

Algebraic Properties of Linear Structural Equation Models



Bohao Yao
Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

*

Abstract

In this thesis, we explore the problems of identifiability and finding constraints in linear structural equation models (SEMs) and explore the applications of such results.

We start by analysing the effects of interventions on linear SEMs and provide sufficient conditions for when the resulting distribution is identifiable. We then demonstrate how we could use a series of interventions to find constraints in linear SEMs, as long as each resulting distribution is identifiable.

Next, we study algebraic properties of a subset of linear SEMs that can be represented by an *HTC-identifiable graph*. HTC-identifiable graphs are a large family of graphs known to be generically identifiable. We prove that a mixed graphs is HTC-identifiable if and only if all the parameters in the model can be recovered from the covariance matrix using straightforward linear algebra operations. We can then exploit this algebraic property to construct an algorithm that recovers the parameters of the model. We also find a set of polynomials, whose indeterminates are entries of the covariance matrix, that generates the zeroes of the model.

Finally, we prove that the maximum likelihood function is strictly convex in a local neighbourhood around the MLE for HTC-identifiable graphs. In particular, for large sample sizes and assuming we specify the correct model, we find a starting point that is in this convex neighbourhood. We then use hill-climbing to find the MLE of Σ for any HTC-identifiable graphical model G . Using this, we can test whether a particular graphical model G is a good fit for a particular data set.

Contents

1	Introduction	1
1.1	Graphical Models	2
1.1.1	Directed Mixed Graphs	3
1.2	Structural Equation Models	4
1.2.1	Treks	7
1.3	Latent Projection	8
1.4	Identifiability	9
1.5	Constraints	11
1.6	Structure of the Thesis	13
2	Regression Identifiability and Edge Interventions in Linear Structural Equation Models	15
2.1	Introduction	15
2.1.1	Interventions	15
2.1.2	Effects of Edge Interventions	18
2.1.3	Outline of this Chapter	18
2.2	Removing a Directed Edge	19
2.2.1	Effect on Covariance	19
2.2.2	Regression Identifiability	23
2.2.3	Effect on Data	26
2.3	Adding a Directed Edge	27
2.4	Limitations	28
2.4.1	Edge Identifiability	29
2.4.2	Cut Vertices	30
2.5	Usage of Removing Edges	31

2.5.1	Constraints	32
2.5.2	Simulating Interventional Data	34
2.6	Conclusion	35
3	Algebraic Properties of HTC-identifiable Graphs	37
3.1	Introduction	37
3.1.1	HTC-identifiability	37
3.1.2	Constraints	39
3.1.3	Outline of this Chapter	40
3.2	Linear Identifiability	41
3.2.1	Properties of Quasi-linearly Identifiable Graphs	45
3.2.2	Properties of Linearly Identifiable Graphs	47
3.2.3	Connection to HTC-identifiable Graphs	49
3.3	Constraints in Linearly Identifiable Graphs	51
3.3.1	Generators of a Model Ideal	51
3.3.2	Minimality of the Generators	55
3.3.3	Time Complexity	57
3.4	Conclusion	59
4	Maximum Likelihood Estimation for HTC-identifiable Graphs	61
4.1	Introduction	61
4.1.1	Preliminaries	61
4.1.2	Outline of this Chapter	63
4.2	Bootstrap	63
4.2.1	Bootstrap Example	64
4.3	Maximum Likelihood Estimation	65
4.3.1	Starting Point	65
4.3.2	Iterative step	66
4.3.3	Local Concavity of the Log-Likelihood Function	67
4.3.4	Hypothesis testing	69
4.4	Limitations of our Algorithm	70
4.4.1	Point of Unidentifiability	70
4.4.2	Small Sample Size and Eigenvalues	72

4.5	Real Data Example	75
4.6	Conclusion	77
5	Conclusion	78
5.1	Future Work	79
5.1.1	Identifiability	79
5.1.2	Constraints	80
5.1.3	Maximum Likelihood	80
	Appendices	81
A	Commutative Algebra	81
B	Appendices for Chapter 2	84
B.1	Calculations for Example 2.30	84
B.2	Removing a Bidirected Edge	85
B.2.1	Effects on Covariance	85
B.2.2	Regression Identifiability	87
B.2.3	Concerns when Removing Bidirected Edges	88
B.3	Vertex Intervention	89
B.3.1	Effect on Data	90
B.3.2	Effect on Covariance	91
B.3.3	Hard Interventions	92
	Bibliography	93

List of Figures

1.1	A mediation model (left) with its latent projection over 1, 2, 3 (right)	5
1.2	The instrumental variables model	10
1.3	The Verma graph.	12
1.4	The ‘gadget’ graph.	12
2.1	Double treatment model (left) with a vertex intervention (right). . .	16
2.2	(a) A DAG representing time series cross sectional data on two countries where country 2 has a trade agreement with country 1; (b) the DAG in (a) after an intervention is performed, severing the alliance between countries 1 and 2 at $t = 2$	17
2.3	Cut-vertex example	30
2.4	Illustration of each step of edge removal in Example 2.30	32
2.5	The ‘double’ Verma graph.	33
2.6	Graphical model for MACS study	34
3.1	Directed Acyclic Graph Example	40
3.2		42
3.3	A generically identifiable graph that is not HTC-identifiable.	50
3.4	Graph that is not HTC-identifiable but can be decomposed into HTC-identifiable mixed components	51
3.5	An HTC-identifiable graph with cycles.	53
3.6	An HTC-identifiable graph that is not HTC-identifiable after adding the dotted edge.	57
3.7	A complete DAG on 4 vertices.	59
4.1	A simple graph that is not globally identifiable.	70

4.2	$n = 10$ using hill-climbing	72
4.3	$n = 20$ using hill-climbing	72
4.4	$n = 10$ using RICE	72
4.5	$n = 20$ using RICE	72
4.6	The ‘gadget’ graph.	73
4.7	$n = 1000$, $\epsilon = 0.1$	74
4.8	$n = 10^6$, $\epsilon = 0.01$	74
4.9	Skeleton inferred using the FCI algorithm with $\alpha = 0.01$ (left) and $\alpha = 10^{-4}$ (right).	76
4.10	Valid graphical models based on the m -separations.	77
B.1	A graphical model where removing $1 \leftrightarrow 3$ may result in a non- positive definite Σ^*	89

Chapter 1

Introduction

Does smoking cause cancer, or does cancer cause you to smoke more? Or perhaps there is a gene that causes you to smoke more and have cancer at the same time? How much would a vaccines or a mask mandate impact the transmission of the coronavirus? The gold-standard method for answering such questions has traditionally been randomised control trials. However, in many cases, it is not feasible to conduct a completely randomised trial; for instance, it is not feasible to get a whole population to be. For a long time, statisticians believed that we could not make causal inferences with only observational data. This changed recently with the potential outcome language, originally introduced by [Neyman \(1923\)](#) and later expanded by [Rubin \(1974\)](#), and causal graphical models, pioneered by [Wright \(1921\)](#) and later formalised by [Pearl \(2000\)](#) and [Spirtes et al. \(1993\)](#).

In a causal graphical model, we represent each random variable as a vertex and each ‘direct effect’ as an edge. The resulting graphical model is a *directed acyclic graph* (DAG). In practice, it is generally not reasonable to assume that all relevant variables have been measured. To deal with DAGs containing hidden variables, it is common to apply a *latent projection* operation to produce a new family of graphs with only the observable variables ([Pearl and Verma, 1992](#)). This new family of graphs are known as the *acyclic directed mixed graphs* (ADMGs). We will give a rigorous definition of this operation in [Definition 1.3](#) and an example in [Figure 1.1](#).

The two causal frameworks, potential outcomes and graphical models, were later unified by [Richardson and Robins \(2013\)](#) who introduced the Single-World

Intervention Graph (SWIG).

1.1 Graphical Models

Graphical models provide a powerful formalism for dealing with uncertainty, by encoding the conditional independence of probability distributions into graph representations. In particular, random variables are represented as vertices of the graph while conditional dependencies are represented by edges. The relation between graphs and probability distributions is usually referred to as the concept of a *Markov property* (Pearl, 1988), which given any graph, determines a set of conditional independence relations satisfied by the underlying probability distribution. These graphs give us an intuitive way of visualising conditional independence, and are also natural data structures for modern computers, allowing us to implement efficient algorithms for inference. Their simplicity and effectiveness provide graphical models with universal applicability.

Originally, only graphs consisting of a single type of edge were used. The origins of graphical models can be traced back to the works by Gibbs (1902) on statistical physics for undirected graphs. The use of graphs as a language for describing causal models then began with the seminal works on genetics by Wright (1921, 1934), who represented the joint probability distribution by a DAG with all the edges directed ($\rightarrow$). For instance, heritable properties of natural species can be represented by an arrow from parent to child in a graph. The idea of “path analysis” introduced in Wright’s work was later taken up in other fields such as econometrics by Haavelmo (1943) and social sciences by Blalock (1971). Eventually, several examples in social sciences required graphs with multiple edge types to be modelled (Lauritzen and Vermuth, 1989).

In this thesis, we shall study the class of *directed mixed graphs*, which contain both directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges. While the use of DAGs is popular in many fields in decision analysis (Olmsted, 1984) and empirical sciences (Pearl, 1988), in many contexts, it is unrealistic to assume that there are no unmeasured confounding variables. However, DAG models with latent variables lack many of the desirable properties of graphical models, such as not being curved exponential families (Geiger et al., 2001). Furthermore, any inference made on these models

may be very sensitive to assumptions about the unobserved variables (Settimi and Smith, 1998). This motivates the study of *directed mixed graphs*, a graph class obtained from directed graphs by removing latent variables and replacing them with bidirected edges via *latent projection* (Verma and Pearl, 1990). Building on earlier works of Pearl (1988) and Verma and Pearl (1990), the problems faced by DAGs with latent variables were partially addressed by Richardson and Spirtes (2002) using ancestral graphs, Wermuth (2011) using summary graphs and Richardson et al. (2017) using nested models.

1.1.1 Directed Mixed Graphs

A *directed mixed graph* is a triple $G = (V, \mathcal{D}, \mathcal{B})$ where V is a set of vertices, $\mathcal{D}$ is the set of directed edges ($\rightarrow$) and $\mathcal{B}$ is the set of bidirected edges ($\leftrightarrow$). Edges in $\mathcal{D}$ are oriented whereas edges in $\mathcal{B}$ have no orientation. A *loop* is an edge joining a vertex to itself. Throughout this thesis, we will only consider graphs without any loops, that is $i \rightarrow i \notin \mathcal{D}$ and $i \leftrightarrow i \notin \mathcal{B}$.

A *walk* of length ℓ is a sequence of $\ell + 1$ adjacent vertices $v_0, \dots, v_\ell$. A *path* is a walk where all vertices are distinct. A *directed path* of length ℓ is a sequence of $\ell + 1$ distinct adjacent vertices $v_0, \dots, v_\ell$, each connected by a directed edge $v_i \rightarrow v_{i+1}$. Hence, a directed path is of the form $v_0 \rightarrow v_1 \rightarrow \dots \rightarrow v_\ell$. Similarly, a *bidirected path* of length ℓ is a sequence of distinct $\ell + 1$ adjacent vertices $v_0, \dots, v_\ell$, each connected by a directed edge $v_i \rightarrow v_{i+1}$, and is of the form $v_0 \leftrightarrow v_1 \leftrightarrow \dots \leftrightarrow v_\ell$. A directed walk of length $\ell \geq 3$ is a *directed cycle* if $v_0 = v_\ell$ and all other vertices v_j are distinct (for $0 < j < \ell$). An *acyclic directed mixed graph* (ADMG) is a directed mixed graph without any directed cycles. A *directed acyclic graph* (DAG) is an ADMG without any bidirected edges. A *topological ordering* of the vertices in a graph is an enumeration of the vertex set $V = \{v_1, \dots, v_k\}$ with $k = |V|$ such that $v_i \in \text{pa}(v_j)$ if and only if $i < j$. This ordering is not unique in general. In an acyclic graph, a topological ordering always exists.

Suppose $x, y \in V$. If $x \rightarrow y$, we say x is a *parent* of y and y is a *child* of x . If either $x = y$ or there is a directed path from x to y , we say that x is an *ancestor* of y and y is a *descendant* of x . Otherwise, we say that y is a *non-descendant* of x . The sets of all parents, children, ancestors, descendants and non-descendants

of x are denoted $\text{pa}(x)$, $\text{ch}(x)$, $\text{an}(x)$, $\text{de}(x)$ and $\text{nd}(x)$ respectively. We can also extend these definitions to a vertex set $S \subseteq V$, e.g. $\text{an}(S) = \bigcup_{v \in S} \text{an}(v)$.

Definition 1.1. Let $i, j \in V$. If there is both a directed and a bidirected edge connecting i and j , we say that i and j form a *bow*. If a mixed graph has no bows, we say that the graph is *bow-free*.

Given a path π , a vertex v is a *collider on π* if two arrowheads of π meet head to head at v (i.e. if π contains a subpath of the form: $\rightarrow v \leftarrow$, $\leftrightarrow v \leftarrow$, $\rightarrow v \leftrightarrow$ or $\leftrightarrow v \leftrightarrow$). Otherwise, v is a *non-collider on π* .

Definition 1.2. Let $S \subseteq V$ be a vertex set. A path π between vertices u and v is said to be *open given S* if every non-collider on π is not in S and every collider on π is an ancestor of S in G . Otherwise, we say the path π is *blocked given S* .

A graph, $G' = (V', E')$, is a *subgraph* of $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$. If E' contains all the edges of G that has both endpoints in V' , then G' is called an *induced subgraph* of G . We denote the induced subgraph on the vertex set $W \subseteq V$ by G_W .

A set of vertices $S \subseteq V$ is *bidirected-connected* if for every $v, w \in S$, there is a bidirected path from v to w in G_S . A maximal bidirected-connected set of vertices is called a *district*. For $v \in V$, let $\text{dis}(v)$ denote the district containing v in G . A vertex v is *fixable* if $\text{de}(v) \cap \text{dis}(v) = \{v\}$ (Richardson et al., 2017). Fixable vertices are crucial as we are always able to identify the causal effect of v on any $d \in \text{de}(v)$. The *Markov blanket* of v is

$$\text{mb}(v) := \text{pa}(\text{dis}(v)) \cup (\text{dis}(v) \setminus \{v\}).$$

We write $\text{dis}_S(v)$ for the district containing v in G_S . Similarly, we write $\text{mb}_S(v)$ for the Markov blanket of v in G_S .

1.2 Structural Equation Models

An important parametric subclass of graphical models are the linear *structural equation models* (SEMs), often used in causal inference (Spirites et al., 1993). In fact, the work by Wright (1921) was originally within the SEM class. In the

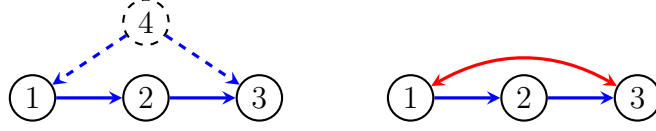


Figure 1.1: A mediation model (left) with its latent projection over 1, 2, 3 (right)

present day, SEMs are applied in subject fields as varied as epidemiology, ecology, behavioural sciences, social sciences and economics ([Grace, 2006](#); [Hershberger, 2003](#); [Rothman and Greenland, 2005](#); [Sobel, 2000](#)). The reader is referred to [Bollen \(1989\)](#); [Pearl \(2000\)](#); [Spirtes et al. \(1993\)](#) for a more thorough background on SEMs.

Given a mixed graph $G = (V, \mathcal{D}, \mathcal{B})$ and the corresponding random variables $\{X_i \mid i \in V\}$, a linear SEM is defined as

$$X_i = \sum_{j \in \text{pa}(i)} \lambda_{ji} X_j + \epsilon_i, \quad i \in V, \quad (1.1)$$

where $\text{pa}(i)$ represents the set of parents of the vertex i , each λ_{ji} is the edge coefficient of $j \rightarrow i$ and $\epsilon = (\epsilon_i)$ is a multivariate Gaussian vector with $\omega_{ij} := \text{Cov}(\epsilon_i, \epsilon_j) \neq 0$ only if $i \leftrightarrow j \in \mathcal{B}$ or $i = j$.

Consider the following example given by [Pearl \(2000\)](#).

Example 1.1. Consider the mediation model in Figure 1.1 where X_4 is a latent variable. We can represent the variables as

- X_1 - Smoking,
- X_2 - Tar in lungs,
- X_3 - Lung cancer,
- X_4 - Genes.

The latent projection on the right can be represented as the following SEM,

$$\begin{aligned} X_1 &= \epsilon_1, \\ X_2 &= \lambda_{12}X_1 + \epsilon_2, \\ X_3 &= \lambda_{23}X_2 + \epsilon_3, \\ \text{Cov}(\epsilon_1, \epsilon_2) &= \text{Cov}(\epsilon_2, \epsilon_3) = 0, \\ \text{Cov}(\epsilon_1, \epsilon_3) &= \omega_{13}. \end{aligned}$$

While SEMs originated from the works of Wright (1921), they have since been developed across several disciplines by academics such as Duncan (1966), Jöreskog (1970), Blalock (1971) and Goldberger (1972). A background of SEMs is provided by Duncan (1975) and Bollen (1989). In the present day, applications of SEMs are published in journals in subject fields as varied as behaviour sciences, social sciences and economics. Examples of such journals include *Econometrica*, *Psychometrika*, *Sociological Methods and Research* and *Multivariate Behavioral Research*. A summary of the growth of SEMs can be found in Hershberger (2003).

Let $\Lambda = (\lambda_{ij})$ be the matrix holding the edge coefficients, we can rewrite (1.1) in matrix form

$$X = \Lambda^T X + \epsilon.$$

We have now constructed a structural equation model from a directed mixed graph where $\lambda_{ij} = 0$ if the edge $i \rightarrow j \notin \mathcal{D}$. This gives us a natural definition for a set containing all possible Λ ,

$$\mathbb{R}^{\mathcal{D}} := \{\Lambda \in \mathbb{R}^{|V| \times |V|} : \lambda_{ij} = 0 \text{ if } i \rightarrow j \notin \mathcal{D}\}.$$

Let $\mathbb{R}_{\text{reg}}^{\mathcal{D}}$ be the subset of matrices $\Lambda \in \mathbb{R}^{\mathcal{D}}$ for which $(I - \Lambda)$ is invertible. Note that if $G = (V, \mathcal{D}, \mathcal{B})$ is an ADMG, then given any topological ordering of V , Λ is strictly upper triangular, hence, $(I - \Lambda)$ is invertible. Working on this subset, we obtain a unique solution to the structural equations:

$$X = (I - \Lambda^T)^{-1} \epsilon.$$

Let $\Omega = (\omega_{ij}) = \text{Cov}[\epsilon] \in \mathbb{R}^{|V| \times |V|}$ be a covariance matrix of ϵ with $\omega_{ij} = 0$ if $i \leftrightarrow j \notin \mathcal{B}$. Similarly, we can define a set containing all possible Ω ,

$$PD(\mathcal{B}) := \{\Omega \in PD_V : \Omega \text{ symmetric, } \omega_{ij} = 0 \text{ if } i \neq j \text{ and } i \leftrightarrow j \notin \mathcal{B}\},$$

where PD_V is the cone of positive definite $|V| \times |V|$ matrices. Hence, X has the covariance matrix

$$\Sigma := \text{Cov}[X] = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}, \quad (1.2)$$

where A^{-T} denote the inverse transpose of A . By Cramer's rule ([Artin, 1991](#)), entries in the covariance matrix in (1.2) are rational functions of the entries in Λ and Ω . We define the linear structural equation model given by a directed mixed graph $G = (V, \mathcal{D}, \mathcal{B})$ to be the family of all multivariate normal distributions on $\mathbb{R}^{|V|}$ with the covariance matrix in the set

$$\mathcal{M}(G) := \{(I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1} : \Lambda \in \mathbb{R}_{\text{reg}}^{\mathcal{D}}, \Omega \in PD(\mathcal{B})\}.$$

1.2.1 Treks

A *trek* from i to j is a walk from i to j without any colliders. Therefore, all treks are of the form

$$v_\ell^L \leftarrow v_{\ell-1}^L \leftarrow \cdots \leftarrow v_1^L \leftarrow v_0^L \leftrightarrow v_0^R \rightarrow v_1^R \rightarrow \cdots \rightarrow v_{r-1}^R \rightarrow v_r^R$$

or

$$v_\ell^L \leftarrow v_{\ell-1}^L \leftarrow \cdots \leftarrow v_1^L \leftarrow v_0^{LR} \rightarrow v_1^R \rightarrow \cdots \rightarrow v_{r-1}^R \rightarrow v_r^R,$$

where $v_\ell^L = i$, $v_r^R = j$. Note that paths on either side may have length zero, so a trek may just be a directed path or a single vertex. For a trek π with no bidirected edges and a source i , we define the trek monomial as

$$\pi(\Lambda, \Omega) = \omega_{ii} \prod_{x \rightarrow y \in \pi} \lambda_{xy}.$$

For a trek π with a bidirected edge $i \leftrightarrow j$, we define the trek monomial as

$$\pi(\Lambda, \Omega) = \omega_{ij} \prod_{x \rightarrow y \in \pi} \lambda_{xy}.$$

Theorem 1.2 (Trek rule). *Let $\mathcal{T}_{vw}$ be the set of all treks from v to w . The covariance matrix Σ that is Markov with respect to a mixed graph G is given by*

$$\sigma_{vw} = \sum_{\pi \in \mathcal{T}_{vw}} \pi(\lambda, \omega). \quad (1.3)$$

Proof. The proof originated from [Wright \(1934\)](#), and is obtained from expanding out the matrix in (1.2) using Taylor series. If G is acyclic, the RHS of (1.2) and (1.3) agrees.

If G is not acyclic, the RHS of (1.2) is a formal power series. Writing $(I - \Lambda)^{-1} = I + \Lambda + \Lambda^2 + \dots$, the RHS of (1.2) and (1.3) again agrees. $\square$

1.3 Latent Projection

Let G be a DAG or a mixed graph with latent variables. The *latent projection* ([Pearl and Verma, 1992](#)) of G is a new mixed graph whose vertices are the observed variables of G .

Definition 1.3. Let G be a mixed graph with vertex set $V \sqcup L$, where V is the set of observable variables, L is the set of latent variables and $\sqcup$ indicates a disjoint union. The *latent projection* of G is a new mixed graph $G^* = (V, \mathcal{D}, \mathcal{B})$, which for each vertex $v, w \in V$,

- there is an edge $v \rightarrow w \in \mathcal{D}$ if and only if there exists a directed path from v to w for which all intermediate vertices are in L , and
- there is an edge $v \leftrightarrow w \in \mathcal{B}$ if and only if there exists a trek, that is not a directed path, from v to w for which all intermediate vertices are in L .

Note that from Definition 1.3, it is implied that the latent projection of G inherits the induced subgraph G_V , with some additional edges added. We have provided an example of latent projection in Figures 1.1, with the DAG with a latent variable on the left and its latent projection on the right.

By introducing latent projection, we simplify the graphical representation by removing latent variables and replacing them with bidirected edges. For example, when performing a search for Markov equivalence classes, we need only consider a finite set of mixed graphs instead of an infinite class of DAGs or mixed graphs with arbitrary number of latent variables.

1.4 Identifiability

A model is *identifiable* if all of its parameters can be uniquely determined from the probability distribution over the model variables. If a model is not identifiable, then its parameters could not be estimated from data as there are multiple (often infinitely many) values for the parameters that are compatible with a given dataset. In linear SEMs, parameters are generally identified from the covariance matrix, Σ . For instance, the parameters of the mediation model on the right of Figure 1.1 are identified with

$$\lambda_{12} = \frac{\sigma_{12}}{\sigma_{11}}, \quad \lambda_{23} = \frac{\sigma_{23 \cdot 1}}{\sigma_{22 \cdot 1}}, \quad \omega_{13} = \frac{\sigma_{13 \cdot 2}}{\sigma_{11 \cdot 2}}.$$

The properties of identifiability in linear SEMs is a topic with a long history. A review of the classical conditions of these properties, which do not take into account the graphical structure, is given by [Bollen \(1989\)](#). We shall start with the definition of *global identifiability*.

Definition 1.4. Let $\phi : \Theta \rightarrow N$ be a rational map defined everywhere on the parameter space $\Theta \subseteq \mathbb{R}^{|V|}$ into the natural parameter space N of an exponential family. The model $\mathcal{M} = \text{im } \phi$ is said to be *globally identifiable* if ϕ is a one-to-one map on Θ .

Sufficient conditions for mixed graphs to induce globally identifiable linear SEMs are given by [McDonald \(2002\)](#) and [Richardson and Spirtes \(2002\)](#). Both of these works assume that the presense of a bidirected edge $i \leftrightarrow j$ implies the absence of a directed path from i to j . Graphs with this property were later commonly referred to as the ancestral graphs. Later, [Drton et al. \(2011\)](#) found the necessary and sufficient conditions for a mixed graph to induce a globally identifiable linear SEM.

In standard literature, however, ‘identifiable’ usually referred to a weaker identifiability requirement known as *almost-everywhere* identifiability or *generic identifiability*.

Definition 1.5. Let $\phi : \Theta \rightarrow N$ be a rational map defined everywhere on the parameter space Θ into the natural parameter space N of an exponential family.

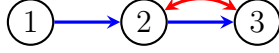


Figure 1.2: The instrumental variables model

The model $\mathcal{M} = \text{im } \phi$ is said to be *generically identifiable* if $\phi^{-1}(\phi(\theta)) = \{\theta\}$ for almost all $\theta \in \Theta$ with respect to the Lebesgue measure.

As we saw in (1.2), for linear SEMs, the model $\mathcal{M}(G)$ is *generically identifiable* if the map

$$\phi : (\Lambda, \Omega) \mapsto (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}$$

is injective almost everywhere on $\Theta := \mathbb{R}_{\text{reg}}^{\mathcal{D}} \times PD(\mathcal{B})$. If at most k parameter values are mapped by ϕ to the same Σ almost everywhere on Θ , then we say that $\mathcal{M}(G)$ is *k-identifiable*.

From the definitions, we see that all globally identifiable models are also generically identifiable. However, not all generically identifiable models are also globally identifiable as we shall see in the following example from Drton (2018, Example 3.1).

Example 1.3. Consider the instrumental variable model in Figure 1.2. We can recover Λ from the observed Σ using

$$\lambda_{12} = \frac{\sigma_{12}}{\sigma_{11}}, \quad \lambda_{23} = \frac{\sigma_{13}}{\sigma_{12}}.$$

Note that the first denominator is always positive since Σ is positive definite and the second denominator is zero if and only if $\lambda_{12} = 0$. In particular, the map ϕ is injective precisely on the set $\{\Lambda \times \Omega \mid \Lambda \in \mathbb{R}_{\text{reg}}^{\mathcal{D}}, \Omega \in PD(\mathcal{B}), \lambda_{12} \neq 0\}$. Therefore, while the instrumental variable model is generically identifiable, it is not globally identifiable as it is not injective on the whole of Θ .

In particular, for any algebraic variety $V \subset \Theta$ over the reals, if $V \neq \Theta$, we have $\dim(V) < \dim(\Theta)$ (Cox et al., 2015) and hence, V has measure zero with respect to the Lebesgue measure on Θ (Okamoto, 1973).

The first class of graphs that were proven to be generically identifiable is the bow-free DAGs by Brito and Pearl (2002a). The concept of the instrumental variable model in Example 1.3 were later generalised by Brito and Pearl (2002b),

who introduced the concept of an *instrumental set*. These resulted in new sufficient graphical conditions for some of the relevant parameters in a mixed graph to be generically identifiable. Instrumental sets demonstrate that algebraic properties of linear independence can be translated into graphical properties that can be used to identify parameters in a model. This method was expanded later by [Brito and Pearl \(2006\)](#) with the introduction of the G-Criterion algorithm, which provides a method to identify the parameters of a linear acyclic SEM. [Foygel et al. \(2012\)](#) later generalised this algorithm with the introduction of a new identifiability criterion known as HTC-identifiability, which we shall provide a formal definition for in Chapter 3. In fact, they proved that all HTC-identifiable graphs are also generically identifiable. This algorithm was further generalised by [Chen et al. \(2014\)](#) with the goal of finding constraints (see Section 1.5) in linear SEMs. We will further study the identifiability properties of HTC-identifiable graphs in Chapter 3.

