DEPARTMENT OF MATHEMATICS

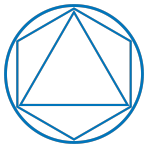
TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis

Confidence in Causal Inference from Interventional Data

Konfidenz in kausaler Analyse interventioneller Daten

Author:	Sanghyun Lee
Supervisor:	Prof. Dr. Mathias Drton
Advisor:	David Strieder
Submission Date:	October 15, 2022



I confirm that this master's thesis is my own work and I have documented all sources and material used.

Munich, 15.10.2022 Sanghyun Lee

1 Introduction

The standard statistical analysis aims to estimate parameters of a distribution from samples. Associations among variables can be inferred by estimating such parameters. However, a general association among variables does not imply a causal relationship among the variables. Causal analysis needs to go one step further. The aim of causal analysis is to infer probabilities under conditions that change. We induce external interventions in order to change probabilities.
Causality is not only interesting but also an essential research topic to understand scientific phenomena and even what happens in our everyday life. Thus causal inference is studied widely in all parts of science. Approaches to causal inference are universally able to be applied across all types of scientific disciplines. The structural causal model describes functional dependency relations among a set of variables. The causal structure of the such model can be represented by a directed graph. The edges in the graph depict the causal dependency between two variables in the set of variables. This intuitive approach to describe the asymmetric relationship among variables was introduced by [1]. In the paper, the directed graph is used to visualize causal structures. A straightforward approach to estimating a causal relation in a set of variables consists of two steps. The graphical structure of data is firstly learned, and the causal effects of each data pair are estimated by statistical methods. This approach enables us to take the uncertainty of the causal effects of each data pair into account, however, not the uncertainty which exists with respect to causal structure. Consequently, this approach leads to an overly optimistic conclusion about the existence and strength of causal effects. Determination of cause and effect from a set of interesting variables is tackled by using asymmetry in a graphical structure of the variables.
This thesis follows the main idea of the paper [2]. As conducted in the paper [2], the confidence intervals for the total causal effect are constructed by using the simplest linear structural equation models in order to infer a causal effect. We consider a simple model called the linear structural equation model with normal errors. In the paper [3], the linear structural equation model with errors that are homoscedastic is used to infer causal effects. Whereas observational data is employed in the paper [2], in this thesis, we additionally access interventional data in the data set and construct a confidence set of a valid hypothetical test for a total causal effect. Interventional data is described as data equipped with intervened observation. Equipped with interventional data in the data set, one can access the conditional probabilities of the set of variables. The following example is a well-known example of interventional data. Suppose that we experiment to investigate the relationship between smoking and teeth color. An interventional data set is obtained by forcing the experiment participants to smoke and checking the yellowness of the participants' teeth. In other words, we intervene the circumstance of the object we observe so that we can restrict the influence of the cause or effect and figure out dependency and causal relationship in data. Our primary interest is to assess the results of a method introduced in [4] by adding interventional data in the data set in the two-variable case and three-variable case as well. By using the method, we can construct a confidence interval of a valid split likelihood ratio test, applying linear models with Gaussian errors, which have non-equal variance. We only focus on the split likelihood ratio

test among the other methods introduced in [2] due to the simplicity of constructing the test without inferring a asymptotic distribution.
In Chapter 2, we will introduce the basic framework of the theory of the graphical model and causal effect inference. The framework is provided by the ideas of [4, 3, 5]. We proceed with the thesis in Chapter 3 by presenting the main idea of linear regression. Accessing interventional data provides us conditional distributions of a variable given other variables. It reduces the problems we have to simple linear regression models since we assume that models follow linear structural equation models. Therefore, we use the linear regression method from Chapter 3 to estimate parameters that arise in the LSEMs. In Chapter 4, we will provide the main idea and methodology from [6] about the universal inference approach. Chapter 4 introduces the concept of the split likelihood ratio test and the mathematical background of this concept. In order to construct a confidence interval, we need to solve a inequality at a fixed confidence level, calculating the so-called profile likelihood function. The definition of the profile likelihood function is also given in this chapter. We continue in Chapter 4 by presenting the calculation of the confidence set and related functions in two- and three-dimensional cases, such as the profile likelihood function. In Chapter 5, we show the experimental result of simulation experiments to compare the result from [2] and evaluate the calculation in Chapter 4.

2 Graphical Model

In this chapter , we will introduce definitions and theorems about the graphical model, which help us understand this thesis. Since we work with interventional data sets to improve outcomes designed in the paper [2], the main methodology of modeling interventions in a system will also be introduced. The definitions and properties in this chapter are initially introduced and provided by [7].

