $\pi : \Theta \rightarrow \Re$ that assigns a real value to each parameter of the model. The pair $\langle M, \pi \rangle$ determines a unique covariance matrix over the observed variables, given by [Bol89]:

$$\Sigma_M(\pi) = \left(I - C(\pi)\right)^{-1} \Psi(\pi) \left[\left(I - C(\pi)\right)^{-1}\right]^T \tag{2.1}$$

where $C(\pi)$ and $\Psi(\pi)$ are obtained by replacing each non-zero entry of C and Ψ by the respective value assigned by π.

Now, we are ready to define formally the problem of Identification in SEM.

Definition 1 (Model Identification) *A structural equation model M is said to be* identified *if, for almost every parametrization π for M, the following condition holds:*

$$\Sigma_M(\pi) = \Sigma_M(\pi') \implies \pi = \pi' \tag{2.2}$$

That is, if we view parametrization π as a point in $\Re^{|\Theta|}$, then the set of points in which condition (2.2) does not hold has Lebesgue measure zero.

The identification status of simple models can be determined by explicitly calculating the covariances between the observed variables, and analyzing if the resulting

11

expressions imply a unique solution for the parameters. This method is illustrated in the following examples.

Consider the model defined by the equations:

$$
\begin{cases}
X &= e_X \\
W &= e_W \\
Y &= aX + e_Y \\
Z &= bY + cW + e_Z
\end{cases}
\qquad
\begin{cases}
Cov(e_X, e_W) = \alpha \\
Cov(e_Y, e_W) = \beta
\end{cases}
\qquad (2.3)
$$

where the covariances of pairs of error terms not listed above are assumed to be zero. We also make the assumption that each of the observed variables has zero mean and is standardized (i.e., has variance 1). This assumption is not important because, if this is not the case, a simple transformation can put the variables in this form. Immediate consequences of this last assumption are: $Cov(X, Y) = E[X \cdot Y]$ and $Var(X) = E[X^2]$.

Calculating the covariances between observed variables, we obtain:

$$
\begin{aligned}
Cov(X, Y) &= E[X \cdot Y] \\
&= E[X \cdot (aX + e_Y)] \\
&= aVar(X) + Cov(e_X, e_Y) \\
&= a
\end{aligned}
\qquad (2.4)
$$

and, by similar derivations,

$$
\begin{aligned}
Cov(X, W) &= \alpha \\
Cov(Y, W) &= a\alpha + \beta \\
Cov(Y, Z) &= b + c\beta \\
Cov(W, Z) &= b\beta + c \\
Cov(X, Z) &= ab + c\alpha
\end{aligned}
\qquad (2.5)
$$

12

Now, it is easy to see that the values of parameters a, α, β are uniquely determined by the covariances $Cov(X, Y)$, $Cov(X, W)$ and $Cov(Y, W)$. Parameters b and c are obtained by solving the system formed by the expressions for $Cov(Y, Z)$ and $Cov(W, Z)$. Hence, if two parametrizations π and π' induce the same covariance matrix they must be identical, and the model is identified.

Note, however, that this argument does not hold if parameter β is exactly 1. In this case, the equations for $Cov(Y, Z)$ and $Cov(W, Z)$ do not allow us to obtain a unique solution for parameters b and c. A unique solution can still be obtained from the expression for $Cov(X, Z)$, but if we also have $a = \alpha = 1$, then the parameters b and c are not uniquely determined by the covariance matrix of the observed variables. This explains why we only require condition (2.2) to hold for almost every parametrization, and allow it to fail in a set of measure zero.

The simplest example of a non-identified model corresponds to:

$$
\begin{cases}
Y_1 & = & e_1 \\
Y_2 & = & aY_1 + e_2 \\
Cov(e_1, e_2) = \beta
\end{cases}
$$

In this case, the covariance matrix for the observed variables Y_1, Y_2 contains only one entry, whose value is given by:

$$
\begin{aligned}
Cov(Y_1, Y_2) & = & E[Y_1 \cdot Y_2] - E[Y_1] \cdot E[Y_2] \\
& = & E[Y_1 \cdot (aY_1 + e_2)] \\
& = & aVar(Y_1) + Cov(e_1, e_2) \\
& = & a + \beta
\end{aligned}
$$

Now, given any parametrization π, it is easy to construct another one parametrization $\pi' \neq \pi$, with $\pi(a) + \pi(\beta) = \pi'(a) + \pi'(\beta)$. But this implies that $\Sigma_M(\pi) = \Sigma_M(\pi')$, and so the model is non-identified.

$$X = e_X$$
$$W = e_W$$
$$Y = aX + e_Y$$
$$Z = bY + cW + e_Z$$
$$Cov(e_X, e_W) = \alpha$$
$$Cov(e_Y, e_W) = \beta$$
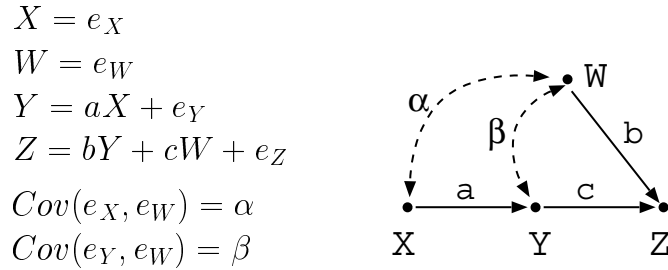
Figure 2.1: A simple structural model and its causal diagram

In general, if a model M is non-identified, for each parametrization π there exists an infinite number of distinct parametrizations π' such that $\Sigma_M(\pi) = \Sigma_M(\pi')$. However, it is also possible that for most parametrizations π only a finite number of distinct parametrizations generate the same covariance matrix. We will return to this issue in Chapter 6, where we study sufficient conditions for non-identification, and will provide an example of this situation.

A few other algebraic methods, like algebra of expectations [Dun75], have been proposed in the literature. However, those techniques are too complicated to analyze complex models. Here, we pursue a different strategy, and study the identification status of SEM models using graphical methods. For this purpose, we introduce the graphical representation of the model, called a causal diagram [Pea00a].

The causal diagram of a model M consists of a directed graph whose nodes correspond to the observed variables $Y_1, \ldots, Y_n$ in the model. A directed edge from Y_i to Y_j indicates that Y_i appears on the right-hand side of the equation for Y_j with a non-zero coefficient. A bidirected arc between Y_i and Y_j indicates that the corresponding error terms, e_i and e_j, have non-zero correlation. The graphical representation can be completed by labeling the directed edges with the respective coefficients of the linear equations, and the bidirected arcs with the non-zero entries of the covariance matrix
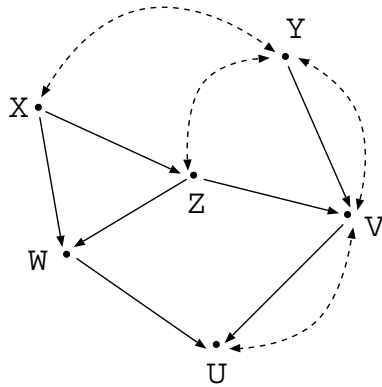
Figure 2.2: A causal diagram

Ψ. Figure 2.1 shows the causal diagram for the example given in Eq. (2.3). Note that the causal diagram of a recursive model does not have any cycle composed only of directed edges.

The next section presents some basic definitions and facts about the type of directed graphs considered here. Then, in section 2.3 we establish the connection between the Identification problem and the graphical representation of the model.

2.2   Graph Background

A *path* between variables X and Y in a causal diagram consists of a sequence of edges $\langle e_1, e_2, \ldots, e_n \rangle$ such that e_1 is incident to X, e_n is incident to Y, and every pair of consecutive edges in the sequence has a common variable. Variables X and Y are called the extreme points of the path, and every other variable appearing in some edge e_i is said to be an intermediate variable in the path. We say that the path points to extreme point X (Y) if the edge e_1 (e_n) has an arrow head pointing to X (Y).

For example, the following are some of the paths between X and U in the causal diagram of Figure 2.2:

- $X \to Z \to V \to U$

- $X \to Z \to V \leftrightarrow U$

- $X \leftrightarrow Y \leftrightarrow Z \to W \to U$

- $X \to Z \to V \leftrightarrow Y \leftrightarrow Z \to W \to U$

Note that only the third path points to variable X, but all of them point to U.

A path $p = \langle e_1, \ldots, e_n \rangle$ between X and Y is *valid* if variable X only appears in e_1, variable Y only appears in e_n, and every intermediate variable appears in exactly two edges in the path. Among the examples above, only the first three are valid. The last one is invalid because variable Z appears in more than two edges.

The special case of a path composed only by directed edges, all of which oriented in the same direction, is called a *chain*. The first example above corresponds to a chain from X to U.

We will also make use of a few family terms to refer to variables in particular topological relationships. Specifically, if the edge $X \to Y$ is present in the causal diagram, then we say that X is a *parent* of Y. Similarly, if there exists a chain from X to Y, then X is said to be an *ancestor* of Y, and Y is a *descendant* of X. Clearly, in a recursive model, we cannot have the situation where X is both an ancestor and a descendant of some other variable Y. In the causal diagram of Figure 2.2, W and V are the parents of variable U, and X is an ancestor of both U and V.

Given a path p between X and Y, and an intermediate variable Z in p, we denote by $p[X..Z]$ the path consisting of the edges of p that appear between X and Z. [1]

Variable Z is a *collider* in path a p between X and Y, if both $p[X..Z]$ and $p[Z..Y]$

---

[1] Here, and in most of the following, we are only concerned about valid paths, so this concept is well-defined

point to Z. A path that does not contain any collider is said to be *unblocked*. Next, we consider a few important facts about unblocked paths.

Define the *depth* of a node Y in a causal diagram as the length (i.e., number of edges) of the longest chain from any ancestor of Y to Y. Nodes with no ancestors have depth 0.

Lemma 1 *Let X and Y be nodes in the causal diagram of a recursive model such that depth(X) $\geq$ depth(Y). Then, every path between X and Y which includes a node Z with depth(Z) $\geq$ depth(X) must have a collider.*

Proof: Consider a path p between X and Y and node Z satisfying the conditions above. We observe that Z cannot be an ancestor of either X or Y, otherwise we would have $depth(Z) < depth(X)$ or $depth(Z) < depth(Y)$.

Now, consider the subpath of p between Z and Y. If this subpath has the form $Z \rightarrow \ldots Y$, then it must contain a collider, since it cannot be a directed path from Z to Y. Similarly, if the subpath of p between X and Z has the form $X \ldots \leftarrow Z$, then it must contain a collider.

In all the remaining cases Z is a collider blocking the path. $\qquad \square$

If p is a path between X and Y, and q is a path between Y and Z, then $p \oplus q$ denotes the path obtained by the concatenation of the sequences of edges corresponding to p and q.

Lemma 2 *Let p be an unblocked path between X and Y, and let Y be an unblocked path between Y and Z. Then, $p \oplus q$ is a valid unblocked path between X and Z if and only if:*

(i) p and q do not have any intermediate variable in common;
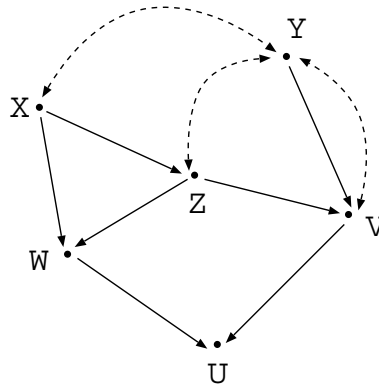
Figure 2.3: A causal diagram

(ii) either p is a chain from Y to X, or q is a chain from Y to Z.

Definition 2 (*d-separation*)

*A set of nodes **Z** d-separates X from Y in a graph, if Z closes every path between X and Y. A path p is closed by a set **Z** (possibly empty) if one of the following holds:*

*(i) p contains at least one non-collider that is in **Z**;*

*(ii) p contains at least one collider that is outside **Z** and has no descendant in **Z**.*

For example, consider the path $X \leftrightarrow Y \rightarrow V \rightarrow U$ in Figure 2.3. This path is closed by any set containing variables Y or V. On the other hand, the path $X \leftrightarrow Y \leftrightarrow Z \rightarrow W \rightarrow U$ is closed by the empty set $\{\}$, but is not closed by any set containing Y or V but not Z or W. It is easy to verify by inspection that the set $\{V, W\}$ closes every path between X and U, and so $\{V, W\}$ d-separates X from U.

2.3 Wright's Method of Path Analysis

The method of path analysis [Wri34] for identification is based on a decomposition of the correlations between observed variables into polynomials on the parameters of the

18
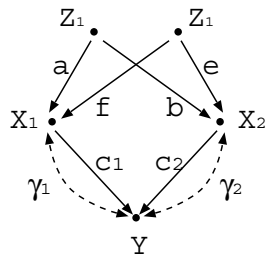
$$\rho_{Z_1X_1} = a \qquad \rho_{Z_1X_2} = b$$
$$\rho_{Z_2X_1} = f \qquad \rho_{Z_2X_2} = e$$

$$\rho_{Z_1Y} = ac_1 + bc_2$$
$$\rho_{Z_2Y} = fc_1 + ec_2$$

$$\rho_{X_1Y} = c_1 + \gamma_1 + (ab + ef)c_2$$
$$\rho_{X_2Y} = (ab + ef)c_+1 + c_2 + \gamma_2$$

Figure 2.4: Wright's equations.

model. More precisely, for variables X and Y in a recursive model, the correlation coefficient of X and Y, denoted by ρ_{XY}, can be expressed as:

$$\rho_{X,Y} = \sum_{\text{paths } p_l} T(p_l) \tag{2.6}$$

where the term $T(p_l)$ represents the product of the parameters of the edges along path p_l, and the summation ranges over all unblocked paths between X and Y. For this equality to hold, the variables in the model must be standardized (i.e., variance equal to 1) and have zero mean. We refer to Eq.(2.6) as Wright's decomposition for ρ_{XY}. Figure 2.4 shows a simple model and the decompositions of the correlations for each pair of variables.

The set of equations obtained from Wright's decompositions summarizes all the statistical information encoded in the model. Therefore, any question about identification can be decided by studying the solutions for this system of equations. However, since this is a system of non-linear equations, it can be very difficult to analyze the identification of large models by directly studying the solutions for these equations.

CHAPTER 3

Auxiliary Sets for Model Identification

3.1 Introduction

In this chapter we investigate sufficient conditions for model identification. Specifically, we want to find graphical conditions on the causal diagram that guarantee the identification of every parameter in the model. One example of the type of result obtained here is the Bow-Free Condition, which states that every model whose causal diagram has at most one edge connecting any pair of variables is identified.

The starting point for our analysis of identification is the set of equations provided by Wright's decompositions of correlations. Then, we make the following important observation.

For an arbitrary variable Y, let S be a set of incoming edges to Y (i.e., edges with an arrow head pointing to Y). Then, any unblocked path in the causal diagram can include at most one edge from S. This follows because if two such edges appear in a valid path, then they must be consecutive. But since both edges point to Y (e.g., $\ldots \to Y \leftrightarrow \ldots$), the path must be blocked.

Now, recall that each term in the polynomial of Wright's decomposition corresponds to an unblocked path in the causal diagram. Thus, the observation above implies that such polynomials are linear in the parameters of the edges in S.

Hence, our approach to the problem of Identification in SEM can be summarized

as follows. First, we partition all the edges in the causal diagram into sets of incoming edges. Then, we study the identification of the parameters associated with each set by analyzing the solution of a system of linear equations.

Two conditions must be satisfied to obtain the identification of the parameters corresponding to a set of edges S. First, there must exist a sufficient number of linearly independent equations. Second, the coefficients of these equations, which are functions of other parameters in the model, must be identified.

To address the first issue, we developed a graphical characterization for linear independence, called the G Criterion. That is, for a fixed variable Y, if a set of variables $\{Z_1, \ldots, Z_k\}$ satisfies the graphical conditions established by the G criterion, then the decompositions of $\rho_{Z_1 Y}, \ldots, \rho_{Z_k Y}$ are linearly independent (with respect to the parameters of edges in S). These conditions are based on the existence of specific unblocked paths between Y and each of the Z_i's.

The second point is addressed by establishing an appropriate order to solve the systems of equations.

The following sections will formally develop this graphical analysis of identification.

3.2   Basic Systems of Linear Equations

We begin by partitioning the set of edges in the causal diagram into sets of incoming edges.

Fix an ordering Δ for the variables in the model, with the only restriction that if $depth(X) < depth(Y)$, then X must appear before Y in Δ. For each variable Y, we define $Inc(Y)$ as the set of edges in the causal diagram that connect Y to any variable appearing before Y in the ordering Δ.

21

It easily follows from this definition that, for each variable Y, $Inc(Y)$ contains all directed edges pointing to Y (i.e., $X \to Y$).

Lemma 3 *Any unblocked path between Y and some variable Z can include at most one edge from $Inc(Y)$. Moreover, if $depth(Z) \leq depth(Y)$, then any such path must include exactly one edge from $Inc(Y)$.*

Proof: The first part of the lemma follows from the argument given in Section 3.1. For the second part, assume that p is an unblocked path between Z and Y, which does not contain any edge from $Inc(Y)$. Let W be the variable adjacent to Y in path p. Clearly, $depth(W) \geq depth(Y)$ (otherwise edge (W,Y) would belong to $Inc(Y)$). But then Lemma 1 saysays that p contains a collider, which is a contradiction. $\qquad\square$

Now, fix an arbitrary variable Y, and let $\lambda_1, \ldots, \lambda_m$ denote the parameters of the edges in $Inc(Y)$. Then, Lemma 3 allows us to express Wright's decomposition of the correlation between Z and Y as a linear equation on the λ_j's:

$$\rho_{Z,Y} = a_0 + \sum_{j=1}^{m} a_j \cdot \lambda_j$$

where $a_0 = 0$ if $depth(Z) < depth(Y)$.

Figure 3.1 shows the linear equations obtained from the correlations between Y and every other variable in a model.