While the algorithms by [Brito and Pearl \(2006\)](#), [Foygel et al. \(2012\)](#) and [Chen et al. \(2014\)](#) identify the parameters in the model by searching for linear independences, a different approach was provided by [Tian and Pearl \(2002a\)](#); [Tian \(2005, 2007, 2009\)](#), who identified the model parameters by decomposing the structural equations into orthogonal partial regression equations. In Tian’s papers, new necessary and sufficient conditions for a mixed graph to be globally identifiable were provided for non-parametric models. The results of [Tian and Pearl \(2002a\)](#) were also built upon by [Shpitser and Pearl \(2006\)](#), who provided a complete identification algorithm for non-parametric models. Identification by regression were also studied by [Tian \(2005\)](#). We refer to this as *regression identifiability* in Chapter 2. Furthermore, we shall provide the necessary and sufficient graphical criteria for a parameter (λ or ω) to be regression identifiable in Chapter 2.

1.5 Constraints

Consider two graphs G and G' . If both G and G' are *directed acyclic graphs* (DAGs), mixed graphs without bidirected edges and directed cycles, then $\mathcal{M}(G) = \mathcal{M}(G')$ if and only if G and G' have the same d-separation relations ([Geiger et al.](#),

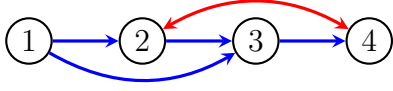


Figure 1.3: The Verma graph.

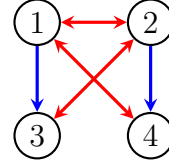


Figure 1.4: The ‘gadget’ graph.

1990). However, this is only helpful in situations where all the variables are observed. If G is a DAG with hidden variables, we can use the latent projection operation (Pearl and Verma, 1992) to transform G into an *acyclic directed mixed graph* (ADMG). In this case, the conditional independence relations on the observed vertices are no longer sufficient to describe $\mathcal{M}(G)$ in general. The ideal generating these algebraic relations with equality constraints is often referred to as the *vanishing ideal*,

$$\mathcal{I}(G) = \langle f \in \mathbb{R}[\Sigma] \mid f(\Sigma) = 0 ; \forall \Sigma \in \mathcal{M}(G) \rangle.$$

Example 1.4. Consider the Verma graph in Figure 1.3. While there are no conditional independences involving only the observed variables X_1, X_2, X_3 and X_4 . However, we do have an equality constraint on the corresponding covariance matrix Σ , in the sense that $\Sigma \in \mathcal{M}(G)$ only if

$$\begin{aligned} f_{\text{Verma}}(\Sigma) = & \sigma_{11}\sigma_{13}\sigma_{22}\sigma_{34} - \sigma_{12}^2\sigma_{13}\sigma_{34} - \sigma_{11}\sigma_{14}\sigma_{22}\sigma_{33} + \sigma_{12}^2\sigma_{14}\sigma_{33} \\ & - \sigma_{11}\sigma_{13}\sigma_{23}\sigma_{24} + \sigma_{11}\sigma_{14}\sigma_{23}^2 + \sigma_{12}\sigma_{13}^2\sigma_{24} - \sigma_{12}\sigma_{13}\sigma_{14}\sigma_{23} = 0. \end{aligned}$$

The graph decomposition result by Tian and Pearl (2002a) that leads to Tian’s algorithm (Richardson et al., 2017; Tian and Pearl, 2002b) provides a method to find constraints in non-parametric graphical models. For instance, the constraint in Example 1.4 can be seen as the independence between X_1 and X_4 after fixing X_2 and X_3 (i.e. changing random vertices 2 and 3 into a fixed vertex by removing incoming all edges into those vertices (Richardson et al., 2017)). Though non-parametrically complete, Tian’s algorithm fails to find the following constraint on the ‘gadget’ graph in Figure 1.4:

$$\sigma_{11}\sigma_{22}\sigma_{34} - \sigma_{13}\sigma_{14}\sigma_{22} + \sigma_{13}\sigma_{12}\sigma_{24} - \sigma_{23}\sigma_{11}\sigma_{24} = 0.$$

In parametric models, the vanishing ideal $\mathcal{I}(G)$ of a model corresponding to an arbitrary mixed graph G is traditionally found using Gröbner bases (Cox et al., 2015). However, this is computationally expensive. Various other attempts, such as that of Drton et al. (2020), aim to find a more efficient algorithm for computing $\mathcal{I}(G)$. However, finding such an algorithm still remains an open problem.

In this thesis, we will explore multiple algorithms to finding constraints. In Chapter 2, we will provide an algorithm for finding constraints by removing all the directed edges. However, this algorithm requires each intermediate interventional distribution to be identifiable, for which we will provide some sufficient graphical conditions. In Chapter 3, we provide a set of polynomials that generates the zeroes of the HTC-identifiable graphical model intersected with the cone of positive definite matrices. In Chapter 4, we will test if a particular dataset satisfies the constraints implied by a HTC-identifiable graph using the bootstrap and a likelihood ratio test.

1.6 Structure of the Thesis

This thesis is structured as follows:

In Chapter 2, we shall first introduce a new identifiability criteria for linear structural equation models, which we call *regression identifiability*. We will then provide necessary and sufficient graphical conditions for a directed edge to be regression identifiable. Suppose Σ^* corresponds to the covariance matrix of the graphical model G^* obtained by performing an edge intervention to G with corresponding covariance matrix Σ . Using regression identifiability, we obtain necessary graphical conditions for Σ^* to be identifiable given Σ . We also identify what would happen to an individual data point if there were such an intervention. Finally, we provide some statistical problems where our methods could be used, such as finding constraints and simulating interventional data from observational data.

In Chapter 3, we shall introduce a graph class, known as the “HTC-identifiable” graph, first defined by Foygel et al. (2012). We will explore the algebraic properties of linear structural equation models that can be represented by an HTC-identifiable graph. In particular, we prove that all mixed graphs are HTC-identifiable if and only if all the regression coefficients can be recovered from the covariance matrix

using straightforward linear algebra operations on specified equations. By exploiting these algebraic properties, we will provide an efficient algorithm to recover the parameters, (Λ, Ω) , of any HTC-identifiable graph. Furthermore, given an HTC-identifiable graph, we show that the set of polynomials provided by [van Ommen and Mooij \(2017\)](#) in fact generates the *model ideal*, the ideal that generates the zeroes of the corresponding graphical model. We further prove that, for a subset of HTC-identifiable graphs, this set of polynomials are the minimal generators of a model ideal. Complexity of our algorithm will also be discussed.

In Chapter 4, we will consider the problem that given a dataset $X \stackrel{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma)$, of testing $H_0 : \Sigma \in \mathcal{M}(G)$ against the general hypothesis $H_1 : \Sigma \in PD_V$, where G is a linear structural equation model represented by a HTC-identifiable graph. For the case where the generators of a model ideal is given, we will show that this could be solved using bootstrapping. However, if the generators of a model ideal are unknown, we will expand on our algorithm in Chapter 3 to produce an algorithm for fitting a HTC-identifiable graph using maximum likelihood. We will first provide a region around the MLE for which the log-likelihood function is locally concave for a constrained covariance matrix Σ under the assumption that we are working with a linear SEM. We will then show how a hill-climbing algorithm can be implemented with the starting point found by our algorithm in Chapter 3. Furthermore, we discuss some limitations of this algorithm, such as fitting graphical models at a point of unidentifiability, having small samples or when the distribution has a small eigenvalue. Finally, we will use likelihood ratio test to decide whether we should reject H_0 .

In Chapter 5, we will summarise our results in this thesis. We also provide a few open problems that extends our results, which may provide a basis for future works in this subject area.

Chapter 2

Regression Identifiability and Edge Interventions in Linear Structural Equation Models

2.1 Introduction

Questions relating to the effects of interventions are central to various topics in scientific research and policy making. For instance, which therapy would be the most effective for treating a particular patient from lung cancer? Would enforcing a mask mandate significantly reduce the spread of Covid?

2.1.1 Interventions

Traditionally, the main focus for the majority of causal inference literature has been on variable interventions. Utilising the frameworks established by [Neyman \(1923\)](#), [Rubin \(1974\)](#) and [Pearl \(2000\)](#), variable intervention has been applied to various fields in empirical and social sciences. The interested reader is referred to works by [Halpern and Pearl \(2005\)](#); [Tian and Pearl \(2001\)](#); [Woodward \(2001\)](#) for an incomplete survey.

In the past two decades, there has been a shift towards a more general notion of intervention, with [Korb et al. \(2004\)](#) proposing several generalisations of interventions. Similar ideas were introduced by [Eberhardt and Scheines \(2007\)](#), where the shortcomings for each of the proposed interventions were discussed. [Malinsky \(2018\)](#) extended the interventionist framework to enable the exploration of the

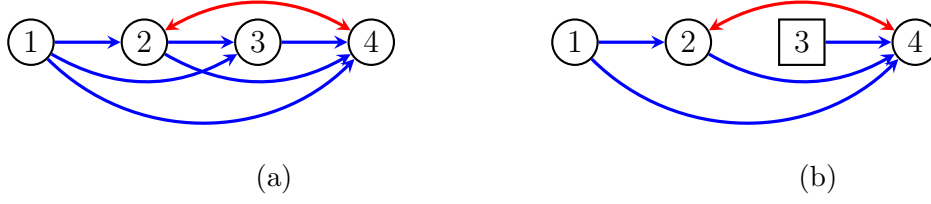


Figure 2.1: Double treatment model (left) with a vertex intervention (right).

consequences of intervention on a “macro-level”. A type of edge intervention in social networks has been proposed by [Ogburn et al. \(2017\)](#) to study changes in network ties. Furthermore, [Shpitser and Tchetgen Tchetgen \(2016\)](#) considered edge intervention in the context of mediation analyses. In addition, [Robins and Richardson \(2010\)](#) proposed hypothetical randomised trials that are essentially a form of edge intervention. These ideas were later formalised as conditional separable effects by [Stensrud et al. \(2022\)](#). Most recently, [Sherman and Shpitser \(2020\)](#) evaluated general intervention on network ties, which allows us to envision the counterfactual world where an edge is removed or added. The works by [Shpitser and Tchetgen Tchetgen \(2016\)](#) and [Sherman and Shpitser \(2020\)](#) are mainly studies of non-parametric models.

The difference between edge and vertex interventions are highlighted in the example below.

Example 2.1. Consider a dynamic treatment model introduced in [Robins \(1986\)](#) with outcome X_4 and two treatment variables X_1 and X_3 , where the second treatment is dependent on both the first treatment and an intermediate outcome X_2 . The final outcome X_4 is dependent on the two treatments, the intermediate outcome, and an unobserved confounder between the two outcomes. The corresponding graphical model is shown in Figure 2.1a.

- **Vertex intervention:** Suppose we intervene on X_3 in order to provide the second treatment to all the patients. This is graphically equivalent replacing the random variable X_3 with a fixed variable and removing all incoming edges into the vertex 3 as shown in Figure 2.1b.

Note that not all interventional distributions are generically identifiable (e.g. if we intervene on the intermediate outcome X_2) as not all vertices are *fixable*.

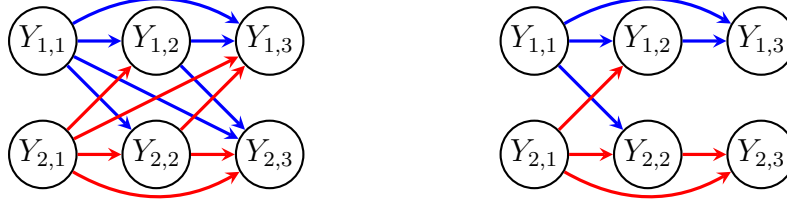


Figure 2.2: (a) A DAG representing time series cross sectional data on two countries where country 2 has a trade agreement with country 1; (b) the DAG in (a) after an intervention is performed, severing the alliance between countries 1 and 2 at $t = 2$.

- **Edge intervention:** Edge interventions involve the removal or the change in strength of direct causal effects. For instance, we may want to change the strength of the direct causal effect of X_2 on X_3 due to a shortage or surplus of available treatments.

We now provide an example from the social sciences where edge interventions could be applied.

Example 2.2. Consider Figure 2.2, a model of trade relations between countries by Sherman and Shpitser (2020). Each country i is represented by temporally sequential observations $\{Y_{i,t} \mid t \in \{1, 2, \dots\}\}$ where each Y is a vector of economic variables (e.g. GDP). We could consider a move from Figure 2.2 (a) to (b) by a severance of trade relations between two countries. Similarly, we could consider a reverse intervention of extending a trade agreement between countries, moving from Figure 2.2 (b) to (a). Once again, we see that these interventions could not simply be performed by an intervention of the variables $Y_{1,3}$ and $Y_{2,3}$ as these will remove all incoming arrows into $Y_{1,3}$ and $Y_{2,3}$.

In this chapter, we will study edge interventions in the parametric case where the model is a linear structural equation model (SEM). A brief introduction of SEMs could be found in Section 1.2. In particular, we will find sufficient graphical conditions for a distribution to be identifiable after an edge intervention by introducing a stronger identifiability condition which we call *regression identifiability*. In particular, if all the parameters in Λ and Ω are regression identifiable, then the graphical model is also globally identifiable.

2.1.2 Effects of Edge Interventions

In addition to studying when the resulting interventional distribution is identifiable, finding the effect of an intervention on a data point has important practical applications. In Example 2.2, an economic advisor could assess the effects of a particular foreign policy or conflict, such as the war in Ukraine, will have on the economy. In the medicinal sciences, questions such as “if a patient has died due to a heart transplant, would he be alive if the heart transplant had not happened?” are also crucial for recommending treatments.

Popular methods used to estimate the total causal effect of a particular treatment on the outcome include adjusting for confounding (Shpitser et al., 2010; Perkovic et al., 2018) and inverse probability weighting (Robins et al., 2000). In the medical literature, common definitions of causal effect are the controlled direct effect (Robins and Greenland, 1992), the principal stratum effect (Frangakis and Rubin, 2002) and the conditional separable effect (Stensrud et al., 2022). In particular, conditional separable effect could be interpreted as the effect of an edge intervention.

In this chapter, we study how an edge intervention would affect individual data points under the assumption that the error terms are stable. This is possible due to the linearity of error terms.

2.1.3 Outline of this Chapter

In this chapter, we will study the effects of edge interventions in Gaussian ADMGs. In particular, we shall study how adding or removing a directed edge will affect the distribution of the model and provide sufficient graphical conditions to identify the new distribution and (under additional assumptions) the counterfactual for individual-level data.

This chapter is structured as follows: In Section 2.2, we will study the effects of removing a directed edge. In particular, we provide sufficient algebraic and graphical conditions for Σ^* to be identifiable given Σ . To achieve the graphical conditions, we will introduce a stronger identifiability condition which we call *regression identifiability*. Necessary and sufficient conditions for an edge and a

directed path to be regression identifiable are also given. We will also identify the effects of removing an edge on individual data points.

In Section 2.3, we will consider the reverse and study the effects of adding a directed edge. Once again, we provide the sufficient algebraic conditions and graphical conditions for Σ^* to be identifiable given Σ . We will also identify the effects of adding an edge on individual data points. In Section 2.4, we will discuss the limitations of our graphical conditions. Finally, in Section 2.5, we will provide some statistical problems, finding constraints and simulating interventional data from observation data, where edge interventions could be used.

2.2 Removing a Directed Edge

Suppose we have an ADMG $G = (V, \mathcal{D}, \mathcal{B})$ with a corresponding distribution $N(0, \Sigma)$. Suppose we want to remove some directed edge $a^* \rightarrow b$. Let $A = \text{nd}(b)$, $B = \{b\}$ and $C = \text{de}(b) \setminus \{b\}$, so that $A \cup B \cup C = V$. Throughout this section, we suppose that after removing the directed edge $a^* \rightarrow b$, we obtain a new ADMG G^* with a corresponding distribution $N(0, \Sigma^*)$.

2.2.1 Effect on Covariance

First, we shall find an expression for each entry of Σ^* .

Lemma 2.3. *Every trek from some vertex $a \in A$ to b containing the directed edge $a^* \rightarrow b$ must end with $\dots a^* \rightarrow b$.*

Proof. Since $a \in \text{nd}(b)$, we do not have a directed path $a \leftarrow \dots \leftarrow b$. Hence, treks from a to b must end with either $\rightarrow$ or $\leftrightarrow$. Suppose there is a trek from a to b containing the edge $a^* \rightarrow b$ but does not end with $\dots a^* \rightarrow b$. Then the edge $a^* \rightarrow b$ must be contained in the middle of the trek. Since all treks from a to b ends with either $\rightarrow$ or $\leftrightarrow$, if $b \leftarrow a^*$ is on the left side of the trek, we can no longer reach a since $a \in \text{nd}(b)$. If $a^* \rightarrow b$ is on the right side of the trek, we will end up with a directed cycle:

$$a \leftarrow \dots a^* \longrightarrow b \longrightarrow \dots \longrightarrow b.$$

Hence, treks from a to b containing $a^* \rightarrow b$ must end with $\dots a^* \rightarrow b$. $\square$

Corollary 2.4. For all $a \in A$, $\sigma_{ab}^* = \sigma_{ab} - \sigma_{aa^*}\lambda_{a^*b}$.

Proof. By Lemma 2.3, all treks from a to b that contain the edge $a^* \rightarrow b$ are simply the treks from a to a^* with the additional edge $a^* \rightarrow b$ appended to the end. By the trek rule, the new covariance is equal to the old covariance with all the treks from a to b containing the edge $a^* \rightarrow b$ subtracted. $\square$

Lemma 2.5. Let $a, a' \in A$. There are no treks between a and a' containing the edge $a^* \rightarrow b$.

Proof. Suppose there is a trek from a to a' containing the edge $a^* \rightarrow b$. Then the trek must be

$$a \leftarrow \dots a^* \rightarrow b \rightarrow \dots \rightarrow a' \quad \text{or} \quad a \leftarrow \dots \leftarrow b \leftarrow a^* \dots \rightarrow a'.$$

But we have $a, a' \in \text{nd}(b)$ which is a contradiction. $\square$

Corollary 2.6. $\Sigma_{AA}^* = \Sigma_{AA}$.

Lemma 2.7. All treks from b to b containing the edge $a^* \rightarrow b$ must either begin with $b \leftarrow a^*$ on the left side or end with $a^* \rightarrow b$ on the right side (or both).

Proof. Otherwise, the edge $a^* \rightarrow b$ is contained in the middle of the trek. Similar to the proof of Lemma 2.3, we will end up with a directed cycle:

$$b \leftarrow \dots \leftarrow b \leftarrow a^* \dots \rightarrow b \quad \text{or} \quad b \leftarrow \dots \leftarrow a^* \rightarrow b \rightarrow \dots \rightarrow b.$$

$\square$

Corollary 2.8. $\sigma_{bb}^* = \sigma_{bb} - 2\lambda_{a^*b}\sigma_{a^*b} + \lambda_{a^*b}^2\sigma_{a^*a^*}$.

Proof. Consider all treks from b to b containing $a^* \rightarrow b$. If a trek starts with $a^* \rightarrow b$ on the left side, this is just a trek from a^* to b with the additional edge $b \leftarrow a^*$ appended to the start. Similarly, a trek ending with $a^* \rightarrow b$ on the right side is a trek from b to a^* with the additional edge $a^* \rightarrow b$ appended at the end. A trek that starts and ends with $a^* \rightarrow b$ is a trek from a^* to a^* with $b \leftarrow a^*$ appended to the start and $a^* \rightarrow b$ appended to the end. The result follows from the inclusion-exclusion principle. $\square$

Lemma 2.9. *Treks from b to $c \in C$ containing the edge $a^* \rightarrow b$ must either start with $b \leftarrow a^*$ on the left side or contain $a^* \rightarrow b$ on the right side (or both).*

Proof. Similar to the proof of Lemma 2.7, if a trek contains $b \leftarrow a^*$ on the left side but does not start with $b \leftarrow a^*$, we have a directed cycle:

$$b \leftarrow \dots \leftarrow b \leftarrow a^* \dots \rightarrow c.$$

□

Corollary 2.10. *For all $c \in C$,*

$$\sigma_{bc}^* = \sigma_{bc} - \lambda_{a^*b}\sigma_{a^*c} - \lambda_{a^*b}\sigma_{a^*b}\sigma(\mathcal{D}_{bc}) + \lambda_{a^*b}^2\sigma_{a^*a^*}\sigma(\mathcal{D}_{bc}), \quad (2.1)$$

where $\mathcal{D}_{bc}$ denotes all directed paths from b to c and $\sigma(\mathcal{D}_{bc}) \in \mathbb{R}[\Lambda]$ is the polynomial function in the form

$$\sigma(\mathcal{D}_{bc}) = \sum_{\pi \in \mathcal{D}_{bc}} \prod_{x \rightarrow y \in \pi} \lambda_{xy}.$$

Proof. Similar to the proof of Corollary 2.8, treks starting with $b \leftarrow a^*$ are of the form:

$$b \leftarrow a^* \dots \rightarrow c$$

which corresponds to the second term of (2.1) while treks containing $a^* \rightarrow b$ on the right side are of the form:

$$b \leftarrow \dots a^* \rightarrow b \rightarrow \dots \rightarrow c$$

which corresponds to the third term of (2.1). Finally, treks containing $a^* \rightarrow b$ on both sides are of the form:

$$b \leftarrow a^* \leftarrow \dots \rightarrow a^* \rightarrow b \rightarrow \dots \rightarrow c$$

which corresponds the last term of (2.1). The result follows from the inclusion-exclusion principle. □

Corollary 2.11. *For all $c, c' \in C$,*

$$\sigma_{cc'}^* = \sigma_{cc'} - \lambda_{a^*b}\sigma_{a^*c}\sigma(\mathcal{D}_{bc'}) - \lambda_{a^*b}\sigma_{a^*c'}\sigma(\mathcal{D}_{bc}) + \lambda_{a^*b}^2\sigma_{a^*a^*}\sigma(\mathcal{D}_{bc})\sigma(\mathcal{D}_{bc'}). \quad (2.2)$$

Proof. Similar to the proof of Corollary 2.10, but the treks can contain $a^* \rightarrow b$ anywhere on either side. Note that the edge $a^* \rightarrow b$ can only be used at most once on each side of a trek as the graph is acyclic. The treks will be of the forms:

$$c \leftarrow \dots a^* \rightarrow b \rightarrow \dots \rightarrow c'$$

$$c \leftarrow \dots \leftarrow b \leftarrow a^* \dots \rightarrow c'$$

$$c \leftarrow \dots \leftarrow b \leftarrow a^* \leftarrow \dots \rightarrow a^* \rightarrow b \rightarrow \dots \rightarrow c'$$

corresponding to the second, third and last terms of (2.2) respectively. $\square$

Lemma 2.12. *Treks from $a \in A$ to $c \in C$ containing the edge $a^* \rightarrow b$ must contain that edge on the right side. In particular, the source of all treks from a to c lies entirely within A .*

Proof. If we have a trek from a to c with b on the left side, it will be of the form $a \leftarrow \dots \leftarrow b \dots$. But $a \in \text{nd}(b)$. $\square$

Corollary 2.13. *For all $a \in A$ and $c \in C$, $\sigma_{ac}^* = \sigma_{ac} - \sigma_{aa^*} \lambda_{a^*b} \sigma(\mathcal{D}_{bc})$.*

Since each entry of Σ^* is a function of Σ , λ_{a^*b} and $\sigma(\mathcal{D}_{bc})$, we obtain the following Corollary.

Proposition 2.14. *Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing the directed edge $a^* \rightarrow b$ from G . Then, Σ^* is identifiable given Σ if and only if both λ_{a^*b} and $\sigma(\mathcal{D}_{bc})$ are identifiable for all $c \in C$.*

Proof. ($\Leftarrow$) Follows from Corollaries 2.4, 2.6, 2.8, 2.10 and 2.11.

($\Rightarrow$) From Corollary 2.4, we have

$$\begin{aligned} \sigma_{a^*b}^* &= \sigma_{a^*b} - \sigma_{a^*a^*} \lambda_{a^*b}, \\ \lambda_{a^*b} &= \frac{\sigma_{a^*b} - \sigma_{a^*b}^*}{\sigma_{a^*a^*}}. \end{aligned}$$

Since $\sigma_{a^*a^*} > 0$, if $\sigma_{a^*b}^*$ is identifiable given Σ , λ_{a^*b} is also identifiable. Similarly, we can rearrange (2.1) into

$$\sigma(\mathcal{D}_{bc}) = \frac{\sigma_{bc}^* - \sigma_{bc} + \lambda_{a^*b}\sigma_{a^*c}}{\lambda_{a^*b}\sigma_{a^*a^*} - \lambda_{a^*b}\sigma_{a^*b}}.$$

Since Σ is positive definite, $\sigma(\mathcal{D}_{bc})$ is also identifiable. □

2.2.2 Regression Identifiability

As we have seen in Proposition 2.14, in order for Σ^* to be identifiable, both λ_{a^*b} and $\sigma(\mathcal{D}_{bc})$ for all $c \in C$ need to be identifiable. We first define the identifiability criterion we will be using in our paper.

Definition 2.1. A directed edge $v \rightarrow w$ is *regression identifiable* if $\lambda_{vw} = \beta_{vw.S}$ for some vertex subset $S \subseteq V \setminus \{v, w\}$, where $\beta_{vw.S} = \sigma_{vw.S} / \sigma_{vv.S}$ is the regression coefficient of X_v on X_w conditional on X_S , and $\sigma_{vw.S} = \sigma_{vw} - \Sigma_{vS}\Sigma_{SS}^{-1}\Sigma_{Sw}$.

The set S in Definition 2.1 is often referred to as the *adjustment set* (Pearl, 2000) and criteria for such a set have been well studied (Shpitser et al., 2010; Perković et al., 2015; Perkovic et al., 2018). We will provide a brief introduction of adjustment sets below.

Definition 2.2. We say that S is a valid *adjustment set* for v to w if

- no vertex in S is a descendent of $\{u \mid u \neq v, u \text{ lies on the causal path from } v \text{ to } w\}$, and
- every non-causal path from v to w is blocked by S .

Theorem 2.15. For a linear SEM, the causal effect of v on w is obtained by regression v on w and a valid adjustment set S . In particular, $\sigma_{\mathcal{D}_{vw}} = \beta_{vw.S}$.

Proof. See (Pearl, 2000, Chapter 5). □

While all edges that are regression identifiable are also globally identifiable, the converse is not true in general. We will discuss more about the limitations when considering identification by regression in Section 2.4. We now define our identifiability criterion for Σ^*

Definition 2.3. Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing the directed edge $a^* \rightarrow b$ from G . We say that Σ^* is *simply identifiable* given Σ if both λ_{a^*b} and $\sigma(\mathcal{D}_{bc})$ are regression identifiable for all $c \in C$.

By definition, if Σ^* is simply identifiable given Σ , then Σ^* is identifiable given Σ . We shall first discuss necessary and sufficient conditions for when the directed path $\sigma(\mathcal{D}_{bc})$ is identifiable by regression for all $c \in C$.

Theorem 2.16. *The path $\sigma(\mathcal{D}_{bc})$ is regression identifiable for some $c \in C$ if and only if b is fixable in $G_{\text{an}(c)}$. In particular, $\sigma(\mathcal{D}_{bc}) = \beta_{bc.S(c)}$ where $S(c) = \text{mb}_{\text{an}(c)}(b)$.*

Proof. ($\Leftarrow$) It suffices to prove that $\sigma(\mathcal{D}_{bc}) = \beta_{bc.S(c)}$ (i.e. $S(c)$ is a valid adjustment set for b to c). Consider all paths from b to c conditioned on $\text{mb}_{\text{an}(c)}(b)$. First consider the case when the path starts with $b \rightarrow \dots$. If it is a directed path, the path is open because $\text{mb}_{\text{an}(c)}(b)$ does not contain any descendants of b . If there is a collider node v on the path, the path is blocked if $v \notin \text{mb}_{\text{an}(c)}(b)$. But if all the collider nodes are in $\text{mb}_{\text{an}(c)}(b)$, then the first collider is a descendant of b . Hence there exists some vertex in the same district as b that is also a descendant of b , contradicting the fact that b is fixable.