2.1 Graphical Structure

Definition 2.1.1. A directed graph is a pair $G = (V,E)$ consisting of a finite set V and a set $E \subseteq V \times V$ The set V is the vertex set of G, and E is the edge set. we will only consider directed graphs that do not contain any self-loops, so $E \cap \{(v,v) : v \in V\} = \phi$

The following terminology and notation are necessary to grasp the mathematical background in this section. Let $G = (V,E)$ be a directed graph.

· Vertices $v, w \in V$ are adjacent if there is an edge between v and w, i.e., if $(v,w) \in E$ or $(w,v) \in E$

· We will also write $v \to w \in E$ to express that $(v,w) \in E$

· The edge $v \to w \in E$ points from v to w, vertex v is the tail of the edge, and w is the head of the edge. We also say that v and w are the endpoints of the edge.

· G is complete if every pair of distinct vertices is adjacent.

Definition 2.1.2 (Subset, walks and paths)**.**

· A subgraph of $G = (V,E)$ is a graph $G' = (V',E')$ such that $V' \subseteq V$ and $E' \subseteq E \cap (V' \times V')$

· The subgraph induced by $A \subseteq V$ is

$$G_A = (A, E \cap (A \times A)).$$

· A walk in $G = (V,E)$ is a sequence

$$\Pi = (v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n_1}, v_n)$$

with $v_1, \ldots v_n \in V$ and $e_1, \ldots, e_n \in E$ such that

$$e_i = (v_i, v_{i+1}) \text{ or } e_i = (v_{i+1}, v_i) \text{ for all } i = 1, \ldots, n-1$$

· We say that v_1 and v_n are the endpoints of Π and that Π is a walk from v_1 to v_n

· Its length is the number of edges, here $n-1$

· If $v_1,\ldots,v_n$ are all distinct, then Π is a path.

· If $e_i=(v_i,v_{i+1})\ \forall i=1,\ldots,n-1$, then Π is a directed walk or path

In the following, we define relations among vertices. Let $v,w\in V$.

· If $v\to w\in E$ then v is a parent of w and w is a child of v. And,
$\mathrm{pa}(v)=\{w\in V: w\to v\in E\}$ is the set of parents of v,
$\mathrm{ch}(v)=\{w\in V: v\to w\in E\}$ is the set of children of v.

· If there is a directed path from v to W in G, then v is an ancestor of w and w is a descendant of v. And,
$\mathrm{an}(v)=\{w\in V:\exists w\to\ldots\to v\in G\}$ is the set of ancestors of v,
$\mathrm{de}(v)=\{w\in V:\exists v\to\ldots\to w\in G\}$ is the set of descendants of v.
We allow paths of length 0, so $v\in\mathrm{an}(v)$ and $v\in\mathrm{de}(v)$

· For $A\subseteq V$, define
$\mathrm{an}(A)=\cup_{v\in A}\mathrm{an}(v)$
$\mathrm{de}(A)=\cup_{v\in A}\mathrm{de}(v)$

Definition 2.1.3 (Directed Acyclic Graphs)**.** Let $G=(V,E)$ be a directed graph. A directed cycle in G is a directed walk from a vertex v to itself.
A directed acyclic graph (DAG) is a directed graph that dose not contain any directed cycles. In a DAG :

$$\mathrm{an}(v)\cap\mathrm{de}(v)=\{v\}.$$

Otherwise, we have directed cycle in the graph.

We investigate causal effects in a directed graphic under a setup of graphical modeling. We introduce the setup of graphical modeling. Given that $X=(X_v: v\in V)$ is random vector, P^X is joint distribution of X and $G=(V,E)$ is a DAG. For a subset $A\subseteq V$,

$$X_A=(X_v: v\in A).$$

For $A,B,C\subseteq V$, we have the following shorthand

$$A\perp\!\!\!\perp B\mid C:\Longleftrightarrow X_A\perp\!\!\!\perp X_B\mid X_C.$$

The main idea of the mathematical concept of causality is to describe the causal relationship between certain random variables by means of a directed acyclic graph (DAG). The edge between vertices represents the dependency. The directions of the edges correspond to the directions of cause to effect in this framework. The direction between two vertices and the random structure of DAG is the main interest of the work. Determining both is a statistical challenge that we focus on. In this chapter, we introduced the definition of the DAG. We will assume that the structure of the graph, which describes both the presence and the directions of dependencies, is acyclic type. In order to understand this idea, one needs to understand the concept.of topological orderings and its relation to DAG [4].