Now, given a set of variables $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$, we let $\Phi_{\mathbf{Z},Y}$ [1] denote the system of equations corresponding to the decompositions of correlations $\rho_{Z_1 Y}, \ldots, \rho_{Z_k Y}$:

$$\begin{cases} \rho_{Z_1 Y} = a_{10} + \sum_{j=1}^{m} a_{1j} \cdot \lambda_j \\ \ldots \\ \rho_{Z_k Y} = a_{k0} + \sum_{j=1}^{m} a_{kj} \cdot \lambda_j \end{cases}$$

---

[1]Whenever clear from the context, we drop the reference to Y and simply write $\Phi_{\mathbf{Z}}$.
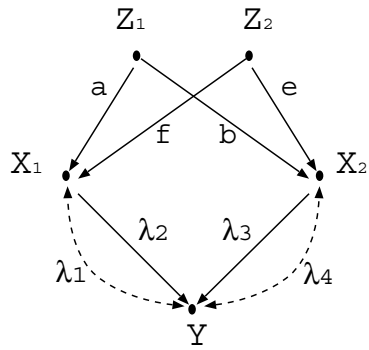
$$\rho_{Z_1 Y} = a\lambda_2 + b\lambda_3$$
$$\rho_{Z_2 Y} = f\lambda_2 + e\lambda_3$$
$$\rho_{X_1 Y} = \lambda_1 + \lambda_2 + (ab + ef)\lambda_3$$
$$\rho_{X_1 Y} = (ab + ef)\lambda_2 + \lambda_3 + \lambda_4$$

Figure 3.1: Wright's equations.

3.3 Auxiliary Sets and Linear Independence

Following the ideas presented in Section 3.1, we would like to find a set of variables that provides a system of linearly independent equations. This motivates the following definition:

Definition 3 (Auxiliary Sets) *A set of variables* $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ *is said to be an Auxiliary Set with respect to* Y *if and only if the system of equations* $\Phi_{\mathbf{Z}, Y}$ *is linearly independent.*

Next, we obtain sufficient graphical conditions for a given set of variables to be an auxiliary set for Y. Since the terms in Wright's decompositions correspond to unblocked paths, it is natural to expect that linear independence between equations translate into properties of such paths. In the following, we explore this connection by analyzing a few examples, and then we introduce the G criterion.

For each of the models in Figure 3.2 we will verify if the set $\mathbf{Z} = \{Z_1, Z_2\}$ qualifies as an Auxiliary Set for Y. In model M_1, the system of equations $\Phi_{\mathbf{Z}}$ is given by:

23
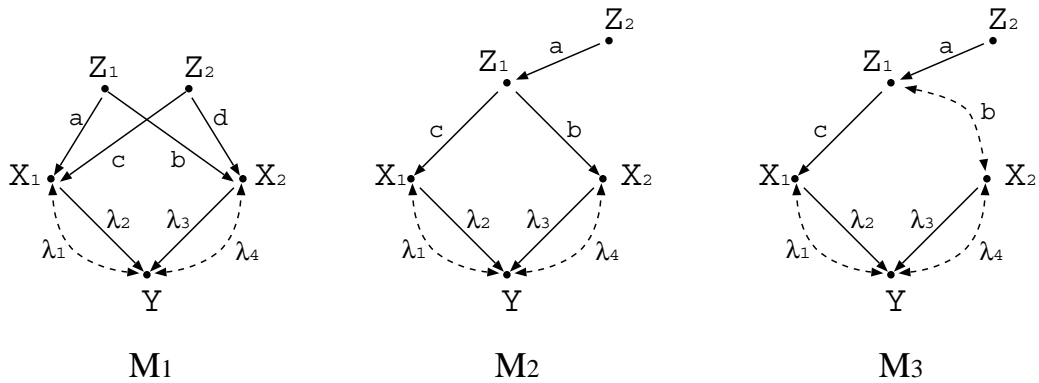
Figure 3.2: Figure

$$
\begin{cases}
\rho_{Z_1 Y} & = \quad a\lambda_2 + b\lambda_3 \\
\rho_{Z_2 Y} & = \quad c\lambda_2 + d\lambda_3
\end{cases}
$$

It is easy to see that the equations are linearly independent [2], and so $\mathbf{Z}$ is an Auxiliary Set for Y. We also call attention to the fact that unblocked paths $p_1 : Z_1 \rightarrow X_1 \rightarrow Y$ and $p_2 : Z_2 \rightarrow X_2 \rightarrow Y$ have no intermediate variables in common.

In model M_2, system $\Phi_{\mathbf{Z}}$ is formed by:

$$
\begin{cases}
\rho_{Z_1 Y} & = \quad a\lambda_2 + b\lambda_3 \\
\rho_{Z_2 Y} & = \quad ca\lambda_2 + cb\lambda_3 \quad = \quad c \cdot [a\lambda_2 + b\lambda_3]
\end{cases}
$$

Clearly, the equations are not linearly independent in this case. This occurs because every unblocked path between Z_1 and Y in M_2 can be extended by the edge $Z_1 \rightarrow Z_2$ to give an unblocked path between Z_2 and Y.

Finally, in model M_3, the system $\Phi_{\mathbf{Z}}$ is given by:

---

[2]This is not true if *ad = bc*, but this condition only holds on a set of measure zero (see discussion in Section 2.1)

24
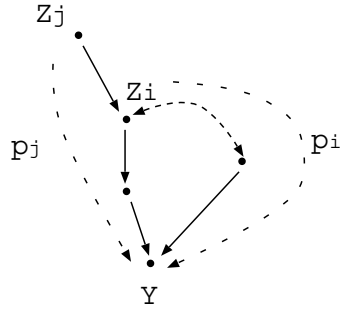
Figure 3.3: Example illustrating condition (ii) in the G criterion

$$\begin{cases} \rho_{Z_1 Y} & = & a\lambda_2 + b\lambda_3 \\ \rho_{Z_2 Y} & = & ca\lambda_2 \end{cases}$$

and again we obtain a pair of linearly independent equations. The important fact to note here is that, if we extend path $Z_1 \leftrightarrow X_2 \rightarrow Y$ by edge $Z_2 \rightarrow Z_1$, we obtain a path blocked by Z_1.

In general, the situation can become much more complicated, with one equation being a linear combination of several others. However, as we will see in the following, the examples discussed above illustrate the essential graphical properties that characterize linear independence.

G Criterion: *A set of variables* $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ *satisfies the* G criterion *with respect to* Y *if there exist paths* $p_1, \ldots, p_k$ *such that:*

  (i) p_i *is an unblocked path between* Z_i *and* Y *including some edge from* $Inc(Y)$;

  (ii) *for* $i < j$, Z_i *is the only possible common variable in paths* p_i *and* p_j *(other than* Y*), and in this case, both* p_i *and* $p_j[Z_j..Z_i]$ *must point to* Z_i *(see Figure 3.3 for an example).*

25

Next, we prove a technical lemma, and then establish the main result of this Chapter.

Let $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ be a set of variables, and assume that paths $p_1, \ldots, p_k$ witness the fact that $\mathbf{Z}$ satisfies the G criterion with respect to Y. Let $\lambda_1, \ldots, \lambda_m$ denote the parameters of the edges in $Inc(Y)$, and, without loss of generality, assume that, for $1 \leq i \leq k$, path p_i contains the edge with parameter λ_i. (It is easy to see that condition (ii) of the G criterion does not allow paths p_i and p_j to have a common edge.)

Lemma 4 *For $j > i$, let p be an unblocked path between Z_j and Y including the edge with parameter λ_i. Then, p must contain an edge that does not appear in any of $p_1, \ldots, p_n$.*

Proof: Without loss of generality, we may assume that, for all $s, t < l \leq k$, if both variables Z_s, Z_t appear in path p_l, and Z_s is an intermediate variable in $p_l[Z_l..Z_t]$, then $s > t$. If this is not the case, then we can always rename the variables such that this condition holds, and condition (ii) of the G criterion is not violated.

Now, let p be a path satisfying the conditions of the lemma, and assume that p contains only edges appearing in $p_1, \ldots, p_k$. In the following we show that p must be blocked by a collider.

Clearly, we can divide the path p into segments $q_1, \ldots, q_h$ such that all the edges in each segment belong to the same path p_l.

Now, note that variable Z_j can appear only in a path p_t for $t \geq j$ (from condition (ii) of the G criterion). On the other hand, the edges of the last segment q_h belong to path p_i.

Since $j > i$, there must exist two consecutive segments q_s, q_{s+1} and indices $a \geq j > b$, such that the edges in q_s belong to p_a and the edges in q_{s+1} belong to p_b.

The common variable of q_s and q_{s+1} appear in both p_a and p_b. Since $a > b$, it must be Z_b, and q_{s+1} must point to it.

If the edges in q_s belong to subpath $p_a[Z_a..Z_b]$, then q_s also points to Z_b. In this case, Z_b is a collider in path p and the lemma follows.

In the other case, q_s cannot be the first segment of path p, and we consider segment q_{s-1} whose edges belong to, say, path p_c. Since we assumed that q_{s+1} is the first segment with edges from a path p_l with $l < j$, we conclude that $c > b$.

But we also have that Z_b appears in subpath $p_a[Z_a..Z_c]$, and the initial assumption is that $b > c$. Thus, we have a contradiction, and the lemma follows. $\square$

Theorem 1 *If the set of variables* $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ *satisfies the G criterion with respect to* Y, *then* $\mathbf{Z}$ *is an Auxiliary Set for* Y.

Proof: The system of equations $\Phi_{\mathbf{Z}}$ can be written in matrix form as:

$$\rho = A \cdot \Lambda$$

where $\rho = [(\rho_{Z_1 Y} - a_{10}) \ldots (\rho_{Z_k Y} - a_{k0})]'$, $A = [a_{ij}]$ is a k by m matrix, and $\Lambda = [\lambda_1 \ldots \lambda_m]'$.

Let A_k denote the submatrix corresponding to the first k columns of A. We will show that $Det(A_k) \neq 0$, which implies that $rank(A) = k$ and the equations in $\Phi_{\mathbf{Z}}$ are linearly independent with respect to the λ_i's.

Applying the definition of determinant, we obtain

$$Det(A_k) = \sum_\sigma (-1)^{|\sigma|} \prod_{j=1}^{k} a_{j\sigma(j)} \tag{3.1}$$

where the summation ranges over all permutations of $\langle 1, \ldots, k \rangle$, and $|\sigma|$ denotes the parity of permutation σ.

First, observe that entry a_{ii} corresponds to unblocked paths between Z_i and Y including the edge from $Inc(Y)$ with parameter λ_i. In particular, p_i is one of these paths, and we can write $a_{ii} = \left(\frac{T(p_i)}{\lambda_i} + a'_{ii} \right)$. This implies that the term $T^* = \left[\prod_i \frac{T(p_i)}{\lambda_i} \right]$ appears in the summand corresponding to permutation $\sigma = \langle 1, \ldots, k \rangle$. Also, note that every factor in T^* is the parameter of an edge in some p_i.

On the other hand, any term in the summand of a permutation distinct from σ must contain a factor from some entry a_{ji}, with $j > i$. Such an entry corresponds to unblocked paths between Z_j and Y including the edge from $Inc(Y)$ with parameter λ_i. But lemma 4 says that those paths must have at least one edge that does not appear in any of $p_1, \ldots, p_k$. This implies that T^* is not cancelled out by any other term in 3.1, and so $Det(A_k)$ does not vanish, completing the proof of the theorem. $\qquad\square$

3.4   Model Identification Using Auxiliary Sets

Assume that for each variable Y there is an Auxiliary Set $\mathcal{A}_Y$, with $|\mathcal{A}_Y| = |Inc(Y)|$.

This implies that for each Y there exists a system of linear equations $\Phi_{\mathcal{A}_Y}$ that can be solved uniquely for the parameters $\lambda_1, \ldots, \lambda_m$ of the edges in $Inc(Y)$. This fact, however, does not guarantee the identification of the λ_i's, because the solution for each λ_i is a function of the coefficients in the linear equations, which may depend on non-identified parameters.

To prove identification we need to find an appropriate order to solve the systems of equations. This order will depend on the variables that compose each auxiliary set. The following theorem gives a simple sufficient condition for identification:

Theorem 2 *Assume that the Auxiliary Set $\mathcal{A}_Y$ of each variable Y satisfies:*

(i) $|\mathcal{A}_Y| = |Inc(Y)|$;

(ii) $depth(Z_i) < depth(Y)$, *for all* $Z_i \in \mathcal{A}_Y$.

Then, the model is identified.

Proof: We prove the theorem by induction on the depth of the variables.

Let Y be a variable at depth 0.

Note that $Inc(Y)$ can only contain bidirected edges connecting Y to another variable at depth 0. Let $(W, Y) \in Inc(Y)$. Observing that the only unblocked path between W and Y consists precisely of edge (W, Y), we get that the parameter of edge (W, Y) is identified and given by ρ_{WY}.

Now, assume that, for every variable X at depth smaller than k, the parameters of the edges in $Inc(X)$ are identified.

Let Y be a variable at depth k, and let $Z_i \in \mathcal{A}_Y$. Lemma 1 implies that every intermediate variable of an unblocked path between Z_i and Y has depth smaller than k. The inductive hypothesis then implies that the coefficient of the linear equations in $\Phi_{\mathcal{A}_Y}$ are identified. Hence, the parameters of the edges in $Inc(Y)$ are identified. $\qquad\square$

In the general case, however, the auxiliary set for some variable Y may contain variables at greater depths than Y, or even descendants of Y. This would force us to solve the systems of equations in a different order than the one established by the depth of the variables.

In the following, we provide two rules that impose restrictions on the order in which the linear systems must be solved. We will see that if these restrictions do not generate a cycle, then the model is identified.

R1: If, for every bidirected edge (V_j, Y) in $Inc(Y)$, variable V_j is not an ancestor of any $Z_i \in \mathcal{A}_Y$, then $\Phi_{\mathcal{A}_Y}$ can be solved at any time.

R2: Otherwise,
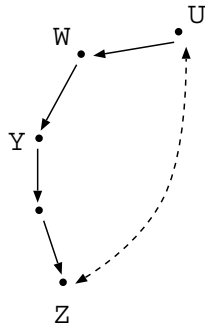
Figure 3.4: Example illustrating rule $\mathbf{R}2(b)$.

a) For every $Z \in \mathcal{A}_Y$, $\Phi_{\mathcal{A}_Z}$ must be solved before $\Phi_{\mathcal{A}_Y}$ is solved.

b) If $Z \in \mathcal{A}_Y$ is a descendant of Y, and (U, Z) is a bidirected edge with U an ancestor of Y, then for every W lying on a chain from U to Y (see Figure 3.4), $\Phi_{\mathcal{A}_W}$ must be solved before $\Phi_{\mathcal{A}_Y}$.

For a model M and a given choice of Auxiliary Sets, the restrictions above can be represented by a directed graph, called the dependence graph D_M, as follows:

- Each node in D_M corresponds to a variable in the model;

- There exists a directed edge from Z to Y in D_M if rule **R**1 does not apply to Y, and rule **R**2 imposes that $\Phi_{\mathcal{A}_Z}$ must be solved before $\Phi_{\mathcal{A}_Y}$.

The next theorem states our general sufficient condition for model identification:

Theorem 3 *Assume that there exist Auxiliary Sets for each variable in model M, such that the associated dependence graph D_M has no directed cycles. Then, model M is identified.*

The proof of the theorem is given in appendix A.

Figure 3.5 shows an example that illustrates the method just described. Apparently, this is a very simple model. However, it actually requires the full generality of Theorem 3. The Figure also shows the auxiliary sets for each variable, and the corresponding dependence graph D_M. The fact that rule $R1$ can be applied to variable Z avoids a dependence of Z on W and eliminates the possibility of a cycle.

3.5   Simpler Conditions for Identification

The sufficient condition for identification presented in the previous section is very general, but complicated to verify by visual inspection of the causal diagram. The

$$\mathcal{A}_X = \emptyset$$
$$\mathcal{A}_Y = \{X, W\}$$
$$\mathcal{A}_Z = \{Y\}$$
$$\mathcal{A}_W = \{Z, X\}$$
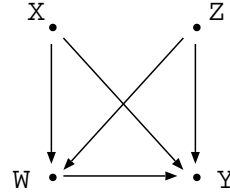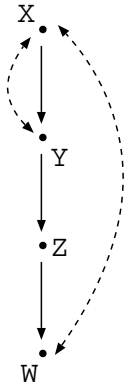
R1 applies to Z.

R2 applies to Y,W.

Figure 3.5: Example illustrating Auxiliary Sets method.

main difficulty resides in finding the required unblocked paths witnessing that a set of variables is an Auxiliary Set. A solution for this problem is provided in Chapter 4, where we develop an algorithm to find an appropriate Auxiliary Set for a fixed variable Y. However, simple conditions for identification still seem to be useful, and this is the subject of this section.

3.5.1 Bow-free Models

A bow-free model is characterized by the property that no pair of variables is connected by more than one edge. Actually, this represents the simplest situation for our method of identification.

Corollary 1 *Every bow-free model is identified.*

32
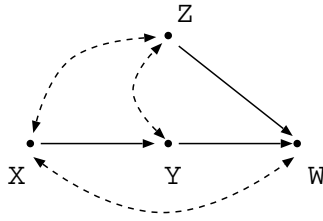
Figure 3.6: Example of a Bow-free model.

Proof: Fix an arbitrary variable Y, and let $\mathbf{X} = \{X_1, \ldots, X_k\}$ be the set of variables such that, for $i = 1, \ldots, k$, edge (X_i, Y) belongs to $Inc(Y)$. Since the model is bow-free, it follows that $|\mathbf{X}| = |Inc(Y)|$. Moreover, the set of paths $\{p_1, \ldots, p_k\}$, where each p_i is the trivial path consisting of the single edge (X_i, Y), witnesses that X is an Auxiliary Set for Y.

Now, let Δ be the ordering used in th construction of the sets of incoming edges $Inc(Y)$. Then, for every Y, all the variables in $\mathcal{A}_Y$ appear before Y in Δ. This implies that the dependence graph D_M is acyclic, and the corollary follows. $\qquad\square$

Figure 3.6 shows an interesting example. A brief examination of this causal diagram will reveal that no conditional independence holds among the variables X, Y, Z and W. As a consequence, most traditional methods for Identification (e.g., Instrumental Variables, Back-door criterion) would fail to classify this model as identified. However, as can be easily verified, this model is bow-free and hence identified.

3.5.2 Instrumental Condition

From the preceding result, it is clear that all the problems for identification arise from the existence of bow-arcs in the causal diagram (i.e., pairs of variables connected by both a directed and a bidirected edge). Let us examine this structure in more detail and try to understand why it represents a problem for identification.