Now, all paths that start with $b \leftarrow a \dots$ are blocked since $a \in \text{pa}(b) \subseteq \text{mb}_{\text{an}(c)}(b)$. Paths starting with $b \leftrightarrow d_1 \leftrightarrow \dots \leftrightarrow d_n \leftarrow a' \dots$ are blocked since $a' \in \text{pa}(d_n) \subset \text{mb}_{\text{an}(c)}(b)$ and paths starting with $b \leftrightarrow d_1 \leftrightarrow \dots \leftrightarrow d_n \rightarrow \dots$ are blocked since $d_n \in \text{mb}_{\text{an}(c)}(b)$. Note that $c \notin \text{dis}(b)$ since b is fixable in $G_{\text{an}(c)}$. So the only open paths are the direct paths.

($\Rightarrow$) Now, suppose that b is not fixable in $G_{\text{an}(c)}$. Hence, there exists some $c' \in \text{de}(b) \cap \text{dis}(b) \subseteq C$. Therefore, there is a bidirected path $b \leftrightarrow d_1 \leftrightarrow \dots \leftrightarrow d_m \leftrightarrow c'$ from b to c' . Since each $d_i \in \text{an}(c)$, we have an open path from b to c' of the form $b \leftarrow \dots \leftarrow d_1 \rightarrow \dots \rightarrow c'$ or $b \leftrightarrow d_1 \rightarrow \dots \rightarrow c'$. If we conditioned on either d_1 or any vertex in $\text{de}(d_1)$, then we have the open path $b \leftarrow \dots \leftarrow d_1 \leftrightarrow d_2 \rightarrow \dots \rightarrow c'$ or $b \leftrightarrow d_1 \leftrightarrow d_2 \rightarrow \dots \rightarrow c'$. Conditioning

on d_2 or any vertex in $\text{de}(d_1)$, we have the open path $b \leftarrow \cdots \leftarrow d_1 \leftrightarrow d_2 \leftrightarrow d_3 \rightarrow \cdots \rightarrow c'$ or $b \leftrightarrow d_1 \leftrightarrow d_2 \leftrightarrow d_3 \rightarrow \cdots \rightarrow c'$ and so on until either the path $b \leftarrow \cdots \leftarrow d_1 \leftrightarrow \cdots \leftrightarrow c'$ or $b \leftrightarrow \cdots \leftrightarrow c'$ is open. Hence, $\sigma(\mathcal{D}_{bc'})$ is not regression identifiable.

□

Next, we shall provide necessary and sufficient conditions for when the edge coefficient λ_{a*b} is identifiable by regression.

Theorem 2.17. λ_{ab} is regression identifiable if and only if

- $a \notin \text{dis}_{\text{an}(b)}(b)$, and
- there is no directed edge from a to d for all $d \in \text{dis}_{\text{an}(b)}(b)$.

In particular, $\lambda_{ab} = \beta_{ab \cdot S(b) \setminus \{a\}}$ where $S(b) = \text{mb}_{\text{an}(b)}(b)$.

Proof. ($\Leftarrow$) Consider all paths from a to b conditioned on $\text{mb}_{\text{an}(b)}(b) \setminus \{a\}$. In $G_{\text{an}(b)}$, all paths ending with $\cdots \rightarrow b$ are blocked unless they end with the edge $a \rightarrow b$. For $d_1, \dots, d_n \in \text{dis}(b)$, all paths ending with $\leftarrow d_1 \leftrightarrow \cdots \leftrightarrow d_n \leftrightarrow b$ are blocked since $d_1 \in \text{mb}(b)$ and paths ending with $a' \rightarrow d_1 \leftrightarrow \cdots \leftrightarrow d_n \leftrightarrow b$ are blocked since $a' \in \text{mb}(b)$ ($a' \neq a$ as there is no directed edge from a to d_i). Paths ending with $\cdots \leftarrow b$ do not exist in the induced subgraph $G_{\text{an}(b)}$. Furthermore, there are no bidirected paths from a to b in the induced subgraph $G_{\text{an}(b)}$ since $a \notin \text{dis}_{\text{an}(b)}(b)$.

($\Rightarrow$) First, consider the case where $a \in \text{dis}_{\text{an}(b)}(b)$. Then there is a bidirected path $a \leftrightarrow d_1 \leftrightarrow \cdots \leftrightarrow d_m \leftrightarrow b$ from a to b . Since each $d_i \in \text{an}(b)$, we have an open trek from a to b of the form $a \leftarrow \cdots \leftarrow d_1 \rightarrow \cdots \rightarrow b$ or $a \leftrightarrow d_1 \rightarrow \cdots \rightarrow b$. If we conditioned on either d_1 or any vertex in $\text{de}(d_1)$, then we have the open trek $a \leftarrow \cdots \leftarrow d_1 \leftrightarrow d_2 \rightarrow \cdots \rightarrow b$ or $a \leftrightarrow d_1 \leftrightarrow d_2 \rightarrow \cdots \rightarrow b$. Conditioning on d_2 or any vertex in $\text{de}(d_1)$, we have the open trek $a \leftarrow \cdots \leftarrow d_1 \leftrightarrow d_2 \leftrightarrow d_3 \rightarrow \cdots \rightarrow b$ or $a \leftrightarrow d_1 \leftrightarrow d_2 \leftrightarrow d_3 \rightarrow \cdots \rightarrow b$ and so on until either the trek $a \leftarrow \cdots \leftarrow d_1 \leftrightarrow \cdots \leftrightarrow b$ or $a \leftrightarrow \cdots \leftrightarrow b$ is open. Hence, λ_{ab} cannot be identified by an ordinary regression.

Now, suppose that we have a directed edge $a \rightarrow d_1$ for some $d_1 \in \text{dis}_{\text{an}(b)}(b)$. Suppose there is a bidirected path $d_1 \leftrightarrow d_2 \leftrightarrow \dots \leftrightarrow b$ in $G_{\text{an}(b)}$. Since $d_1 \in \text{an}(b)$, we have the open trek $a \rightarrow d_1 \rightarrow \dots \rightarrow b$. Similarly to the previous case, if we conditioned on either d_1 or any vertex in $\text{de}(d_1)$, then we have the open trek $a \rightarrow d_1 \leftrightarrow d_2 \rightarrow \dots \rightarrow b$ and so on until the trek $a \rightarrow d_1 \leftrightarrow \dots \leftrightarrow b$ is open. Hence, λ_{ab} cannot be identified by an ordinary regression.

□

Now, we have a necessary and sufficient criterion for Σ^* to be simply identifiable given Σ .

Corollary 2.18. *Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing the directed edge $a^* \rightarrow b$ from G . Then, Σ^* is simply identifiable given Σ if and only if*

- $a^* \notin \text{mb}_{\text{an}^*(b)}(b)$, where $\text{mb}_{\text{an}^*(b)}(b)$ denotes the Markov blanket of b in $G_{\text{an}(b)}^*$, and
- b is fixable.

2.2.3 Effect on Data

We will now explore the effects that removing a directed edge will have on specific data point assuming multivariate Gaussian distribution and that the error terms are stable. This is highly practical as finding the counterfactual distribution allows us to evaluate the impact of various policy, business and medical decisions before their implementation. This could answer crucial questions such as “how much would vaccination reduce the risk of Covid” or “will undergoing a particular treatment increase the probability of survival” for a particular individual.

Suppose we have a dataset $X_V \sim N(0, \Sigma)$ and we would like to remove the directed edge $a^* \rightarrow b$.

Theorem 2.19. Suppose λ_{a^*b} and $\sigma(\mathcal{D}_{bc})$ are identifiable for all $c \in \text{de}(b) \setminus \{b\}$, then removing the directed edge $a^* \rightarrow b$ is equivalent to replacing X_V by

$$X_V^* = X_V - \lambda_{a^*b} \Sigma(\mathcal{D}_{bV}) X_{a^*}.$$

Proof. Since $a \in \text{nd}(b)$, $\sigma(\mathcal{D}_{ba}) = 0$. Hence, $X_A^* = X_A$. Similarly, since $\sigma(\mathcal{D}_{bb}) = 1$, we can check that for all $a \in A$ and $c \in C$,

$$\begin{aligned} \text{Cov}(X_a^*, X_b^*) &= \text{Cov}(X_a, X_b) - \lambda_{a^*b} \text{Cov}(X_a, X_{a^*}) \\ &= \sigma_{ab} - \lambda_{a^*b} \sigma_{a^*a}, \\ \text{Cov}(X_b^*, X_b^*) &= \text{Cov}(X_b, X_b) - 2\lambda_{a^*b} \text{Cov}(X_{a^*}, X_b) + \lambda_{a^*b}^2 \text{Cov}(X_{a^*}, X_b) \\ &= \sigma_{bb} - 2\lambda_{a^*b} \sigma_{a^*b} + \lambda_{a^*b}^2 \sigma_{a^*a^*}, \\ \text{Cov}(X_a^*, X_c^*) &= \text{Cov}(X_a, X_c) - \lambda_{a^*b} \sigma(\mathcal{D}_{bc}) \text{Cov}(X_a, X_{a^*}) \\ &= \sigma_{ac} - \lambda_{a^*b} \sigma(\mathcal{D}_{bc}) \sigma_{a^*a}, \end{aligned}$$

which satisfy the Corollaries found in Section 2.2.1. Similarly, we can check that $\text{Cov}(X_b^*, X_c^*)$ and $\text{Cov}(X_c^*, X_{c'}^*)$ satisfies the Corollaries in Section 2.2.1. $\square$

2.3 Adding a Directed Edge

Instead of removing an edge, we now consider the reverse intervention whereby an edge is added to an ADMG G . In this scenario, we have our own choice for the edge coefficients λ_{a^*b} and hence there is no need to identify it.

Once again, suppose that we have an ADMG $G = (V, \mathcal{D}, \mathcal{B})$ with covariance matrix Σ and $a^* \rightarrow b \notin \mathcal{D}$. This time, we want to add the directed edge $a^* \rightarrow b$ to G . As in Section 2.2, let $A = \text{nd}(b)$, $B = \{b\}$ and $C = \text{de}(b) \setminus \{b\}$ and suppose we obtain a new ADMG G^* with covariance matrix Σ^* by adding the directed edge $a^* \rightarrow b$ with edge coefficient λ_{a^*b} . By the trek rule, we have

$$\begin{aligned} \Sigma_{AA}^* &= \Sigma_{AA}, \\ \Sigma_{Ab}^* &= \Sigma_{Ab} + \lambda_{a^*b} \Sigma_{Aa^*}, \\ \sigma_{bb}^* &= \sigma_{bb} + 2\lambda_{a^*b} \sigma_{a^*b} - \lambda_{a^*b}^2 \sigma_{a^*a^*}, \\ \Sigma_{AC}^* &= \Sigma_{AC} + \lambda_{a^*b} \Sigma_{Aa^*} \Sigma(\mathcal{D}_{bC}), \\ \Sigma_{bC}^* &= \Sigma_{bC} + \lambda_{a^*b} \Sigma_{a^*C} + \lambda_{a^*b} \sigma_{a^*b} \Sigma(\mathcal{D}_{bC}) - \lambda_{a^*b}^2 \sigma_{a^*a^*} \Sigma(\mathcal{D}_{bC}), \\ \Sigma_{CC}^* &= \Sigma_{CC} + \lambda_{a^*b} \Sigma_{a^*C}^T \Sigma(\mathcal{D}_{bC}) + \lambda_{a^*b} \Sigma_{a^*C}^T \Sigma(\mathcal{D}_{bC}) - \lambda_{a^*b}^2 \sigma_{a^*a^*} \Sigma^T(\mathcal{D}_{bC}) \Sigma(\mathcal{D}_{bC}), \end{aligned} \tag{2.3}$$

where $\Sigma(\mathcal{D}_{bC}) = [\sigma(\mathcal{D}_{bc})]_{c \in C}$ is the vector of directed treks. Note that the equations are identical to those in Section 2.2.1, with the signs flipped. Hence, in order to add the directed edge $a^* \rightarrow b$ with edge coefficient λ_{a^*b} , we need to be able to identify $\Sigma(\mathcal{D}_{bC})$.

Proposition 2.20. *Suppose that the covariance matrix Σ^* is obtained by adding a directed edge $a^* \rightarrow b$ to G . Then, Σ^* is identifiable given Σ if and only if $\sigma(\mathcal{D}_{bc})$ is identifiable for all $c \in C$.*

Proof. Rearrange (2.3) to obtain $\Sigma(\mathcal{D}_{bC})$ as a function of Σ , Σ^* and λ_{a^*b} . $\square$

Definition 2.4. Suppose the covariance matrix Σ^* is obtained from Σ by the addition of the directed edge $a^* \leftrightarrow b$. We say that Σ^* is *simply identifiable* given Σ if $\sigma(\mathcal{D}_{bc})$ is regression identifiable for all $c \in C$.

By Theorem 2.16, we have the following Corollary.

Corollary 2.21. *Suppose that we obtain an ADMG G^* with covariance matrix Σ^* by adding a directed edge $a^* \rightarrow b$ with edge coefficient λ_{a^*b} to G . Then, Σ^* is simply identifiable given Σ if and only if b is fixable.*

Theorem 2.22. *Given λ_{a^*b} and suppose $\sigma(\mathcal{D}_{bc})$ are identifiable for all $c \in \text{de}(b) \setminus \{b\}$, then adding the directed edge $a^* \rightarrow b$ is equivalent to replacing X_V by*

$$X_V^* = X_V + \lambda_{a^*b} \Sigma(\mathcal{D}_{bV}) X_{a^*}.$$

Proof. Similar to the proof of Theorem 2.19. $\square$

2.4 Limitations

For our main results on edge interventions, we relied heavily on simple identifiability and assumed that the edge coefficients λ_{ab} and path coefficients $\sigma(\mathcal{D}_{bc})$ are regression identifiable. However, for Σ^* to be identifiable given Σ , we only require λ_{ab} and $\sigma(\mathcal{D}_{bc})$ to be generically identifiable. In this section, we shall provide some other graphical criteria for which λ_{ab} or $\sigma(\mathcal{D}_{bc})$ is generically identifiable but not regression identifiable.

2.4.1 Edge Identifiability

In this section, we shall provide two other methods for finding sufficient conditions for λ_{ab} to be generically identifiable. These methods are not exhaustive as the question of finding necessary and sufficient conditions for λ_{ab} to be generically identifiable remains an open problem.

Simple Graphs

A mixed graph $G = (V, \mathcal{D}, \mathcal{B})$ is simple if between any two vertices $v_1, v_2 \in V$, there is at most one edge between them. That is, we cannot have both $v_1 \rightarrow v_2$ and $v_1 \leftrightarrow v_2$.

Theorem 2.23. *If G is simple, then $\mathcal{M}(G)$ is generically identifiable.*

Proof. See [Brito and Pearl \(2002a\)](#). □

Using the Theorem above, we can obtain a similar result to Corollary 2.18:

Proposition 2.24. *Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing the directed edge $a^* \rightarrow b$ from G . Then, Σ^* is generically identifiable given Σ if*

- G is a simple graph, and
- b is fixable.

Generalised Instrumental Sets

In the instrumental variable model shown in Figure 1.2, the directed edge λ_{23} generically identifiable even though it is not regression identifiable. This lead to the study of an *instrumental set* by [Brito and Pearl \(2002b\)](#), where the instrumental variable model was generalised.

Theorem 2.25. *If $\mathbf{z} = \{z_1, \dots, z_n\}$ is an instrumental set relative to causes $\mathbf{a} = \{a_1, \dots, a_n\}$ and effect b , then the parameters of the edges $a_1 \rightarrow b, \dots, a_n \rightarrow b$ are identified almost everywhere, and can be computed by solving a system of linear equations.*

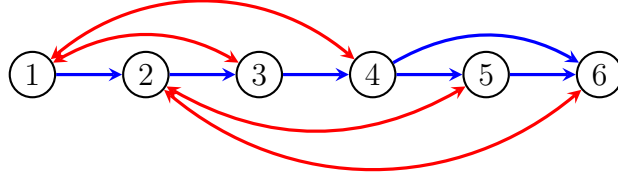


Figure 2.3: Cut-vertex example

If we are only interested in identifying the parameter of the edge $a^* \rightarrow b$, we only need the existence of an instrumental variable z that is instrumental relative to a^* and b . Hence,

Corollary 2.26. *Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing the directed edge $a^* \rightarrow b$ from G . Then, Σ^* is generically identifiable given Σ if*

- *If there exists an vertex z that is instrumental relative to a^* and b , and*
- *b is fixable.*

2.4.2 Cut Vertices

Previously, we dealt with cases where the edge coefficients are generically identifiable but not regression identifiable. Now, we shall provide some cases where directed path coefficients are generically identifiable but not regression identifiable.

Example 2.27. Consider the graph in Figure 2.3. Here, we cannot use Corollary 2.18 to remove the edge $1 \rightarrow 2$ as 2 is not fixable. However, we can observe that all directed paths contributing to $\sigma(\mathcal{D}_{25})$ and $\sigma(\mathcal{D}_{26})$ pass through the vertex 4. Hence, we are able to identify those using $\sigma(\mathcal{D}_{25}) = \sigma(\mathcal{D}_{24})\sigma(\mathcal{D}_{45}) = \beta_{24 \cdot 1}\beta_{45 \cdot 123}$ and $\sigma(\mathcal{D}_{26}) = \sigma(\mathcal{D}_{24})\sigma(\mathcal{D}_{46}) = \beta_{24 \cdot 1}\beta_{46 \cdot 123}$. In fact, we see that it is possible to remove the edge $1 \rightarrow 2$ and identify the new covariance matrix Σ^* . We shall refer to the vertex 4 as a *cut vertex* which we shall rigorously define.

Definition 2.5. For an ADMG, G , let $G^{(d)}$ be the subgraph of G with all the bidirected edges removed and let $G_{a \rightarrow b}^{(d)}$ be the induced subgraph $G_{\text{an}(b) \cap \text{de}(a)}$ with all bidirected edges removed. We say that a vertex v is a *cut vertex* of $G_{a \rightarrow b}^{(d)}$ if $G_{a \rightarrow b}^{(d)} \setminus \{v\}$ is disconnected.

Lemma 2.28. Let $b \in V$ and suppose $c \in \text{de}(b)$. Denote the directed paths from b to c by $\tau_1, \dots, \tau_n$. Further partition each τ_i into directed subpaths $\tau_i^1, \dots, \tau_i^m$. Then

$$\sigma(\mathcal{D}_{bc}) = \sum_{i=1}^n \prod_{j=1}^m \sigma(\tau_i^j).$$

In particular, if $u_1, \dots, u_n$ is a sequence of cut-vertices in $G_{b \rightarrow c}^{(d)}$, then

$$\sigma(\mathcal{D}_{bc}) = \sigma(\mathcal{D}_{bu_1}) \sigma(\mathcal{D}_{u_1 u_2}) \dots \sigma(\mathcal{D}_{u_n c}).$$

Proof. Follows from the trek rule. $\square$

Corollary 2.29. Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing the directed edge $a^* \rightarrow b$ from G . Then, Σ^* is identifiable given Σ if

- $a^* \notin \text{mb}_{\text{an}^*(b)}(b)$, where $\text{mb}_{\text{an}^*(b)}(b)$ denotes the Markov blanket of b in $G_{\text{an}(b)}^*$, and
- For all $c \in C$, we can find a sequence of vertices $v_0, v_1, \dots, v_n$ such that $v_0 = b$, $v_n = c$, $v_1, \dots, v_{n-1}$ is a sequence of cut-vertices in $G_{b \rightarrow c}^{(d)}$ and v_i is fixable in $G_{\text{an}(v_{i+1})}$ for all $0 \leq i \leq n-1$.

2.5 Usage of Removing Edges

In this section, we shall showcase some statistical problems where our identifiability results could be applied. We first show that our results may be used to find constraints in ADMGs, provided that we can either remove all the directed or bidirected edges (see Appendix B.2). We shall then provide an example from epidemiology where we have two treatments, and we make an intervention on conditions required for the first treatment.

2.5.1 Constraints

In a DAG G , the set of conditional independences yields an implicit description of $\mathcal{M}(G)$ (Garcia et al., 2005; Kiiveri et al., 1984). However, in the class of ADMGs, Robins (1986) noted a new form of constraint similar to that of conditional independence, sometimes referred to as the *functional constraints*. A full introduction of constraints could be found in Section 1.5. In this section, we will demonstrate how removing edges could allow us to recover some of the constraints of an ADMG.

Suppose we can iteratively remove bidirected or directed edges from an ADMG G to obtain a new graph G^* with covariance matrix Σ^* . Then, the constraints implied by G correspond to the non-trivial polynomials of the form σ_{ij}^* for all vertex pairs i, j without a trek between them in G^* . If G^* has no directed edges, this will simply correspond to non-adjacent vertex pairs. If G^* has no edges left, this will correspond to the strictly upper triangular entries of Σ^* . We shall provide examples where we remove all directed edges.

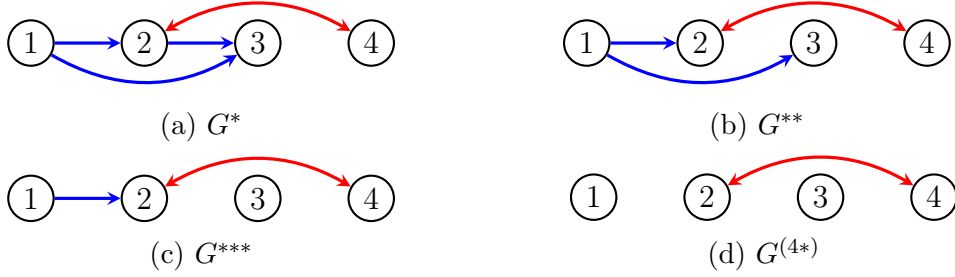


Figure 2.4: Illustration of each step of edge removal in Example 2.30

Example 2.30 (Removing directed edges from Verma graph). Consider the Verma graph in Figure 1.3. Note that we may not remove the directed edge $1 \rightarrow 2$ since the vertex 2 is not fixable. Suppose instead we first remove the directed edge $3 \rightarrow 4$ to obtain a new graph G^* . We shall illustrate each step of this example in Figure 2.4. Then, $A = \{1, 2, 3\}$, $B = \{4\}$ and $C = \emptyset$. Now, since $\lambda_{34} = \sigma_{34 \cdot 12} / \sigma_{33 \cdot 12}$ by Theorem 2.17, the covariance matrix of the new graph, Σ^* , becomes

$$\Sigma^* = \begin{bmatrix} \Sigma_{123,123} & \Sigma_{123,4} - \Sigma_{123,3} \frac{\sigma_{34 \cdot 12}}{\sigma_{33 \cdot 12}} \\ \sigma_{44} - 2 \frac{\sigma_{34} \sigma_{34 \cdot 12}}{\sigma_{33 \cdot 12}} + \sigma_{33} \frac{\sigma_{34 \cdot 12}^2}{\sigma_{33 \cdot 12}^2} \end{bmatrix}.$$

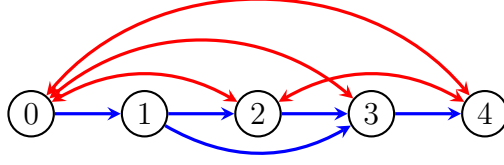


Figure 2.5: The ‘double’ Verma graph.

In this new graph with the edge $3 \rightarrow 4$ removed, the vertex 2 is now fixable. Now, removing the edge $2 \rightarrow 3$, $\lambda_{23}^* = \frac{\sigma_{23,1}}{\sigma_{22,1}}$, $\sigma^*(\mathcal{D}_{34}) = 0$.

$$\Sigma^{**} = \begin{bmatrix} \Sigma_{12,12} & \Sigma_{12,3} - \Sigma_{12,2} \frac{\sigma_{23,1}}{\sigma_{22,1}} & \Sigma_{12,4}^* \\ \sigma_{33} - \frac{2\sigma_{23}\sigma_{23,1}}{\sigma_{22,1}} + \frac{\sigma_{23,1}^2}{\sigma_{22,1}} & \sigma_{34}^* - \frac{\sigma_{23,1}\sigma_{24}^*}{\sigma_{22,1}} & \sigma_{44}^* \end{bmatrix}.$$

Removing edge $1 \rightarrow 3$, $\lambda_{13}^{**} = \frac{\sigma_{13}^{**}}{\sigma_{11}^{**}}$, $\sigma^{**}(\mathcal{D}_{34}) = 0$.

$$\Sigma^{***} = \begin{bmatrix} \Sigma_{12,12} & \Sigma_{12,3}^{**} - \Sigma_{12,1} \frac{\sigma_{13}^{**}}{\sigma_{11}^{**}} & \Sigma_{12,4}^* \\ \sigma_{33}^{**} - 2\frac{\sigma_{13}^{**2}}{\sigma_{11}^{**}} + \frac{\sigma_{13}^{**2}}{\sigma_{11}^{**}} & \sigma_{34}^{**} - \frac{\sigma_{13}^{**}\sigma_{14}^*}{\sigma_{11}^{**}} & \sigma_{44}^* \end{bmatrix}.$$

Removing edge $1 \rightarrow 2$, $\lambda_{12}^{***} = \frac{\sigma_{12}}{\sigma_{11}}$, $\sigma^{***}(\mathcal{D}_{23}) = \sigma^{***}(\mathcal{D}_{24}) = 0$.

$$\Sigma^{(4*)} = \begin{bmatrix} \sigma_{11} & 0 & \sigma_{13}^{***} & \sigma_{14}^* \\ \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}} & \sigma_{23}^{***} - \frac{\sigma_{12}\sigma_{13}^{***}}{\sigma_{11}} & \sigma_{24}^* - \frac{\sigma_{12}\sigma_{14}^*}{\sigma_{11}} \\ \sigma_{33}^{***} & \sigma_{34}^{***} & \sigma_{44}^* \end{bmatrix}.$$

Note $\sigma_{13}^{***} = 0$ since there are no treks from 1 to 3 in G^{***} (see Figure 2.4c). Now, since $\Sigma^{(4*)}$ corresponds to the graph on 4 vertices with only the bidirected edge $2 \leftrightarrow 4$, we can equate the strictly upper triangular part of $\Sigma^{(4*)}$, apart from $\sigma_{24}^{(4*)}$ to be zero to obtain the constraints of the original graph.

Since there are no treks from vertex 3 to any of the other vertices in G^{***} , σ_{13}^{***} , $\sigma_{23}^{(4*)}$ and σ_{34}^{***} are identically zero by the trek rule. Hence, equating the entries of $\Sigma^{(4*)}$ corresponding to vertex pairs in $G^{(4*)}$ without a trek between them to zero, we have four expressions that trivially vanish and the equation $\sigma_{14}^* = 0$. Equating $\sigma_{14}^* = 0$, we recover $f_{\text{Verma}} = 0$. We show this computation in full in the appendix.

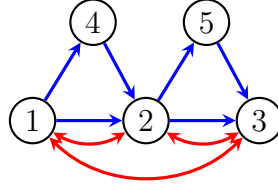


Figure 2.6: Graphical model for MACS study

Example 2.31 (The ‘double’ Verma graph). Consider the ‘double’ Verma graph in Figure 2.5 with a doubly nested constraint where we need to fix two vertices, with a marginalisation in between, to obtain the conditional independence constraints.

In this graph, the vertex 1 is fixable and the vertex 0 is not in the Markov blanket of 1 with the edge $0 \rightarrow 1$ removed. Hence, we can remove the edge $0 \rightarrow 1$ to obtain a new ADMG G^* . Now, we have a graph similar to the Verma example, where we are able to remove the directed edge $3 \rightarrow 4$ to obtain a new ADMG G^{**} since 3 is fixable in $G_{\text{an}(4)}^*$. Now, we can remove edges $2 \rightarrow 3$ and $1 \rightarrow 3$ as vertices 1 and 2 are fixable in $G_{\text{an}(3)}^{**}$. Similarly, we are able to remove the edge $1 \rightarrow 2$. Hence, we are able to recover the constraint in the double Verma graph.

2.5.2 Simulating Interventional Data

We will now provide a real world data example where we want to simulate an interventional distribution from observational data.