Definition 2.1.4. Let $G = (V,E)$ be a DAG. A topological ordering of G is a mapping $\sigma : V \to \{1,2,\ldots,d\}$ such that for all $j,k \in V$ we have that $k \in \mathrm{de}(j)$ implies $\sigma(j) < \sigma(k)$.

In the book [8], a theorem is provided which shows that the existence of a topological ordering exactly chracterizes the class of DAGs.

Theorem 2.1.1. Let $G = (V,E)$ be a directed graph. Then G is acyclic if and only if there exists a topological ordering σ of G.

Proof. $\Leftarrow$ Suppose that G has a cycle C and the vertex $j \in C$ satisfies the property of topological ordering, that is, $\sigma(j) < \sigma(k)$ for all $k \in C$. Now we choose $i \in C$ such that $(i,j) \in E$. Then $j \in \mathrm{de}(i)$, therefore $\sigma(i) < \sigma(j)$. However this contradicts the choice of i.

$\Rightarrow$ Due to acyclic property of a DAG, DAG has a node without parents. To check this, we choose any $i \in G$ and follow a path of edges backward from i. Let j be the first vertex which the path passes through. This path is traversed between two visits of j. This is a cycle, and this contradicts that G is a DAG.

□

The Markov property is a widely used assumption for graphical models. Once a distribution satisfies the Markov property with respect to a graph, the graph encodes independence in the distribution. In the following, we provide the definitions and examples of the Markov properties.

Definition 2.1.5 (Local Markov Property)**.** The joint distribution p^X satisfies the local Markov property relative to the DAG G if

$$\forall v \in V : v \perp\!\!\!\perp V \setminus (\mathrm{pa}(v) \cup \mathrm{de}(v)) | \mathrm{pa}(v)$$

Definition 2.1.6 (Pairwise Markov Property)**.** The joint distribution P^X satisfies the pairwise Markov property relative to the DAG G if

$$\forall v,w \text{ not adjacent with } w \notin \mathrm{de}(v) : v \perp\!\!\!\perp w | (V \setminus \mathrm{de}(v)) \setminus \{w\}$$

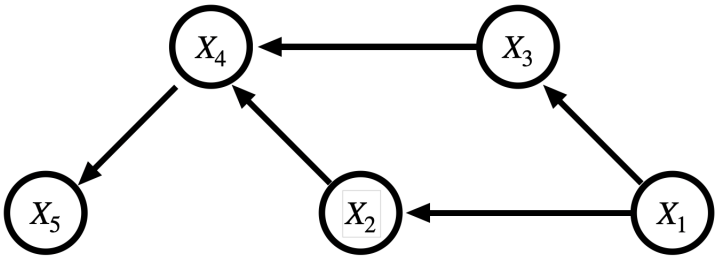


Figure 2.1: The joint distribution with respect to the graph satisfies local M.P. if the condition (2.1) is fulfilled in Example 2.1.1. If the condition (2.2) is fulfilled, the joint distribution satisfies pairwise Markov property.

Example 2.1.1. Consider the graph in Fig.2.1. The local M.P. for the graph is satisfied if

$$\begin{aligned} &X_2 \perp\!\!\!\perp X_4|X_1 \\ &X_4 \perp\!\!\!\perp X_1|\{X_2,X_3\} \\ &X_5 \perp\!\!\!\perp \{X_1,X_2,X_3\}|X_4. \end{aligned} \tag{2.1}$$

The graph satisfies pairwise Markov property if

$$\begin{aligned} &X_2 \perp\!\!\!\perp X_3|X_1 \\ &X_4 \perp\!\!\!\perp X_1|\{X_2,X_3\} \\ &X_5 \perp\!\!\!\perp X_1|\{X_2,X_3,X_4\} \\ &X_5 \perp\!\!\!\perp X_2|\{X_1,X_3,X_4\} \\ &X_5 \perp\!\!\!\perp X_3|\{X_1,X_2,X_4\} \end{aligned} \tag{2.2}$$

There is one more Markov property, which is called the global Markov property. In fact, both local and global Markov properties are equivalent. We will not prove it in this thesis. We will introduce, however, the further definition and theorem.

Definition 2.1.7 (Global Markov Property)**.** The joint distribution P^X satisfies the global Markov property relative to the DAG G if

$$A \perp\!\!\!\perp B|C, \ \ \forall A,B,C \subset V \text{ disjoint}, A,B \notin \emptyset,$$

such that C d-separates A and B. In other word, A and B are d-separate given C.