Assume that there is a bow-arc between variables X and Y, and let δ, λ denote the parameters of the directed and bidirected edges in the bow, respectively. Then, Wright's decomposition for correlation ρ_{XY} gives [3]:

$$\rho_{XY} \;=\; \delta + \lambda$$

Now, if this is the only constraint on the values of δ and λ, then there exists an infinite number of solutions for δ and λ that are consistent with the observed correlation ρ_{XY}. This situation occurs, for example, when both edges in the bow point to Y, and there is no other edge in the model with an arrow head pointing to X. In this case, by analyzing the decomposition of the correlations between any pair of variables, we observe that either they do not depend on the values of δ, λ at all, or they only depend on their sum (See Figure 3.7(a) for an example). From this observation it is possible to derive a proof that any such model is non-identified. Similar techniques will be explored in Chapter 6 to obtain graphical conditions for non-identification.

Now, consider the situation where there exists a third variable Z with a directed edge pointing to X. A variable Z with such properties is sometimes called an Instrumental Variable. Figure 3.7(b) shows an example of this situation, with the respective decompositions of correlations. It is easy to verify that there exists a unique solution for parameters a, δ, λ in terms of the observed correlations.

Hence, the basic idea for our next sufficient condition for identification is to find, for each bow-arc between X_i and Y, a distinct variable Z_i with an edge pointing to X_i. The condition is precisely stated as follows.

For a fixed variable Y let $Bow(Y) = \{X_1, \ldots, X_l\}$ be the set of variables that are connected to Y by a directed and a bidirected edges, both pointing to Y.

---

[3]Here, we are assuming that there is no other unblocked path between X and Y.

34
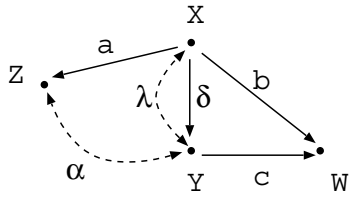
$$\begin{cases} \rho_{XZ} = a \\ \rho_{XW} = b \\ \rho_{XY} = \delta + \lambda \\ \rho_{YZ} = \alpha + a(\delta + \lambda) \\ \rho_{YW} = c + b(\delta + \lambda) \end{cases}$$
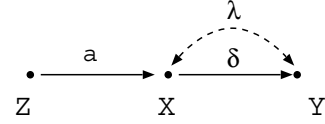
$$\begin{cases} \rho_{ZX} = a \\ \rho_{ZY} = a\delta \\ \rho_{XY} = \delta + \lambda \end{cases}$$

(a)   (b)

Figure 3.7: Examples illustrating the instrumental condition.

Instrumental Condition: We say that a variable Y satisfies the Instrumental Condition if, for each variable $X_i \in Bow(Y)$, there exists a unique variable Z_i satisfying:

(i) $depth(Z_i) < depth(Y)$;

(ii) Z_i is not connected to Y by any edge;

(iii) there exists an edge between Z_i and X_i that points to X_i.

Corollary 2 *Assume that the Instrumental Condition holds for every variable in model M. Then model M is identified.*

CHAPTER 4

Algorithm

In some elaborate models, it is not an easy task to check if a set of variables satisfies the GAV criterion. Moreover, the criterion itself does not provide any guidance to find a set of variables satisfying its conditions. In this chapter we present an algorithm that, finds an Auxiliary Set $\mathcal{A}_Y$ for a fixed variable Y, if any such set exists.

The basic idea is to reduce the problem to an instance of the maximum flow problem in a network.

Cormen *et al* [CCR90] define the maximum flow problem as follows. A flow network $G = (V, E)$ is a directed graph in which each edge $(u, v) \in E$ has a non-negative capacity $c(u, v) \geq 0$. We distinguish two vertices in the flow network: a source s and a sink t. A flow in G is a real-valued function $F : V \times V \to \Re$, satisfying:

- $F(u, v) \leq c(u, v)$, for all $u, v \in V$;

- $F(u, v) = -F(v, u)$, for all $u, v \in V$;

- $\sum_{v \in V} F(u, v) = 0$, for all $u \in V - \{s, t\}$.

That is, condition (i) states that the amount of flow on any edge cannot exceed its capacity; condition (ii) says that the amount of flow running on one direction of an edge is the same as the flow in the other direction, but with opposite sign; and condition (iii) establishes that the amount of flow entering any vertex must be the same as the amount of flow leaving the vertex.

Intuitively, the value of a flow F is the amount of flow that is transfered from the source s to the sink t, and can be formally defined as

$$|F| = \sum_{v \in V} F(s, v)$$

In the maximum flow problem, we are given a flow network G, with source s and sink t, and we wish to find a flow of maximum value from s to t.

Before describing the construction of the flow network, we make a few observations. Fix an arbitrary variable Y.

Lemma 5 *Assume that the set of variables $\mathbf{Z}$ satisfies the conditions of the GAV criterion with respect to Y. Then, there exists a set of variables $\mathbf{Z}'$ and paths $p_1, \ldots, p_{|\mathbf{Z}|}$ such that*

(i) $|\mathbf{Z}| = |\mathbf{Z}'|$

(ii) $p_1, \ldots, p_{|\mathbf{Z}|}$ witness that $\mathbf{Z}'$ satisfies the GAV criterion;

(iii) for $i = 1, \ldots, |\mathbf{Z}|$, every intermediate variable in p_i belongs to $\mathbf{Z}'$.

Proof: Let $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$, and let $q_1, \ldots, q_k$ be paths witnessing that $\mathbf{Z}$ satisfies the GAV criterion. The lemma follows from the next observation.

Let $W \notin \mathbf{Z}$ be an intermediate variable in the path q_i associated with variable $Z_i \in \mathbf{Z}$. Then the paths $q_1, \ldots, q_{i-1}, q_i[W..Y], q_{i+1}, \ldots, q_k$ witness that $\{Z_1, \ldots, Z_{i-1}, W, Z_{i+1}, \ldots, Z_k\}$ satisfies the GAV criterion with respect to Y. □

Let $\mathbf{Z}$ be a set of variables, and $p_1, \ldots, p_k$ be paths satisfying the conditions of Lemma 5. Then, it follows that each of the paths p_i must be either a bidirected edge $(Z_i \leftrightarrow Y)$, or a chain $(Z_i \rightarrow \ldots \rightarrow Y)$, or the concatenation of a bidirected edge with a chain $(Z_i \leftrightarrow Z_j \rightarrow \ldots \rightarrow Y)$.

From these observations we conclude that, in the search for an Auxiliary Set for Y, we only need to consider:

- ancestors of Y;

- variables connected by a bidirected edge to either Y or an ancestor of Y;

Now, from condition (ii) of the GAV criterion, we get that an ancestor Z_i of Y can appear in at most two paths: (1) the path p_i between Z_i and Y; and (2) some path p_j, as an intermediate variable. To allow this possibility, for each ancestor of Y, we create two vertices in the flow network.

Since non-ancestors of Y can appear in at most one path, there will be only one vertex in the flow network corresponding to each such variable.

Directed edges between ancestors of Y in the causal diagram are represented by directed edges between the corresponding vertices in the flow network. Bidirected edges incident to Y and non-ancestors of Y are also represented by directed edges. Bidirected edges between ancestors Z_i and Z_j require special treatment, because they can appear as the first edge of either p_i or p_j, but not in both of them. To enforce this restriction we make use of an extra vertex.

Next, we define the flow network G_Y that will be used to find an Auxiliary Set for Y.

The set of vertices of G_Y consists of:

- for each ancestor Z of Y, we include two vertices, denoted $V_{\overline{Z}}$ and $V_{\hat{Z}}$;

- for each non-ancestor W, we include vertex V_W;

- for each bidirected edge $Z \leftrightarrow U$ connecting ancestors of Y, we include the vertex V_{ZU};

- a source vertex s;

- a sink vertex t, corresponding to variable Y.

The set of edges of G_Y is defined as follows:

- for each ancestor Z of Y, we include the edge $V_{\overline{Z}} \to V_{\hat{Z}}$;

- for each directed edge $Z \to U$ in the causal diagram, connecting ancestors of Y, we include the edge $V_{\overline{Z}} \to V_{\overline{U}}$;

- for each directed edge $Z \to Y$, we include the edge $V_{\overline{Z}} \to t$;

- for each bidirected edge $Z \leftrightarrow$, where Z is an ancestor of Y, we include the edge $V_{\hat{Z}} \to t$;

- for each bidirected edge $W \leftrightarrow Z$, where W is a non-ancestor and Z is an ancestor of Y, we include the edge $V_W \to V_{\overline{Z}}$;

- for each bidirected edge $W \leftrightarrow Y$, where W is a non-ancestor, we include the edge $V_W \to t$;

- for each bidirected edge $Z \leftrightarrow U$, where both Z and U are ancestors of Y, we include the edges: $V_{\hat{Z}} \to V_{ZU}, V_{\hat{U}} \to V_{ZU}, V_{ZU} \to V_{\overline{Z}}, V_{ZU} \to V_{\overline{U}}$;

- for each ancestor Z of Y, we include the edge $s \to V_{\hat{Z}}$;

- for each non-ancestor W, we include the edge $s \to V_W$.

To solve the maximum flow problem on the flow network G_Y defined above, we assign capacity 1 to every edge in G_Y. We also impose the additional constraint of maximum incoming flow capacity of 1 to every vertex in G_Y (this can be implemented

39

by splitting each vertex into two and connecting them by an edge of capacity 1), except for vertices s and t.

We solve the Max-flow problem using the Ford-Fulkerson algorithm. From the integrality theorem ([CCR90], p.603), the computed flow F allocates a non-negative integer amount of flow to each edge in G_Y. Since we assign capacity 1 to every edge, this solution corresponds to disjoint directed paths from s to t. The Auxiliary Set returned by the algorithm is simply the set of variables corresponding to the first vertex in each path.

Theorem 4 *The algorithm described above is sound and complete. That is, the set of variables returned by the algorithm is an Auxiliary Set for Y with maximum size.*

Proof: Fix a variable Y in model M, and let G_Y be the corresponding flow network. Let F be the flow computed by the Ford-Fulkerson algorithm on G_Y, and let $k = |F|$.

As described above, we interpret the flow solution F as a set of edge disjoint paths $P = \{p_1, \ldots, p_k\}$ from the source s to the sink t.

First, we show that each such path corresponds to an unblocked path in the causal diagram of M. Let $p_i = s \rightarrow v_i \rightarrow u_1 \rightarrow \ldots \rightarrow u_l \rightarrow t$ be one of the paths in P. We make the following observations:

- v_i is either a vertex V_W corresponding to a non-ancestor of Y, or a vertex $V_{\hat{Z}}$ corresponding to an ancestor Z of Y in the causal diagram.

- for $j = 1, \ldots, l$, each u_j is either a vertex $V_{\overline{Z}}$ corresponding to ancestor U of Y, or a vertex V_{ZU} corresponding to a bidirected edge between ancestors Z and U of Y, but the later can only occur if $j = 1$ and $V_{\hat{Z}}$.

40

Now, we establish a correspondence between the edges of p_i and the edges of the causal diagram.

1. The directed edge $u_l \to t$ corresponds to the directed edge from the respective ancestor U_l to variable Y.

2. If two vertices u_j and u_{j+1} are both vertices of the type $V_{\overline{U_j}}$ and $V_{\overline{U_{j+1}}}$, then the edge $u_j \to u_{j+1}$ in p_i corresponds to the edge $U_j \to U_{j+1}$ in the causal diagram.

3. If v_i is a vertex of type V_W, then the edge $v_i \to u_1$ corresponds to the edge $W \leftrightarrow U_1$, and path p_i corresponds to the path $W \leftrightarrow U_1 \to \ldots \to U_l \to Y$ in the causal diagram.

4. If v_i is a vertex of type $V_{\hat{V}}$ and u_1 is a vertex of type $V_{\overline{U_1}}$, then the edge $v_i \to u_1$ corresponds to the edge $Z \to U_1$, and path p_i corresponds to the path $Z \to U_1 \to \ldots \to U_l \to Y$ in the causal diagram.

5. If v_i is a vertex of type $V_{\hat{V}}$ and u_1 is of type V_{ZU}, then the edge $v_i \to u_1$ corresponds to the edge $Z \leftrightarrow U_2$, and path p_i corresponds to the path $Z \leftrightarrow U_2 \to \ldots \to U_l \to Y$ in the causal diagram.

Now, let $v_1, \ldots, v_k$ be the first vertices in each of the paths in P, and let $Z_1, \ldots, Z_k$ be the variables associated with those vertices in the causal diagram.

Note that the constraint of maximum incoming flow capacity of 1 on the vertices of G_Y implies that the paths $p_1, \ldots, p_k$ are vertex disjoint (and also implies that the variables $Z_1, \ldots, Z_k$ are all distinct). However, this constraint does not imply that the corresponding paths in the causal diagram are vertex disjoint.

For an ancestor Z of Y, it is possible that vertices $V_{\overline{Z}}$ and $V_{\hat{Z}}$ appear in distinct paths p_i and p_j, and so Z would appear in two of the corresponding paths on the causal
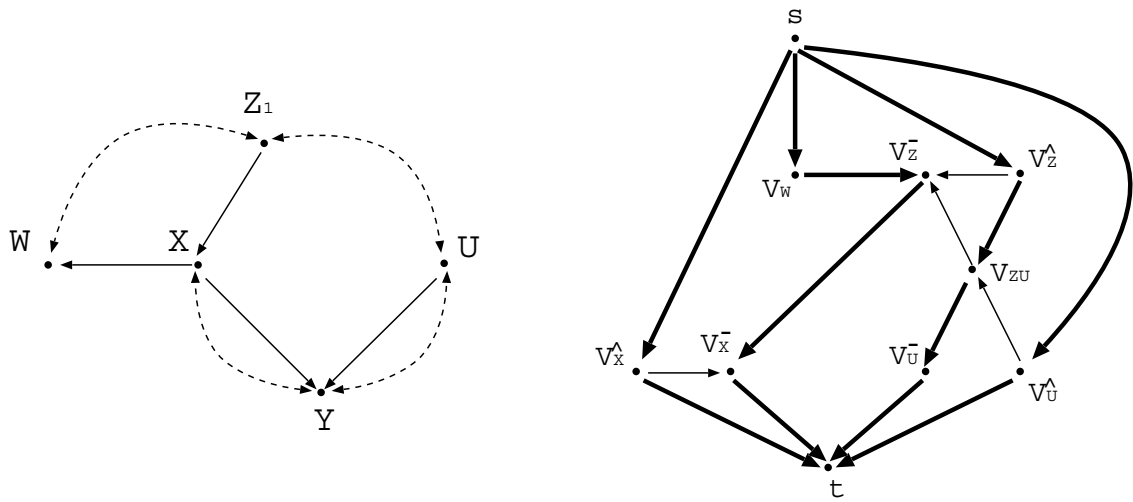
Figure 4.1: A causal diagram and the corresponding flow network

diagram. In fact, this is the only possibility for a variable to appear in more than one such paths, and in this case it is easy to verify that the conditions of the G criterion hold. This proves that the algorithm is sound.

Completeness easily follows from the construction of the flow network, and the optimality of Ford-Fulkerson algorithm □

Figure 4.1 shows an example of a simple causal diagram and the corresponding flow network G_Y.

Theorem 5 *The time complexity of the algorithm described above is $O(n^3)$.*

Proof: The theorem easily follows from the facts that the number of vertices in the flow network is proportional to the variables in the model, and that Ford-Fulkerson algorithm runs in $O(m^3)$, where m is the size of the flow network. □

CHAPTER 5

Correlation Constraints

It is a well-known fact that the set of structural assumptions defining a SEM model may impose constraints on the covariance matrix of the observed variables [McD97]. That is, a given set of structural assumptions may imply that the value of a particular entry in Σ_M is a function of some other entries in this matrix.

An immediate consequence of this observation is that a model M may not be compatible with an observed covariance matrix $\hat{\Sigma}$, in the sense that for every parametrization π we have $\Sigma_M(\pi) \neq \hat{\Sigma}$. This allows to test the quality of the model. That is, by verifying if the constraints imposed by the structural assumptions are satisfied in the observed covariance matrix (at least approximately), we either increase our confidence in the model, or decide to modify (or discard) it.

The first type of constraint imposed by a SEM model is associated with d-separation conditions. Such condition is equivalent to a conditional independence statement [Pea00a], and implies that a corresponding correlation coefficient (or partial correlation) must be zero. Figure 5.1 shows two examples illustrating how we can obtain correlation constraints from d-separation conditions. In the causal diagram (a), it is easy to see that variables X and Y are d-separated (the only path between them is blocked by variable Z). This immediately gives that $\rho_{XY} = 0$. For the model in (b), it follows that variables X and W are d-separated when conditioning on Z. This implies that $\rho_{XW.Z} = 0$. Applying the recursive formula for partial correlations, we get:
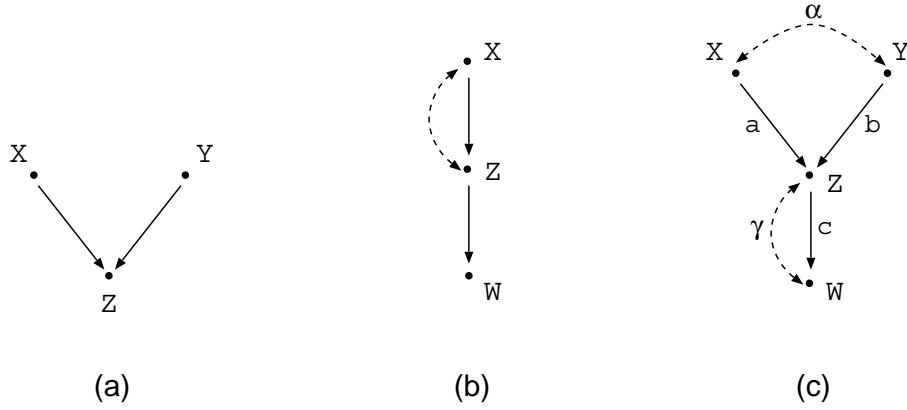
Figure 5.1: D-separation conditions and correlation constraints.

$$\rho_{XW.Z} \quad = \quad \frac{\rho_{XW} - \rho_{XZ} \cdot \rho_{ZW}}{\sqrt{1 - (\rho_{XZ})^2} \cdot \sqrt{1 - (\rho_{WZ})^2}} \quad = \quad 0$$

and so, $\rho_{XW} = \rho_{XZ} \cdot \rho_{ZW}$.