Example 2.32. We consider the data from the Multicenter AIDS Cohort Study (MACS) (Kaslow et al., 1987), a longitudinal observational study of HIV-infected men. Suppose that we are interested in the effect that treating HIV patients with AZT has on their CD4 cell count in 6 months (1 follow-up period). We restrict ourselves to a subset of the MACS study consisting of 1210 men for which there are no missing data for age, medication, AZT doses (if treated) and CD4 count. The mean (interquartile range) of these men is 43.2 (38–49) years old. The graphical model is provided in Figure 2.6 with variables:

- X_1 - Baseline CD4 cell count;
- X_2 - CD4 cell count after first treatment;

- X_3 - CD4 cell count after second treatment;
- X_4 - AZT dosage of first treatment (log scale with 1 added);
- X_5 - AZT dosage of second treatment (log scale with 1 added).

In particular, we hypothesised that the patients age will affect both the treatment proposed and the change in CD4 cell count given any antiviral treatment. Standardising the dataset and assuming that our dataset follows a linear SEM, we obtain $\lambda_{14} = -0.0252$, $\lambda_{42} = -0.0176$, $\lambda_{25} = -0.0287$, $\lambda_{53} = -0.0378$, $\sigma(\mathcal{D}_{42}) = -0.0387$, $\sigma(\mathcal{D}_{43}) = -0.0580$ and $\sigma(\mathcal{D}_{45}) = 0.891$ with

$$\Sigma = \begin{bmatrix} 1.000 & 0.835 & 0.794 & -0.025 & -0.004 \\ 0.835 & 1.000 & 0.853 & -0.039 & -0.029 \\ 0.794 & 0.853 & 1.000 & -0.058 & -0.051 \\ -0.025 & -0.039 & -0.058 & 1.000 & 0.891 \\ -0.004 & -0.029 & -0.051 & 0.891 & 1.000 \end{bmatrix}.$$

Suppose we want to simulate the dataset for which we no longer determine the AZT dosage of the first treatment based on a patient's baseline CD4 cell count. Performing an edge intervention by removing λ_{14} , we have $A = \{1\}$, $B = \{4\}$ and $C = \{2, 3, 5\}$ with

$$\Sigma^* = \begin{bmatrix} 1.000 & 0.835 & 0.793 & -0.025 & 0.018 \\ 0.835 & 0.999 & 0.852 & -0.018 & -0.010 \\ 0.793 & 0.852 & 0.999 & -0.038 & -0.033 \\ -0.025 & -0.018 & -0.038 & 0.999 & 0.891 \\ 0.018 & -0.010 & -0.033 & 0.891 & 1.000 \end{bmatrix}.$$

We could then simulate the distribution of an edge intervention using $N(0, \Sigma^*)$ and reversing the data standardisation.

2.6 Conclusion

In this chapter, we studied the effects that removing or adding an edge will have on the covariance matrix Σ in a Gaussian ADMG.

We first defined a identification criterion, stronger than global identifiability, called *regression identifiability*. The necessary and sufficient graphical conditions

for λ_{ab} , $\sigma(\mathcal{D}_{bc})$ and ω_{ab} to be regression identifiable are given in Theorems 2.16, 2.17 and B.4 respectively. Since regression identifiability implies generic identifiability, we now have sufficient conditions for Σ^* to be identifiable given Σ , where Σ^* is the covariance matrix of the modified graph. We have also identified the effect of an edge intervention on an individual data point in Theorems 2.19 and 2.22.

However, even if either λ_{ab} or $\sigma(\mathcal{D}_{bc})$ is not regression identifiable, it may still be generically identifiable. We studied some cases of the limitations of our results in Section 2.4. If the graphical conditions that are necessary and sufficient for λ_{ab} and $\sigma(\mathcal{D}_{bc})$ to be generically identifiable were to be found, then we would have the graphical conditions which are necessary and sufficient for Σ^* to be identifiable given Σ .

Finally, we have provided some statistical problems, such as finding constraints and simulating interventional data from observation data, where our results can be used.

Chapter 3

Algebraic Properties of HTC-identifiable Graphs

3.1 Introduction

It is often natural to model the joint distribution of a random vector $X = (X_1, \dots, X_n)^T$ as a collection of noisy linear interdependencies. In particular, we may assume that each variable X_i can be expressed as a linear function of the remaining variables and a stochastic noise term ϵ_i ,

$$X_i = \sum_{j \neq i} \lambda_{ji} X_j + \epsilon_i. \quad (3.1)$$

As seen Section 1.2, models of this form are known as *linear structural equation models* (SEMs). We often represent SEMs using graphical models, where each vertex corresponds to a random variable, each directed edge corresponds to the linear dependencies and each bidirected edge corresponds to possible correlations among noise terms. From Section 1.2, we see that X has the covariance matrix

$$\Sigma := \text{Cov}[X] = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}. \quad (3.2)$$

3.1.1 HTC-identifiability

Recall that a *trek* from i to j is a walk from i to j without any colliders, and that all treks are of the form

$$v_\ell^L \leftarrow v_{\ell-1}^L \leftarrow \dots \leftarrow v_1^L \leftarrow v_0^L \leftrightarrow v_0^R \rightarrow v_1^R \rightarrow \dots \rightarrow v_{r-1}^R \rightarrow v_r^R$$

or

$$v_\ell^L \leftarrow v_{\ell-1}^L \leftarrow \cdots \leftarrow v_1^L \leftarrow v_0^{LR} \rightarrow v_1^R \rightarrow \cdots \rightarrow v_{r-1}^R \rightarrow v_r^R,$$

where $v_\ell^L = i$, $v_r^R = j$. A *half-trek* is a trek with $|\pi^L| = 0$. Note that the vertex v_0^{LR} in the second case is a part of both π^L and π^R . Let $\text{htr}(v)$ be the set of nodes in $V \setminus (\{v\} \cup \text{sib}(v))$ that can be reached from v via a half-trek. A set of treks $\Pi = \{\pi_1, \pi_2, \dots\}$ is said to have no *sided intersection* if $\pi_i^L \cap \pi_j^L = \pi_i^R \cap \pi_j^R = \emptyset$ for all $i \neq j$.

Definition 3.1. A set of nodes $Y \subset V$ satisfies the *half-trek criterion* with respect to $v \in V$ if

1. $|Y| = |\text{pa}(v)|$,
2. $Y \cap (\{v\} \cup \text{sib}(v)) = \emptyset$, and
3. there is a system of half-treks with no sided intersection from Y to $\text{pa}(v)$.

In Section 1.4, we provided a survey on identifiability in SEMs and introduced the definitions for *global identifiability* and a weaker *generic identifiability*. While the necessary and sufficient graphical conditions for a linear SEM to be globally identifiable are found later Drton et al. (2011), the graphical conditions for a linear SEM to be generically identifiable is still an open problem. In Brito and Pearl (2002a), it was proven that all bow-free acyclic graphs are generically identifiable. This result was later expanded by Foygel et al. (2012) for HTC-identifiable graphs.

Definition 3.2. We say that $\mathcal{M}(G)$ is *HTC-identifiable* if there exists a family of subsets $\{Y_v : v \in V\}$ of the vertex set V such that for each node $v \in V$, the vertex set Y_v satisfies the half-trek criterion with respect to v and there is a total ordering such that $w \prec v$ whenever $w \in Y_v \cap \text{htr}(v)$.

Note that all acyclic bow-free graphs are HTC-identifiable as we can simply choose $Y_v = \text{pa}(v)$. Furthermore, Foygel et al. (2012) proved that all HTC-identifiable graphs are also generically identifiable. Hence, Foygel's result generalises the previous work by Brito and Pearl (2002a). Furthermore, Foygel et al. (2012) also shown that HTC-identifiable graphs can be decided in polynomial time. While all HTC-identifiable graphs are also generically identifiable, it has been

shown through simulation that the converse is not true in general. In the original paper, Foygel et al. classify graphs into three classes: HTC-identifiable, HTC-infinite-to-one and HTC-inconclusive. In this thesis, we will consider both HTC-infinite-to-one and HTC-inconclusive graphs simply as ‘not HTC-identifiable’.

In this chapter, we will explore the algebraic structure of linear SEMs associated with HTC-identifiable graphs by introducing an identifiability criterion which we call *linear identifiability*. SEMs associated with graphs that are linearly identifiable satisfy certain algebraic structures which could be used to recover Λ from Σ efficiently using linear algebra. We will then prove that linear identifiability and HTC-identifiability are equivalent, hence, we are not able to recover Λ from Σ using linear algebra from graphs that are not HTC-identifiable. In fact, we will give an example of an HTC-inconclusive graph that is generically identifiable in Example 3.14 and show that we have to solve a higher order polynomial equation in order to recover Λ .

3.1.2 Constraints

In a DAG G , the set of conditional independences yields an implicit description of $\mathcal{M}(G)$ (Garcia et al., 2005; Kiiveri et al., 1984). However, in the class of ADMGs, Robins (1986) noted a new form of constraint similar to that of conditional independence, sometimes referred to as the *functional constraints*. A full introduction of constraints could be found in Section 1.5.

van Ommen and Mooij (2017) provided a set of polynomials that is generated by the vanishing ideal, $\mathcal{I}(G)$, for any HTC-identifiable graph G . In this chapter, we are interested in exploiting the algebraic properties of HTC-identifiable graphs to find the set of polynomials that generate $\mathcal{I}(G)$ for any HTC-identifiable graph G . To do this, we will first consider a subideal $J(G)$ of $\mathcal{I}(G)$, which we call the *model ideal*, such that

$$\overline{V(J(G)) \cap PD_V} = \overline{\mathcal{M}(G)},$$

where $V(I)$ refers to the *variety* defined by an ideal I ,

$$V(I) = \{a \in \mathbb{S}_{|V|} : f(a) = 0 \text{ for all } f \in I\}$$

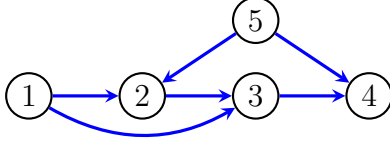


Figure 3.1: Directed Acyclic Graph Example

for symmetric $|V| \times |V|$ matrices $\mathbb{S}_{|V|}$ and $\bar{V}$ denotes the Zariski closure of the variety V (i.e. by removing all inequality constraints). A model ideal $J(G)$ gives a more practical description of $\mathcal{M}(G)$ compared to the vanishing ideal $\mathcal{I}(G)$ as $\mathcal{M}(G) \subseteq PD_V$ and we are dealing with varieties in PD_V . Hence, we only have to check if the constraints in $J(G)$ are satisfied to verify that $\Sigma \in \mathcal{M}(G)$. In particular, we will prove that the constraints found by [van Ommen and Mooij \(2017\)](#) in fact generates the *model ideal* $J(G)$.

The following example illustrates how a model ideal $J(G)$ differs from a vanishing ideal $\mathcal{I}(G)$.

Example 3.1. Consider the graph in Figure 3.1 studied in ([Sullivant, 2008](#), Example 2.13). Here, the conditional independences corresponds to the following statements: $X_1 \perp\!\!\!\perp X_5$, $X_3 \perp\!\!\!\perp X_5 \mid \{X_1, X_2\}$ and $\{X_1, X_2\} \perp\!\!\!\perp X_4 \mid \{X_3, X_5\}$. Let $C(G)$ be the conditional independence ideal generated by the determinantal cubics. In the space of $\mathbb{R}(\Sigma)$, the vanishing ideal $\mathcal{I}(G)$ is generated by $\mathcal{I}(G) = C(G) + \langle f_{\text{Verma}} \rangle$. However, $f_{\text{Verma}}(\Sigma) = 0$ for every matrix Σ in $V(C(G)) \cap PD_V$.

In this example, $C(G)$ is playing the role of a model ideal $J(G)$.

3.1.3 Outline of this Chapter

This chapter is structured as follows: In Section 3.2, we will prove that linearly identifiable and HTC-identifiable are equivalent. In Section 3.3, we will exploit the algebraic properties of HTC-identifiable graphs we found in Section 3.2 to provide an efficient algorithm to recover the parameters, (Λ, Ω) , of any HTC-identifiable graphs. Finally, in Section 3.4, we will use these parameters to find a model ideal $J(G)$ of any HTC-identifiable graph G , where $J(G)$ defines the model $\mathcal{M}(G)$ in the positive definite cone. We will further prove that this ideal is minimal for a

subset of HTC-identifiable graphs and discuss some complexity issues with our algorithm.

3.2 Linear Identifiability

In this section, we will first define linear identifiability and then explore the properties of linearly identifiable graphs. In particular, we will show that linearly identifiable graphs are equivalent to HTC-identifiable graphs.

First, notice that equation (3.2) can be rearranged into

$$(I - \Lambda)^T \Sigma (I - \Lambda) = \Omega. \quad (3.3)$$

By equating each entry in $(I - \Lambda)^T \Sigma (I - \Lambda)$ that corresponds to a zero entry in Ω (i.e. a missing bidirected edge), we obtain a system of equations in terms of the unknowns λ_{ij} only. Since Σ is a covariance matrix, it is symmetric. As any matrix congruent to a symmetric matrix is also symmetric, $(I - \Lambda)^T \Sigma (I - \Lambda)$ is also a symmetric matrix. By assumption, Ω is a positive definite covariance matrix, and is hence symmetric with strictly positive diagonal entries. Therefore, it suffices to equate only the strictly upper or strictly lower triangular entries of both matrices.

Define the matrices $A = (a_{ij}) = (I - \Lambda)^T \Sigma$ and $B = (b_{ij}) = (I - \Lambda)^T \Sigma (I - \Lambda)$, where each entry is a polynomial in which the indeterminates are entries in both Λ and Σ . Expanding the matrix multiplication, we obtain

$$a_{ij} = \sigma_{ij} - \sum_{\ell \in \text{pa}(i)} \lambda_{\ell i} \sigma_{\ell j}, \quad (3.4)$$

$$b_{ij} = a_{ij} - \sum_{\ell \in \text{pa}(j)} \lambda_{\ell j} a_{i\ell}. \quad (3.5)$$

Observe that $a_{ij} \in \mathbb{R}[\Sigma, \Lambda]$ is the polynomial obtained by summing all the covariances for half-treks from i to j , and $b_{ij} \in \mathbb{R}[\Sigma, \Lambda]$ is the polynomial expression for the bidirected edge between i and j . Equating each entry of both matrices in (3.3), we obtain $b_{ij} = \omega_{ij}$. Therefore, for each missing bidirected edge between i and j , we have the equation $b_{ij} = b_{ji} = 0$. We immediately have a necessary condition for generic identifiability.

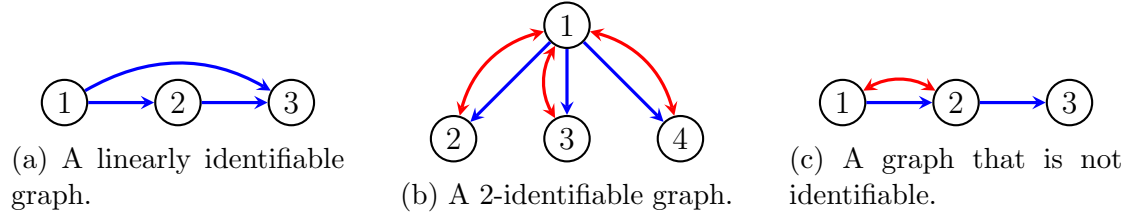


Figure 3.2

Proposition 3.2. *If G has more than $\binom{|V|}{2}$ edges, then $\mathcal{M}(G)$ is not generically identifiable.*

Proof. Each missing bidirected edge corresponds to an equation $b_{ij} = 0$ while each directed edge corresponds to an unknown λ_{ij} . If we have more than $\binom{|V|}{2}$ edges, then we have more unknowns than equations. Hence, $\mathcal{M}(G)$ is not identifiable. $\square$

We shall now provide the definition for a graphical model to be linearly identifiable.

Definition 3.3. Let G be a mixed graph. We say that $\mathcal{M}(G)$ is *linearly identifiable* if for each vertex v , the system of $m = |\text{pa}(v)|$ linearly independent equations of the form $\{b_{jv} = 0 \mid j \leftrightarrow v \notin \mathcal{B}\}$ (defined in (3.5)) that is expressed as a polynomial (e.g. through substitution) that is linear for the m indeterminates $\Lambda_{\text{pa}(v),v} := \{\lambda_{iv} \mid i \in \text{pa}(v)\}$.

If a model is linearly identifiable, then we solve for Λ (as polynomials in Σ) recursively using linear algebra. We can then recover Ω using equation (3.3). Hence, all linearly identifiable graphs are generically identifiable. We shall demonstrate linear identifiability with a few examples.

Example 3.3 (A linearly identifiable graph). Consider the saturated graph G with three vertices in Figure 3.2(a). In this graph, the missing bidirected edges give us the equations

$$\begin{aligned} b_{21} &= \sigma_{12} - \lambda_{12}\sigma_{11} = 0, \\ b_{31} &= \sigma_{13} - \lambda_{13}\sigma_{11} - \lambda_{23}\sigma_{12} = 0, \\ b_{32} &= \sigma_{23} - \lambda_{13}\sigma_{12} - \lambda_{23}\sigma_{22} - \lambda_{12}(\sigma_{13} - \lambda_{13}\sigma_{11} - \lambda_{23}\sigma_{12}) = 0. \end{aligned}$$

The vertex 2 has one parent and we have that the first equation is linear in λ_{12} . Solving this, we obtain $\lambda_{12} = \sigma_{12}/\sigma_{11}$. The vertex 3 has two parents. Substituting λ_{12} into the last two equations, we have a system of two linearly independent equations with unknowns $\lambda_{13}, \lambda_{23}$ and coefficients in Σ . Hence, $\mathcal{M}(G)$ is linearly solvable.

However, not all graphs have such a system of equations, which we shall see in next example.

Example 3.4 (A graph that is not linearly identifiable). Consider the graph G in Figure 3.2(b) with three bows. The missing bidirected edges give us the equations

$$\begin{aligned} b_{32} &= \sigma_{23} - \lambda_{13}\sigma_{12} - \lambda_{12}(\sigma_{13} - \lambda_{13}\sigma_{11}) = 0, \\ b_{42} &= \sigma_{24} - \lambda_{14}\sigma_{12} - \lambda_{12}(\sigma_{14} - \lambda_{14}\sigma_{11}) = 0, \\ b_{43} &= \sigma_{34} - \lambda_{14}\sigma_{13} - \lambda_{13}(\sigma_{14} - \lambda_{14}\sigma_{11}) = 0. \end{aligned}$$

$\mathcal{M}(G)$ is not linearly identifiable as we cannot express any subset of $\lambda_{12}, \lambda_{13}$ or λ_{14} linearly in terms of Σ . In fact, from the first two equations, we have

$$\lambda_{13} = \frac{\sigma_{23} - \lambda_{12}\sigma_{13}}{\sigma_{12} - \lambda_{12}\sigma_{11}} \quad \text{and} \quad \lambda_{14} = \frac{\sigma_{24} - \lambda_{12}\sigma_{14}}{\sigma_{12} - \lambda_{12}\sigma_{11}}.$$

Substituting these into the third equation, we obtain the quadratic equation $a\lambda_{12}^2 + b\lambda_{12} + c = 0$ with

$$\begin{aligned} a &= \sigma_{11}^2\sigma_{34} - \sigma_{11}\sigma_{13}\sigma_{14}, \\ b &= 2\sigma_{12}\sigma_{13}\sigma_{14} - 2\sigma_{11}\sigma_{12}\sigma_{34}, \\ c &= \sigma_{12}^2\sigma_{34} - \sigma_{12}\sigma_{13}\sigma_{24} - \sigma_{12}\sigma_{14}\sigma_{23} + \sigma_{11}\sigma_{23}\sigma_{24}. \end{aligned}$$

Since $b^2 - 4ac \neq 0$, if $\Sigma \in \mathbb{C}^{|V| \times |V|}$ then $\mathcal{M}(G)$ is not generically identifiable. In fact, if $\Sigma \in \mathbb{C}^{|V| \times |V|}$ and we have to solve a k th order equation without repeated roots, then $\mathcal{M}(G)$ is k -identifiable.

However, determining whether a graph is linearly identifiable is not straightforward. For example, there are instances where the coefficient of one of the unknowns, λ_{ij} , is zero.

Example 3.5. Consider the graph in Figure 3.2(c). The missing bidirected edges correspond to the equations

$$b_{31} = \sigma_{13} - \lambda_{23}\sigma_{12} = 0, \quad (3.6)$$

$$b_{32} = \sigma_{23} - \lambda_{23}\sigma_{22} - \lambda_{12}(\sigma_{13} - \lambda_{23}\sigma_{12}) = 0. \quad (3.7)$$

At first glance, these equations might seem to suggest that $\mathcal{M}(G)$ is linearly identifiable as we can use (3.6) to solve for λ_{23} and then substitute that solution into (3.7) to solve for λ_{12} . However, the coefficient of λ_{12} is precisely $b_{31} = 0$. Hence, we cannot find an equation that is linear in λ_{12} , so $\mathcal{M}(G)$ is not linearly identifiable. Indeed, this graph is not even generically identifiable.

As linearly identifiable graphs are not straightforward to classify, we will first introduce a weaker definition that specifically excludes the case where the coefficients of the unknowns might be zero.

Definition 3.4. Let G be a mixed graph. We say that $\mathcal{M}(G)$ is *quasi-linearly identifiable* if there exists a total ordering $\prec$ such that for all vertex v , we can find a system of $m = |\text{pa}(v)|$ linearly independent equations of the form $b_{ij} = 0$ for the m indeterminates $\Lambda_{\text{pa}(v),v} := \{\lambda_{iv} \mid i \in \text{pa}(v)\}$ where each equation is expressed as a polynomial (e.g. through substitution) linear in terms of $\Lambda_{\text{pa}(v),v}$ with coefficients that are functions of Σ and $\Lambda_{\text{pa}(k),k}$ where $k \prec v$.

The main difference between quasi-linearly identifiable models and linearly identifiable models is that in the former, we do not worry about the specific value of each λ_{ij} at each iteration even if it might cause the coefficients of future unknowns to be zero after substitution.

Example 3.6. In Example 3.5, vertex 1 satisfies the quasi-linear condition as $|\text{pa}(1)| = 0$, and a system of zero equations is trivial. Furthermore, the vertex 3 satisfies the quasi-linear condition as (3.6) is a system of $|\text{pa}(3)| = 1$ equation with indeterminates in Σ that is linear in $\{\lambda_{i3} \mid i \in \text{pa}(3)\}$. Finally, the vertex 2 also satisfies the quasi-linear condition as (3.7) is a system of $|\text{pa}(2)| = 1$ equation with indeterminates in Σ and λ_{j3} , where the vertex 3 also satisfies the quasi-linear condition, that is linear in $\{\lambda_{i2} \mid i \in \text{pa}(2)\}$. Hence, the graph in Example 3.5 is quasi-linearly identifiable.

If a quasi-linearly identifiable model is also linearly identifiable, we could solve for λ_{jk} symbolically for all $j \in \text{pa}(k)$ and substitute this back into the equations $b_{\ell v} = 0$ to obtain $|\text{pa}(v)|$ linearly independent equations in λ_{iv} .

3.2.1 Properties of Quasi-linearly Identifiable Graphs

Consider the equation

$$\begin{aligned} b_{ij} &= a_{ij} - \sum_{k \in \text{pa}(j)} \lambda_{kj} a_{ik} \\ &= \sigma_{ij} - \sum_{\ell \in \text{pa}(i)} \lambda_{\ell i} \sigma_{\ell j} - \sum_{k \in \text{pa}(j)} \lambda_{kj} a_{ik}. \end{aligned} \quad (3.8)$$

Suppose that v is the first vertex that satisfies the quasi-linear condition. Then, we want to have a set of $m = |\text{pa}(v)|$ linear equations of the form

$$\sigma_{ij} - \sum_{p \in \text{pa}(v)} \lambda_{pv} \sigma_{hv} = 0$$

for some values of i, j, h , so we can solve for λ_{pv} for all $p \in \text{pa}(v)$.

We see that this can be achieved by picking m equations of the form $b_{vj} = 0$ (i.e. no bidirected edge between v and j) such that the last summation in (3.8) is zero. In other words, for all $k \in \text{pa}(j)$, we have $a_{vk} = 0$ (i.e. no half-trek from v to k).

Writing this in terms of graph properties, we want to find a vertex v and a set of vertices S such that:

1. $|S| = |\text{pa}(v)|$;
2. $S \cap (\text{sib}(v) \cup \{v\}) = \emptyset$; and
3. for each vertex $j \in S$, for all $k \in \text{pa}(j)$, there are no half-treks from v to k .

We can now define the quasi-linear properties above for all vertices recursively using the following algorithm.

Algorithm 3.1.**Input:** Mixed graph G .**Output:** Set of recursively quasi-linear vertices**Initialise:** $Q = \emptyset$;**while** $\exists v \in V \setminus Q$ such that we can find a set of vertices S_v with properties

1. $|S_v| = |\text{pa}(v)|$,
2. $S_v \cap (\text{sib}(v) \cup \{v\}) = \emptyset$ and
3. For each vertex $j \in S_v$, either $j \in Q$, or for all $k \in \text{pa}(j)$, there are no half-treks from v to k .

do| $Q = Q \cup \{v\}$ **end****return** Q .

Definition 3.5. Suppose G is a mixed graph. We say that a vertex v is *recursively quasi-linear in G* if it is contained in the output of Algorithm 3.1.

Lemma 3.7. Let $v \in V$ be an arbitrary vertex. Suppose $|S_v| = |\text{pa}(v)|$ and $S_v \cap (\text{sib}(v) \cup \{v\}) = \emptyset$. Then, for all $j \in S_v$, there are no half-treks from v to k for all $k \in \text{pa}(j)$ if and only if there are no half-treks from v to j .

Proof. ($\Leftarrow$) Suppose we have no half-treks from v to j . Since k is a parent of j , we will not have a half-trek from v to k since otherwise we can append $k \rightarrow j$ at the end of this half-trek to form a half-trek from v to j .

($\Rightarrow$) Suppose that there are no half-treks between v and the parents of j . Since $S_v \cap (\text{sib}(v) \cup \{v\}) = \emptyset$, v is not a sibling of j . Hence, there are no half-treks from v to j . $\square$

Theorem 3.8. Let G be a mixed graph. Then $\mathcal{M}(G)$ is quasi-linearly identifiable if and only if we have recursively defined that all vertices are recursively quasi-linear, i.e. if Algorithm 3.1 outputs V .

Proof. ($\Leftarrow$) It suffices to show that if v is recursively quasi-linear, then we can find $|\text{pa}(v)|$ equations that are linear in $\{\lambda_{iv} \mid i \in \text{pa}(v)\}$. Suppose that for each vertex v we can find such an S_v satisfying the conditions in Algorithm 3.1. From condition 2, for each $j \in S_v$, we have an equation of the form $b_{vj} = 0$. Condition

1 states that we can find $|\text{pa}(v)|$ such equations. Expanding all such equations $b_{vj} = 0$ into (3.8), if $j \in S$ is not recursively quasi-linear, condition 3 guarantees that the last summand in (3.8) is zero. If $j \in S$ is recursively quasi-linear, then by condition 3 the last summand in (3.8) is linear in $\lambda_{\ell i}$ with indeterminates in Σ and λ_{kj} .

($\Rightarrow$) Suppose that a vertex v is not recursively quasi-linear. If either of the first two conditions in Algorithm 3.1 fails, we no longer have $|\text{pa}(v)|$ linearly independent equations in any subset of $\{\lambda_{iv} \mid i \in \text{pa}(v)\}$. Hence, $\mathcal{M}(G)$ is not quasi-linearly identifiable. Now, if the third condition fails, then by definition we have at least one other vertex that is not recursively quasi-linear. So we must have at least one other vertex that fails to satisfy the first two conditions in Algorithm 3.1. $\square$

Example 3.9. In Example 3.5, vertex 1 is recursively quasi-linear trivially. Furthermore, we have $S_3 = \{1\}$ so the vertex 3 is also recursively quasi-linear. Moreover, $S_2 = \{3\}$, so vertex 2 is also recursively quasi-linear. Hence, the graph is quasi-linearly identifiable.