To understand this definition, we need to know the definition of d-separation and collider and non-collider on paths.

Definition 2.1.8 (Collider and Non-collider on Paths)**.** Let $\Pi = (v_1, e_1, v_2, \ldots, v_{n-1}, e_{n-1}, v_n)$ be a path in a DAG $G = (V,E)$. A non-end point vertex v_i, $2 \leq i \leq n-1$, is a collider on the path if v_i is the head of both e_{i-1} and e_i. The graphical structure of the collider is drawn as

$$v_{i-1} \rightarrow v_i \leftarrow v_{i+1}.$$

Otherwise, if v_i is the tail of e_{i-1} or of e_i, then v_i is a non-collider on the path. The graphical structure of the non-collider is drawn as

$$\begin{aligned} &v_{i-1} \rightarrow v_i \rightarrow v_{i+1}, \text{ or} \\ &v_{i-1} \leftarrow v_i \rightarrow v_{i+1}, \\ &v_{i-1} \leftarrow v_i \leftarrow v_{i+1}. \end{aligned}$$

Definition 2.1.9 (d-Separation)**.** Let $v,w \in V$, and $C \subset V \setminus \{v,w\}$. Then v and w are d-connected given C if there exists a path Π from v to w such that

i) every collider on Π is in $\text{an}(C)$, and

ii) every non-collider on Π is not in C.

Example 2.1.2. Consider the directed graphs in Fig.2.2. In the graph a) X_2 d-separates X_1, and X_3, that is, X_2 is a collider and therefore the joint distribution satisfies the global Markov property. However, in graphs b),c), and d), X_2 is non-collider. Thus, the joint distributions do not satisfy the global Markov property relative to the DAGs. In this case, we can conclude that these three DAGs are Markov equivalent.

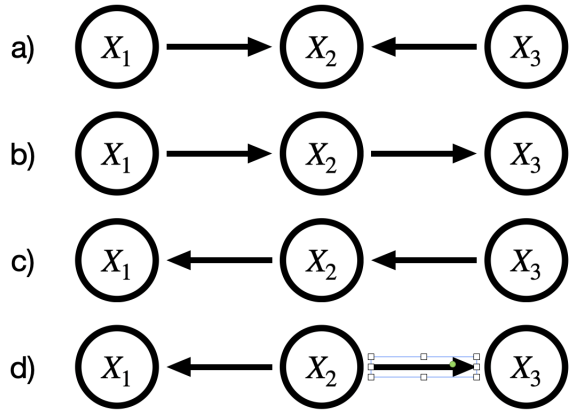


Figure 2.2: The example of the directed graphs which are Markov equivalent to each other.

Remark. As mentioned previously, the local Markov property and global Markov property are equivalent. In other words, The joint distribution P^X satisfies the local Markov property relative to a graph if and only if P^X satisfies the global Markov property relative to the graph. We do not provide proof of this equivalence in the thesis. The detailed proof can be found in chapter 6 [3].

2.2 Density Factorization

We assume that $G = (V,E)$ is a directed acyclic graph (DAG) where V is the vertices and E is the edges of the graph G and P^X is the joint distribution for $X = (X_v, v \in V)$ with density f with respect to a product measure $\mu = \otimes_{v\in V}\mu_v$.

Definition 2.2.1. The distribution P^X factorizes according to the DAG G if there exist non-negative kernel functions $(k_v(x_v, x_{\mathrm{pa}(a)}))_{v\in V}$ with

$$\int k_v(x_v, x_{\mathrm{pa}(v)}) d\mu_v(x_v) = 1 \quad \forall v \in V, \ \ \forall x_{\mathrm{pa}(v)}$$

such that

$$f(x) = \prod_{v\in V} k_v(x_v, x_{\mathrm{pa}(v)}) \quad [\mu - a.s.]$$

The following proposition provides that conditional functions given the parents are the kernel functions.