The d-separation conditions, however, do not capture all the constraints imposed by the structural assumptions. In the model of Figure 5.1(c) no d-separation condition holds, but the following algebraic analysis shows that there is a constraint involving the correlations $\rho_{XZ}, \rho_{YZ}, \rho_{XW}$ and ρ_{YW}.

Applying Wright's decomposition to these correlation coefficients, we obtain the following equations:

$$\begin{cases} \rho_{XZ} & = & a + \alpha b \\ \rho_{YZ} & = & \alpha a + b \\ \rho_{XW} & = & ac + \alpha bc \\ \rho_{YW} & = & \alpha ac + bc \end{cases}$$

Observe that factoring c out in the right-hand side of the equation for ρ_{XW}, we obtain the expression for ρ_{XZ}. Thus, we can write

$$\rho_{XW} = c \cdot \rho_{XZ}$$

Similarly, we obtain

$$\rho_{YW} = c \cdot \rho_{YZ}$$

Now, it is easy to see that the following constraint is implied by the two equations above

$$\rho_{XW} = \frac{\rho_{YW} \cdot \rho_{XZ}}{\rho_{YZ}}$$

Clearly, this type of analysis is not appropriate for complex models. In the following we show how to use the concept of Auxiliary Sets to compute constraints in a systematic way.

5.1 Obtaining Constraints Using Auxiliary Sets

The idea is simple, but gives a general method for computing correlation constraints. Fix a variable Y, and let $\lambda_1, \ldots, \lambda_m$ denote the parameters of the edges in $Inc(Y)$. Recall that if $\mathcal{A}_Y = \{Z_1, \ldots, Z_k\}$ is an Auxiliary Set for Y, then the decomposition of the correlations $\rho_{Z_1 Y}, \ldots, \rho_{Z_k Y}$ gives a linearly independent system of equations, with respect to the λ_i's. Let us denote this system by $\Phi_{\mathcal{A}_Y}$.

Now, if $|\mathcal{A}_Y| = |Inc(Y)|$, then this system of equations is maximal, in the sense that any linear equation on parameters $\lambda_1, \ldots, \lambda_m$ can be expressed as a linear combination of the equations in $\Phi_{\mathcal{A}_Y}$. Hence, computing this linear combination for the decomposition of ρ_{WY}, where $W \notin \mathcal{A}_Y$, gives a constraint involving the correlations among $\rho_{WY}, \rho_{Z_1 Y}, \ldots, \rho_{Z_k Y}$.

45

In the following, we describe the method in more detail.

Let $\mathcal{A}_Y = \{Z_1, \ldots, Z_m\}$ be an Auxiliary Set for Y, and assume that $|\mathcal{A}_Y| = |Inc(Y)|$. The decomposition of the correlations $\rho_{Z_1Y}, \ldots, \rho_{Z_kY}$ can be written as:

$$
\begin{cases}
\rho_{Z_1 Y} & = & \sum_{i=1}^{m} a_{1i} \cdot \lambda_i \\
\ldots \\
\rho_{Z_m Y} & = & \sum_{i=1}^{m} a_{mi} \cdot \lambda_i
\end{cases}
\tag{5.1}
$$

Or, in matrix form,

$$\rho = A \cdot \Lambda$$

Since $\mathcal{A}_Y$ is an Auxiliary Set and $|\mathcal{A}_Y| = |Inc(Y)|$, it follows that A is an m by m matrix with rank m.

Now, let $W \notin \mathcal{A}_Y \cup \{Y\}$, and write the decomposition of ρ_{WY} as

$$\rho_{WY} = \sum_{i=1}^{m} b_i \cdot \lambda_i \tag{5.2}$$

Clearly, the vector $B = [b_1 \ldots b_m]$ can be expressed as a linear combination of the rows of matrix A as

$$B = \sum_{i=1}^{m} d_i \cdot A_i$$

where A_i denotes the ith row of matrix A, and the d_i's are the coefficients of the linear combination.

But this implies that we can express the correlation ρ_{WY} as

$$\rho_{WY} = \sum_{i=1}^{m} d_i \cdot \rho_{Z_iY}$$

by considering the left-hand side of Equations 5.1 and 5.2.
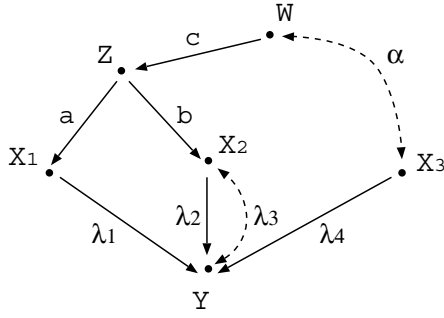
Figure 5.2: Example of a model that imposes correlation constraints

We also note that the coefficients d_i are functions of the parameters of the model. But if the model is identified, then each of these parameters can be expressed as a function of the correlations among the observed variables. Hence, we obtain an expression only in terms of those correlations.

To illustrate the method, let us consider the model in Figure 5.2. It is not difficult to check that $\mathcal{A}_Y = \{Z, X_1, X_2, X_3\}$ is an Auxiliary Set for variable Y. Next, we obtain a constraint involving the correlations $\rho_{WY}, \rho_{WZ}, \rho_{ZY}$ and ρ_{WX_3}.

The decompositions of the correlations between Y and each variable in $\mathcal{A}_Y$ gives the following system of equations:

$$
\Phi_{\mathcal{A}_Y} = \begin{cases}
\rho_{ZY} & = & a\lambda_1 + b\lambda_2 + c\alpha\lambda_4 \\
\rho_{X_1 Y} & = & \lambda_1 + ab\lambda_2 + ac\alpha\lambda_4 \\
\rho_{X_2 Y} & = & ab\lambda_1 + \lambda_2\lambda_3 + bc\alpha\lambda_4 \\
\rho_{X_3 Y} & = & ac\alpha\lambda_1 + bc\alpha\lambda_2 + \lambda_4
\end{cases}
$$

and the decomposition of ρ_{WY} gives the equation

$$
\rho_{WY} = ca\lambda_1 + cb\lambda_2 + \alpha\lambda_4
$$

The corresponding matrix A and vector B are then given by:

47

$$A = \begin{bmatrix} a & b & 0 & c\alpha \\ 1 & ab & 0 & ac\alpha \\ ab & 1 & 1 & bc\alpha \\ ac\alpha & bc\alpha & 0 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} ca & cb & 0 & \alpha \end{bmatrix}$$

After some calculations, one can verify that vector B can be expressed as a linear combination of the first and fourth rows of A as:

$$B = \left(\frac{c - \alpha^2 c}{1 - c^2\alpha^2} \right) A_1 + \left(\frac{\alpha - c^2\alpha}{1 - c^2\alpha^2} \right) A_4 \tag{5.3}$$

Since the model is identified, we can express the parameters α and c in terms of the correlation coefficients. In this case, we obtain $\alpha = \rho_{WX_3}$ and $c = \rho_{WZ}$. Substituting these expressions in Equation 5.3 and performing some algebraic manipulations, we obtain the following constraint:

$$\rho_{WY}[1 - \rho_{WX_3}^2 \cdot \rho_{WZ}^2] = \rho_{WZ} \cdot \rho_{ZY}[1 - \rho_{WX_3}^2] + \rho_{WX_3} \cdot \rho_{X_3Y}[1 - \rho_{WZ}^2]$$

CHAPTER 6

Sufficient Conditions For Non-Identification

The ultimate goal of this research is to solve the problem of identification in SEM. That is, to obtain a necessary and sufficient condition for identification, based only on the structural assumptions of the model.

In Chapter 3, we introduced our graphical approach for identification, and provided a very general sufficient condition for model identification. Indeed, we are not aware of any example of an identified model that cannot be proven to be so using the method of auxiliary sets. Here, we present the results of our initial efforts on the other side of the problem. That is, we investigate necessary graphical conditions for the identification of a SEM model.

The method of auxiliary sets imposes two main conditions to classify a model M as identified. First, for each variable Y in M we must find a sufficiently large auxiliary set. Second, given the choice of auxiliary sets, a precedence relation is established among the variables, and represented by a dependence graph D_M. This graph cannot contain any directed cycle.

A natural strategy, then, is to assume that one of these conditions does not hold, and try to prove that the model is non-identified. The proof of non-identification is conceptually simple. We begin with an arbitrary parametrization π for model M. Then, we show that, under the specified conditions, it is possible to construct another parametrization $\pi' \neq \pi$ such that $\Sigma_M(\pi) = \Sigma_M(\pi')$. This proves that the model is
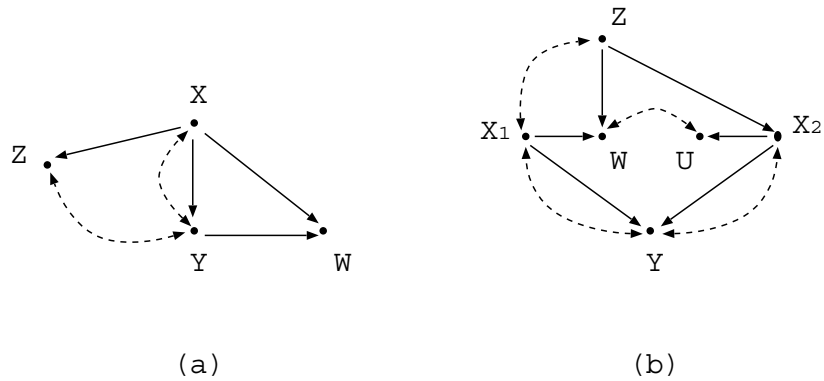
Figure 6.1: Figure

non-identified.

6.1 Violating the First Condition

In this section we analyze two sets of conditions that prevent the existence of a sufficiently large Auxiliary Set for a given variable Y. In both cases we can show that they imply the non-identification of the underlying model. These conditions, however, do not provide a complete characterization of the situation where the first condition of the Auxiliary Sets method does not hold. At the end of this section, we give an example that illustrates this fact. At this point, such a characterization is the major difficulty to obtaining a more general condition for non-identification.

The simplest example of non-identification, briefly discussed in Section 3.5.2, consists of a model in which a pair of variables X, Y is connected by a directed and a bidirected edge, both pointing to Y, and there is no other edge pointing to X in the causal diagram (see Figure 6.1(a)).

The next theorem provides a simple generalization of this situation. We assume that, for a fixed variable Y, there exists an Auxiliary Set $\mathcal{A}_Y$, with not enough variables.

We also assume that every edge connecting a variable $Z \in \mathcal{A}_Y$ and a variable outside $\mathcal{A}_Y \cup \{Y\}$ does not point to Z (see Figure 6.1(b)). Under these conditions, it is possible to show that there exists no Auxiliary Set for Y which is larger than $\mathcal{A}_Y$. The theorem proves that any such model is non-identified.

Theorem 6 *Let Y be an arbitrary variable in model M, and let $\mathcal{A}_Y = \{Z_1, \ldots, Z_k\}$ be an Auxiliary Set for Y. Assume that*

(i) $|\mathcal{A}_Y| < |Inc(Y)|$;

(ii) $\mathcal{A}_Y$ *only contains non-descendants of Y;*

(iii) *There exists no edge between some $Z_i \in \mathcal{A}_Y$ and a variable $W \notin \mathcal{A}_Y \cup \{Y\}$ pointing to Z_i.*

Then, model M is non-identified.

Proof: Let $\lambda_1, \ldots, \lambda_m$ denote the parameters of the edges in $Inc(Y)$. First, we establish that, for every pair of variables U, V the correlation ρ_{UV} only depends on the values of the λ_i's through the correlations $\rho_{Z_1Y}, \ldots, \rho_{Z_kY}$. That is, any modification of the values of the λ_i's that does not change the values of $\rho_{Z_1Y}, \ldots, \rho_{Z_kY}$, also does not change the value of ρ_{UV}. This is formally stated as follows:

Lemma 6 *Let U and V be arbitrary variables in model M. Then, the correlation ρ_{UV} can be expressed as*

$$\rho_{UV} = c_0 + \sum_{i=1}^{k} c_i \cdot \rho_{Z_iY}$$

where the independent term c_0 and the coefficients c_i's do not depend on the parameters $\lambda_1, \ldots, \lambda_m$.

The proof of the lemma is given in Appendix B.

Now, fix an arbitrary parametrization π for model M. This parametrization induces a unique correlation coefficient for each pair of variables U, V, that we denote by $\rho_{UV}(\pi)$.

According to Wright's decomposition, the correlations $\rho_{Z_1Y}(\pi), \ldots, \rho_{Z_kY}(\pi)$ depend on the values of parameters λ_i's through the linear equations:

$$\begin{cases} \rho_{Z_1Y}(\pi) & = & \sum_{i=1}^{m} a_{1i}(\pi) \cdot \lambda_i \\ \ldots \\ \rho_{Z_kY}(\pi) & = & \sum_{i=1}^{m} a_{ki}(\pi) \cdot \lambda_i \end{cases}$$

Since $|\mathcal{A}_Y| < |Inc(Y)|$, this system of equations has more variables than equations, and so there exists an infinite number of solutions for the λ_i's. The values assigned to the λ_i's by parametrization π corresponds to one of these solutions. Any other solution gives a parametrization π' with $\Sigma_M(\pi) = \Sigma_M(\pi')$. $\square$

The next theorem considers a much weaker set of assumptions that still prevent the existence of an Auxiliary Set for Y with enough variables. The theorem shows that any such model is non-identified.

Let Y be an arbitrary variable in model M, and let $\mathcal{A}_Y = \{Z_1, \ldots, Z_k\}$ be an Auxiliary Set for Y. Define the boundary of variables of $\mathcal{A}_Y$, denoted $\overline{\mathcal{A}_Y}$, as the subset of variables Z_i in $\mathcal{A}_Y$ for which there exists a variable $W \notin \mathcal{A}_Y \cup \{Y\}$ and an edge (W, Z_i) pointing to Z_i.

Theorem 7 *Assume that for a given variable Y in model M, the following conditions hold:*

(i) $|\mathcal{A}_Y| < |Inc(Y)|$;

(ii) For any $W \notin \mathcal{A}_Y$, there exists no edge (W, Y) pointing to Y;
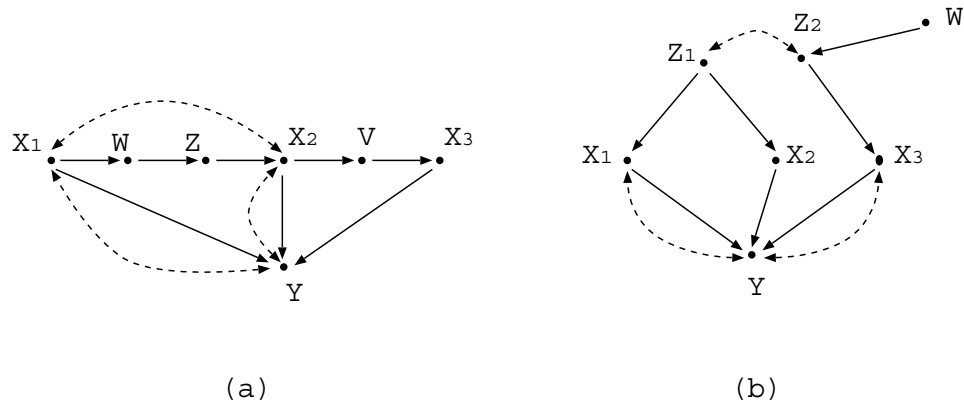
Figure 6.2: Figure

(iii) For any $Z_i \in \overline{\mathcal{A}_Y}$ and $Z_j \in (\mathcal{A}_Y - \overline{\mathcal{A}_Y}) \cup \{Y\}$, there is no edge (Z_i, Z_j) pointing to Z_i.

Then, model M is non-identified.

The proof of this theorem in given in Appendix B.

Figure 6.2(a) shows a causal diagram satisfying the conditions in Theorem 7. In this model, the Auxiliary Set for Y is given by $\mathcal{A}_Y = \{X_1, X_2, X_3, Z\}$, with boudary $\overline{\mathcal{A}_Y} = \{X_3, Z\}$. Note that $X_3 \rightarrow Y$ and $Z \rightarrow X_2$ are the only edges between a variable in $\overline{\mathcal{A}_Y}$ and a variable in $(\mathcal{A}_Y - \overline{\mathcal{A}_Y}) \cup \{Y\}$, and they do not point to Z and X_3.

Figure 6.2(b) shows a causal diagram where the conditions of Theorem 7 do not hold. In this case, the Auxiliary Set for Y is given by $\mathcal{A}_Y = \{X_1, X_2, X_3, Z_1, Z_2\}$, with boudary $\overline{\mathcal{A}_Y} = \{Z_2\}$. Note that edge $Z_1 \leftrightarrow Z_2$ violates condition (iii). However, it is still possible to show that there is no Auxiliary Set for Y with more than 5 variables.

6.2 Violating the Second Condition

In the previous section we studied the problem of identification under two sets of assumptions. In both cases we could show that, for every parametrization π for model M, there exists an infinite number of distinct parametrizations for M that generate the same covariance matrix $\Sigma_M(\pi)$. We conjecture that this will be the case whenever the first condition of the method of Auxiliary Sets fails. Equivalently, we believe that the conditions that characterize an Auxiliary Set are also necessary for linear independence among the equations given by Wright's decompositions.

Surprisingly, the situation seems to be very different when the second condition fails. Next, we discuss an example in which every variable has a large enough auxiliary set, but the corresponding dependence graph D_M contains a cycle. A simple algebraic analysis shows that, for almost every parametrization π, there exists exactly one other distinct parametrization π' such that $\Sigma_M(\pi) = \Sigma_M(\pi')$.

At this point, we still cannot provide conditions for the existence of only a finite number of parametrizations generating the same covariance matrix. So, we leave the problem open. However, this seems to be an important question. According to the definition in section 2.1, any such model would be considered non-identified. On the other hand, if there exists only a small number of parametrizations compatible with the observed correlations, the investigator can decide to proceed with the SEM analysis, and decide which parametrization is more appropriate for the case in hand, based on domain knowledge.