3.2.2 Properties of Linearly Identifiable Graphs

Previously, we found the necessary and sufficient conditions for a graph to be quasi-linearly identifiable. The following Theorem will determine when a quasi-linearly identifiable graph is also linearly identifiable.

Theorem 3.10. *Let G be a mixed graph such that $\mathcal{M}(G)$ is quasi-linearly identifiable. Then $\mathcal{M}(G)$ is linearly identifiable if and only if for each vertex v , we have a set of vertices S_v satisfying the half-trek criterion with respect to v in addition to the three conditions in Algorithm 3.1.*

Proof. Since $\mathcal{M}(G)$ is quasi-linearly identifiable, we have a system of $|\text{pa}(v)|$ linear equations in $\{\lambda_{iv} \mid i \in \text{pa}(v)\}$ for each vertex v . We can rewrite the equations $b_{sv} = a_{sv} - \sum_{\ell \in \text{pa}(v)} \lambda_{\ell v} a_{s\ell} = 0$ for some choice of $s \in S_v$ into the following matrix

equation

$$\begin{bmatrix} a_{s_1 p_1} & \cdots & a_{s_1 p_m} \\ \vdots & & \vdots \\ a_{s_m p_1} & \cdots & a_{s_m p_m} \end{bmatrix} \begin{bmatrix} \lambda_{p_1 v} \\ \vdots \\ \lambda_{p_m v} \end{bmatrix} = \begin{bmatrix} a_{s_1 v} \\ \vdots \\ a_{s_m v} \end{bmatrix}$$

where $S_v = \{s_1, \dots, s_m\}$ and $\text{pa}(v) = \{p_1, \dots, p_m\}$. Note that the leftmost matrix is precisely $A_{S_v, \text{pa}(v)}$. Hence the system of equations are linearly independent if and only if $A_{S_v, \text{pa}(v)}$ is of full rank.

Now, if $A_{S_v, \text{pa}(v)}$ has full rank, by Lemma 3 of [Foygel et al. \(2012\)](#), S_v satisfies the half-trek criterion. On the other hand, if S_v satisfies the half-trek criterion, then we have a system of half-treks with no sided intersection from S_v to $\text{pa}(v)$. Since each entry of $A_{S_v, \text{pa}(v)}$ is a polynomial obtained from summing all the covariances for half-trek between $s_i \in S_v$ and $p_j \in \text{pa}(v)$, $A_{S_v, \text{pa}(v)}$ has full rank. $\square$

Example 3.11. In Example 3.5, we have that $\mathcal{M}(G)$ is quasi-linearly identifiable. However, $\mathcal{M}(G)$ is not linearly identifiable since $S_2 = \{3\}$ but there are no half-treks from 3 to 2, therefore S_2 does not satisfy the half-trek criterion with respect to 2.

We now have the following necessary and sufficient conditions for linear identifiability.

Definition 3.6. We say that a vertex v satisfies the *linear identifiability criterion* if there exists a set of vertices S_v such that:

- S_v satisfies the half-trek criterion with respect to v ;
- for each vertex $j \in S_v$, either j satisfies the linear identifiability criterion or there are no half-treks from v to j .

Proposition 3.12. *Let G be a mixed graph. Then $\mathcal{M}(G)$ is linearly identifiable if and only if we can recursively check that all vertices in G satisfy the linear identifiability criterion.*

Proof. From Lemma 3.7 and Theorem 3.8, a graphical model is quasi-linearly identifiable if and only if for each vertex $v \in V$, we can find a set of vertices S_v such that

- $|S_v| = |\text{pa}(v)|$,
- $S_v \cap (\text{sib}(v) \cup \{v\}) = \emptyset$ and
- for each vertex $j \in S_v$, either j is recursively quasi-linear or there are no half-treks from v to j .

From Theorem 3.10, a quasi-linearly identifiable graphical model is linearly identifiable if and only if the sets S_v defined earlier further satisfies the half-trek criterion with respect to v for each vertex $v \in V$. $\square$

3.2.3 Connection to HTC-identifiable Graphs

Note that from Definition 3.2, we can rewrite the definition of HTC-identifiability as follows.

Definition 3.7. Let G be an HTC-identifiable graph. Then, $\mathcal{M}(G)$ is HTC-identifiable if and only if for each vertex v , there is a total ordering $\prec$ such that we can find a vertex set S_v satisfying both

1. S_v satisfies the half-trek criterion with respect to v , and
2. For each vertex $j \in S_v$, either $j \prec v$ or there are no half-treks from v to j .

Theorem 3.13. For any mixed graph G , $\mathcal{M}(G)$ is linearly identifiable if and only if $\mathcal{M}(G)$ is HTC-identifiable.

Proof. ($\Rightarrow$) Define a total ordering on the vertex set where $j \prec v$ if we defined j to satisfy the linear identifiability criterion before v in the recursion. Hence, the conditions in Definition 3.6 are the same as those in Definition 3.7.

($\Leftarrow$) We shall proceed by induction. First, let 1 be the first vertex in the total ordering $\prec$. Then for each vertex $j \in S_1$, there are no half-treks from 1 to j , hence, 1 satisfies the linear identifiability criterion.

Now, suppose that vertices $1 \prec \dots \prec k$ satisfy the linear identifiability criterion. We want to show that the $k+1^{\text{st}}$ vertex in the total ordering $\prec$ also satisfies the linear identifiability criterion. Since $\mathcal{M}(G)$ is HTC-identifiable, there exists some S_{k+1} satisfying the half-trek criterion with respect to $k+1$ and if $j \in S_{k+1}$,

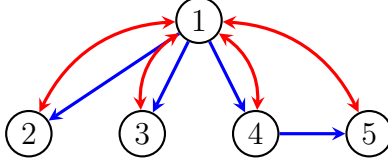


Figure 3.3: A generically identifiable graph that is not HTC-identifiable.

either $j \prec k + 1$ or there are no half-treks from v to j . But if $j \prec k + 1$, by our induction hypothesis, j satisfies the linear identifiability criterion. Hence, $k + 1$ also satisfies the linear identifiability criterion. $\square$

Linear identifiability can be used to explain the algebraic properties of HTC-identifiable graphs and graphs that are not HTC-identifiable.

Example 3.14. The graph in Figure 3.3 that is modified from Example 3.4 with an additional vertex added. This graph is not linearly identifiable as we cannot find an S_v satisfying both conditions of Definition 3.6 for any vertex v except for $v = \{1\}$ as we can take $S_1 = \emptyset$. Hence, it is not HTC-identifiable. Suppose we have

$$\Lambda = \begin{bmatrix} 0 & 0.25 & 0.5 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \Omega = \begin{bmatrix} 1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 1 & 0 & 0 & 0 \\ 0.1 & 0 & 1 & 0 & 0 \\ 0.1 & 0 & 0 & 1 & 0 \\ 0.1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

As we have seen in Example 3.4, equating $b_{32} = b_{42} = b_{43} = 0$, we obtain a quadratic equation in λ_{12} . Solving this, we obtain $\lambda_{12} = 0.25$ or 0.45 . If we equate $b_{32} = b_{52} = b_{53} = 0$ and solve the corresponding quadratic equation, we also obtain $\lambda_{12} = 0.25$ or 0.45 . However, equating $b_{42} = b_{52} = b_{54} = 0$ and solving the corresponding quadratic equation, we obtain $\lambda_{12} = 0.25$ or 0.354 . Since only $\lambda_{12} = 0.25$ is a common solution, this graph is generically identifiable, even though it is not linearly identifiable.

However, graphs that are not HTC-identifiable may be decomposed, using Tian's decomposition (Tian and Pearl, 2002a), into mixed components that are HTC-identifiable. Hence, the parameters for each of the mixed components could be recovered using linear algebra.

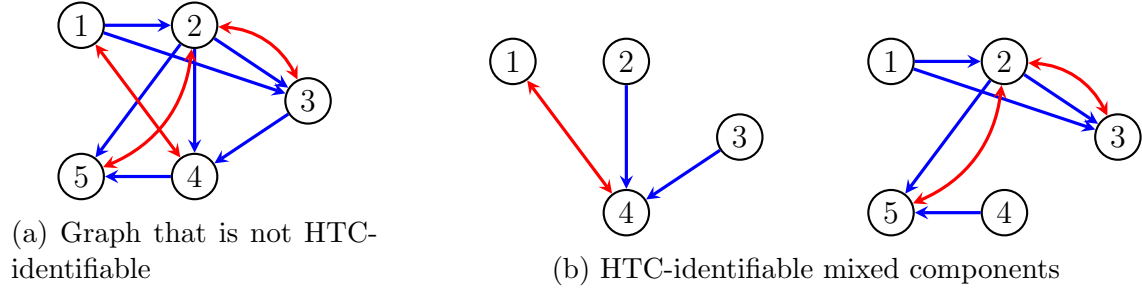


Figure 3.4: Graph that is not HTC-identifiable but can be decomposed into HTC-identifiable mixed components

Example 3.15. Consider the graph $G = (V, \mathcal{D}, \mathcal{B})$ in Figure 7(a) which is not HTC-identifiable. Suppose we decompose G into its mixed components $G_1 = (V_1, \mathcal{D}_1, \mathcal{B}_1)$ and $G_2 = (V_2, \mathcal{D}_2, \mathcal{B}_2)$ shown in Figure 7(b). We see that both G_1 and G_2 are HTC-identifiable. Hence, we are able to use linear algebra to recover the parameters Λ_1 and Λ_2 of G_1 and G_2 respectively. However, for a directed edge $i \rightarrow j \in \mathcal{D}$, the value of λ_{ij} is identical to that of $[\Lambda_k]_{ij}$ for which $i \rightarrow j \in \mathcal{D}_k$. So even if G is not HTC-identifiable, its parameters Λ could be recovered using linear algebra operations, but not from the system of equations $\{b_{ij} = 0 \mid i \leftrightarrow j \notin \mathcal{B}\}$.

3.3 Constraints in Linearly Identifiable Graphs

In this section, we will demonstrate some applications of our result by exploiting the algebraic properties of HTC-identifiable graphs to compute Λ and the *model ideal* $J(G)$. While algorithms for HTC-identifiable graphs for finding Λ already exist (Foygel et al., 2012), our argument will be purely algebraic in nature. We will also prove that the generators we find for a model ideal are minimal for some subset of HTC-identifiable graphs.

3.3.1 Generators of a Model Ideal

Let v be a vertex in a linearly identifiable graph G . Suppose $S_v = \{s_1, \dots, s_k\}$ and $\text{pa}(v) = \{p_1, \dots, p_k\}$. Then,

$$b_{s_i v} = a_{s_i v} - \sum_{\ell \in \text{pa}(v)} \lambda_{\ell v} a_{s_i \ell} = 0$$

for each $1 \leq i \leq k$. We can rewrite this into a matrix

$$\begin{bmatrix} a_{s_1 p_1} & \dots & a_{s_1 p_k} \\ \vdots & & \vdots \\ a_{s_k p_1} & \dots & a_{s_k p_k} \end{bmatrix} \begin{bmatrix} \lambda_{p_1 v} \\ \vdots \\ \lambda_{p_k v} \end{bmatrix} = \begin{bmatrix} a_{s_1 v} \\ \vdots \\ a_{s_k v} \end{bmatrix}. \quad (3.9)$$

By definition of linear identifiability, each of the a_{ij} can be expressed solely in terms of Σ , and the square matrix on the left is invertible. Since linearly identifiable graphs are also HTC-identifiable, we have a total ordering on the vertices of the graph. This leads us to an algorithm to recover Λ , Ω and the generators for a model ideal $J(G)$.

Algorithm 3.2.

Input: HTC-identifiable graph G , the associated sets $\{S_v \mid v \in V\}$ satisfying the half-trek criterion wrt $v \in V$ and the total ordering $1 \prec \dots \prec |V|$.

Output: Symbolic values of all regression coefficients as a matrix $\Lambda = (\lambda_{ij})$.

Initialise: $v = 1$;

while $v \preceq |V|$ **do**

for $s_1, \dots, s_k \in S_v$ and $p_1, \dots, p_k \in \text{pa}(v)$ **do**

 Solve for $\lambda_{p_i v}$ for all $1 \leq i \leq k$ using the equation:

$$\begin{bmatrix} \lambda_{p_1 v} \\ \vdots \\ \lambda_{p_k v} \end{bmatrix} = \begin{bmatrix} a_{s_1 p_1} & \dots & a_{s_1 p_k} \\ \vdots & & \vdots \\ a_{s_k p_1} & \dots & a_{s_k p_k} \end{bmatrix}^{-1} \begin{bmatrix} a_{s_1 v} \\ \vdots \\ a_{s_k v} \end{bmatrix},$$

 where $a_{ij} = \sigma_{ij} - \sum_{\ell \in \text{pa}(i)} \lambda_{\ell i} \sigma_{\ell j}$.

end

$v = v + 1$.

end

return $\Lambda = (\lambda_{ij})$.

Given Σ , Algorithm 3.2 could also be used to compute the values of Λ numerically. After obtaining either the numeric or the symbolic values of Λ , we can compute the values for Ω using the equation

$$\Omega = (I - \Lambda)^T \Sigma (I - \Lambda). \quad (3.10)$$

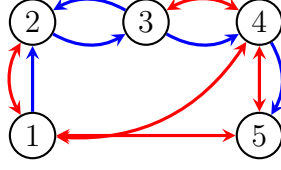


Figure 3.5: An HTC-identifiable graph with cycles.

If both Λ and Ω were computed symbolically, we could proceed to find the generators a model ideal.

Example 3.16. Consider the graph in Figure 3.5 with the total ordering $1 \prec 3 \prec 5 \prec 2 \prec 4$ and sets $S_1 = \emptyset, S_3 = \{1\}, S_5 = \{3\}, S_2 = \{3, 5\}, S_4 = \{2\}$. Note that these sets satisfy the conditions in Definition 3.7.

Applying Algorithm 3.2, for the first vertex, there is nothing to do as $\text{pa}(1) = \emptyset$ and hence we have no λ 's to solve for. For the next vertex, we have $\lambda_{23} = a_{12}^{-1}a_{13} = \sigma_{12}^{-1}\sigma_{13}$. Similarly, $\lambda_{45} = a_{34}^{-1}a_{35} = (\sigma_{34} - \lambda_{23}\sigma_{24})/(\sigma_{35} - \lambda_{23}\sigma_{25}) = (\sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24})/(\sigma_{12}\sigma_{35} - \sigma_{13}\sigma_{25})$ and so forth.

Theorem 3.17. Suppose that Λ is a matrix of rational functions with indeterminates Σ found by Algorithm 3.2. For an acyclic HTC-identifiable graph G ,

$$\left\langle [(I - \Lambda)^T \Sigma (I - \Lambda)]_{ij} \mid i \leftrightarrow j \notin \mathcal{B}, i \notin S_j \text{ and } j \notin S_i \right\rangle$$

generate a model ideal. For a cyclic HTC-identifiable graph, a model ideal can be found by saturating the above ideal by $\det(I - \Lambda)$.

Proof. Equating entries in (3.3), the equality constraints are generated by

$$J(G) = \left\langle [(I - \Lambda)^T \Sigma (I - \Lambda)]_{ij} \mid i \leftrightarrow j \notin \mathcal{B} \right\rangle.$$

We claim that if either $i \in S_j$ or $j \in S_i$, $[(I - \Lambda)^T \Sigma (I - \Lambda)]_{ij}$ is the zero polynomial. From Algorithm 3.2, we used the system of equations $\{b_{sv} = 0 \mid s \in S_v\}$ to solve for λ_{pv} for all $p \in \text{pa}(v)$. Hence if either $i \in S_j$ or $j \in S_i$, substituting these λ 's back into the polynomial b_{ij} will result in a zero polynomial. Note that for all other choices of (i, j) , substituting Λ will not result in b_{ij} being a zero polynomial as the structural equations are linearly independent.

In cyclic graphs, we remove the infeasible solution $\det(I - \Lambda) = 0$ by saturating the ideal (see section 8 of Drton (2018)). $\square$

Theorem 3.17 could be interpreted as having each missing bidirected edge being either a part of the system of equations used in Algorithm 3.2 to solve for Λ , or is a generator for a model ideal $J(G)$. In particular, our choice for $J(G)$ is generated by $\binom{|V|}{2} - |\mathcal{D}| - |\mathcal{B}|$ polynomials. Note that the result of Theorem 3.17 is very similar to (van Ommen and Mooij, 2017, Theorem 1) where it was proven that $b_{ij} \in \mathcal{I}(G)$ if $i \leftrightarrow j \notin \mathcal{B}$, $i \notin S_j$ and $j \notin S_i$. However, in Theorem 3.17, we showed that these polynomials in fact generate a model ideal.

Example 3.18. Let G be the Verma graph in Figure 1.3, we have $S_v = \text{pa}(v)$ for all vertex v and

$$\Lambda = \begin{bmatrix} 0 & \lambda_{12} & \lambda_{13} & 0 \\ 0 & 0 & \lambda_{23} & 0 \\ 0 & 0 & 0 & \lambda_{34} \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \Omega = \begin{bmatrix} \omega_{11} & 0 & 0 & 0 \\ 0 & \omega_{22} & 0 & \omega_{24} \\ 0 & 0 & \omega_{33} & 0 \\ 0 & \omega_{24} & 0 & \omega_{44} \end{bmatrix}.$$

In the first iteration of our algorithm, we obtain $\lambda_{12} = \sigma_{11}^{-1} \sigma_{12}$. In the second iteration, we have

$$\begin{bmatrix} \lambda_{13} \\ \lambda_{23} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}$$

where $a_{ij} = \sigma_{ij} - \sum_{k \in \text{pa}(i)} \lambda_{ki} \sigma_{kj}$. Solving this matrix equation, we obtain

$$\lambda_{13} = \frac{\sigma_{13 \cdot 2}}{\sigma_{11 \cdot 2}} \quad \text{and} \quad \lambda_{23} = \frac{\sigma_{23 \cdot 1}}{\sigma_{22 \cdot 1}}$$

where the conditional covariance matrix of a multivariate normal distribution is given by

$$\Sigma_{AC \cdot B} = \Sigma_{AC} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BC}.$$

In the last iteration, we have $\lambda_{34} = a_{33}^{-1} a_{34} = \sigma_{34 \cdot 12} / \sigma_{33 \cdot 12}$. Note that we could have computed Ω symbolically by equation (3.3). Finally, since the only pair of non-adjacent vertices are 1 and 4, we have

$$\begin{aligned} J(G) = \langle b_{41} \rangle &= \left\langle \sigma_{14} - \sum_{k=1}^3 \lambda_{k4} \sigma_{1k} \right\rangle = \langle \sigma_{14} - \lambda_{34} \sigma_{13} \rangle \\ &= \langle \sigma_{11} \sigma_{13} \sigma_{22} \sigma_{34} - \sigma_{12}^2 \sigma_{13} \sigma_{34} - \sigma_{11} \sigma_{14} \sigma_{22} \sigma_{33} + \sigma_{12}^2 \sigma_{14} \sigma_{33} - \\ &\quad \sigma_{11} \sigma_{13} \sigma_{23} \sigma_{24} + \sigma_{11} \sigma_{14} \sigma_{23}^2 + \sigma_{12} \sigma_{13}^2 \sigma_{24} - \sigma_{12} \sigma_{13} \sigma_{14} \sigma_{23} \rangle \end{aligned}$$

which is exactly what we have in Example 1.4. Note that the second to last equality comes from the fact that both $\lambda_{14} = \lambda_{24} = 0$.

3.3.2 Minimality of the Generators

Now that we can find the generators of $J(G)$ for any HTC-identifiable G , it is natural to ask if generators of the model ideal found in Theorem 3.17 are in fact the minimal generators in the following sense:

Definition 3.8. We say that the polynomials $f_1, \dots, f_n$ are *minimal generators* for I if we have $\langle f_1, \dots, f_n \rangle = I$ but $\langle f_1, \dots, f_{i-1}, f_{i+1}, \dots, f_n \rangle \neq I$ for any $1 \leq i \leq n$. In other words, the polynomials $f_1, \dots, f_n$ generate I but if we remove any of the f_i for $1 \leq i \leq n$, the remaining polynomials no longer generate I .

First, we will define a subset of HTC-identifiability:

Definition 3.9. The sets $\{S_v \mid v \in V\}$ in an HTC-identifiable graph has *no subset cycles* if it satisfies the criterion in Definition 3.7 and there do not exist vertices $v_1, \dots, v_n$ with $v_1 \in S_{v_2}, v_2 \in S_{v_3}, \dots, v_n \in S_{v_1}$.

Lemma 3.19. Suppose that we have an HTC-identifiable graph $G = (V, \mathcal{D}, \mathcal{B})$ with a collection of sets $\{S_v \mid v \in V\}$ that has no subset cycles. Further suppose that there exist vertices $i, j \in V$ such that $i \leftrightarrow j \notin \mathcal{B}, i \notin S_j$ and $j \notin S_i$. Then, we can add the bidirected edge $i \leftrightarrow j$ to obtain another HTC-identifiable graph with the sets $\{S_v \mid v \in V\}$ unchanged.

Proof. Firstly, we check that each S_v still satisfies the half-trek criterion with respect to v .

1. $|S_v| = |\text{pa}(v)|$ still holds as adding bidirected edges does not impact the number of parents of any vertex v while each S_v remains unchanged.
2. Since $S_i \cap (\{i\} \cup \text{sib}(i)) = \emptyset$, adding the bidirected edge $i \leftrightarrow j$ where $j \notin S_i$ preserves that equality. By symmetry, $S_j \cap (\{j\} \cup \text{sib}(j)) = \emptyset$. For any other vertices w , the equation $S_w \cap (\{w\} \cup \text{sib}(w)) = \emptyset$ is preserved as $\text{sib}(w)$ and S_w remains unchanged.
3. There still exists a system of half-treks with no sided intersection from S_v to $\text{pa}(v)$, as adding the bidirected edge does not impact half-treks.

Recall that a graph is HTC-identifiable if S_v satisfies the half-trek criterion with respect to each vertex $v \in V$ and $w \prec v$ whenever $w \in S_v \cap \text{htr}(v)$. For the second condition, by definition by the total ordering in HTC-identifiable graphs, it is equivalent to view the total ordering as a topological ordering $<$ on the directed graph $G' = (V', \mathcal{D}', \mathcal{B}')$ with $V' = V$, $\mathcal{D}' = \{w \rightarrow v \mid w \in S_v \cap \text{htr}(v)\}$, $\mathcal{B}' = \emptyset$ since $w \prec v$ in G if and only if $w < v$ in G' . If the sets S_v have no subset cycles, G' is a DAG and a topological ordering exists. Therefore, leaving S_v unchanged, the new graph satisfies both conditions of HTC-identifiability. $\square$

Theorem 3.20. *If G is an HTC-identifiable graph with no subset cycles, then the generators of the model ideal $J(G)$ found in Theorem 3.17 are minimal.*

Proof. Suppose that $J(G) = \langle f_1, \dots, f_n \rangle$ is not a minimal generator of $J(G)$. Without loss of generality, suppose that the constraint f_1 corresponding to the missing edge between v and w is redundant such that $J(G) = \langle f_2, \dots, f_n \rangle$. Now, consider the graph G' obtained by adding the bidirected edge $v \leftrightarrow w$ to G . Note that G' is also an HTC-identifiable graph by Lemma 3.19, with $J(G') = \langle f_2, \dots, f_n \rangle$.

However, since the constraint f_1 can be generated by polynomials generating $J(G')$, the bidirected edge $v \leftrightarrow w$ in G' can only take one value $\omega_{vw} = 0$, which is a contradiction. $\square$

Corollary 3.21. *If G is a bow-free acyclic graph, then the generators of the model ideal $J(G)$ found in Theorem 3.17 are minimal.*

Proof. In a bow-free acyclic graph, we have $S_v = \text{pa}(v)$. Since the graph is acyclic, these sets contain no subset cycles. The result follows from Theorem 3.20. $\square$

Example 3.22. If we have an HTC-identifiable graph that does not satisfy the conditions of Lemma 3.19, it is possible that it is no longer HTC-identifiable if we add a bidirected edge.

Consider the graph in Figure 3.6. Before adding the bidirected edge $1 \leftrightarrow 4$, we must have $S_5 = \{3\}$. Now, for S_3 , we can have either $S_3 = \{1\}$ or $S_3 = \{5\}$. However, if we pick the latter, we have $3 \in S_5$ and $5 \in S_3$ which is HTC-nonidentifiable (Foygel et al., 2012, Theorem 2). Similarly, we have either $S_1 = \{3\}$ or $S_1 = \{5\}$ but by picking the former we have $3 \in S_1$ and $1 \in S_3$ which is HTC-nonidentifiable. Hence, the only choice is $S_1 = \{5\}$, $S_3 = \{1\}$ and $S_5 = \{3\}$.

first solve a linear equation with at most one unknown, then two linear equations with at most two unknowns and so on until we solved $|V| - 1$ linear equations with at most $|V| - 1$ unknowns. Since we this process is iterated $|V|$ times, the complexity of finding the values of λ_{ij} 's is at most $O(|V|^4)$. Finally, since the complexity of matrix multiplication is $O(|V|^3)$, we can compute both Λ and Ω in $O(|V|^4)$. $\square$

In particular, using our algorithm, the bound of $O(|V|^4)$ is attained by a complete DAG. However, there are faster methods to handle complete DAGs such as Cholesky decomposition. Furthermore, in graphs where S_v is known, Proposition 3.24 could be useful. Otherwise, if S_v is unknown, the complexity is at most $O(|V|^2(|V| + r)^3)$.

One might notice that each unknown λ_{ij} corresponds precisely to a directed edge in G . In a sparse graph where $|\mathcal{D}|$ is small, it might be beneficial to express the complexity in terms of $|\mathcal{D}|$ instead.

Proposition 3.25. *Suppose G is HTC-identifiable and S_v is known for each $v \in V$. Then the complexity of finding Λ numerically is at most $O(|\mathcal{D}|^3)$.*

Proof. This follows from the fact that we have at least $|\mathcal{D}|$ equations with exactly $|\mathcal{D}|$ unknowns. Observe that each time when we apply the Gauss-Jordan elimination, we can solve for any subset of the $|\mathcal{D}|$ unknowns. Therefore, we obtain the following optimization problem:

$$\begin{aligned} & \underset{k_1, \dots, k_{|\mathcal{D}|}}{\text{maximise}} && k_1^3 + \dots + k_{|\mathcal{D}|}^3 \\ & \text{subject to} && k_1 + \dots + k_{|\mathcal{D}|} = |\mathcal{D}| \\ & \text{where} && k_1, \dots, k_{|\mathcal{D}|} \in \mathbb{Z}_{\geq 0}, \end{aligned}$$

where k_i is the number of unknowns in the i^{th} set of equations. Since the cubic term grows faster than the linear term, the maximum is obtained when $k_i = |\mathcal{D}|$ for some i and $k_j = 0$ for all $j \neq i$. $\square$

Corollary 3.26. *Suppose G is an HTC-identifiable graph and S_v is known for each $v \in V$. Denote $P = \max_{v \in V} |\text{pa}(v)|$. Then the complexity of finding Λ numerically is at most $O(|V|P^3)$.*

Proof. We use the Gauss-Jordan elimination at most v times, solving for at most P unknowns each time. $\square$

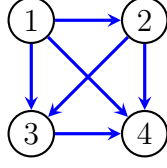


Figure 3.7: A complete DAG on 4 vertices.

Symbolic Complications

Now, we will show that our algorithm does not run symbolic calculations in polynomial time by providing a simple counter-example. Since finding the generators of $J(G)$ requires symbolic computation, we can not compute $J(G)$ in polynomial time in general.