Consider the model illustrated by Figure 6.3. Applying Wright's decomposition to the correlations between each pair of variables, we obtain the following equations:
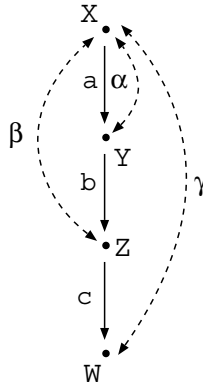
Figure 6.3: Figure

$$
\begin{cases}
\rho_{XY} &= a + \alpha \\
\rho_{XZ} &= (a + \alpha)b + \beta \\
\rho_{XW} &= (a + \alpha)bc + \beta c\gamma \\
\rho_{YZ} &= b + a\beta \\
\rho_{YW} &= (b + a\beta)c + a\gamma \\
\rho_{ZW} &= c + ab\gamma
\end{cases}
$$

Using the equations for ρ_{XZ} and ρ_{YZ}, we can express parameters b and β in terms of a:

$$
b = \frac{\rho_{YZ} - a \cdot \rho_{XZ}}{1 - a \cdot \rho_{XY}} \qquad \beta = \frac{\rho_{XZ} - \rho_{XY} \cdot \rho_{YZ}}{1 - a \cdot \rho_{XY}}
$$

Similarly, we use the equations for ρ_{XW} and ρ_{YW}, to obtain expressions for parameters c and γ in terms of a:

$$
c = \frac{\rho_{YW} - a \cdot \rho_{XW}}{\rho_{YZ} - a \cdot \rho_{XZ}} \qquad \gamma = \frac{\rho_{YZ} \cdot \rho_{XW} - \rho_{XZ} \cdot \rho_{YW}}{\rho_{YZ} - a \cdot \rho_{XZ}}
$$

Now, substituting the expressions for b, c and γ in the equation for ρ_{ZW}, we get

$$\rho_{ZW} = \frac{\rho_{YW} - a\rho_{XW}}{\rho_{YZ} - a\rho_{XZ}} + a\left[\frac{\rho_{YZ} - a\rho_{XZ}}{1 - a\rho_{XY}}\right]\left[\frac{\rho_{YZ}\rho_{XW} - \rho_{XZ}\rho_{YW}}{\rho_{YZ} - a\rho_{XZ}}\right]$$

After some algebraic manipulation, we obtain the following quadratic equation in terms of a:

$$a^2\left[\rho_{ZW}\rho_{XZ}\rho_{XY} - \rho_{XY}\rho_{XW} + \rho_{XZ}\rho_{YZ}\rho_{XW} - \rho_{XZ}^2\rho_{YX}\right]$$
$$+a\left[\rho_{XY}\rho_{YW} + \rho_{XW} + \rho_{YZ}\rho_{XZ}\rho_{YW} - \rho_{ZW}\rho_{YZ}\rho_{XY} - \rho_{ZW}\rho_{XZ} - \rho_{YZ}^2\rho_{XW}\right]$$
$$+\rho_{ZW}\rho_{YZ} - \rho_{YW} = 0$$

where it is possible to show that the coefficient of the quadratic term does not vanish.

This implies that there exist exactly two distinct parametrizations for this model that generate the same covariance matrix. For example, it is not hard to verify that the following two parametrizations generate the same covariance matrix:

$$\begin{cases} a = 2 \\ b = 3 \\ c = 1 \end{cases} \qquad \begin{cases} \alpha = 5 \\ \beta = 2 \\ \gamma = 3 \end{cases}$$

$$\begin{cases} a = 0.3225 \\ b = 0.3333 \\ c = -11 \end{cases} \qquad \begin{cases} \alpha = 6.6774 \\ \beta = 20.666 \\ \gamma = 279 \end{cases}$$

CHAPTER 7

Instrumental Sets

So far we have concentrated our efforts on questions related to the entire model, such as: "Is the model identified?", or "Does the model impose any constraint on the correlations among observed variables?".

In this chapter we focus our attention on the identification of specific subsets of parameters of the model. This goal is based on the observation that, even when we cannot prove the identification of the model (or, when the model is actually non-identified), we may still be able to show that the value of some parameters is uniquely determined by the structural assumptions in the model and the observed data.

This type of result may be valuable in situations where, even though a model is specified for all the observed variables, the main object of study consists of the relations among a small subset of those variables.

We also restrict ourselves to the identification of parameters associated with directed edges in the causal diagram. Those parameters represent the strength of (direct) causal relationships, and are usually more important that the spurious correlations (associated with bidirected edges).

The main result of the chapter is a sufficient condition for the identification of the parameters of a subset of directed incoming edges to a variable Y (i.e., directed edges with arrow head pointing to Y). An important characteristic of this condition is that it does not depend on the identification of any other parameter in the model.

This result actually generalizes the graphical version of the method of Instrumental Variables [Pea00a]. According to this method, the parameter of edge $X \rightarrow Y$ is identified if we can find a variable Z and a set of variables $\mathbf{W}$ satisfying specific d-separation conditions. A more detailed explanation of this method is given in the next section.

Our contribution is to extend this method to allow the use of multiple pairs $(Z_1, \mathbf{W}_1), \ldots, (Z_k, \mathbf{W}_k)$ to prove, simultaneously, the identification of the parameters of directed edges $X_1 \rightarrow Y, \ldots, X_k \rightarrow Y$. As we will see later, there exist examples in which our method of Instrumental Sets proves the identification of a subset of parameters, but the application of the original method to the individual parameters would fail.

Before proceeding, we would like to call attention to a few aspects of the proof technique developed in this chapter, which seem to be of independent interest. The method imposes two main conditions for a set of variables to be an Instrumental Set. First, we require the existence of a number of unblocked paths with specific properties. This condition is very similar to the one in the GAV Criterion, and basically ensures the linear independence of a system of equations. The second consists of d-separation conditions between Y and each Z_i, given the variables in $\mathbf{W}_i$.

Now, to be able to take advantage of these d-separation conditions, we have to work with partial correlations instead of the standard correlation coefficients used in Chapter 3. The problem, however, is that no technique similar to Wright's decomposition was available to express partial correlation as linear expressions on the parameters of interest.

This difficulty is overcome by developing a new decomposition of the partial correlation $\rho_{XY.Z_1...Z_k}$ into a linear expression in terms of the correlations among the variables $X, Y, Z_1, \ldots, Z_k$. This leads to a linear equation on parameters of the model, by applying Wright's decomposition to each of the correlation coefficients. The d-

separation conditions then imply that only the parameters under study appear in these equations. The result finally follows by showing that all the coefficients in the linear equations can be estimated from data.

7.1 Causal Influence and the Components of Correlation

This section introduces some concepts that will be used in the description of Instrumental Variables methods.

Intuitively, we say that variable X has causal influence on variable Y, if changes in the value of X lead to changes in the value of Y. This notion is captured in the graphical language as follows.

An observed variable X has causal influence on variable Y if there exists at least one path in the causal diagram, consisting only of directed edges, going from X to Y. This path reflects the existence of variables $Z_1, \ldots, Z_k$ such that

- X appears in the r.h.s. of the equation for Z_1;

- for $i = 1, \ldots, k - 1$, each Z_i appears on the r.h.s. of the equation for Z_{i+1};

- Z_k appears in the r.h.s. of the equation for Y.

Now, it is easy to see that any change on the value of X would propagate through the equations for the Z_i's and affect the value of Y.

Similarly, an error term e_j has causal influence on variable Y if either there exists a directed path from X_j to Y, or a path consisting of a bidirected edge (X_j, U) concatenated with a directed path from U to Y (the error term e_j can be viewed as sitting on the top of bidirected edge (X_j, U)). Note that observed variables do not have causal influence on error terms, and the causal relationships among error terms are not specified by the model.

We also say that the causal influence of X (or some error term e_j) on Y is mediated by a variable W, if W appears in every directed path from X to Y (and the unblocked paths starting with bidirected edges, in the case of error terms).

The correlation between two observed variables X, Y can then be explained as the sum of two components:

- the causal influence of X on Y; and,

- the correlation created by other variables (or error terms) which have causal influence on both X and Y (sometimes called spurious correlation).

The problem of identification consists in assessing the strength of the (direct) causal influences of one variable on another, given the structure of the model and the correlations between observed variables. This requires the ability to separate the portions of the correlation contributed by each of the components above.

7.2 Instrumental Variable Methods

The traditional definition qualifies variable Z as instrumental, relative to a cause X and effect Y if [Pea00a]:

1. Every error term with causal influence on Y not mediated by X is independent of Z (i.e., uncorrelated);

2. Z is not independent of X.

Property (1) implies that the correlation between Z and Y is created by the correlation between Z and X, and the causal influence of X on Y, represented by parameter c. This implies that $\rho_{ZY} = c \cdot \rho_{ZX}$. Property (2), allows us to obtain the value of c by writing $c = \rho_{ZY} \, \rho_{ZX}$. Hence, parameter c is identified.
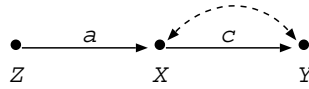
Figure 7.1: Typical Instrumental Variable

Figure 7.1 shows a typical example of the use of an instrumental variable. In this model, it is easy to verify that variable Z has properties (1) and (2). Note that this causal diagram could be just a portion of a larger model. But, as long as properties (1) and (2) hold, variable Z can be used to obtain the identification of parameter c.

A generalization of the method of Instrumental Variables (IV) is offered through the use of conditional IV's. A conditional IV is a variable Z that does not have properties (1) and (2) above, but after conditioning on a set of variables $\mathbf{W}$ such properties hold (with independence statements replaced by conditional independence). When such a pair $(Z, \mathbf{W})$ is found, the causal influence of X on Y is identified and given by $\rho_{ZY.\mathbf{W}}/\rho_{ZX.\mathbf{W}}$.

Now, given the graphical interpretation of causal influence provided in Section 7.1, we can express the properties of a conditional IV in graphical terms using d-separation [Pea00b]:

Let M be a model, and let G_M denote its causal diagram. Variable Z is a conditional IV relative to edge $X \rightarrow Y$ if there exists a set of variables $\mathbf{W}$ (possibly empty) satisfying:

1. $\mathbf{W}$ contains only non-descendents of Y;

2. $\mathbf{W}$ d-separates Z from Y in the subgraph $\overline{G}$ obtained from G_M by removing edge $X \rightarrow Y$;

3. $\mathbf{W}$ does not d-separate Z from X in $\overline{G}$.

61
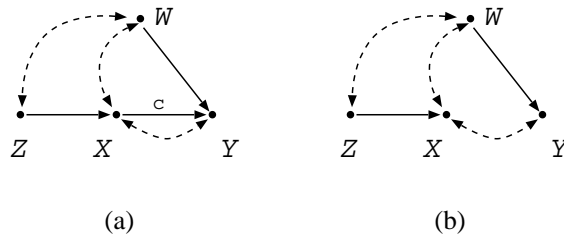
Figure 7.2: Conditional IV Examples

As an example, let us verify if Z is a conditional IV for the edge $X \to Y$ in the model of Figure 7.2(a). The first step consists in removing the edge $X \to Y$ to obtain the subgraph $\overline{G}$, shown in Figure 7.2(b). Now, it is easy to see that, after conditioning on variable W, Z becomes d-separated from Y but not from X in $\overline{G}$.

7.3   Instrumental Sets

Before introducing our method of Instrumental Sets, let us analyze an example where the procedure above fails.

Consider the model in Figure 7.3(a). A visual inspection of this causal diagram shows that variable Z_1 does not qualify as a conditional IV for the edge $X_1 \to Y$. The problem here is that Z_1 cannot be d-separated from Y, no matter how we choose the set $\mathbf{W}$. If $\mathbf{W}$ does not include variable X_2, then the path $Z_1 \to X_2 \to Y$ remains open. However, including variable X_2 in $\mathbf{W}$ would open the path $Z_1 \to X_2 \leftrightarrow Y$.

By symmetry, we also conclude that Z_1 does not qualify as a conditional IV for the edge $X_2 \to Y$, and the situation is exactly the same for Z_2. Hence, the identification of parameters c_1 and c_2 cannot be proved using the graphical criterion for the method of conditional IV.

However, observe the subgraph in Figure 7.3(b), obtained by deleting edges $X_1 \to$
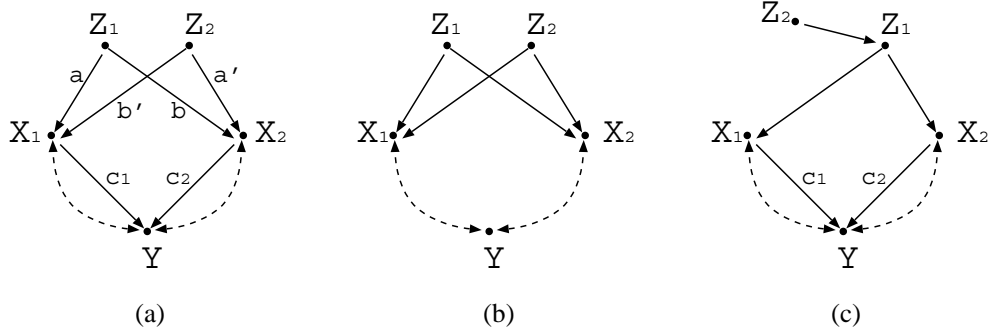
Figure 7.3: Simultaneous use of two IVs

Y and $X_2 \to Y$. It is easy to verify that properties $(1) - (3)$ hold for Z_1 and Z_2 in this subgraph (by taking $\mathbf{W} = \phi$ for both Z_1 and Z_2).

Thus, the idea is to let Z_1 and Z_2 form an Instrumental Set, and try to prove the identification of parameters c_1 and c_2 simultaneously.

Note that the d-separation conditions are not sufficient to guarantee the identification of the parameters. In the model of Figure 7.3(c), variables Z_1 and Z_2 also become d-separated from Y after removing edges $X_1 \to Y$ and $X_2 \to Y$. However, parameters c_1 and c_2 are non-identified in this case. Similarly to the GAV criterion, we have to require the existence of specific paths between each of the Z_i's and Y.

Next, we give a precise definition of Instrumental Sets, in terms of graphical conditions, and state the main result of this chapter.

Definition 4 (Instrumental Sets) *Let* $\mathbf{X} = \{X_1, \dots, X_k\}$ *be an arbitrary subset of parents of* Y. *The set* $\mathbf{Z} = \{Z_1, \dots, Z_n\}$ *is said to be an Instrumental Set relative to* $\mathbf{X}$ *and* Y *if there exist triplets* $(Z_1, \mathbf{W}_1, p_1), \dots, (Z_n, \mathbf{W}_n, p_n)$ *such that:*

(i) for $i = 1, \dots, k$, *variable* Z_i *and the elements of* $\mathbf{W}_i$ *are non-descendents of* Y;

(ii) Let $\overline{G}$ *be the subgraph obtained by deleting edges* $X_1 \to Y, \dots, X_n \to Y$ *from*

63
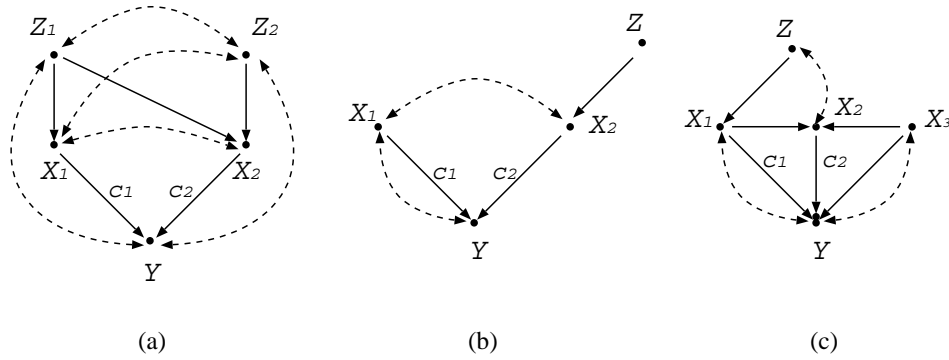
Figure 7.4: More examples of Instrumental Sets

the causal diagram of the model. Then, for $i = 1, \ldots, k$, the set $\mathbf{W}_i$ d-separates Z_i from Y in $\overline{G}$; but $\mathbf{W}_i$ does not block path p_i;

(iii) for $i = 1, \ldots, k$, p_i is an unblocked path between Z_i and Y including the edge $X_i \to Y$.

(iv) for $i < j$, the only possible common variable in paths p_i and p_j (other than Y) is variable Z_i, and, in this case, both p_i and $p_j[Z_j..Z_i]$ must point to Y;

Theorem 8 *If $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ is an Instrumental Set relative to variable Y and set of parents $\mathbf{X}$, then the parameters of edges $X_1 \to Y, \ldots, X_n \to Y$ are identified, and can be computed by solving a system of linear equations.*

The proof the theorem is given in Appendix C.

Figure 7.4 shows more examples in which the method of conditional IV's fails, but our new criterion is able to prove the identification of parameters c_i's. In particular, model (a) is a bow-free model, and thus is completely identifiable. Model (b) illustrates an interesting case in which variable X_2 is used as the instrument for $X_1 \to Y$, while Z is the instrument for $X_2 \to Y$. Finally, in model (c) we have an example

in which the parameter of edge $X_3 \to Y$ is nonidentifiable, and still the method can prove the identification of c_1 and c_2.

CHAPTER 8

Discussion and Future Work

An important contribution of this work was to offer a new approach to the problem of identification in SEM, based on the graphical analysis of the causal diagram of the model. This approach allowed us to obtain powerful sufficient conditions for identification, and a new method of computing correlation constraints.

This new approach has important advantages over existing techniques. Traditional methods [Dun75, Fis66] are based on algebraic manipulation of the equations that define the model. As a consequence, they have to handle the two types of structural assumptions contained in the model (e.g., (a) which variables appear in each equation; and, (b) how the error terms are correlated with each other) in different ways, which makes the analysis more complicated. The language of graphs, on the other hand, allows us to represent both types of assumptions in the same way, namely, by the presence or absence of edges in the causal diagram. This permits a uniform treatment in the analysis of identification.

Another important advantage of our approach relates to conditional independencies implied by the model. Most existing methods [Fis66, BT84, Pea00b] strongly rely on conditional independencies to prove that a model is identified. As a consequence, such methods are not very informative when the model has few such relations. Since we do not make direct us of conditional independencies in our derivations, we can prove identification in many cases where most methods fail. Even the method of instrumental sets, which involves a d-separation test, is not very sensitive to condi-

tional independencies, because the tests are performed on a potentially much sparser graph.

The sufficient conditions for identification obtained in this work correspond to the most general criteria current available to SEM investigators. To the best of our knowledge, this is also the first work that provides necessary conditions for identification. Although these results answer many questions in practical applications of SEM, finding a necessary and sufficient condition for identification is still an outstanding open problem from a theoretical perspective.

Another important question that remains open is the application of our graphical methods to non-recursive models. Since the basic tool for our analysis (i.e., Wright's decomposition of correlations) only applies to recursive models, this problem may require the development of new techniques.

APPENDIX A

(Proofs from Chapter 3)

Proof of Theorem 3:

Fix an arbitrary variable Y with Auxiliary Set $\mathcal{A}_Y = \{Z_1, \ldots, Z_m\}$.

Let $\mathbf{X} = \{X_1, \ldots, X_d\}$ be the set of parents of Y. For $i = 1, \ldots, d$, let δ_i denote the parameter of edge $X_i \to Y$.

Similarly, let $\mathbf{V} = \{V_1, \ldots, V_k\}$ be the set of variables such that bidirected edge $V_j \leftrightarrow Y$ belongs to $Inc(Y)$. For $j = 1, \ldots, k$, let λ_j denote the parameter of edge $V_j \leftrightarrow Y$. [Note that we may have $\mathbf{X} \cap \mathbf{V} \neq \phi$.]

Assume first that condition **C1** can be applied to the pair $(Y, \mathcal{A}_Y)$. Let Z be any variable from $\mathcal{A}_Y$. Then, we can write the decomposition of ρ_{ZY} as:

$$\rho_{ZY} = \sum_{i=1}^{d} c_i \delta_i + \sum_{j=1}^{k} b_j \lambda_j$$

and we show that coefficients c_i's and b_j's are identified.

First, note that every unblocked path between Z and Y including and edge $X_i \to Y$ can be decomposed into an unblocked path between Z and X_i and the edge $X_i \to Y$. Moreover, every unblocked path between Z and X_i can be extendend by edge $X_i \to Y$ to give an unblocked path between Z and Y. These factes imply that c_i is identified and given by ρ_{ZX_i}.

Now, let (V_j, Y) be an arbitrary bidirected edge from $Inc(Y)$. If $Z = V_j$ then $b_j = 1$, because there is only one unblocked path between Z and Y including (V_j, Y)

which consists precisely of this edge. In the other case, we have that $b_j = 0$, because any unblocked path between Z and Y including (V_j, Y) should consist of a chain from V_j to Z concatenated to the edge (V_j, Y), but no such chain exists. Hence, b_j is identified.

Assume now that condition **C2** is applied to $(Y, \mathcal{A}_Y)$. Let $Z \in \mathcal{A}_Y$, and let $\{W_1, \ldots, W_l\}$ be the set of parents of Z. For $s = 1, \ldots, l$, let α_s denote the parameter of edge $W_s \rightarrow Z$.

The idea is to replace the equation corresponding to the decomposition of ρ_{ZY} in $\Phi_{\mathcal{A}_Y, Y}$, by the following linear combination:

$$\rho_{ZY} - \sum_{s=1}^{l} \alpha_s \rho_{W_s Y} = c_0 + \sum_{i=1}^{d} c_i \delta_i + \sum_{j=1}^{k} b_j \lambda_j$$

This equation is also linearly independent of all the other ones in $\Phi_{\mathcal{A}_Y, Y}$ (*has to show that*). However, since many of the unblocked paths between Z and Y can be obatined by concatenation of some edge $W_s \rightarrow Z$ and an unblocked path between W_s and Y, many terms are cancelled out in the r.h.s.. We also observe that term c_0 in the r.h.s. corresponds to pahts that do not included edges from $Inc(Y)$, and is zero when Z is a non-descendant of Y (?).

Next lemma proves the identification of coefficients b_j's.

Lemma 7 *Each coefficient b_j in the r.h.s. of (*) is identified and has value either 0 or 1.*

Proof:

Fix a bidirected edge (V_j, Y) with parameter λ_j. We consider a few separate cases: (a) $Z \neq V_j$ and Z is a non-descendant of V_j.

Any unblocked path beween Z and Y including edge (V_j, Y) must consist of a chain from V_j to Z concatenaed with the edge (V_j, Y). Since Z is a non-descendant

of V_j, no such chain exists. Thus, the coefficient of λ_j in the decomposition of ρ_{ZY} is zero. Since each parent W_s of Z is also a non-descendant of V_j, the coefficient of λ_j in the decomposition of ρ_{W_sY} is also zero. Hence, $b_j = 0$.

(b) $Z = V_j$

Note that, in this case, there is only one unblocked path between Z and Y including (V_j, Y), which consists precisely of the edge (V_j, Y). Thus, the coefficient of λ_j in the decomposition of ρ_{ZY} is 1. Again, each parent W_s of Z is a non-descendant of Y, so the coefficient of λ_j in the decomposition of ρ_{W_sY} is zero. Hence, $b_j = 1$.

(c) Z is a descendant of V_j.

Let C_Z denote the set of chains from V_j to Z that do not include Y. (*For non-descendants of Y this is the set of all chains from V_j to Z*)

Clearly, each of these chains must include a parent of Z. Thus, we can partition the chains in C_Z according to the last parent of Z appearing in each of them. More precisely, let C_{Z,W_s} denote the set of chains in C_Z that include edge $W_s \to Z$. Now, we can write

$$
\begin{aligned}
T(C_Z) &= \textstyle\sum_{W_s} T(C_{Z,W_s}) \\
&= \textstyle\sum_{W_s} \alpha_s [T(C_{Z,W_s})/\alpha_s]
\end{aligned}
$$

Observing that $T(C_Z)$ is the coefficient of λ_j in the decomposition of ρ_{ZY}, and that $[T(C_{Z,W_s})/\alpha_s]$ is the coefficient of λ_j in the decomposition of ρ_{W_sY}, we conclude that $b_j = 0$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Now, it just remain to analyze the coefficients c_i's. The next lemma takes care of the case when Z is a non-descendant of Y.

Lemma 8 *Let U be a non-descendant of Y. Then, the coefficient of δ_i in the decomposition of ρ_{UY} is identified and given by ρ_{UX_i}.*

Proof:

Recall that δ_i is th parameter of directed edge $X_i \to Y$. Then, the lemma follows from the facts that every unblocked path between U and X_i can be extended by edge $X_i \to Y$ to give an unblocked path between U and Y, and those are all the unblocked paths between U and Y including edge $X_i \to Y$. $\qquad\square$

The identification of coefficients c_i's when Z is a non-descendant of Y, then follows from the identification of the α_s's.

Now, consider the identification of the c_i's when Z is a descendant of Y.

First, we observe that any unblocked path between Z and Y that contains a directed edgepointing to Z can be decomposed into an unblocked path between W_s and Y, and edge $W_s \to Z$ (for some W_s). Thus, all terms associated with such path are cancelled out in (*).

This immediately gives that $c_0 = 0$, since every unblocked path between Y and Z must either end with a directed edge pointing to Z, or be a bidirected edge between Y and Z. However, in the later case, Y would be in the Auxiliary Set for Z, which would create a cycle in the dependency graph.

For $i = 1, \ldots, k$, each term in coefficient c_i corresponds to a path consisting of the concatenation of bidirected edge (U, Z) and a chain from U to X_i. Since $\Phi_{\mathcal{A}_Z, Z}$ and all $\Phi_{\mathcal{A}_W, W}$, where W is an intermediated variable in the chain from U to Z, are solved brfore $\Phi_{\mathcal{A}_Y, Y}$, all paramters in this path are identified. But this imples that C_i is identified. $\qquad\square$

APPENDIX B

(Proofs from Chapter 6)

Proof of Lemma 6:

Let $\lambda_1, \ldots, \lambda_m$ denote the parameters of the edges in $Inc(Y)$.

Let Q denote the set of unblocked paths between U and V that do not include any edge from $Inc(Y)$. Then, c_0 is given by $T(Q)$, and clearly does not depend on the values of the λ_i's.

Now, let P denote the set of unblocked paths p between U and V that include some edge from $Inc(Y)$, with such edge appearing in subpath $p[U..Y]$ (the other case is similar).

Let p be an arbitrary path from P. Then, it follows from assumption (iii) in the theorem that $p[U..Y]$ must contain at least one variable from $\mathcal{A}_Y$. We let Z_i denote the first (i.e., closest to U) variable from $\mathcal{A}_Y$ to appear in p, and divide the path into three segments:

- $p_1 = p[U..Z_i]$;

- $p_2 = p[Z_i..Y]$;

- $p_3 = p[Y..V]$;

where both p_1 and p_3 can be null, if $U = Z_i$ or $Y = V$, respectively. We also note that p_1 is a chain from Z_i to U (by assumption (iii)), and p_3 is a chain from Y to V, because every edge from $Inc(Y)$ points to Y and p is unblocked.

Now, for each $Z_i \in \mathcal{A}_Y$, let C_i denote the set of all chains from Z_i to U that do not contain any variable from $\mathcal{A}_Y \cup \{Y\}$; and, let B denote the set of all chains from Y to V.

Also, for each $Z_i \in \mathcal{A}_Y$, let R_i denote the set of unblocked paths between Z_i and Y.

Proposition 1 *For any $Z_i \in \mathcal{A}_Y$, let $p_1 \in C_i$, let $p_2 \in R_i$, and let $p_3 \in B$. Then, the concatenation of p_1, p_2, and p_3 gives a valid unblocked path between U and V if and only if p_1 and p_3 do not have any intermediate variable in common.*

Proof: If p_1 and p_3 have a common variable, then clearly the concatenation gives an invalid path.

In the other case, the proposition follows by observing that every intermediate variable in p_2 belongs to $\mathcal{A}_Y$ (follows from assumption (iii)), and that intermediate variables in p_1 and p_3 cannot belong to $\mathcal{A}_Y$ (by definition, and by assumption (ii)). $\square$

The arguments above give that we can write

$$T(P) = \sum_i c_i \cdot T(R_i)$$

where

$$c_i = \sum_{p_1 \in C_i, p_3 \in B, p_1 \cap p_3 = \phi} T(p_1) \cdot T(p_3)$$

Clearly, each of the c_i is independent of the values of parameters λ_i's. The lemma follows by observing that $T(R_i) = \rho_{Z_i Y}$. $\square$

Proof of Theorem 7:

For each variable $Z_j \in \mathcal{A}_Y - \overline{\mathcal{A}_Y}$, let R_j denote the set composed by the following unblocked paths between Z_j and Y:

a) all chains from Z_j to Y whose intermediate variables belong to $\mathcal{A}_Y - \overline{\mathcal{A}_Y}$;

b) all unblocked paths p that point to Z_j and do not have a variable $W \notin (\mathcal{A}_Y - \overline{\mathcal{A}_Y}) \cup \{Y\}$ which is a descendant of Z_j and $p[Z_j..W]$ point to W.

For each variable $Z_i \in \overline{\mathcal{A}_Y}$, let R_i denote the set composed by all chains from Z_i to Y whose intermediate variables belong to $\mathcal{A}_Y - \overline{\mathcal{A}_Y}$.

We first show that the correlations of each variable $Z \in \mathcal{A}_Y$ and Y can be expressed as a linear combination of the terms $T(R_j)$'s and $T(R_i)$'s.

Case 1: $Z_j \in \mathcal{A}_Y - \overline{\mathcal{A}_Y}$.

Let $\mathcal{P}_j$ be the set of unblocked paths between Z_j and Y, so that $\rho_{Z_j Y} = T(\mathcal{P}_j)$. Clearly, we can write $\rho_{Z_j Y} = T(R_j) + T(\mathcal{P}_j - R_j)$. Thus, we just need to show that the second term on the right-hand side is a linear combination of the $T(R_j)$'s and $T(R_i)$'s.

There are two types of paths in $\mathcal{P}_j - R_j$:

1) chains from Z_j to Y with some intermediate variable that does not belong to $\mathcal{A}_Y - \overline{\mathcal{A}_Y}$;

2) unblocked paths p that include a variable $W \notin (\mathcal{A}_Y - \overline{\mathcal{A}_Y}) \cup \{Y\}$ which is a descendant of Z_j and $p[Z_j..W]$ point to W.

Let us consider paths of type (1) first.

Proposition 2 *Every path $p \in \mathcal{P}_j - R_j$ of type (1) contains a variable from $\overline{\mathcal{A}_Y}$.*

Proof: Assume that the proposition does not hold for some path p of type (1). Then, p must include at least one variable that does not belong to $\mathcal{A}_Y$. Let W be the last such variable in p (i.e., closest to Y), and let V be the variable adjacent to W in $p[W..Y]$. Since p is a chain from Z_j to Y, it follows from condition (iii) in the theorem that $V \neq Y$, and so we must have $V \in \mathcal{A}_Y$. But then edge $(W \to V)$ witnesses that $V \in \overline{\mathcal{A}_Y}$, which contradicts the initial assumption. $\qquad\square$

Fix a path $p \in \mathcal{P}_j - R_j$ of type (1), and let Z_i be the last variable from $\overline{\mathcal{A}_Y}$ in p. Let q be an arbitrary path from R_i.

Proposition 3 *The concatenation $p[Z_j..Z_i] \oplus q$ gives a valid chain from Z_j to Y.*

Proof: Since $p[Z_j..Z_i]$ is a chain from Z_j to Z_i, and q is a chain from Z_i to Y, it follows that the concatenation of $p[Z_j..Z_i] \oplus q$ is a chain from Z_j to Y. Now, such chain must be a valid one, otherwise we would have a contradicion to the recursiveness of the model. $\qquad\square$

It follows from the two propositions above that $T(\mathcal{P}_j - R_j|_{(1)})$ can be expressed as a linear combination of the $T(R_i)$'s.

Now, we consider the paths of type (2).

Proposition 4 *Every path $p \in \mathcal{P}_j - R_j$ of type (2) contains a variable $Z_i \in \overline{\mathcal{A}_Y}$ such that*

(i) Z_i is a descendant of Z_j;

(ii) $p[Z_j..Z_i]$ points to Z_i.

Proof: Assume that the proposition does not hold for some path p of type (2). Then, p must include at least one variable that does not belong to $\mathcal{A}_Y$ and satisfies (i) and (ii) above. Let W be the last such variable in p (i.e., closest to Y), and let V be the variable

75

adjacent to W in $p[W..Y]$. Since p is unblocked and $p[Z_j..W]$ points to W, we get that $p[W..Y]$ must be a chain from W to Y. This fact, together with condition (iii) gives that $V \neq Y$, and so we must have $V \in \mathcal{A}_Y$. But then edge $(W \to V)$ implies that $V \in \overline{\mathcal{A}_Y}$, which contradicts the initial assumption. $\square$

Fix a path $p \in \mathcal{P}_j - R_j$ of type (2), and let Z_i be the last variable from $\overline{\mathcal{A}_Y}$ in p.

Proposition 5 *Every intermediate variable in $p[Z_j..Z_i]$ is an ancestor of Z_i.*

Proof: Let Z_l be the last variable in $p[Z_j..Z_i]$ such that $p[Z_l..Z_i]$ is not a chain from Z_l to Z_i. Clearly, any intermediate variable in $p[Z_l..Z_i]$ is an ancestor of Z_i. Moreover, subpath $p[Z_l..Z_i]$ must point to Z_l. Since p is unblocked, it follows that $p[Z_j..Z_l]$ must be a chain from Z_l to Z_j. But this implies that Z_l and every intermediate variable in $p[Z_j..Z_l]$ is an ancestor of Z_j. Since Z_j is an ancestor of Z_i the proposition holds. $\square$

Let q be an arbitrary path from R_i.

Proposition 6 *The concatenation $p[Z_j..Z_i] \oplus q$ gives a valid unblocked path between Z_j and Y.*

Proof: Since q is a chain from Z_i to Y, the path obtained from the concatenation is unblocked. The validity of the path follows from the facts that every intermediate variable in $p[Z_j..Z_i]$ is an ancestor of Z_i, every intermediate variable in q is a descendant of Z_i, and the recursiveness of the model. $\square$

It follows from the two propositions above that $T(\mathcal{P}_j - R_j|_{(2)})$ can be expressed as a linear combination of the $T(R_i)$'s.

Case 2: $Z_i \in \overline{\mathcal{A}_Y}$.

Let $\mathcal{P}_i$ be the set of unblocked paths between Z_i and Y, so that $\rho_{Z_i Y} = T(\mathcal{P}_i)$. Clearly, we can write $\rho_{Z_i Y} = T(R_i) + T(\mathcal{P}_i - R_i)$. Thus, we only need to show

that the second term on the right-hand side is a linear combination of the $T(R_i)$'s and $T(R_j)$'s.

There are two types of paths in $\mathcal{P}_i - R_i$:

1) chains from Z_j to Y with some intermediate variable that does not belong to $\mathcal{A}_Y - \overline{\mathcal{A}_Y}$;

2) unblocked paths between Z_i and Y that point to Z_i.

Let us consider first paths of type (1).

Proposition 7 *Every path $p \in \mathcal{P}_i - R_i$ of type (1) contains an intermediate variable from $\overline{\mathcal{A}_Y}$.*

Proof: (same as proof of proposition 2) □

Fix a path $p \in \mathcal{P}_i - R_i$ of type (1), and let Z_l be the last intermediate variable from $\overline{\mathcal{A}_Y}$ in p. Let q be an arbitrary path from R_l.

Proposition 8 *The concatenation of $p[Z_i..Z_l] \oplus q$ gives a valid chain from Z_j to Y.*

Proof: (same as proof of proposition 3) □

It follows from the two propositions above that $T(\mathcal{P}_i - R_i|_{(1)})$ can be expressed as a linear combination of the $T(R_l)$'s.

Now, we consider paths of type (2), and further divide them into:

a) paths p that have an intermediate variable Z_j from $\mathcal{A}_Y - \overline{\mathcal{A}_Y}$, such that $p[Z_i..Z_j]$ is a chain from Z_j to Z_i;

b) all the remaining paths.

77

Let $p \in \mathcal{P}_i - R_i$ be of type $(2a)$, and let Z_j be the first variable from $\mathcal{A}_Y - \overline{\mathcal{A}_Y}$ in p (i.e., closest to Z_i). Let q be an arbitrary path from R_j.