Consider a complete DAG such as the graph shown in Figure 3.7. Using our algorithm, solving for Λ symbolically, we obtain

$$\lambda_{14} = \sigma_{14 \cdot 23} / \sigma_{11 \cdot 23}, \quad \lambda_{24} = \sigma_{24 \cdot 13} / \sigma_{22 \cdot 13}, \quad \lambda_{34} = \sigma_{34 \cdot 12} / \sigma_{33 \cdot 12}.$$

In particular, the regression coefficient λ_{ij} is precisely $\sigma_{ij \cdot S} / \sigma_{ii \cdot S}$ in a DAG, where $S = \text{pa}(j) \setminus \{i\}$ (Pearl, 2000). Recall that the conditional covariance matrix of a multivariate normal distribution is given by

$$\Sigma_{AA \cdot B} = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA}.$$

Now, there are $|B|!$ different terms in $\det(\Sigma_{BB})$. Since $\sigma_{ij \cdot V} \propto \det(\Sigma_{-i, -j})$ which has $(|V| - 1)!$ terms, the number of terms in λ_{ij} increases factorially with respect to $|\text{pa}(i)|$. As Algorithm 3.2 requires us to invert matrices with entries as functions of λ , we might not be able to solve for Λ symbolically in polynomial time. However, the expression may in principle factorise, and therefore we might only have to deal with $O(\log(n!)) = O(n \log(n))$ terms.

3.4 Conclusion

In this paper, we have defined a subset of generically identifiable graphs called *linearly identifiable graphs*. Graphs are linearly identifiable if their model parameters can be recovered from the covariance matrix with straightforward linear algebra

operations, given in Algorithm 3.2 and (3.10). We have also proven that a graph is linearly identifiable if and only if it is HTC-identifiable.

Furthermore, we have shown that, for some graphs, there is a bijection between the generators of the model ideal, generated by all equality constraints of $\mathcal{M}(G) \cap PD_V$, and the vertex pairs $\{(i, j) \mid i \leftrightarrow j \notin \mathcal{B}, i \notin S_j \text{ and } j \notin S_i\}$. In graphs where the collection of sets $\{S_v \mid v \in V\}$ has no subset cycles, the generators of the vertex pairs are minimal. To prove this, we utilise Lemma 3.19, which fails to hold in all HTC-identifiable graphs as demonstrated in Example 3.22. However, we do conjecture that the generators we found are minimal for all HTC-identifiable graphs.

Conjecture 3.1. *For any HTC-identifiable graph G , the generators of the model ideal $J(G)$ found in Theorem 3.17 are minimal.*

Finally, we have shown that the complexity to recover the parameters of an HTC-identifiable graph G , given that we know $\{S_v \mid v \in V\}$ is at most $O(|V|^4)$. Meanwhile, it is known that the complexity of finding S_v for each vertex v given an HTC-identifiable graph is at most $O(|V|^2(|V| + r)^3)$ where $r \leq |\mathcal{D}|/2$ is the number of reciprocal edge pairs in $\mathcal{D}$. Unfortunately, we do not have an efficient algorithm to find the sets S_v without subset cycles, if such a collection of sets exists.

Chapter 4

Maximum Likelihood Estimation for HTC-identifiable Graphs

Let G be a HTC-identifiable graph and X be an i.i.d. dataset of size n generated from $N(\mathbf{0}, \Sigma)$. Suppose we want to test $H_0 : \Sigma \in \mathcal{M}(G)$ against the general hypothesis $H_1 : \Sigma \in PD_V$. One simple method is to test if each generator of the constraint $J(G)$ found in the previous chapter indeed evaluates to zero. This could be done by bootstrapping the dataset X to generate multiple sample covariance matrices and construct a confidence interval.

An alternative method is to use a likelihood ratio test. This involves finding the MLE of Λ and Ω in a model given by an HTC-identifiable graph G . In the previous chapter, we are able to use Algorithm 3.2 and equation (3.10) to recover parameters Λ and Ω symbolically from Σ . Now, suppose we are further given a dataset $X \stackrel{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma)$, we could replace the input of Algorithm 3.2 from Σ to the sample covariance matrix S . Now, given a sample covariance matrix S , we would like to find the *maximum likelihood estimates* (MLEs) for the parameters Λ and Ω .

4.1 Introduction

4.1.1 Preliminaries

In this chapter, we address the problem of testing $H_0 : \Sigma \in \mathcal{M}(G)$ against the general hypothesis $H_1 : \Sigma \in PD_V$.

If the generators of the constraint $J(G)$ of the HTC-identifiable graph G are known, we could simply use bootstrap methods (Efron, 1979) to construct a confidence interval for each generator and test if zero is in all the confidence intervals. However, as we have seen in the previous chapter, finding the generators of the constraints $J(G)$ symbolically is computationally expensive. Hence, an alternative method for testing if $\Sigma \in \mathcal{M}(G)$ is desirable.

One alternative method is to use the likelihood ratio test. Recall that the likelihood ratio statistic is

$$2 \left(\sup_{\Sigma \in PD_V} \ell(\Sigma) - \sup_{\Sigma \in \mathcal{M}(G)} \ell(\Sigma) \right),$$

where ℓ is the log-likelihood. Hence, in order to utilise the likelihood ratio test, we need to first compute the MLEs of Λ and Ω in linear SEMs represented by HTC-identifiable graphs. There have been several algorithms that solve this problem using quasi-Newton methods for nonlinear optimisation. These include R packages such as `sem` (Fox, 2006) and `lavaan` (Rosseel, 2012), in addition to the commercial software by Narayanan (2012). However, these methods often face problems with convergence and often requires a careful choice of starting values (Steiger, 2001). This is particularly problematic when the graphical model G is a poor fit for the given dataset X (Drton et al., 2009).

An alternative method for finding MLEs in linear SEMs is by cycling through all the considered variables, updating the parameters related to those variables at each step. This algorithm was first proposed by Chaudhuri et al. (2007) and was known as *Iterative Conditional Fitting* (ICF), aimed at computing the MLE of a constraint covariance matrix with no directed edges (i.e. $\Lambda = \mathbf{0}$ and some entries of Ω are zero). Drton et al. (2009) later generalised the ICF algorithm to the case where $\Lambda \neq \mathbf{0}$. The resulting *Residual Iterative Conditional Fitting* (RICF) algorithm is in finding the MLEs of Λ and Ω for identifiable ADMGs. This result was later extended by Drton et al. (2019) who introduced the block-coordinate descent (BCD) algorithm which works on all identifiable mixed graphs (including cyclic graphs).

In this chapter, we will use a hill-climbing algorithm (also known as gradient ascent) to find the MLEs for Λ and Ω in HTC-identifiable graphs. We also address

some problems faced by the previous quasi-Newton methods, such as using Algorithm 3.2 to find a starting value and proving that the log-likelihood is concave for a constrained covariance matrix Σ under the assumption that we are working with a linear SEM.

4.1.2 Outline of this Chapter

This chapter is structured as follows: in Section 4.2, we assume that the generators of the constraints $J(G)$ are known. Here, we can simply use bootstrap to test if zero is in the confidence interval for each generator of $J(G)$. In Section 4.3, we consider the case where the constraints of G are not known. As finding the constraints of a mixed graphical model G is computationally hard, we shall devise a more efficient algorithm for finding the MLE of Σ . We will first prove the concavity of the log-likelihood function under the constraints of any linear SEM. We will then illustrate how a hill-climbing algorithm could be implemented. The code of this algorithm has been published on GitHub¹. In Section 4.4, we will discuss some limitations of our algorithm in finding the MLE, especially at points of unidentifiability, if the sample size is small, or if the true value of Σ has a small eigenvalue. Finally, we shall apply our algorithm to a real data example in Section 4.5.

4.2 Bootstrap

One way of testing $H_0 : \Sigma \in \mathcal{M}(G)$ against the general hypothesis $H_1 : \Sigma \in PD_V$ is to test if each generator of $J(G)$ evaluates to zero. In particular, given $J(G) = \langle f_1, \dots, f_n \rangle$, we want to test if $f_1 = 0, \dots, f_n = 0$. This could be tested using bootstrapping by generating multiple sample covariance matrices, constructing a $1 - \alpha$ confidence interval for each generator f_i and checking if zero is contained in the confidence interval.

In the previous chapter, we provided a way of finding the generators of a model ideal, $J(G)$ given any HTC-identifiable graph G . Given the generators of these constraints, we will show that we are able to use bootstrap to test for $\Sigma \in \mathcal{M}(G)$ in Example 4.1.

¹<https://github.com/bohaoyao/fitting-Gaussian-graphical-models>

4.2.1 Bootstrap Example

Suppose we have a dataset X generated from $N(\mathbf{0}, \Sigma)$. Given such a data set X and a graphical model G , we want to test $H_0 : \Sigma \in \mathcal{M}(G)$ against $H_1 : \Sigma \notin \mathcal{M}(G)$. In the next example, we shall test for H_0 given that we know a generating set of $J(G)$.

Example 4.1. Suppose we have four variables. We shall first fix Σ by picking the following values for Λ and Ω :

$$\Lambda = \begin{bmatrix} 0 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \Omega = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.25 \\ 0 & 0 & 1 & 0 \\ 0 & 0.25 & 0 & 1 \end{bmatrix}.$$

First, we generate a dataset X where we sample 1000 data points from $N(\mathbf{0}, \Sigma)$. Now, suppose that we only observe X (and not Σ) and want to test if Σ fits our Verma model $\mathcal{M}(G)$, where G is the Verma graph in Figure 1.3.

For each generator f_i of $J(G)$, we can use bootstrap to generate multiple sample covariance matrices and test whether $f_i = 0$ is in the $1 - \alpha$ confidence interval. In the Verma model, we only have one such polynomial f_{Verma} that generates $J(G)$. In our case, the 95% confidence interval of f_{Verma} is $(-0.125, 0.0675)$. Hence, we do not have sufficient evidence to reject H_0 .

Repeating the above simulation 1000 times where we generate a different dataset X at each iteration, we found that in 940 of the cases, 0 is contained in the 95% confidence interval.

On the other hand, we can also provide an example where Σ is not in the model. For instance, if we generate some $\Sigma \notin \mathcal{M}(G)$, where G is still the Verma graph in Figure 1.3, by changing the numerical values of Λ to be

$$\Lambda = \begin{bmatrix} 0 & 0.5 & 0.25 & 0.2 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Clearly, $\Sigma \notin \mathcal{M}(G)$ since there is a missing edge between 1 and 4 but $\lambda_{14} \neq 0$. Simulating from this covariance matrix, we obtain that the 95% confidence interval

for λ_{14} is $(-0.409, -0.179)$. In particular, 0 is no longer in the confidence interval. This is expected since our dataset X is now generated from $N(\mathbf{0}, \Sigma)$ for some $\Sigma \notin \mathcal{M}(G)$. Hence, we have strong evidence to reject H_0 . In fact, repeating this simulation 1000 times as before, we see that 0 is not contained in the 95% confidence interval for any of those cases.

4.3 Maximum Likelihood Estimation

While the bootstrap method as seen in Example 4.1 does allow us to test for $\Sigma \in \mathcal{M}(G)$, it requires us to have prior knowledge of the set $J(G)$ and run the bootstrap for each polynomial $f \in J(G)$. This could be extremely time-consuming. In this section, we shall first fit a dataset X to a given graphical model G using maximum likelihood.

In the previous chapter, we have applied Algorithm 3.2 to recover Λ symbolically from Σ . Substituting into equation (3.10), we are able to recover Ω as well. However, when we are given a sample covariance S , we could replace the input of Algorithm 3.2 from Σ to S . Together with the equation (3.10), we are able to provide numeric estimates of Λ and Ω , which we shall denote $\tilde{\Lambda}$ and $\tilde{\Omega}$. We will then use hill-climbing to find the MLEs $\hat{\Lambda}$, $\hat{\Omega}$ and $\hat{\Sigma}$. Finally, we can use the likelihood ratio statistic to test $H_0 : \Sigma \in \mathcal{M}(G)$ against the general hypothesis $H_1 : \Sigma \in PD_V$.

4.3.1 Starting Point

First, to run the hill-climbing algorithm, we need a starting point for Λ and Ω . We can obtain a starting point for Λ , denoted $\tilde{\Lambda}$, from applying the sample covariance matrix S to Algorithm 3.2. In particular, $\tilde{\Lambda} \in \mathbb{R}^{\mathcal{D}}$. Now, from $\tilde{\Lambda}$, we can also obtain a starting point for Ω , denoted $\tilde{\Omega}$ using the equations from Section 3:

$$\tilde{\omega}_{ij} = b_{ij} = a_{ij} - \sum_{\ell=1}^{j-1} \tilde{\lambda}_{\ell j} a_{i\ell}, \quad \text{where} \quad a_{ij} = S_{ij} - \sum_{\ell=1}^{i-1} \tilde{\lambda}_{\ell i} S_{\ell j}. \quad (4.1)$$

We will see in Section 4.4.2 that while $\tilde{\Omega}$ might not be positive definite, it is positive definite for a large enough sample size n .

We shall first prove that if $\Sigma \in \mathcal{M}(G)$, then $\tilde{\Lambda}$ and $\tilde{\Omega}$ are consistent estimators of Λ and Ω .

Proposition 4.2. *If $\Sigma \in \mathcal{M}(G)$ and both Λ and Ω are well-defined, then $\tilde{\Lambda}$ and $\tilde{\Omega}$ are $\sqrt{n}$ -consistent estimators of Λ and Ω respectively.*

Proof. Since S is a consistent estimator of Σ , S converges to Σ in probability. The symbolic values of both $\tilde{\Lambda}$ and $\tilde{\Omega}$ are obtained by applying Algorithm 3.2 to S and Σ respectively. In the former application, $\tilde{\Lambda}$ is expressed symbolically as additions and multiplications of the random variables S_{ij} , each converging in probability to σ_{ij} . Since Λ is defined (i.e. no entry is divided by zero), $\tilde{\Lambda}$ converges to Λ in probability. Similarly, $\tilde{\Omega}$ is also expressed symbolically as additions and multiplications of $\tilde{\lambda}_{ij}$ and S_{ij} by (4.1), each converging in probability to λ_{ij} and σ_{ij} respectively. Hence, $\tilde{\Omega}$ converges to Ω in probability by Slutsky's theorem (Slutsky, 1925).

Now since S is $\sqrt{n}$ -consistent for Σ and the map from S to $\tilde{\Lambda}$ and $\tilde{\Omega}$ is differentiable and non-singular, by the Delta method, we get that $\tilde{\Lambda}$ and $\tilde{\Omega}$ are also $\sqrt{n}$ -consistent for Λ and Ω respectively. $\square$

Corollary 4.3. *If $\Sigma \in \mathcal{M}(G)$, then $\tilde{\Sigma} = (I - \tilde{\Lambda})^{-T} \tilde{\Omega} (I - \tilde{\Lambda})^{-1}$ is a $\sqrt{n}$ -consistent estimator for Σ .*

Hence, if $\Sigma \in \mathcal{M}(G)$, $\tilde{\Lambda}$ and $\tilde{\Omega}$ are good starting points for Λ and Ω for large sample sizes.

4.3.2 Iterative step

Now, for the iterative step, we want to find the derivative of the log likelihood $\ell(\Lambda, \Omega)$ with respect to Λ and Ω . The log-likelihood function for the covariance matrix Σ is given by

$$\ell(\Sigma) = -\frac{n}{2} \log \det \Sigma - \frac{n}{2} \text{tr}(S \Sigma^{-1}). \quad (4.2)$$

Now, in the iterative step, we want to maintain $\Sigma \in \mathcal{M}(G)$. To achieve this, we substitute $\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}$ into the log-likelihood function (4.2) to ensure that $\Sigma \in \mathcal{M}(G)$. The objective function hence becomes

$$\ell(\Lambda, \Omega) = -\frac{n}{2} \log \det \Omega - \frac{n}{2} \text{tr}(SK) \quad (4.3)$$

where $K = \Sigma^{-1} = (I - \Lambda)\Omega^{-1}(I - \Lambda)^T$. Differentiating, we obtain

$$\partial\ell(\Lambda, \Omega) = \langle 2S(I - \Lambda)\Omega^{-1}, \partial\Lambda \rangle + \langle \Omega^{-1}(I - \Lambda)^T S(I - \Lambda)\Omega^{-1} - \Omega^{-1}, \partial\Omega \rangle$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product with $\langle M_1, M_2 \rangle = \text{tr}(M_1^T M_2)$.

Hence, our partial derivatives are

$$\begin{aligned} \frac{\partial\ell}{\partial\Lambda} &= 2S(I - \Lambda)\Omega^{-1}, \\ \frac{\partial\ell}{\partial\Omega} &= \Omega^{-1}(I - \Lambda)^T S(I - \Lambda)\Omega^{-1} - \Omega^{-1}. \end{aligned}$$

Now, for $i = 1, 2, 3, \dots$ with $\Lambda^{(1)} = \tilde{\Lambda}$ and $\Omega^{(1)} = \tilde{\Omega}$, let

$$\begin{aligned} \Lambda^{(i+1)} &= \Lambda^{(i)} + \alpha \frac{\partial\ell}{\partial\Lambda} \\ \Omega^{(i+1)} &= \Omega^{(i)} + \alpha \frac{\partial\ell}{\partial\Omega} \end{aligned}$$

for some training rate α . We repeat this iterative step until convergence, that is, at some $i = t$ for which

$$\ell(\Lambda^{(t-1)}, \Omega^{(t-1)}) - \ell(\Lambda^{(t)}, \Omega^{(t)}) < \epsilon$$

for some convergence tolerance $\epsilon \geq 0$. The resulting output is $(\hat{\Lambda}, \hat{\Omega}) = (\Lambda^{(t)}, \Omega^{(t)})$.

4.3.3 Local Concavity of the Log-Likelihood Function

We shall now prove that $\hat{\Sigma} = (I - \hat{\Lambda})^{-T} \hat{\Omega} (I - \hat{\Lambda})^{-1}$ is in fact the MLE if $\Sigma \in \mathcal{M}(G)$ by proving the concavity of the log-likelihood function locally around the MLE. Furthermore, we will show that our starting point $\tilde{\Sigma}$ will lie in this concave neighbourhood almost surely as $n \rightarrow \infty$.

The strict concavity of the log-likelihood function for an unconstrained Gaussian model is well known (Zwiernik et al., 2017). We shall adapt the proof to show that this function is also locally strictly concave with high probability starting from a reasonable estimate in $\mathcal{M}(G)$.

Let $A \in \mathbb{R}^{\mathcal{D}}$ and B be a symmetric matrix. Differentiating (4.3), we have

$$\begin{aligned} \nabla\ell(\Lambda, \Omega) &= -\frac{n}{2}\text{tr}(\Omega^{-1}B) - \frac{n}{2}\text{tr}(S\nabla K) \\ \nabla^2\ell(\Lambda, \Omega) &= \frac{n}{2}\text{tr}(\Omega^{-1}B\Omega^{-1}B) - \frac{n}{2}\text{tr}(S\nabla^2 K) \end{aligned}$$

where we shorten $\nabla_{A \times B}$ to ∇ for the sake of cleanliness and

$$\begin{aligned}\nabla K &= -A\Omega^{-1}(I - \Lambda)^T - (I - \Lambda)\Omega^{-1}B\Omega^{-1}(I - \Lambda)^T - (I - \Lambda)\Omega^{-1}A^T \\ \nabla^2 K &= 2A\Omega^{-1}B\Omega^{-1}(I - \Lambda)^T + 2A\Omega^{-1}A^T + 2(I - \Lambda)\Omega^{-1}B\Omega^{-1}B\Omega^{-1}(I - \Lambda)^T + \\ &\quad 2(I - \Lambda)\Omega^{-1}B\Omega^{-1}A^T.\end{aligned}$$

Lemma 4.4. *Both ∇K and $\nabla^2 K$, are symmetric matrices.*

Proof. First, note that B is a symmetric matrix. Next, notice that the matrix

$$(I - \Lambda)\Omega^{-1}B\Omega^{-1}(I - \Lambda)^T = (I - \Lambda)\Omega^{-1}B[(I - \Lambda)\Omega^{-1}]^T$$

is the conjugate transpose of a symmetric matrix is hence also symmetric. Finally, the first and last terms of ∇K are transposes of each other. Since $C + C^T$ is a symmetric matrix for any matrix C , ∇K is a symmetric matrix. Using similar arguments, one can also deduce that $\nabla^2 K$, is symmetric. $\square$

Theorem 4.5. *If $\nabla K \neq \mathbf{0}$, then there exists an $\alpha > 1$ such that the log-likelihood function $\ell(\Sigma)$ is strictly concave on the region*

$$\{\Sigma \in \mathcal{M}(G) : 0 \prec \Sigma \prec \alpha S\}.$$

In particular, as $n \rightarrow \infty$, the probability that $\tilde{\Sigma}$ lies in the concave region converges to 1.

Proof. Note that the second derivative can be rewritten as

$$\begin{aligned}\nabla^2 \ell(\Lambda, \Omega) &= -\frac{n}{2} \text{tr}((\nabla K)^{1/2} K^{-1} (\nabla K)^{1/2} (\nabla K)^{1/2} K^{-1} (\nabla K)^{1/2}) + \\ &\quad \frac{n}{2} \text{tr}((\nabla^2 K)^{1/2} K^{-1} (\nabla^2 K)^{1/2}) - \frac{n}{2} \text{tr}((\nabla^2 K)^{1/2} S (\nabla^2 K)^{1/2}).\end{aligned}$$

Since both $K^{-1} = \Sigma$ and S are positive definite matrices and both ∇K and $\nabla^2 K$ are symmetric, each of the matrices $(\nabla K)^{1/2} \Sigma (\nabla K)^{1/2}$, $(\nabla^2 K)^{1/2} \Sigma (\nabla^2 K)^{1/2}$ and $(\nabla^2 K)^{1/2} S (\nabla^2 K)^{1/2}$ are positive definite. Therefore, we have both $\text{tr}(\Sigma \nabla^2 K) > 0$ and $\text{tr}(S \nabla^2 K) > 0$.

Furthermore, the matrix $U = (\nabla K)^{1/2} \Sigma (\nabla K)^{1/2}$ is positive definite and hence has positive eigenvalues. Since $\nabla K \neq \mathbf{0}$ and Σ is a full-rank (as it is positive definite), there is at least one strictly positive eigenvalue. Pick $\varepsilon = \text{tr}(U^2) > 0$.

Now, pick $\alpha > 1$ to be the largest such that if $\Sigma \prec \alpha S$, we have $\text{tr}(\Sigma \nabla^2 K) - \text{tr}(S \nabla^2 K) < \varepsilon$. Since S is a $\sqrt{n}$ -consistent estimator for Σ , hence, $\nabla^2 \ell < 0$ with a rate of $\sqrt{n}$ and therefore, ℓ is concave on this region.

Finally, notice that our choice of $\varepsilon > 0$ is not affected by the sample size n . Hence, as $n \rightarrow \infty$, we have $\text{tr}(\Sigma \nabla^2 K) - \text{tr}(S \nabla^2 K) < \varepsilon$ almost surely. $\square$

Corollary 4.6. *For a large enough sample size, we can find the supremum of $\ell(\Sigma)$ for $\Sigma \in \mathcal{M}(G)$ using a hill-climbing algorithm with the starting point $\tilde{\Sigma}$.*

Proof. Fix $\alpha > 1$ to be the constant in Theorem 4.5. For a large enough sample, we have that $0 \prec \Sigma \prec \alpha S$. Furthermore, in each iteration of the hill climbing, $\nabla K \neq 0$ as we will be terminating otherwise. The Corollary then follows from Theorem 4.5. $\square$

4.3.4 Hypothesis testing

Given $\Sigma \in \mathcal{M}(G)$, we are now able find the MLE of Σ by the hill-climbing algorithm through maximising Λ and Ω . We will now show that we can use a likelihood ratio test to decide if we should reject $H_0 : \Sigma \in \mathcal{M}(G)$. In this case, our test statistic is

$$2 \left(\sup_{\Sigma \in PD_V} \ell(\Sigma) - \sup_{\Sigma \in \mathcal{M}(G)} \ell(\Sigma) \right) = 2(\ell(S) - \ell(\hat{\Sigma}))$$

where $\hat{\Sigma}$ is the value found by our hill-climbing algorithm and S is the sample covariance.

Theorem 4.7 (Likelihood ratio statistic). *If $\Sigma \in \mathcal{M}(G)$, then*

$$2(\ell(S) - \ell(\hat{\Sigma})) \xrightarrow{D} \chi_p^2$$

where $p = \dim(PD_V) - \dim(\mathcal{M}(G))$.

Proof. Follows from Wilks (1938). $\square$

Now, if $\Sigma \notin \mathcal{M}(G)$, since our estimate $\hat{\Sigma}$ lies in $\mathcal{M}(G)$, $\hat{\Sigma}$ would not converge to Σ when $n \rightarrow \infty$, unlike the case when $\Sigma \in \mathcal{M}(G)$. Hence, for each observation, the expected value of the likelihood ratio statistic $2(\ell(S) - \ell(\hat{\Sigma}))$ is a strictly positive constant. Hence, as $n \rightarrow \infty$, this likelihood ratio statistic will diverge. In particular, it will not converge to a chi-squared distribution.

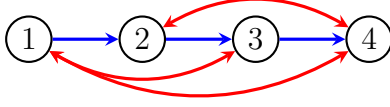


Figure 4.1: A simple graph that is not globally identifiable.

Example 4.8. Let G be the Verma graph and Σ to be the same as in Example 4.1. We first generate a dataset X by drawing 1000 samples from $N(\mathbf{0}, \Sigma)$. We then compute the MLE $\hat{\Sigma}$ for $\Sigma \in \mathcal{M}(G)$ given the dataset X using the hill-climbing algorithm above. This allows us to compute the likelihood ratio statistic. Iterating this 1000 times, we obtain the empirical CDF of the likelihood ratio statistic.

In particular, when comparing to the CDF of χ_1^2 , the Kolmogorov-Smirnov test gives us a p-value of 0.191 which suggests that there is no evidence that this distribution does not come from χ_1^2 .

Finally, given MLEs $\hat{\Lambda}$, $\hat{\Omega}$ and $\hat{\Sigma}$, we can numerically check that the the second derivative of the likelihood function in (4.3) is indeed negative.

4.4 Limitations of our Algorithm

Unfortunately, while our algorithm can be used to compute the MLE for the parameters of an HTC-identifiable graph in general, there are special cases where the computation may fail numerically. In this section, we will explore what would happen if there is a point of unidentifiability, if the sample size is too small or if the distribution has a small eigenvalue.

4.4.1 Point of Unidentifiability

One common type of failure for all such algorithms estimating parameters of a graphical model is when the model has a point of unidentifiability. This could occur in an HTC-identifiable graph that is not globally identifiable such as the instrumental variable model. We will provide an example where a point of unidentifiability could cause problems in a graphical model that is not globally identifiable.

Example 4.9 (Point of unidentifiability). Consider the graph in Figure 4.1 that is not globally identifiable. From Algorithm 3.2, we obtain

$$\begin{aligned}\lambda_{12} &= \frac{\sigma_{12}}{\sigma_{11}}, \\ \lambda_{23} &= \frac{\sigma_{23}\sigma_{11} - \sigma_{12}\sigma_{13}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} = \frac{\sigma_{23.1}}{\sigma_{22.1}}, \\ \lambda_{34} &= \frac{\sigma_{11}\sigma_{22}\sigma_{34} - \sigma_{12}^2\sigma_{34} - \sigma_{11}\sigma_{23}\sigma_{24} + \sigma_{12}\sigma_{13}\sigma_{24}}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{12}^2\sigma_{33} - \sigma_{11}\sigma_{23}^2 + \sigma_{12}\sigma_{13}\sigma_{23}}.\end{aligned}$$

Observe that while the denominator of λ_{23} is strictly positive (as it is in the form of $\sigma_{aa.B}$), this is not the case for λ_{34} . In particular, we can find a positive definite Σ such that the denominator of λ_{34} is zero. Let

$$\Sigma = \begin{bmatrix} 1 & 0.8 & 0.4375 & 0.76 \\ 0.8 & 1 & 0.8 & 0.5 \\ 0.4375 & 0.8 & 1 & 0.1 \\ 0.76 & 0.5 & 0.1 & 1 \end{bmatrix}.$$

The minimum eigenvalue of Σ is 0.0664, hence Σ is positive definite. However, with this particular value of Σ , the denominator of λ_{34} is zero. Hence, we are unable to run Algorithm 1 numerically using this particular value of Σ as we will be attempting to invert a singular matrix.

On the other hand, if we were to observe some sample covariance matrix S , the denominator for λ_{23} is almost surely non-zero as there is zero probability of obtaining specific values in a continuous distribution. Hence, we should, in theory, be almost surely able to use Algorithm 3.2 to compute for Λ numerically. Unfortunately, as the sample sizes increases, the denominator gets closer to zero, which will result in huge standard errors and an error on most computational software.