Proposition 9 *The concatenation $p[Z_i..Z_j] \oplus q$ gives a valid unblocked path between Z_j and Y.*

Proof: Since $p[Z_i..Z_j]$ is a chain from Z_j to Z_i, the path obtained from the concatenation is unblocked. Now, let V be an intermediate variable in $p[Z_i..Z_j]$. Then, $V \notin (\mathcal{A}_Y - \overline{\mathcal{A}_Y}) \cup \{Y\}$ and V is a descendant of Z_j. But, by definition, no path in R_j contain such variables. Thus, it follows that the concatenation produces a valid path. $\square$

It follows from the two propositions above that $T(\mathcal{P}_i - R_i|_{(2a)})$ can be expressed as a linear combination of the $T(R_j)$'s.

Now, consider paths p of type $(2b)$.

Proposition 10 *Every path $p \in \mathcal{P}_i - R_i$ be of type $(2b)$ contains an intermediate variable Z_l from $\overline{\mathcal{A}_Y}$ such that $p[Z_l..Y]$ is a chain from Z_l to Y.*

Proof: Let p be a path of type $(2b)$, and let W be the first variable in p such that $p[Z_i..W]$ is not a chain. Clearly, such variable must exist, otherwise p would be a chain from Y to Z_i that does not include any edge from $Inc(Y)$. Moreover, $p[Z_i..W]$ points to W, and so $p[W..Y]$ must be a chain from W to Y.

If $W \in \overline{\mathcal{A}_Y}$, then we are done. So, we consider the two remaining cases.

Assume that $W \notin \mathcal{A}_Y$. Then, the proposition easily follows from the facts that $p[W..Y]$ is a chain from W to Y, and that there is no edge (U, Z) pointing to Z, where $U \notin \mathcal{A}_Y$ and $Z \in (\mathcal{A}_Y - \overline{\mathcal{A}_Y}) \cup \{Y\}$.

Finally, if $W \in (\mathcal{A}_Y - \overline{\mathcal{A}_Y})$, then let V be the variable adjacent to W in $p[Z_i..W]$. Since edge (V, W) points to W, it follows that $V \in \mathcal{A}_Y$. But V cannot belong to

$(\mathcal{A}_Y - \overline{\mathcal{A}_Y})$, because $p[Z_i..V]$ is a chain from V to Z_i, and then p would be of type $(2a)$. Thus, we must have $V \in \overline{\mathcal{A}_Y}$, and condition (ii) of the theorem gives that $p[V..W]$ is a directed edge from V to W. Hence, $p[V..Y]$ is a chain from V to Y. $\quad\square$

Fix a path $p \in \mathcal{P}_i - R_i$ of type $(2b)$, and let Z_l be the last intermediate variable from $\overline{\mathcal{A}_Y}$ in p. Let q be an arbitrary path from Z_l.

Proposition 11 *The concatenation $p[Z_i..Z_l] \oplus q$ gives a valid unblocked path between Z_j and Y.*

Proof: Since q is a chain from Z_l to Y, the path obtained from the concatenation is unblocked. Now, let V be the last variable in $p[Z_i..Z_l]$ such that $p[Z_i..V]$ is a chain from V to Z_i (if there is no such variable, we take $V = Z_i$). It follows that, since p is of type $(2b)$, $p[Z_i..V]$ does not contain any variable from $(\mathcal{A}_Y - \overline{\mathcal{A}_Y})$. Moreover, every intermediate variable in $p[V..Z_l]$ is an ancestor of Z_l. But, by definition, intermediate variables in any path from R_l must belong to $(\mathcal{A}_Y - \overline{\mathcal{A}_Y})$ and be descendants of Z_l. Hence, the concatenation produces a valid path. $\quad\square$

It follows from the two propositions above that $T(\mathcal{P}_i - R_i|_{(2b)})$ can be expressed as a linear combination of the $T(R_l)$'s.

Case 3: $W \notin \mathcal{A}_Y$.

Let $\mathcal{P}_w$ denote the set of unblocked paths between W and Y, so that $\rho_{WY} = T(\mathcal{P}_w)$.

There are two types of paths in $\mathcal{P}_w$:

1) chains from W to Y;

2) unblocked paths between W and Y that point to W.

Let us consider paths of type (1) first.

Proposition 12 *Every path $p \in \mathcal{P}_w$ of type (1) contains an intermediate variable from* $\overline{\mathcal{A}_Y}$.

Proof: Follows from the fact that there is no edge (U, Z) pointing to Z, where $U \notin \mathcal{A}_Y$ and $Z \in (\mathcal{A}_Y - \overline{\mathcal{A}_Y} \cup \{Y\}$. $\qquad\qquad\square$

Fix a path $p \in \mathcal{P}_w$ of type (1), and let Z_i be the last intermediate variable from $\overline{\mathcal{A}_Y}$ in p. Let q be an arbitrary path from R_i.

Proposition 13 *The concatenation of $p[W..Z_i] \oplus q$ gives a valid chain from W to Y.*

Proof: (same as proof of proposition 3) $\square$

It follows from the two propositions above that $T(\mathcal{P}_w|_{(1)})$ can be expressed as a linear combination of the $T(R_i)$'s.

Now, we consider paths of type (2), and further divide them into:

a) paths p that have an intermediate variable Z_j from $(\mathcal{A}_Y - \overline{\mathcal{A}_Y})$ such that $p[W..Z_j]$ is a chain from Z_j to W;

b) all the remaining paths.

Fix a path $p \in \mathcal{P}_w$ of type $(2a)$, and let Z_j be the first variable from $(\mathcal{A}_Y - \overline{\mathcal{A}_Y})$ in p. Let q be an arbitrary path from R_j.

Proposition 14 *The concatenation $p[W..Z_j] \oplus q$ gives a valid unblocked path between W and Y.*

Proof: (same as proof of proposition 9) $\square$

It follows from the two propositions above that $T(\mathcal{P}_w|_{(2a)})$ can be expressed as a linear combination of the $T(R_j)$'s.

Now, we consider paths of type $(2b)$.

Proposition 15 *Every path $p \in \mathcal{P}_w$ of type $(2b)$ contains an intermediate variable Z_i from $\overline{\mathcal{A}_Y}$ such that $p[Z_i..Y]$ is a chain from Z_i to Y.*

Proof: (same as proof of proposition 10) □

Fix a path $p \in \mathcal{P}_w$ of type $(2b)$, and let Z_i be the last intermediate variable from $\overline{\mathcal{A}_Y}$ in p. Let q be an arbitrary path from R_i.

Proposition 16 *The concatenation $p[W..Z_i] \oplus q$ gives a valid unblocked path between W and Y.*

Proof: (same as proof of proposition 11) □

It follows from the two propositions above that $T(\mathcal{P}_w|_{(2b)})$ can be expressed as a linear combination of the $T(R_i)$'s. □

APPENDIX C

(Proofs from Chapter 7)

C.1   Preliminary Results

C.1.1   Partial Correlation Lemma

Next lemma provides a convenient expression for the partial correlation coefficient of Y_1 and Y_2, given $Y_3, \ldots, Y_n$, denoted $\rho_{12.3\ldots n}$. The proof is given in Section C.3.

Lemma 9 *The partial correlation $\rho_{12.3\ldots n}$ can be expressed as the ratio:*

$$\rho_{12.3\ldots n} = \frac{\phi(1, 2, \ldots, n)}{\psi(1, 3, \ldots, n) \cdot \psi(2, 3, \ldots, n)} \tag{C.1}$$

where ϕ and ψ are functions of the correlations among $Y_1, Y_2, \ldots, Y_n$, satisfying the following conditions:

(i) $\phi(1, 2, \ldots, n) = \phi(2, 1, \ldots, n)$.

(ii) $\phi(1, 2, \ldots, n)$ *is linear on the correlations $\rho_{12}, \rho_{32}, \ldots, \rho_{n2}$, with no constant term.*

(iii) *The coefficients of $\rho_{12}, \rho_{32}, \ldots, \rho_{n2}$, in $\phi(1, 2, \ldots, n)$ are polynomials on the correlations among the variables $Y_1, Y_3, \ldots, Y_n$. Moreover, the coefficient of ρ_{12} has the constant term equal to 1, and the coefficients of $\rho_{32}, \ldots, \rho_{n2}$, are linear on the correlations $\rho_{13}, \rho_{14}, \ldots, \rho_{1n}$, with no constant term.*

(iv) $(\psi(i_1, \ldots, i_{n-1}))^2$, *is a polynomial on the correlations among the variables* $Y_{i_1}, \ldots,$

$Y_{i_{n-1}}$, *with constant term equal to 1.*

C.1.2 Path Lemmas

The following lemmas explore some consequences of the conditions in the definition of Instrumental Sets.

Lemma 10 *W.l.o.g., we may assume that, for* $1 \leq i < j \leq n$, *paths* p_i *and* p_j *do not have any common variable other than (possibly)* Z_i.

Proof: Assume that paths p_i and p_j have some variables in common, different from Z_i. Let V be the closest variable to X_i in path p_i which also belongs to path p_j.

We show that after replacing triple $(Z_i, \mathbf{W}_i, p_i)$ by triple $(V, \mathbf{W}_i, p_i[V..Y])$, conditions $(i) - (iii)$ still hold.

It follows from condition (iii) that subpath $p_i[V..Y]$ must point to V. Since p_i is unblocked, subpath $p_i[Z_i..V]$ must be a directed path from V to Z_i.

Now, variable V cannot be a descendent of Y, because $p_i[Z_i..V]$ is a directed path from V to Z_i, and Z_i is a non-descendent of Y. Thus, condition (i) still holds.

Consider the causal graph $\overline{G}$. Assume that there exists a path p between V and Y witnessing that $\mathbf{W}_i$ does not d-separate V from Y in $\overline{G}$. Since $p_i[Z_i..V]$ is a directed path from V to Z_i, we can always find another path witnessing that $\mathbf{W}_i$ does not d-separate Z_i from Y in $\overline{G}$ (for example, if p and $p_i[Z_i..V]$ do not have any variable in common other than V, then we can just take their concatenation). But this is a contradiction, and thus it is easy to see that condition (ii) still holds.

Condition (iii) follows from the fact that $p_i[V..Y]$ and $p_j[Z_j..V]$ point to V. $\quad\square$

In the following, we assume that the conditions of lemma 10 hold.

Lemma 11 *For all* $1 \leq i \leq n$, *there exists no unblocked path between* Z_i *and* Y, *different from* p_i, *which includes edge* $X_i \rightarrow Y$ *and is composed only by edges from* $p_1, \ldots, p_i$.

Proof: Let p be an unblocked path between Z_i and Y, different from p_i, and assume that p is composed only by edges from $p_1, \ldots, p_i$.

According to condition (iii), if Z_i appears in some path p_j, with $j \neq i$, then it must be that $j > i$. Thus, p must start with some edges of p_i.

Since p is different from p_i, it must contain at least one edge from $p_1, \ldots, p_{i-1}$. Let (V_1, V_2) denote the first edge in p which does not belong to p_i.

From lemma 10, it follows that variable V_1 must be a Z_k for some $k < i$, and by condition (iii), both subpath $p[Z_i..V_1]$ and edge (V_1, V_2) must point to V_1. But this implies that p is blocked by V_1, which contradicts our assumptions. $\square$

The proofs for the next two lemmas are very similar to the previous one, and so are omitted.

Lemma 12 *For all* $1 \leq i \leq n$, *there is no unblocked path between* Z_i *and some* W_{i_j} *composed only by edges from* $p_1, \ldots, p_i$.

Lemma 13 *For all* $1 \leq i \leq n$, *there is no unblocked path between* Z_i *and* Y *including edge* $X_j \rightarrow Y$, *with* $j < i$, *composed only by edges from* $p_1, \ldots, p_i$.

C.2   Proof of Theorem 8

C.2.1   Notation and Basic Linear Equations

Fix a variable Y in the model. Let $\mathbf{X} = \{X_1, \ldots, X_k\}$ be the set of all non-descendents of Y which are connected to Y by an edge (directed, bidirected, or both). Define the

following set of edges with an arrowhead at Y:

$$Inc(Y) = \{(X_i, Y) : X_i \in \mathbf{X}\}$$

Note that for some $X_i \in \mathbf{X}$ there may be more than one edge between X_i and Y (one directed and one bidirected). Thus, $|Inc(Y)| \geq |\mathbf{X}|$. Let $\lambda_1, \ldots, \lambda_m$, $m \geq k$, denote the parameters of the edges in $Inc(Y)$.

It follows that edges $X_1 \to Y, \ldots, X_n \to Y$, belong to $Inc(Y)$, because $X_1, \ldots, X_n$, are clearly non-descendents of Y. W.l.o.g., let λ_i be the parameter of edge $X_i \to Y$, $1 \leq i \leq n$, and let $\lambda_{n+1}, \ldots, \lambda_m$ be the parameters of the remaining edges in $Inc(Y)$.

Let Z be any non-descendent of Y. Wright's equation for the pair (Z, Y), is given by

$$\rho_{Z,Y} = \sum_{\text{paths } p_l} T(p_l) \tag{C.2}$$

where each term $T(p_l)$ corresponds to an unblocked path between Z and Y. Next lemma proves a property of such paths.

Lemma 14 *Let Y be a variable in a recursive model, and let Z be a non-descendent of Y. Then, any unblocked path between Z and Y must include exactly one edge from $Inc(Y)$.*

Lemma 14 allows us to write Eq. (C.2) as

$$\rho_{Z,Y} = \sum_{j=1}^{m} a_j \cdot \lambda_j \tag{C.3}$$

Thus, the correlation between Z and Y can be expressed as a linear function of the parameters $\lambda_1, \ldots, \lambda_m$, with no constant term.

Consider a triple $(Z_i, \mathbf{W}_i, p_i)$, and let $\mathbf{W}_i = \{W_{i_1}, \ldots, W_{i_k}\}$ [1]. From lemma 9, we can express the partial correlation of Z_i and Y given $\mathbf{W}_i$ as:

---

[1] To simplify the notation, we assume that $|\mathbf{W}_i| = k$, for $i = 1, \ldots, n$

$$\rho_{Z_i Y . \mathbf{W}_i} = \frac{\phi_i(Z_i, Y, W_{i_1}, \ldots, W_{i_k})}{\psi_i(Z_i, W_{i_1}, \ldots, W_{i_k}) \cdot \psi_i(Y, W_{i_1}, \ldots, W_{i_k})} \tag{C.4}$$

where function ϕ_i is linear on the correlations $\rho_{Z_i Y}$, $\rho_{W_{i_1} Y}$, ..., $\rho_{W_{i_k} Y}$, and ψ_i is a function of the correlations among the variables given as arguments. We abbreviate $\phi_i(Z_i, Y, W_{i_1}, \ldots, W_{i_{k_i}})$ by $\phi_i(Z_i, Y, \mathbf{W}_i)$, and $\psi_i(V, W_{i_1}, \ldots, W_{i_k})$ by $\psi_i(V, \mathbf{W}_i)$.

We have seen that the correlations $\rho_{Z_i Y}$, $\rho_{W_{i_1} Y}$, ..., $\rho_{W_{i_k} Y}$, can be expressed as linear functions of the parameters $\lambda_1, \ldots, \lambda_m$. Since ϕ_i is linear on these correlations, it follows that we can express ϕ_i as a linear function of the parameters $\lambda_1, \ldots, \lambda_m$.

Formally, by lemma 9, $\phi_i(Z_i, Y, \mathbf{W}_i)$ can be written as:

$$\phi_i(Z_i, Y, \mathbf{W}_i) = b_{i_0} \rho_{Z_i Y} + b_{i_1} \rho_{W_{i_1} Y} + \ldots + b_{i_k} \rho_{W_{i_k} Y} \tag{C.5}$$

Also, for each $V_j \in \{Z_i\} \cup \mathbf{W}_i$, we can write:

$$\rho_{V_j Y} = a_{i_j 1} \lambda_1 + \ldots + a_{i_j m} \lambda_m \tag{C.6}$$

Replacing each correlation in Eq.(C.5) by the expression given by Eq. (C.6), we obtain

$$\phi_i(Z_i, Y, \mathbf{W}_i) = q_{i1} \lambda_1 + \ldots + q_{im} \lambda_m \tag{C.7}$$

where the coefficients q_{il}'s are given by:

$$q_{il} = \sum_{j=0}^{k} b_{i_j} a_{i_j l} \qquad , l = 1, \ldots, m \tag{C.8}$$

Lemma 15 *The coefficients $q_{i,n+1}, \ldots, q_{im}$ in Eq. (C.7) are identically zero.*

Proof: The fact that $\mathbf{W}_i$ d-separates Z_i from Y in $\overline{G}$, implies that $\rho_{Z_i Y . \mathbf{W}_i} = 0$ in any probability distribution compatible with $\overline{G}$ ([Pea00a], pg. 142). Thus, $\phi_i(Z_i, Y, \mathbf{W}_i)$ must vanish when evaluated in $\overline{G}$. But this implies that the coefficient of each of the λ_i's in Eq. (C.7) must be identically zero.

Now, we show that the only difference between evaluations of $\phi_i(Z_i, Y, \mathbf{W}_i)$ on the causal graphs $\overline{G}$ and G, consists on the coefficients of parameters $\lambda_1, \ldots, \lambda_n$.

First, observe that coefficients $b_{i_0}, \ldots, b_{i_k}$ are polynomials on the correlations among the variables $Z_i, W_{i_1}, \ldots, W_{i_k}$. Thus, they only depend on the unblocked paths between such variables in the causal graph. However, the insertion of edges $X_1 \to Y$, $\ldots$, $X_n \to Y$, in $\overline{G}$ does not create any new unblocked path between any pair of $Z_i, W_{i_1}, \ldots, W_{i_k}$ (and obviously does not eliminate any existing one). Hence, the coefficients $b_{i_0}, \ldots, b_{i_k}$ have exactly the same value in the evaluations of $\phi_i(Z_i, Y, \mathbf{W}_i)$ on $\overline{G}$ and G.

Now, let λ_l be such that $l > n$, and let $V_j \in \{Z_i\} \cup \mathbf{W}_i$. Note that the insertion of edges $X_1 \to Y, \ldots, X_n \to Y$, in $\overline{G}$ does not create any new unblocked path between V_j and Y including the edge whose parameter is λ_l (and does not eliminate any existing one). Hence, coefficients $a_{i_j l}$, $j = 0, \ldots, k$, have exactly the same value on $\overline{G}$ and G.

From the two previous facts, we conclude that, for $l > n$, the coefficient of λ_l in the evaluations of $\phi_i(Z_i, Y, \mathbf{W}_i)$ on $\overline{G}$ and G have exactly the same value, namely zero. Next, we argue that $\phi_i(Z_i, Y, \mathbf{W}_i)$ does not vanish when evaluated on G.

Finally, let λ_l be such that $l \leq n$, and let $V_j \in \{Z_i\} \cup \mathbf{W}_i$. Note that there is no unblocked path between V_j and Y in $\overline{G}$ including edge $X_l \to Y$, because this edge does not exist in $\overline{G}$. Hence, the coefficient of λ_l in the expression for the correlation $\rho_{V_j Y}$ on $\overline{G}$ must be zero.

On the other hand, the coefficient of λ_l in the same expression on G is not necessarily zero. In fact, it follows from the conditions in the definition of Instrumental sets that, for $l = i$, the coefficient of λ_i contains the term $T(p_i)$. $\qquad \square$

From lemma 15, we get that $\phi_i(Z_i, Y, \mathbf{W}_i)$ is a linear function only on the parameters $\lambda_1, \ldots, \lambda_n$.

C.2.2 System of Equations Φ

Rewriting Eq.(C.4) for each triple $(Z_i, \mathbf{W}_i, p_i)$, we obtain the following system of linear equations on the parameters $\lambda_1, \ldots, \lambda_n$:

$$
\Phi = \begin{cases}
\phi_1(Z_1, Y, \mathbf{W}_1) = & \rho_{Z_1 Y . \mathbf{W}_1} \\
& \cdot \psi_1(Z_1, \mathbf{W}_1) \cdot \psi_1(Y, \mathbf{W}_1) \\
\ldots & \\
\phi_n(Z_n, Y, \mathbf{W}_n) = & \rho_{Z_n Y . \mathbf{W}_n} \\
& \cdot \psi_n(Z_n, \mathbf{W}_n) \cdot \psi_n(Y, \mathbf{W}_n)
\end{cases}
$$

where the terms on the right-hand side can be computed from the correlations among the variables $Y, Z_i, W_{i_1}, \ldots, W_{i_k}$, estimated from data.

Our goal is to show that Φ can be solved uniquely for the λ_i's, and so prove the identification of $\lambda_1, \ldots, \lambda_n$. Next lemma proves an important result in this direction. Let Q denote the matrix of coefficients of Φ.

Lemma 16 $Det(Q)$ *is a non-trivial polynomial on the parameters of the model.*

Proof: From Eq.(C.8), we get that each entry q_{il} of Q is given by

$$
q_{il} = \sum_{j=0}^{k} b_{i_j} \cdot a_{i_j l}
$$

where b_{i_j} is the coefficient of $\rho_{W_{i_j} Y}$ (or $\rho_{Z_i Y}$, if $j = 0$), in the linear expression for $\phi_i(Z_i, Y, \mathbf{W}_i)$ in terms of correlations (see Eq.(C.5)); and $a_{i_j l}$ is the coefficient of λ_l in the expression for the correlation $\rho_{W_{i_j} Y}$ in terms of the parameters $\lambda_1, \ldots, \lambda_m$ (see Eq.(C.6)).

From property (iii) of lemma 9, we get that b_{i_0} has constant term equal to 1. Thus, we can write $b_{i_0} = 1 + \hat{b}_{i_0}$, where $\hat{b}_{i_0}$ represent the remaining terms of b_{i_0}.

Also, from condition (i) of Theorem 8, it follows that $a_{i_0 i}$ contains term $T(p_i)$. Thus, we can write $a_{i_0 i} = T(p_i) + \hat{a}_{i_0 i}$, where $\hat{a}_{i_0 i}$ represents all the remaining terms of $a_{i_0 i}$.

Hence, a diagonal entry q_{ii} of Q, can be written as

$$q_{ii} = T(p_i)[1 + \hat{b}_{i_0}] + \hat{a}_{i_0 i} \cdot b_{i_0} + \sum_{j=1}^{k} b_{i_j} \cdot a_{i_j i} \qquad \text{(C.9)}$$

Now, the determinant of Q is defined as the weighted sum, for all permutations π of $\langle 1, \ldots, n \rangle$, of the product of the entries selected by π (entry q_{il} is selected by permutation π if the ith element of π is l), where the weights are 1 or (-1), depending on the parity of the permutation. Then, it is easy to see that the term

$$T^* = \prod_{j=1}^{n} T(p_j)$$

appears in the product of permutation $\pi = \langle 1, \ldots, n \rangle$, which selects all the diagonal entries of Q.

We prove that $det(Q)$ does not vanish by showing that T^* appears only once in the product of permutation $\langle 1, \ldots, n \rangle$, and that T^* does not appear in the product of any other permutation.

Before proving those facts, note that, from the conditions of lemma 10, for $1 \leq i < j \leq n$, paths p_i and p_j have no edge in common. Thus, every factor of T^* is distinct from each other.

Proposition 17 *Term T^* appears only once in the product of permutation $\langle 1, \ldots, n \rangle$.*

Proof: Let τ be a term in the product of permutation $\langle 1, \ldots, n \rangle$. Then, τ has one factor corresponding to each diagonal entry of Q.

A diagonal entry q_{ii} of Q can be expressed as a sum of three terms (see Eq.(C.9)).

Let i be such that for all $l > i$, the factor of τ corresponding to entry q_{ll} comes from the first term of q_{ll} (i.e., $T(p_l)[1 + \hat{b}_{l_0}]$).

Assume that the factor of τ corresponding to entry q_{ii} comes from the second term of q_{ii} (i.e., $\hat{a}_{i_0 i} \cdot b_{i_0}$). Recall that each term in $\hat{a}_{i_0 i}$ corresponds to an unblocked path between Z_i and Y, different from p_i, including edge $X_i \to Y$. However, from lemma 11, any such path must include either an edge which does not belong to any of $p_1, \ldots, p_n$, or an edge which appears in some of $p_{i+1}, \ldots, p_n$. In the first case, it is easy to see that τ must have a factor which does not appear in T^*. In the second, the parameter of an edge of some p_l, $l > i$, must appear twice as a factor of τ, while it appears only once in T^*. Hence, τ and T^* are distinct terms.

Now, assume that the factor of τ corresponding to entry q_{ii} comes from the third term of q_{ii} (i.e., $\sum_{j=1}^{k} b_{i_j} \cdot a_{i_j i}$). Recall that b_{i_j} is the coefficient of $\rho_{W_{i_j} Y}$ in the expression for $\phi_i(Z_i, Y, \mathbf{W}_i)$. From property (iii) of lemma 9, b_{i_j} is a linear function on the correlations $\rho_{Z_i W_{i_1}}, \ldots, \rho_{Z_i W_{i_k}}$, with no constant term. Moreover, correlation $\rho_{Z_i W_{i_l}}$ can be expressed as a sum of terms corresponding to unblocked paths between Z_i and W_{i_l}. Thus, every term in b_{i_j} has the term of an unblocked path between Z_i and some W_{i_l} as a factor. By lemma 12, we get that any such path must include either an edge that does not belong to any of $p_1, \ldots, p_n$, or an edge which appears in some of $p_{i+1}, \ldots, p_n$. As above, in both cases τ and T^* must be distinct terms.

After eliminating all those terms from consideration, the remaining terms in the product of $\langle 1, \ldots, n \rangle$ are given by the expression:

$$T^* \cdot \prod_{i=1}^{n} (1 + \hat{b}_{i_0})$$

Since $\hat{b}_{i_0}$ is a polynomial on the correlations among variables $W_{i_1}, \ldots, W_{i_k}$, with no constant term, it follows that T^* appears only once in this expression. $\square$

Proposition 18 *Term T^* does not appear in the product of any permutation other than*

$\langle 1, \ldots, n \rangle$.

Proof: Let π be a permutation different from $\langle 1, \ldots, n \rangle$, and let τ be a term in the product of π.

Let i be such that, for all $l > i$, π selects the diagonal entry in the row l of Q. As before, for $l > i$, if the factor of τ corresponding to entry q_{ll} does not come from the first term of q_{ll} (i.e., $T(p_l)[1 + \hat{b}_{l_0}]$), then τ must be different from T^*. So, we assume that this is the case.

Assume that π does not select the diagonal entry q_{ii} of Q. Then, π must select some entry q_{il}, with $l < i$. Entry q_{il} can be written as:

$$q_{il} = b_{i_0} a_{i_0 l} + \sum_{j=1}^{k_i} b_{i_j} a_{i_j l}$$

Assume that the factor of τ corresponding to entry q_{il} comes from term $b_{i_0} \cdot a_{i_0 l}$. Recall that each term in $a_{i_0 l}$ corresponds to an unblocked path between Z_i and Y including edge $X_l \to Y$. Thus, in this case, lemma 13 implies that τ and T^* are distinct terms.

Now, assume that the factor of τ corresponding to entry q_{il} comes from term $\sum_{j=1}^{k} b_{i_j} a_{i_j l}$. Then, by the same argument as in the previous proof, terms τ and T^* are distinct. $\square$

Hence, term T^* is not cancelled out and the lemma holds. $\square$

C.2.3 Identification of $\lambda_1, \ldots, \lambda_n$

Lemma 16 gives that $det(Q)$ is a non-trivial polynomial on the parameters of the model. Thus, $det(Q)$ only vanishes on the roots of this polynomial. However, [Oka73] has shown that the set of roots of a polynomial has Lebesgue measure zero. Thus, system Φ has unique solution almost everywhere.

It just remains to show that we can estimate the entries of the matrix of coefficients

of system Φ from data.

Let us examine again an entry q_{il} of matrix Q:

$$q_{il} = \sum_{j=0}^{k} b_{i_j} \cdot a_{i_j l}$$

From condition (iii) of lemma 9, the factors b_{i_j} in the expression above are polynomials on the correlations among the variables $Z_i, W_{i_1}, \ldots, W_{i_k}$, and thus can be estimated from data.

Now, recall that $a_{i_0 l}$ is given by the sum of terms corresponding to each unblocked path between Z_i and Y including edge $X_l \rightarrow Y$. Precisely, for each term t in $a_{i_0 l}$, there is an unblocked path p between Z_i and Y including edge $X_l \rightarrow Y$, such that t is the product of the parameters of the edges along p, except for λ_l.

However, notice that for each unblocked path between Z_i and Y including edge $X_l \rightarrow Y$, we can obtain an unblocked path between Z_i and X_l, by removing edge $X_l \rightarrow Y$. On the other hand, for each unblocked path between Z_i and X_l we can obtain an unblocked path between Z_i and Y, by extending it with edge $X_l \rightarrow Y$.

Thus, factor $a_{i_0 l}$ is nothing else but $\rho_{Z_i X_l}$. It is easy to see that the same argument holds for $a_{i_j l}$ with $j > 0$. Thus, $a_{i_j l} = \rho_{W_{i_j} X_l}, j = 0, \ldots, k$.

Hence, each entry of matrix Q can be estimated from data, and we can solve the system of equations Φ to obtain the parameters $\lambda_1, \ldots, \lambda_n$.

C.3   Proof of Lemma 9

Functions $\phi(1, \ldots, n)$ and $\psi(i_1, \ldots, i_{n-1})$ are defined recursively. For $n = 3$,

$$\begin{cases} \phi^3(1, 2, 3) & = \rho_{12} - \rho_{13}\rho_{23} \\ \psi^2(i_1, i_2) & = \sqrt{(1 - \rho_{i_1, i_2}^2)} \end{cases}$$

For $n > 3$, we have

$$
\begin{cases}
\phi^n(1, \ldots, n) &= \left(\psi^{n-2}(n, 3, \ldots, n-1)\right)^4 \\
&\quad \cdot \phi^{n-1}(1, 2, 3, \ldots, n-1) \\
&\quad - \left(\psi^{n-2}(n, 3, \ldots, n-1)\right)^2 \\
&\quad \cdot \phi^{n-1}(1, n, 3, \ldots, n-1) \\
&\quad \cdot \phi^{n-1}(2, n, 3, \ldots, n-1) \\
\psi^{n-1}(i_1, \ldots, i_{n-1}) &= \Big[\Big(\psi^{n-2}(i_1, i_2, \ldots, i_{n-2}) \\
&\quad \cdot \psi^{n-2}(i_{n-1}, i_2, \ldots, i_{n-2})\Big)^2 \\
&\quad - \left(\phi^{n-1}(i_1, i_{n-1}, i_2, \ldots, i_{n-2})\right)^2\Big]^{\frac{1}{2}}
\end{cases}
$$

Using induction and the recursive definition of $\rho_{12.3\ldots n}$, it is easy to check that:

$$
\rho_{12.3\ldots N} = \frac{\phi^N(1,2,\ldots,N)}{\psi^{N-1}(1,N,3,\ldots,N-1) \cdot \psi^{N-2}(N,3,\ldots,N-1)}
$$

Now, we prove that functions ϕ^n and ψ^{n-1} as defined satisfy the properties $(i) - (iv)$. This is clearly the case for $n = 3$. Now, assume that the properties are satisfied for all $n < N$.

Property (i) follows from the definition of $\phi^N(1, \ldots, N)$ and the assumption that it holds for $\phi^{N-1}(1, \ldots, N-1)$.

Now, $\phi^{N-1}(1, \ldots, N-1)$ is linear on the correlations $\rho_{12}, \ldots, \rho_{N-1,2}$. Since $\phi^{N-1}(2, N, 3, \ldots, N-1)$ is equal to $\phi^{N-1}(N, 2, 3, \ldots, N-1)$, it is linear on the correlations $\rho_{32}, \ldots, \rho_{N,2}$. Thus, $\phi^N(1, \ldots, N)$ is linear on $\rho_{12}, \rho_{32}, \ldots, \rho_{N,2}$, with no constant term, and property (ii) holds.

Terms $(\psi^{N-2}(N, 3, \ldots, N-1))^2$ and $\phi^{N-1}(1, N, 3, \ldots, N-1)$ are polynomials on the correlations among the variables $1, 3, \ldots, N$. Thus, the first part of property (iii) holds. For the second part, note that correlation ρ_{12} only appears in the first term of $\phi^N(1, \ldots, N)$, and by the inductive hypothesis $(\psi^{N-2}(N, 3, \ldots, N-1))^4$ has constant term equal to 1. Also, since $\phi^N(1, 2, 3, \ldots, N) = \phi^N(2, 1, 3, \ldots, N)$ and the later one

93

is linear on the correlations $\rho_{12}, \rho_{13}, \ldots, \rho_{1N}$, we must have that the coefficients of $\phi^N(1, 2, \ldots, N)$ must be linear on these correlations. Hence, property (iv) holds.

Finally, for property (iv), we note that by the inductive hypothesis, the first term of $\left(\psi^{N-2}(N, 3, \ldots, N-1)\right)^2$ has constant term equal to 1, and the second term has no constant term. Thus, property (iv) holds. $\qquad\square$

REFERENCES

[BMW94] P.A. Bekker, A. Merckens, and T.J. Wansbeek. *Identification, equivalent models, and computer algebra*. Academic Press, Boston, 1994.

[Bol89] K.A. Bollen. *Structural Equations with Latent Variables*. John Wiley, New York, 1989.

[BP02] C. Brito and J. Pearl. "A Graphical Criterion for the Identification of Causal Effects in Linear Models." In *Proceedings of the AAAI Conference*, pp. 533–538, Edmonton, 2002.

[BT84] R.J. Bowden and D.A. Turkington. *Instrumental Variables*. Cambridge University Press, Cambridge, England, 1984.

[BW80] P. M. Bentler and D. G. Weeks. "Linear structural equations with latent variables." *Psychometrika*, pp. 289–307, 1980.

[CCK83] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey. *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA, 1983.

[CCR90] T. Cormen, C.Leiserson, and R. Rivest. *Introduction to Algorithms*. The MIT Press, 1990.

[Dun75] O.D. Duncan. *Introduction to Structural Equation Models*. Academic Press, New York, 1975.

[Fis66] F.M. Fisher. *The Identification Problem in Econometrics*. McGraw-Hill, New York, 1966.

[Ger83] V. Geraci. "Errors in variables and the indivicual structure model." *International Economic Review*, 1983.

[HD96] C. Huang and A. Darwiche. "Inference in belief networks: A procedural guide." *International Journal of Approximate Reasoning*, **15**(3):225–263, 1996.

[KKB98] D.A. Kenny, D.A. Kashy, and N. Bolger. "Data Analysis in Social Psychology." In D. Gilbert, S. Fiske, and G. Lindzey, editors, *The Handbook of Social Psychology*, pp. 233–265. McGraw-Hill, Boston, MA, 1998.

[KS80] R. Kindermann and J.L. Snell. *Markov Random Fields and their Applications*. American Mathematical Society, Providence, RI, 1980.

[LS88]   S. L. Lauritzen and D. J. Spiegelhalter. "Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems." *Journal of the Royal Statistical Society (B)*, **50**:157–224, 1988.

[McD97]   R. McDonald. "Haldane's lungs: A case study in path analysis." *Multiv. Behav. Res.*, pp. 1–38, 1997.

[Oka73]   M. Okamoto. "Distinctness of the Eigenvalues of a Quadratic form in a Multivariate Sample." *Annals of Statistics*, **1**(4):763–765, July 1973.

[Pea88]   J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[Pea98]   J. Pearl. "Graphs, causality, and structural equation models." *Sociological Methods and Research*, **27**(2):226–284, 1998.

[Pea00a]   J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

[Pea00b]   J. Pearl. "Parameter identification: A new perspective." Technical Report R-276, UCLA, August 2000.

[Rig95]   E.E. Rigdon. "A necessary and sufficient identification rule for structural models estimated in practice." *Multivariate Behavioral Research*, **30**(3):359–383, 1995.

[SRM98]   P. Spirtes, T. Richardson, C. Meek, R. Scheines, , and C. Glymour. "Using path diagrams as a structural equation modelling tool." *Socioligical Methods and Research*, pp. 182–225, November 1998.

[TEM93]   K. Tambs, L.J. Eaves, T. Moum, J. Holmen, M.C. Neale, S. Naess, and P.G. Lund-Larsen. "Age-specific genetic effects for blood pressure." *Hypertension*, **22**:789–795, 1993.

[WL83]   N. Wermuth and S.L. Lauritzen. "Graphical and recursive models for contingency tables." *Biometrika*, **70**:537–552, 1983.

[Wri34]   S. Wright. "The method of path coefficients." *Ann. Math. Statist.*, **5**:161–215, 1934.