For our choice of Σ and graph G , we have encountered 2 errors out of 1000 trial runs if we draw $n = 1000$ samples from $N(\mathbf{0}, \Sigma)$. If we increase this sample size to $n = 1,000,000$, we have encountered 104 errors out of the 1000 trial runs.

The reader could consult the study by Drton (2009) on how points of identifiability could affect the convergence of algorithms in practice.

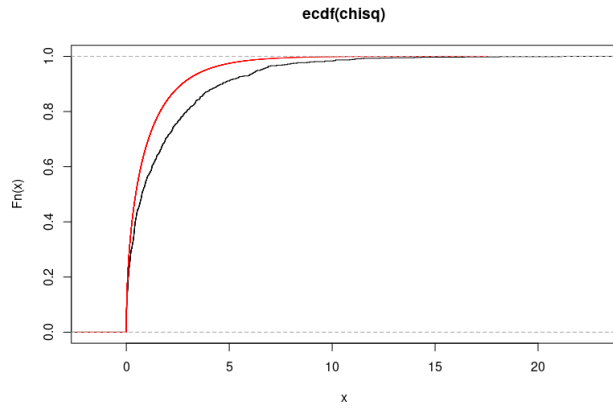


Figure 4.2: $n = 10$ using hill-climbing

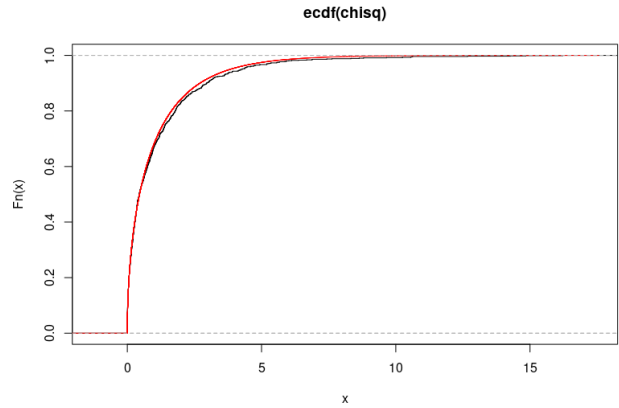


Figure 4.3: $n = 20$ using hill-climbing

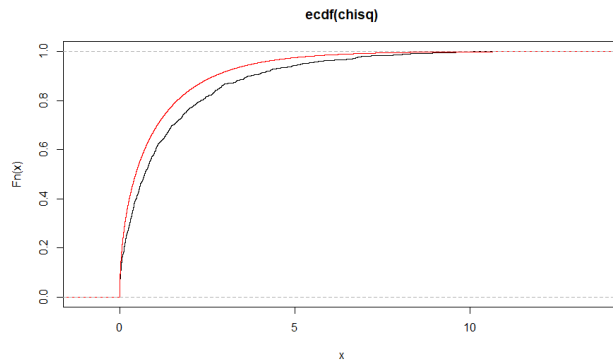


Figure 4.4: $n = 10$ using RICF

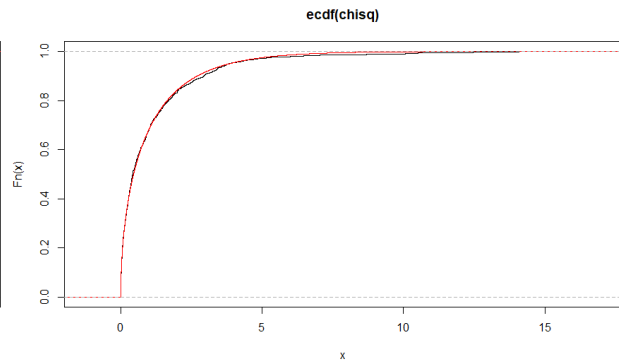


Figure 4.5: $n = 20$ using RICF

4.4.2 Small Sample Size and Eigenvalues

From Corollary 4.6, we observe that in order for us to be able to find the MLE using the hill-climbing algorithm, we must have a large enough sample size. We will now examine the complications that arises if we have a small sample size.

Example 4.10. In this example, we shall repeat the setup of Example 4.8 for smaller sample sizes of $n = 10$ and $n = 20$.

In the $n = 10$ case, comparing the empirical CDF to the CDF of χ_1^2 , the Kolmogorov-Smirnov test gives us a p-value below 2.2×10^{-16} which suggests strong evidence that the empirical distribution does not come from χ_1^2 . On the other hand, when $n = 20$, the Kolmogorov-Smirnov test has a p-value of 0.657

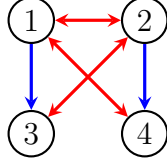


Figure 4.6: The ‘gadget’ graph.

when comparing the empirical CDF to that of χ_1^2 .

Figures 4.2 and 4.3 show the plots of the empirical CDF (black) against the CDF of χ_1^2 , for sample sizes $n = 10$ and $n = 20$ respectively using our hill-climbing algorithm.

Running the same dataset using RICF algorithms, we see that when $n = 10$ (Figure 4.4), the Kolmogorov-Smirnov test gives us a p-value of 1.02×10^{-11} . By inspection, we also see that the empirical CDF generated by the RICF algorithm is much close to the CDF of χ_1^2 compared to that generated by our hill-climbing algorithm. On the other hand, when $n = 20$ (Figure 4.5), the Kolmogorov-Smirnov test gives us a p-value of 0.775, which suggests no evidence that the empirical CDF does not come from a χ_1^2 distribution.

We have run a comparison test of RICF against our hill-climbing algorithm for which $n = 10$ and when $n = 1000$. In both cases, the log-likelihood produced by RICF is higher than that of our hill-climbing algorithm. This difference is negligible when $n = 1000$ but significant when $n = 10$.

On even smaller sample sizes (e.g. $n = 5$), we might run into the additional problem that $\tilde{\Omega}$ is not positive definite. In this case, our algorithm is unable to even run the hill-climbing algorithm since the log-likelihood will be undefined. In this case, our hill-climbing algorithm will start at $(\tilde{\Lambda}, \tilde{\Omega}' = \tilde{\Omega} + cI)$ where $c = -(\tilde{\epsilon} - 1)$ where $\tilde{\epsilon}$ is the minimum eigenvalue for $\tilde{\Omega}$. This ensures that $\tilde{\Omega}'$ is positive definite. However, the new choice of starting point $\tilde{\Sigma}$ may not lie in the concave region and therefore, the output $\hat{\Sigma}$ may not be the MLE. Note that the same issue will arise when the minimum eigenvalue for the true value of Σ approaches zero.

We shall now address the problem when Σ has a small eigenvalue.

Example 4.11. Consider the ‘gadget’ graph in Figure 4.6. Suppose the true value

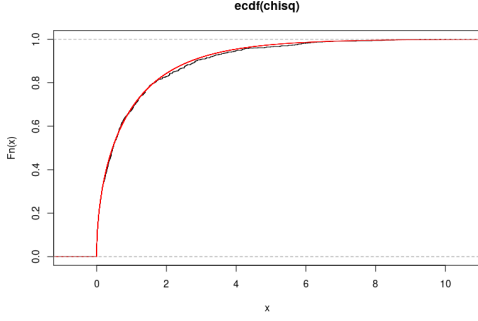


Figure 4.7: $n = 1000$, $\epsilon = 0.1$.

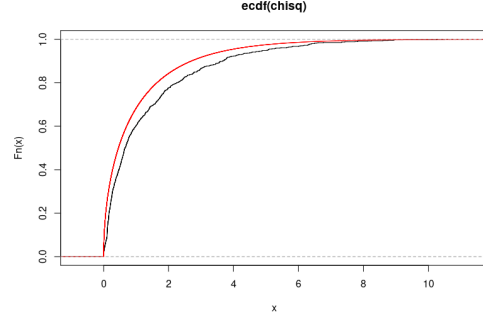


Figure 4.8: $n = 10^6$, $\epsilon = 0.01$.

of Ω is

$$\Omega = \begin{bmatrix} 1 & \rho & 0 & \rho \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & 0 \\ \rho & 0 & 0 & 1 \end{bmatrix}$$

where $\rho = \frac{2-2\epsilon}{1+\sqrt{5}}$ and ϵ is very small. Then ϵ is the minimum eigenvalue of Ω . Further suppose that the values of all entries in Λ are at most one. Then the minimum eigenvalue of Σ will be very small as well. In this example, we shall let $\lambda_{13} = \lambda_{24} = 0.25$.

Let $\epsilon = 0.1$. After simulating data with a sample size of $n = 1000$ and iterating 1000 times, the empirical CDF we obtained closely matches to that of χ_1^2 (p-value 0.854 by KS test) as shown in Figure 4.7. However, for $\epsilon = 0.01$, simulated data with a sample size of $n = 100,000$ might produce an estimate $\tilde{\Omega}$ that is not positive definite. Even when we repeat the simulation a sample size of $n = 1,000,000$, the empirical CDF does not come from the χ_1^2 distribution as seen in Figure 4.8.

Proposition 4.12. *Let $\epsilon \ll 1$ be the minimum eigenvalue for Σ . Then for every $0 < \alpha < 1$, the approximate sample size required for the estimate $\tilde{\Omega}$ to be positive definite with probability $1 - \alpha$ is $O(1/\epsilon^2)$.*

Proof. First note that the eigenvalues of a matrix with continuous entries are themselves continuous functions (Lax, 2007). Furthermore, the minimum of two continuous functions is continuous. Therefore, the minimum eigenvalue of a matrix with continuous entries is a continuous function. But $\tilde{\Omega}$ is a $\sqrt{n}$ -consistent estimator of Ω by Proposition 4.2, the empirical value of ϵ comes from either a

Gaussian or sub-Gaussian distribution with mean ϵ and variance $O(1/n)$ which we denote $D(\epsilon, O(1/n))$. If $n = O(1/\epsilon^2)$, we can draw a positive number from $D(\epsilon, O(\epsilon^2))$ with probability $1 - \alpha$ for any $0 < \alpha < 1$. $\square$

4.5 Real Data Example

In this section, we showcase an application of our algorithm to validating a graphical model given a real dataset. In the study of American occupational structure by [Blau and Duncan \(1967\)](#), the following covariates were measure with a sample size of $n = 20700$:

V : Father's educational attainment,

X : Father's occupational status,

U : Son's educational attainment,

W : Son's status of first job,

Y : Son's status of occupation in 1962.

The data can be summarised as the following correlation matrix of (V, X, U, W, Y) :

$$S = \begin{bmatrix} 1.000 & 0.516 & 0.453 & 0.332 & 0.322 \\ 0.516 & 1.000 & 0.438 & 0.417 & 0.405 \\ 0.453 & 0.438 & 1.000 & 0.538 & 0.596 \\ 0.332 & 0.417 & 0.538 & 1.000 & 0.541 \\ 0.322 & 0.405 & 0.596 & 0.541 & 1.000 \end{bmatrix}.$$

At the significance level of $\alpha = 0.01$, we find that we have the conditional independence $Y \perp\!\!\!\perp V \mid U, X$. Hence, the FCI algorithm ([Spirtes et al., 1993](#)) removes the edge between V and Y . At significance level $\alpha = 0.0001$, the FCI algorithm removes an additional edge between V and W as we have another conditional independence $W \perp\!\!\!\perp V \mid U, X$.

The skeletons identified by the FCI algorithm are shown in Figure 4.9 where the $\circ$ — edgemark can be either a tail or an arrowhead.

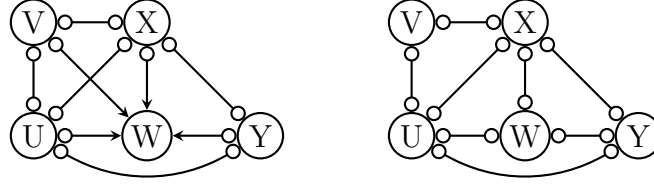


Figure 4.9: Skeleton inferred using the FCI algorithm with $\alpha = 0.01$ (left) and $\alpha = 10^{-4}$ (right).

Now, we can obtain the following valid graphical models based on the skeletons in Figure 4.9:

1. In the first model, we assume that the father's education attainment directly affects his occupational status, the son's educational attainment and the son's status of first job. We further assume that the father's occupational status directly affects the son's educational attainment, the son's status of first job and the son's status of occupation in 1962. Furthermore, we assume that the son's education attainment directly affects his status of first job and his status of occupation in 1962. Finally, there is an unobserved confounder between the son's status of first job and the son's status of occupation in 1962.
2. In the second model, we have the same assumption as the first model, but the father's educational status no longer directly affects the son's status of first job.

Note that the first model satisfies the conditional independence $Y \perp\!\!\!\perp V \mid U, X$ and the second model satisfies both conditional independence $Y \perp\!\!\!\perp V \mid U, X$ and $W \perp\!\!\!\perp V \mid U, X$. These models are graphically represented in Figure 4.10. Notice that our second graphical model is identical to the first with the edge $V \rightarrow W$ removed.

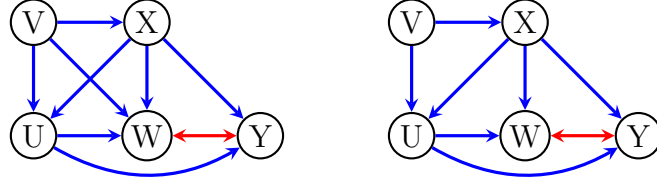


Figure 4.10: Valid graphical models based on the m -separations.

Using our algorithm to fit the correlation matrix to the first model, we obtain a p-value of 0.308. This shows that there is no evidence that the data collected did not come from the first model. However, comparing the MLE of the first and second model and running a likelihood ratio test, we found a p-value of 3.58×10^{-5} , indicating strong evidence of causal effect the variable V has on W in the context of the model. The reader could refer to the works of [Guo and Richardson \(2020\)](#) and [Spirtes et al. \(1993\)](#) for a different application of this data set.

4.6 Conclusion

In this chapter, we consider the problem of testing $H_0 : \Sigma \in \mathcal{M}(G)$ against the general hypothesis $H_1 : \Sigma \in PD_V$. In the first case, we assumed that the generators of $J(G)$ are all known. For this case, we have shown that we can simply use bootstrap to construct confidence intervals of the generators of $J(G)$ to test if $\Sigma \in \mathcal{M}(G)$.

On the other hand, if we do not have a set of polynomials that generates $J(G)$, This involves finding the MLE for Λ and Ω . From the previous Chapter, we have an algorithm that generates a starting point for Λ and Ω given a sample covariance S . We proved that this starting point is a $\sqrt{n}$ -consistent estimator. Furthermore, we proved that likelihood function is strictly concave in a local neighbourhood around the MLE. In particular, if $\Sigma \in \mathcal{M}(G)$, our starting points for Λ and Ω will lie in this concave neighbourhood. Hence, we can simply use hill-climbing to obtain the MLE.

Chapter 5

Conclusion

In this thesis, we have studied the properties of identifiability and the problem of finding constraints in linear structural equation models.

In Chapter 2, we first study the concept of edge interventions, and explore how an edge intervention could affect the covariance matrix and the data set. Throughout this chapter, we computed each entry of Σ^* as a function of Σ , where Σ^* could be obtained from adding or removing a directed edge. We further obtained sufficient conditions for Σ^* to be identifiable given Σ . We also studied a simpler and more computationally efficient identification criterion called *regression identifiability* and proved some graphical conditions which are necessary and sufficient for the directed edge coefficient λ_{ab} , and the directed path covariance $\sigma(\mathcal{D}_{bc})$, to satisfy this criterion. By removing all the directed edges in G and equating the non-trivial polynomials corresponding to vertex pairs that are not siblings, we obtain the constraints generating the model ideal. However, this requires that at each step, we are able to identify the new covariance matrix $\Sigma^{(i+1)}$ given $\Sigma^{(i)}$, where $\Sigma^{(i)}$ corresponds to the underlying distribution after removing i directed edges.

In Chapter 3, we explored the algebraic properties of HTC-identifiable graphs. In particular, we prove that all mixed graphs are HTC-identifiable if and only if all the regression coefficients can be recovered from the covariance matrix using straightforward linear algebra operations on specified equations. By exploiting these algebraic properties, we provided an efficient algorithm to recover the parameters, (Λ, Ω) , of any HTC-identifiable graph. Furthermore, we showed that the ideal $J(G)$ is in fact generated by the set of polynomials provided by [van Ommen](#)

and Mooij (2017). Similar to Chapter 2, our result only provides the generators of $J(G)$ under certain graphical conditions. In this case, we require G to be a HTC-identifiable graph.

Given the generators of $J(G)$, we showed that we could test if a particular data set $X \sim N(\mathbf{0}, \Sigma)$ fits a graphical model G (i.e. $\Sigma \in \mathcal{M}(G)$) by bootstrapping in Chapter 4. However, this method is often not feasible as finding the generators of the model ideal symbolically is computationally intensive. In the absence of the generators of $J(G)$, we introduced a hill-climbing algorithm to compute the MLE of Λ and Ω , which we can then use to test for goodness of fit using a likelihood ratio test. We start by first proving the likelihood function is strictly concave in a local neighbourhood around the MLE, and showed that the estimates of Λ and Ω using Algorithm 3.2 will lie in this concave neighbourhood if $\Sigma \in \mathcal{M}(G)$. This result was an extension of works by Zwiernik et al. (2017), and is not present in other works on finding MLEs. Hence, if $\Sigma \in \mathcal{M}(G)$, we showed that with high probability, our estimate from hill-climbing is in fact the MLE and not a local maxima.

5.1 Future Work

5.1.1 Identifiability

While we have provided several weaker identifiability conditions in this thesis, two outstanding problems remain:

1. What are the sufficient and necessary graphical conditions for a graphical model to be generically identifiable?
2. Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing a directed edge from G . What are the sufficient and necessary graphical conditions for Σ^* to be identifiable given Σ ?

While this thesis did not make much headway in the first question, we provided sufficient conditions for the second. From the conclusion of Chapter 2, we see that if we answer the first question, we would also be able to answer the second. Knowing

necessary and sufficient conditions for when Σ^* is identifiable given Σ is useful as it will provide us with the conditions required for finding constraints by iteratively removing edges.

5.1.2 Constraints

We have provided two methods for finding the generators of $J(G)$ in this thesis. In Chapter 2, we proposed that we can find the generators of $J(G)$ by iteratively removing directed edges, and then equating polynomials to vertex pairs that are not siblings. This requires that we are able to identify the new covariance matrix $\Sigma^{(i+1)}$ given $\Sigma^{(i)}$, where $\Sigma^{(i)}$ corresponds to the underlying distribution after removing i directed edges. In Chapter 3, we showed that the generators of $J(G)$ is in fact generated by the set of polynomials provided by [van Ommen and Mooij \(2017\)](#). However, this requires G to be HTC-identifiable. In both methods, finding the generators of $J(G)$ symbolically is infeasible for large $|V|$.

However, the method of finding constraints using Gröbner basis is even more computationally expensive. Finding an efficient algorithm to compute the generators for either $J(G)$ or $\mathcal{I}(G)$ is important it would allow us to assess precisely which constraint(s) are violated in a model that does not fit the data.

5.1.3 Maximum Likelihood

In Chapter 4, we have designed a hill-climbing algorithm to find the MLEs of Λ and Ω for any HTC-identifiable graph G . In particular, we have provided theoretical guarantees that this will converge to the MLE, and not some local maxima, for a large enough n . We believe that this result could be extended to any arbitrary graphical model G given that most graphical models are curved exponential families.

Appendix A

Commutative Algebra

Here, we shall introduce the basics of commutative algebra and algebraic geometry required to understand the paper. We will mainly be referring to Atiyah-Macdonald [Atiyah and Macdonald \(1969\)](#) for commutative algebra and Cox et al. [Cox et al. \(2015\)](#) for algebraic geometry.

Definition A.1. A *ring* R is a set with two binary operations (addition and multiplication) such that:

1. R is an abelian group with respect to addition, that is
 - (a) $(a+b)+c = a+(b+c)$ for every $a, b, c \in R$ (i.e. addition is associative).
 - (b) $a+b = b+a$ for every $a, b \in R$ (i.e. addition is commutative).
 - (c) There is an additive identity $0 \in R$ such that $a+0 = a$ for all $a \in R$.
 - (d) For each $a \in R$, there is an additive inverse $-a \in R$ such that $a+(-a) = 0$.
2. Multiplication is associative.
3. Multiplication is distributive over addition, that is $a(b+c) = ab+ac$ and $(b+c)a = ba+ca$ for all $a, b, c \in R$.
4. There is a multiplicative identity $1 \in R$ such that $a1 = 1a = a$ for all $a \in R$.

A *unit* in a ring R is an element a which “divides 1” (i.e. there exists $b \in R$ such that $ab = 1$). A *field* is a ring R where $1 \neq 0$ and every non-zero element is a unit.

Definition A.2. We say that I is an ideal of a ring R , denoted $I \triangleleft R$, if

1. for every $x, y \in I$, $x + y \in I$.
2. for every $x \in I$ and $a \in R$, $ax \in I$.

Let $\langle \dots \rangle$ denote the generators of an ideal. In particular, we have $\langle 1 \rangle = R$. An ideal $\mathfrak{p}$ is *prime* if $\mathfrak{p} \neq \langle 1 \rangle$ and $xy \in \mathfrak{p} \implies$ either $x \in \mathfrak{p}$ or $y \in \mathfrak{p}$. An ideal $\mathfrak{m}$ is *maximal* if $\mathfrak{m} \neq \langle 1 \rangle$ and there do not exist an ideal $\mathfrak{a}$ such that $\mathfrak{m} \subsetneq \mathfrak{a} \subsetneq \langle 1 \rangle$. Note that all maximal ideals are prime, but the converse does not hold in general.

Definition A.3. We say that a ring R is Noetherian if every ideal $I \triangleleft R$ is finitely generated (i.e. $\exists f_1, \dots, f_N$ such that $I = \langle f_1, \dots, f_N \rangle$).

Theorem A.1 (Hilbert Basis Theorem). *If R is a Noetherian ring, then $R[X]$ is also a Noetherian ring.*

Proof. See Theorem 7.5 of [Atiyah and Macdonald \(1969\)](#). □

For any field k , the only ideals are $\langle 0 \rangle$ and $\langle 1 \rangle$. Hence, every field is Noetherian. In particular, $\mathbb{R}$ is Noetherian. By the Hilbert Basis Theorem, $\mathbb{R}(\Sigma)$ is a Noetherian ring.

Definition A.4. We say that a *chain of prime ideals* has length m if there is a strict chain of inclusions $\mathfrak{p}_0 \subsetneq \mathfrak{p}_1 \subsetneq \dots \subsetneq \mathfrak{p}_m$ where each $\mathfrak{p}_i$ is a prime ideal. The *height* $\text{ht}(\mathfrak{p})$ of a prime ideal $\mathfrak{p}$ is the supremum of the lengths of all chains of prime ideals contained in $\mathfrak{p}$.

$$\text{ht}(\mathfrak{p}) = \sup_m (\mathfrak{p}_0 \subsetneq \mathfrak{p}_1 \subsetneq \dots \subsetneq \mathfrak{p}_m = \mathfrak{p}).$$

Theorem A.2 (Krull's Height Theorem). *For any Noetherian ring R , let $\mathfrak{p} \triangleleft R$ be a prime ideal. Then, the height $\mathfrak{p}$ is at most the number of generators of $\mathfrak{p}$. Conversely, if $\mathfrak{p}$ has height m , then $\mathfrak{p}$ is a minimal prime ideal over an ideal $I \subseteq \mathfrak{p}$ generated by m elements. In particular, $\mathfrak{p}$ is generated by at least m elements.*

Definition A.5. Let k be a field and $I \triangleleft k[x_1, \dots, x_n]$ be an ideal. The *variety* defined by I is the set

$$V(I) = \{a \in k^n : f(a) = 0 \text{ for all } f \in I\}.$$

On the other hand, given a set $X \subseteq k^n$, the *vanishing ideal* of X is defined as

$$I(X) = \{f \in k[x_1, \dots, x_n] : f(a) = 0 \text{ for all } a \in X\}.$$

A variety X is *reducible* if we can write $X = X_1 \cup X_2$ for proper subsets $X_i \subsetneq X$. Otherwise, X is *irreducible*.

Theorem A.3. *The variety X is irreducible if and only if the vanishing ideal $I(X)$ is prime.*

Proof. See Proposition 3 in Section 4.5 of [Cox et al. \(2015\)](#). □

Proposition A.4. *If k is an infinite field and V is a variety defined by the rational parametrization*

$$\begin{aligned} x_1 &= \frac{f_1(t_1, \dots, t_m)}{g_1(t_1, \dots, t_m)} \\ &\vdots \\ x_n &= \frac{f_n(t_1, \dots, t_m)}{g_n(t_1, \dots, t_m)} \end{aligned}$$

where $f_1, \dots, f_n, g_1, \dots, g_n \in k[t_1, \dots, t_m]$, then V is irreducible.

Proof. See Proposition 6 in Section 4.5 of [Cox et al. \(2015\)](#). □

Appendix B

Appendices for Chapter 2

B.1 Calculations for Example 2.30

In this section, we will complete the working for Example 2.30 and show that $\sigma_{14}^* = f_{\text{Verma}}$. Starting off, we have

$$\begin{aligned}\sigma_{14}^* &= 0 \\ \sigma_{14} - \frac{\sigma_{13}\sigma_{34 \cdot 12}}{\sigma_{33 \cdot 12}} &= 0 \\ \sigma_{14}\sigma_{33 \cdot 12} - \sigma_{13}\sigma_{34 \cdot 12} &= 0\end{aligned}\tag{B.1}$$

where

$$\sigma_{33 \cdot 12} = \sigma_{33} - \frac{\sigma_{13}^2\sigma_{22} - \sigma_{12}\sigma_{13}\sigma_{23} + \sigma_{11}\sigma_{23}^2 - \sigma_{12}\sigma_{13}\sigma_{23}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\tag{B.2}$$

$$\sigma_{34 \cdot 12} = \sigma_{34} - \frac{\sigma_{13}\sigma_{14}\sigma_{22} - \sigma_{12}\sigma_{14}\sigma_{23} + \sigma_{11}\sigma_{23}\sigma_{24} - \sigma_{12}\sigma_{13}\sigma_{24}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}.\tag{B.3}$$

Substituting (B.2) and (B.3) into (B.1) and clearing denominators, we obtain

$$\begin{aligned}&\sigma_{14}[\sigma_{33}(\sigma_{11}\sigma_{22} - \sigma_{12}^2) - (\sigma_{13}^2\sigma_{22} - \sigma_{12}\sigma_{13}\sigma_{23} + \sigma_{11}\sigma_{23}^2 - \sigma_{12}\sigma_{13}\sigma_{23})] - \\ &\sigma_{13}[\sigma_{34}(\sigma_{11}\sigma_{22} - \sigma_{12}^2) - (\sigma_{13}\sigma_{14}\sigma_{22} - \sigma_{12}\sigma_{14}\sigma_{23} + \sigma_{11}\sigma_{23}\sigma_{24} - \sigma_{12}\sigma_{13}\sigma_{24})]\end{aligned}$$

which simplifies to f_{Verma} .

B.2 Removing a Bidirected Edge

Let us consider what would occur in theory if we want to remove a bidirected edge. Once again, suppose that $G = (V, \mathcal{D}, \mathcal{B})$ is an ADMG with a corresponding distribution $N(0, \Sigma)$ and we want to remove the bidirected edge $a^* \leftrightarrow b$. Assume without loss of generality that $a^* \in \text{nd}(b)$. Again, let $A = \text{nd}(b)$, $B = \{b\}$ and $C = \text{de}(b) \setminus \{b\}$, then $A \cup B \cup C = V$. Throughout this section, we suppose that after removing the bidirected edge $a^* \leftrightarrow b$, we obtain a new ADMG G^* with a corresponding distribution $N(0, \Sigma^*)$ with the assumption that $\Sigma_{AA}^* = \Sigma_{AA}$.

B.2.1 Effects on Covariance

As before, we shall first compute Σ^* given Σ . The majority of the proofs here are shorten as they are similar to the directed edge case.

Lemma B.1. *For all $a \in A$, we have*

$$\begin{aligned}\sigma_{ab}^* &= \sigma_{ab} - \sigma(\mathcal{D}_{a^*a})\omega_{a^*b}, \\ \sigma_{bb}^* &= \sigma_{bb} - \sigma(\mathcal{D}_{a^*b})\omega_{a^*b}, \\ \Sigma_{AA}^* &= \Sigma_{AA}.\end{aligned}$$

Proof. Since $a \in \text{nd}(b)$, every trek from a to b containing the bidirected edge $a^* \leftrightarrow b$ must end with $a^* \leftrightarrow b$. By the same reasoning, all treks from b to b containing the bidirected edge $a^* \leftrightarrow b$ must either start with (or by symmetry end with) $a^* \leftrightarrow b$. Furthermore, there cannot be any treks from a to another $a' \in A$ containing the edge $a^* \leftrightarrow b$. $\square$

Lemma B.2. *For all $a \in A$ and $c, c' \in C$,*

$$\begin{aligned}\sigma_{ac}^* &= \sigma_{ac} - \sigma(\mathcal{D}_{a^*a})\omega_{a^*b}\sigma(\mathcal{D}_{bc}), \\ \sigma_{bc}^* &= \sigma_{bc} - \omega_{a^*b}\sigma(\mathcal{D}_{a^*c}) - \sigma(\mathcal{D}_{a^*b})\omega_{a^*b}\sigma(\mathcal{D}_{bc}), \\ \sigma_{cc'}^* &= \sigma_{cc'} - \sigma(\mathcal{D}_{a^*c})\omega_{a^*b}\sigma(\mathcal{D}_{bc'}) - \sigma(\mathcal{D}_{a^*c'})\omega_{a^*b}\sigma(\mathcal{D}_{bc}).\end{aligned}$$

Proof. Treks from a to c containing $a^* \leftrightarrow b$ are of the form

$$a \longleftarrow \cdots \longleftarrow a^* \longleftrightarrow b \longrightarrow \cdots \longrightarrow c.$$

Treks from b to c containing $a^* \leftrightarrow b$ are of the form

$$b \longleftarrow \cdots \longleftarrow a^* \longleftrightarrow b \longrightarrow \cdots \longrightarrow c \quad \text{or} \quad b \longleftrightarrow a^* \longrightarrow \cdots \longrightarrow c.$$

Treks from c to c' containing $a^* \leftrightarrow b$ are of the form

$$c \longleftarrow \cdots \longleftarrow a^* \longleftrightarrow b \longrightarrow \cdots \longrightarrow c' \quad \text{or} \quad c \longleftarrow \cdots \longleftarrow b \longleftrightarrow a^* \longrightarrow \cdots \longrightarrow c'.$$

□

Once again, we see that each entry of Σ^* is a function of Σ , ω_{a^*b} , $\sigma(\mathcal{D}_{a^*v})$ and $\sigma(\mathcal{D}_{bc})$ for all $v \in V$ and $c \in C$. Hence, we obtain the following Corollary.

Proposition B.3. *Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing the bidirected edge $a^* \leftrightarrow b$ from G . Then, Σ^* is identifiable given Σ if and only if ω_{a^*b} , $\sigma(\mathcal{D}_{a^*v})$ and $\sigma(\mathcal{D}_{bc})$ are identifiable for all $v \in V$ and $c \in C$.*

Proof. ($\Leftarrow$) Follows from Lemmas B.1 and B.2.

($\Rightarrow$) From Lemma B.1, we have

$$\begin{aligned} \sigma_{a^*b}^* &= \sigma_{a^*b} - \sigma_{a^*a^*} \omega_{a^*b}, \\ \omega_{a^*b} &= \frac{\sigma_{a^*b} - \sigma_{a^*b}^*}{\sigma_{a^*a^*}}. \end{aligned}$$

Hence, ω_{a^*b} is identifiable if Σ^* is identifiable given Σ . Rearranging the first two equations in Lemma B.1, we obtain

$$\begin{aligned} \sigma(\mathcal{D}_{a^*a}) &= \frac{\sigma_{ab}^* - \sigma_{ab}}{\omega_{a^*b}}, \\ \sigma(\mathcal{D}_{a^*b}) &= \frac{\sigma_{bb}^* - \sigma_{bb}}{\omega_{a^*b}}. \end{aligned}$$

Since ω_{a^*b} is identifiable, $\sigma(\mathcal{D}_{a^*a})$ and $\sigma(\mathcal{D}_{a^*b})$ are also identifiable. Now, the first equation of Lemma B.2 can be rearranged to

$$\sigma(\mathcal{D}_{bc}) = \frac{\sigma_{ac} - \sigma_{ac}^*}{\sigma(\mathcal{D}_{a^*a}) \omega_{a^*b}}.$$

Since both ω_{a^*b} and $\sigma(\mathcal{D}_{a^*a})$ are identifiable, so is $\sigma(\mathcal{D}_{bc})$. Finally, the second equation of Lemma B.2 can be rearranged to

$$\sigma(\mathcal{D}_{a^*c}) = \frac{\sigma_{bc} - \sigma_{bc}^* - \sigma(\mathcal{D}_{a^*b})\omega_{a^*b}\sigma(\mathcal{D}_{bc})}{\omega_{a^*b}}.$$

Since ω_{a^*b} , $\sigma(\mathcal{D}_{a^*b})$ and $\sigma(\mathcal{D}_{bc})$ are all identifiable, $\sigma(\mathcal{D}_{a^*c})$ is also identifiable. $\square$

Here, we see that a different identifiability criteria for removing bidirected edges is required.

Definition B.1. Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing the bidirected edge $a^* \leftrightarrow b$ from G . We say that Σ^* is *simply identifiable* given Σ if ω_{a^*b} , $\sigma(\mathcal{D}_{a^*v})$ and $\sigma(\mathcal{D}_{bc})$ are all regression identifiable for all $v \in V$ and $c \in C$.

B.2.2 Regression Identifiability

Once again, we shall find the graphical criterion under which Σ^* is identifiable. From Theorem 2.16, we know that $\sigma(\mathcal{D}_{a^*v})$ is identifiable by regression for all $v \in V$ if and only if a^* is fixable in G with $\sigma(\mathcal{D}_{a^*v}) = 0$ if $v \in \text{nd}(a^*)$, and $\sigma(\mathcal{D}_{bc})$ is identifiable by regression for all $c \in C$ if and only if b is fixable in G . Hence, it suffices to find conditions for which the edge coefficient ω_{a^*b} is identifiable where both a^* and b are fixable.

Theorem B.4. Suppose we have a bidirected edge $a^* \leftrightarrow b$ where a^* and b are fixable. Then ω_{a^*b} is identifiable by regression if and only if there are no bidirected paths from a^* to b in $G_{\text{an}(b)}^*$. In particular, $\omega_{a^*b} = \sigma_{a^*b, \xi}$ where $\xi = \text{mb}_{\text{an}(b)}(b)$.

Proof. ($\Leftarrow$) It suffices to prove that $\omega_{a^*b} = \sigma_{a^*b, \xi}$ if there are no bidirected paths from a^* to b in $G_{\text{an}(b)}^*$. First, note that we do not have a directed edge $a^* \rightarrow b$ since a^* is fixable. Similar to Theorem 2.17, it is sufficient to consider the induced subgraph $G_{\text{an}(b)}$, as any path containing a vertex $c \notin G_{\text{an}(b)}$ has some collider $c' \notin G_{\text{an}(b)}$.

Now, paths of the form $a^* \cdots \rightarrow b$ are blocked by $\text{pa}(b) \subset \text{mb}_{\text{an}(b)}(b)$. Paths of the form $a^* \cdots \rightarrow d_1 \leftrightarrow \cdots \leftrightarrow d_n \leftrightarrow b$ and $a^* \cdots \leftarrow d_1 \leftrightarrow \cdots \leftrightarrow d_n \leftrightarrow b$ are blocked by $\text{mb}_{\text{an}(b)}(b)$. So the only paths that are not blocked are the bidirected paths from a^* to b . But the only path from a^* to b is the bidirected edge $a^* \leftrightarrow b$.

($\Rightarrow$) Suppose there is a bidirected path from a^* to b in $G_{\text{an}(b)}^*$, say $a^* \leftrightarrow d_1 \leftrightarrow \cdots \leftrightarrow d_n \leftrightarrow b$. Since $d_1 \in \text{an}(b)$, there is an open path $a^* \leftrightarrow d_1 \rightarrow \cdots \rightarrow b$. If we condition on any intermediate vertices on that path, we have another open path $a^* \leftrightarrow d_1 \leftrightarrow d_2 \rightarrow \cdots \rightarrow b$ since $d_2 \in \text{an}(b)$. By iteratively conditioning on intermediate vertices, we will end up with the open path $a^* \leftrightarrow d_1 \leftrightarrow \cdots \leftrightarrow d_n \leftrightarrow b$.

□

Corollary B.5. *Suppose that $N(0, \Sigma)$ is a distribution corresponding to the graphical model G and the distribution $N(0, \Sigma^*)$ corresponds to the new graphical model G^* obtained by removing the bidirected edge $a^* \leftrightarrow b$ from G . Then, Σ^* is simply identifiable given Σ if and only if*

- both a^* and b are fixable, and
- $a^* \notin \text{dis}_{\text{an}^*(b)}(b)$, where $\text{dis}_{\text{an}^*(b)}(b)$ denotes the Markov blanket of b in $G_{\text{an}(b)}^*$.

B.2.3 Concerns when Removing Bidirected Edges

While we could identify the new distribution when removing a bidirected edge theoretically, several issues may arise in practice:

1. **Interpretation.** While removing a directed edge $i \rightarrow j$ simply implies removing a direct effect of X_i on X_j , the removal of a bidirected edge $i \leftrightarrow j$ is slightly more complicated to interpret as we are removing the covariance between the residuals ϵ_i and ϵ_j , where ϵ_ℓ is the residual obtained when we regress X_ℓ on its parents. One such interpretation is the removal of all unobserved confounders of X_i and X_j , which may be impractical. In practice, the removal of a bidirected edge is often a byproduct of a vertex intervention.



Figure B.1: A graphical model where removing $1 \leftrightarrow 3$ may result in a non-positive definite Σ^* .

2. **Distribution of the new model.** At the start of this section, we have assumed that after removing a bidirected edge $a^* \leftrightarrow b$, the distribution of X_{a^*} remains fixed. This is fine for removing a directed edge $a^* \rightarrow b$ as X_{a^*} is not affected by the removal of the causal effect of X_{a^*} on X_b . However, when removing the bidirected edge $a^* \leftrightarrow b$, the distribution of both X_{a^*} and X_b may be affected which would in turn have an impact on their descendants.
3. **Identifiability.** While we have the theoretical result to recover the new distribution $N(0, \Sigma^*)$ if Σ^* is simply identifiable given Σ , the new covariance matrix Σ^* may not be positive definite; for instance, consider the graphical model in Figure B.1 with a large value for ω_{13} . If we remove $1 \leftrightarrow 3$, we are setting ω_{13}^* to be zero, which may result in a non-positive definite matrix for Σ^* .
4. **Effect on Data.** There is no general expression for X_V^* that satisfies $\text{Cov}(X_V^*, X_V^*) = \Sigma^*$, the covariance matrix found in Section B.2.1.

B.3 Vertex Intervention

Suppose we have an ADMG $G = (V, \mathcal{D}, \mathcal{B})$ with a corresponding distribution $N(0, \Sigma)$. Let $A = \text{nd}(B)$ and $C = \text{de}(B) \setminus B$, so that $A \cup B \cup C = V$. Throughout this section, we suppose that after performing a vertex intervention on B , we obtain a new ADMG G^* with a corresponding distribution $N(0, \Sigma^*)$.

In general, performing a vertex intervention involves replacing the probability distribution p with

$$p^*(x_V) := \frac{p(x_V)}{p(x_B | x_A)} = p(x_A) \cdot p(x_C | x_A, x_B)$$

which is well defined if *either* $p(x_A) = 0$ or $p(x_B | x_A) > 0$. In the Gaussian case, all densities are positive and these manipulations are always uniquely defined. The

new probability distribution, p^* , is nominally a *conditional kernel* over X_A, X_C given X_B , with the following properties:

- $p^*(x_A) = p(x_A)$, i.e. the marginal distribution of X_A is preserved;
- $X_B \perp\!\!\!\perp X_A$ under p^* ;
- $p^*(x_C | x_A, x_B) = p(x_C | x_A, x_B)$, i.e. the conditional distribution of X_C given X_A, X_B is preserved.

For the multivariate Gaussian case with mean 0, the properties above suggest that in the new covariance matrix Σ^* , which corresponds to the ADMG after the vertex intervention, we will have

- $\Sigma_{AA}^* = \Sigma_{AA}$ and $\Sigma_{AB}^* = 0$;
- $K_{CV}^* = K_{CV}$, where K denotes the concentration (or precision) matrix $K = \Sigma^{-1}$.

Note that this leaves open a choice of value for Σ_{BB}^* , and that this choice will affect the values of Σ_{CV}^* required to fulfil the restriction on K_{CV}^* .

Because all effects are linear, we can generally identify the individual edges in a Gaussian SEM and remove their effects. This means that it is well-defined to talk about fixing data, rather than only for a distribution.

B.3.1 Effect on Data

Theorem B.6. *Suppose the vertex set B is fixable. Then, intervening on the vertex set B is equivalent to replacing X_V by*

$$X_V^* = X_V - \Sigma_{VB \cdot A} \Sigma_{BB \cdot A}^{-1} \Sigma_{BA} \Sigma_{AA}^{-1} X_A,$$

where $A = \text{pa}(B) \cup \text{sib}(B)$.

Proof. Note that when intervening on the vertex set, we are removing all incoming arrowheads into B . This meant that, in the new covariance matrix Σ^* , we have

$$\begin{aligned} \Sigma_{AA}^* &= \Sigma_{AA} \\ \Sigma_{AB}^* &= 0 \\ K_{CV}^* &= K_{CV} \end{aligned}$$

where $K = \Sigma^{-1}$. We can check that $X_A^* = X_A$ as $\Sigma_{AB \cdot A} = 0$, and

$$\begin{aligned}\text{Cov}(X_B^*, X_A^*) &= \text{Cov}(X_B, X_A) - \Sigma_{BA}(\Sigma_{AA})^{-1}\text{Cov}(X_A, X_A) \\ &= \Sigma_{BA} - \Sigma_{BA}(\Sigma_{AA})^{-1}\Sigma_{AA} = 0.\end{aligned}$$

Now, the conditional covariance of $X_C^* \mid X_A, X_B^*$ is the same as that of X_C , since the only difference is subtraction of a conditionally constant term. Writing X_C in the form of a linear regression on X_A, X_B , we have

$$\begin{aligned}\mathbb{E}[X_C^* \mid X_A^*, X_B^*] &= \mathbb{E}[X_C^* \mid X_A, X_B] \\ &= \mathbb{E}[X_C - MX_A \mid X_A, X_B] \\ &= \Sigma_{CA \cdot B}(\Sigma_{AA \cdot B})^{-1}X_A + \Sigma_{CB \cdot A}(\Sigma_{BB \cdot A})^{-1}X_B - MX_A \\ &= \Sigma_{CA \cdot B}(\Sigma_{AA \cdot B})^{-1}X_A + \Sigma_{CB \cdot A}(\Sigma_{BB \cdot A})^{-1}(X_B^* + \Sigma_{BA}(\Sigma_{AA})^{-1}X_A) - MX_A \\ &= \Sigma_{CA \cdot B}(\Sigma_{AA \cdot B})^{-1}X_A + \Sigma_{CB \cdot A}(\Sigma_{BB \cdot A})^{-1}X_B^*\end{aligned}$$

where $M = \Sigma_{CB \cdot A}(\Sigma_{BB \cdot A})^{-1}\Sigma_{BA}(\Sigma_{AA})^{-1}$. Hence, the conditional distribution of $X_C^* \mid X_A, X_B^*$ is the same as that of $X_C \mid X_A, X_B$. $\square$

B.3.2 Effect on Covariance

Writing $\Sigma = X_V^T X_V$, we can ask what is the effect of vertex intervention on this covariance matrix:

$$\begin{aligned}X_V^{*T} X_V^* &= \Sigma - \Sigma_{VA}(\Sigma_{AA})^{-1}\Sigma_{AB}(\Sigma_{BB \cdot A})^{-1}\Sigma_{BV \cdot A} - ((\Sigma_{AA})^{-1}\Sigma_{AB}(\Sigma_{BB \cdot A})^{-1}\Sigma_{BV \cdot A})^T \Sigma_{AV} \\ &\quad + ((\Sigma_{AA})^{-1}\Sigma_{AB}(\Sigma_{BB \cdot A})^{-1}\Sigma_{BV \cdot A})^T \Sigma_{AA}(\Sigma_{AA})^{-1}\Sigma_{AB}(\Sigma_{BB \cdot A})^{-1}\Sigma_{BV \cdot A} \\ &= \Sigma - \Sigma_{VA}(\Sigma_{AA})^{-1}\Sigma_{AB}(\Sigma_{BB \cdot A})^{-1}\Sigma_{BV \cdot A} - \Sigma_{VB \cdot A}(\Sigma_{BB \cdot A})^{-1}\Sigma_{BA}(\Sigma_{AA})^{-1}\Sigma_{AV} \\ &\quad + \Sigma_{VB \cdot A}(\Sigma_{BB \cdot A})^{-1}\Sigma_{BA}(\Sigma_{AA})^{-1}\Sigma_{AB}(\Sigma_{BB \cdot A})^{-1}\Sigma_{BV \cdot A} \\ &= \Sigma - \Sigma_{VA}\beta_{AB}\beta_{BV \cdot A} - (\beta_{AB}\beta_{BV \cdot A})^T \Sigma_{AV} + (\beta_{AB}\beta_{BV \cdot A})^T \Sigma_{AA}\beta_{AB}\beta_{BV \cdot A} \\ &= \Sigma - \Sigma_{VA}(\beta_{AV} - \beta_{AV \cdot B}) - (\beta_{AV} - \beta_{AV \cdot B})^T \Sigma_{AV} + (\beta_{AV} - \beta_{AV \cdot B})^T \Sigma_{AA}(\beta_{AV} - \beta_{AV \cdot B})\end{aligned}$$

where $\beta_{AB \cdot C} = (\Sigma_{AA \cdot C})^{-1}\Sigma_{AB \cdot C}$ denotes the regression coefficients of X_B on X_A conditional on X_C . Note that if $\beta_{AB} = 0$ then intervening on the vertex set B does not change anything. The three terms after Σ in the final equation could be interpreted as:

1. removing all treks that use the edges $A \rightarrow v$ on the right side;
2. removing all treks that use the edges $A \rightarrow v$ on the left side;
3. adding back in those that pass through such an edge on both sides and were therefore removed twice.

This interpretation is similar the equation we found in Corollaries 2.8, 2.10 and 2.11. Note that if $v \in A$, then $\beta_{Av.C}$ is invariant to the conditioning set C . Hence values of Σ_{AV}^* are just

$$\Sigma_{AV}^* = \Sigma_{AV} - \Sigma_{AA}(\beta_{AV} - \beta_{AV.B}).$$

B.3.3 Hard Interventions

Suppose we instead performed a hard intervention on X_B by assigning $X_B = x_B$. The probability distribution for this new distribution, $p^{**}(x)$, is the probability distribution of the vertex intervention conditioned on $X_B = x_B$ (i.e. $p^{**}(x_A, x_C) = p^*(x_A, X_C \mid x_B)$). Hence we have

$$\Sigma_{UU}^{**} = \Sigma_{UU.B}^* = \begin{bmatrix} \Sigma_{AA}^* & \Sigma_{AC}^* \\ \Sigma_{AC}^* & \Sigma_{CC.B}^* \end{bmatrix}$$

since $\Sigma_{AV.B}^* = \Sigma_{AV}^* - \Sigma_{AB}^*(\Sigma_{BB}^*)^{-1}\Sigma_{BV}^* = \Sigma_{AV}^*$ as $\Sigma_{AB}^* = 0$. Note that when reversing the data standardisation, the mean of X_B is now x_B , and the mean of X_C has been changed to

$$\overline{X_C} - \sum_{b \in B} \sigma_{bb}^* \Sigma(\mathcal{D}_{bC}).$$

Bibliography

- M. Artin. *Algebra*. Pearson Prentice Hall, 1991.
- M. F. Atiyah and I. G. Macdonald. *Introduction to commutative algebra*. Addison-Wesley Publishing Co., 1969.
- J. Blalock. *Causal models in the social sciences*. Routledge, 1971.
- P. M. Blau and O. D. Duncan. *The American occupational structure*. ERIC, 1967.
- K. A. Bollen. *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1989. A Wiley-Interscience Publication.
- C. Brito and J. Pearl. A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9(4):459–474, 2002a.
- C. Brito and J. Pearl. Generalized instrumental variables. In *Uncertainty in Artificial Intelligence*, pages 85–93, 2002b.
- C. Brito and J. Pearl. Graphical condition for identification in recursive SEM. In *Uncertainty in Artificial Intelligence*, pages 47–54, 2006.
- S. Chaudhuri, M. Drton, and T. S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216, 2007.
- B. Chen, J. Tian, and J. Pearl. Testable implications of linear structural equation models. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- D. A. Cox, J. Little, and D. O’Shea. *Ideals, varieties, and algorithms*. Undergraduate Texts in Mathematics. Springer, Cham, fourth edition, 2015.

- M. Drton. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2): 979–1012, 2009.
- M. Drton. Algebraic problems in structural equation modeling. In *The 50th anniversary of Gröbner bases*, pages 35–86. Mathematical Society of Japan, 2018.
- M. Drton, M. Eichler, and T. S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(Oct):2329–2348, 2009.
- M. Drton, R. Foygel, S. Sullivant, et al. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.
- M. Drton, C. Fox, and Y. S. Wang. Computation of maximum likelihood estimates in cyclic structural equation models. *The Annals of Statistics*, 47(2):663–690, 2019.
- M. Drton, E. Robeva, and L. Weihs. Nested covariance determinants and restricted trek separation in Gaussian graphical models. *Bernoulli*, 26(4):2503–2540, 2020.
- O. D. Duncan. Path analysis: Sociological examples. *American Journal of Sociology*, 72(1):1–16, 1966.
- O. D. Duncan. *Introduction to structural equation models*. Academic Press, 1975.
- F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.
- B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- J. Fox. Teacher’s corner: structural equation modeling with the sem package in R. *Structural equation modeling*, 13(3):465–486, 2006.
- R. Foygel, J. Draisma, M. Drton, et al. Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, 40(3):1682–1713, 2012.

- C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.
- L. D. Garcia, M. Stillman, and B. Sturmfels. Algebraic geometry of Bayesian networks. *Journal of Symbolic Computation*, 39(3-4):331–355, 2005.
- D. Geiger, T. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.
- D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: graphical models and model selection. *Annals of statistics*, pages 505–529, 2001.
- J. W. Gibbs. *Elementary principles in statistical mechanics*. Courier Corporation, 1902.
- A. S. Goldberger. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001, 1972.
- J. B. Grace. *Structural equation modeling and natural systems*. Cambridge University Press, 2006.
- F. R. Guo and T. S. Richardson. On testing marginal versus conditional independence. *Biometrika*, 107(4):771–790, 2020.
- T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1943.
- J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005.
- S. L. Hershberger. The growth of structural equation modeling: 1994-2001. *Structural Equation Modeling*, 10(1):35–46, 2003.
- K. G. Jöreskog. A general method for analysis of covariance structures. *Biometrika*, 57(2):239–251, 1970.

- R. A. Kaslow, D. G. Ostrow, R. Detels, J. P. Phair, B. F. Polk, and C. R. Rinaldo Jr. The multicenter AIDs cohort study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology*, 126(2):310–318, 1987.
- H. Kiiveri, T. P. Speed, and J. B. Carlin. Recursive causal models. *Journal of the Australian Mathematical Society*, 36(1):30–52, 1984.
- K. B. Korb, L. R. Hope, A. E. Nicholson, and K. Axnick. Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*, pages 322–331. Springer, 2004.
- S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, pages 31–57, 1989.
- P. D. Lax. *Linear Algebra and its Applications*. Wiley-Interscience, New York, USA, 2007.
- D. Malinsky. Intervening on structure. *Synthese*, 195(5):2295–2312, 2018.
- R. P. McDonald. What can we learn from the path equations?: Identifiability, constraints, equivalence. *Psychometrika*, 67(2):225–249, 2002.
- A. Narayanan. A review of eight software packages for structural equation modeling. *The American Statistician*, 66(2):129–138, 2012.
- J. Neyman. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.
- E. L. Ogburn, O. Sofrygin, I. Diaz, and M. J. Van der Laan. Causal inference for social network data. *arXiv preprint arXiv:1705.08527*, 2017.
- M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, pages 763–765, 1973.
- S. M. Olmsted. *On representing and solving decision problems*. PhD thesis, Stanford University, 1984.

- J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.
- J. Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.
- J. Pearl and T. S. Verma. A statistical semantics for causation. *Statistics and Computing*, 2(2):91–95, 1992.
- E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. A complete generalized adjustment criterion. In *Uncertainty in Artificial Intelligence*, page 682, 2015.
- E. Perkovic, J. Textor, M. Kalisch, and M. H. Maathuis. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 2018.
- T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.
- T. S. Richardson and J. M. Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.
- T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.
- J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.
- J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.
- J. M. Robins and T. S. Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, 84:103–158, 2010.

- J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.
- Y. Rosseel. lavaan: An R package for structural equation modeling. *Journal of statistical software*, 48:1–36, 2012.
- K. J. Rothman and S. Greenland. Causation and causal inference in epidemiology. *American Journal of Public Health*, 95(S1):S144–S150, 2005.
- D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.
- R. Settini and J. Q. Smith. On the geometry of Bayesian graphical models with hidden variables. In *Uncertainty in artificial intelligence*, pages 472–479, 1998.
- E. Sherman and I. Shpitser. Intervening on network ties. In *Uncertainty in Artificial Intelligence*, pages 975–984, 2020.
- I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219, 2006.
- I. Shpitser and E. Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433, 2016.
- I. Shpitser, T. VanderWeele, and J. M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Uncertainty in Artificial Intelligence*, pages 527–536, 2010.
- E. Slutsky. Über stochastische asymptoten und grenzwerte. *Metron*, 5(3):3–89, 1925.
- M. E. Sobel. Causal inference in the social sciences. *Journal of the American Statistical Association*, 95(450):647–651, 2000.
- P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, NY, 1993.

- J. H. Steiger. Driving fast in reverse. *Journal of the American Statistical Association*, 96(453):331–338, 2001.
- M. J. Stensrud, J. M. Robins, A. Sarvet, E. J. Tchetgen Tchetgen, and J. G. Young. Conditional separable effects. *Journal of the American Statistical Association*, (just-accepted):1–29, 2022.
- S. Sullivant. Algebraic geometry of Gaussian Bayesian networks. *Advances in Applied Mathematics*, 40(4):482–513, 2008.
- J. Tian. Identifying direct causal effects in linear models. In *AAAI*, pages 346–353, 2005.
- J. Tian. A criterion for parameter identification in structural equation models. In *Uncertainty in Artificial Intelligence*, pages 392–399, 2007.
- J. Tian. Parameter identification in a class of linear structural equation models. In *International Joint Conference on Artificial Intelligence*, 2009.
- J. Tian and J. Pearl. Causal discovery from changes. In *Uncertainty in Artificial Intelligence*, pages 512–521, 2001.
- J. Tian and J. Pearl. A general identification condition for causal effects. In *AAAI*, pages 567–573, 2002a.
- J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. In *Uncertainty in Artificial Intelligence*, pages 519–527, 2002b.
- T. van Ommen and J. M. Mooij. Algebraic equivalence of linear structural equation models. In *Uncertainty in Artificial Intelligence*, 2017.
- T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence*, pages 255–270, 1990.
- N. Wermuth. Probability distributions with summary graph structure. *Bernoulli*, 17(3):845–879, 2011.
- S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

- J. Woodward. Causation and manipulability. *The Stanford Encyclopedia of Philosophy*, 2001.
- S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7): 557–585, 1921.
- S. Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.
- P. Ziernik, C. Uhler, and D. Richards. Maximum likelihood estimation for linear Gaussian covariance models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1269–1292, 2017.