Proposition 1. For all $v \in V$, the kernel function k_v in the factorization are the conditional densities, i.e.,

$$k_v(x_v, x_{\mathrm{pa}(v)}) = f(x_v|x_{\mathrm{pa}(v)}) \quad [\mu - a.s.].$$

Proof. We prove it by the mathematical induction on $m = |V|$.

m = 1 clearly, the claim is true.

m $\to$ m+1 Choose a terminal vertex t. Then for all $v \neq t$ we know $t \notin \mathrm{pa}(v)$. Therefore,

$$\int f(x) d\mu_t(x_t) = \prod_{v \in V \setminus \{t\}} k_v(x_v, x_{\mathrm{pa}(v)}) \cdot \int k_t(x_t, x_{\mathrm{pa}(t)}) d\mu_t(x_t)$$

where the first term is obtained by the factorization according to $G_{V\setminus\{t\}}$ and $k_t(x_t, x_{\mathrm{pa}(t)}) = 1$ since t is a terminal vertex. By induction assumption,

$$k_v(x_v, x_{\mathrm{pa}(v)}) = f(x_v|x_{\mathrm{pa}(v)}) \ \ a.s. \ \text{ for all } v \neq t$$

Furthermore,

$$k_t(x_t, x_{\mathrm{pa}(t)}) = \frac{f(x)}{\int f(x) d\mu_t(x_t)} = f(x_t|x_{V\{t\}})$$

This yields that

$$f(x_t|x_{V\setminus\{t\}}) = f(x_t|x_{\mathrm{pa}(t)})$$

□

Theorem 2.2.1. Let P^X have a density with respect to a product measure μ. Then P^X factorizes according to the DAG G if and only if P^X satisfies the local Markov property for G.

Proof. First, we prove the claim that local M.P. implies factorization of the density. This is proven again by the mathematical induction on $m = |V|$.

$m = 1$ Trivially, the claim is true.

$m \to m+1$ We choose a terminal vertex t and consider the density

$$f(x) = f(x_t|x_{V\setminus\{t\}}) f(x_{V\setminus\{t\}}).$$

The marginal distribution of $X_{V\setminus\{t\}}$ satisfies the local M.P. for $G_{V\setminus\{t\}}$, since $V \setminus \{t\}$ is ancestral. By the induction assumption, we have

$$\begin{aligned} f(X_{V\setminus\{t\}}) &= \prod_{v \in V\setminus\{t\}} f(x_v|x_{\mathrm{pa}_{G_{V\setminus\{t\}}}(v)}) \\ &= \prod_{v \in V\setminus\{t\}} f(x_v|x_{\mathrm{pa}_G}(v)) \end{aligned}$$

Additionally, the local M.P. for G yields

$$t \perp\!\!\!\perp V \setminus [\mathrm{pa}_G(t) \cup \{t\}] | \mathrm{pa}_G(t) \Rightarrow f(x_t | x_{V\setminus\{t\}}) = f(x_t | x_{\mathrm{pa}_G(t)})$$

Hence, we conclude

$$f(x) = \prod_{v \in V} f(x_v | x_{\mathrm{pa}_G(v)})$$

Now, we prove the opposite direction of the implication. This proof is also done by the mathematical induction on $m = |V|$

$\underline{m = 1}$ This is trivial.

$\underline{m \to m+1}$ We choose a terminal vertex t again. As in the Prop. 1,

$$f(x_{V\setminus\{t\}}) = \prod_{v \in V\setminus\{t\}} f(x_v | x_{\mathrm{pa}(v)}) \tag{2.3}$$

factors according to $G_{V\setminus\{t\}}$. Therefore,

$$f(x_t | x_{V\setminus\{t\}}) \frac{f(x)}{f(x_{V\setminus\{t\}})} = f(x_t | x_{\mathrm{pa}(t)})$$

and we obtain that

$$t \perp\!\!\!\perp V \setminus [\mathrm{pa}(t) \cup \{t\}] | \mathrm{pa}(t). \tag{2.4}$$

This is the statement about t, which is made by the local Markov property for G. We still have to show that

$$v \perp\!\!\!\perp V \setminus [\mathrm{pa}(v) \cup \mathrm{de}(v)] | \mathrm{pa}(v), \quad \text{for all } v \in V \setminus \{t\}. \tag{2.5}$$

By Eq.(2.3) and the induction hypothesis, $X_{V\setminus\{t\}}$ satisfies local Markov property for $G_{V\setminus\{t\}}$. In other words, for $v \in V \setminus \{t\}$,

$$v \perp\!\!\!\perp (V \setminus \{t\}) \setminus [\mathrm{pa}_{G_{V\setminus\{t\}}}(v) \cup \mathrm{de}_{G_{V\setminus\{t\}}}(v)] | \mathrm{pa}_{G_{V\setminus\{t\}}}(v) \tag{2.6}$$

Since t is terminal vertex,

$$\mathrm{pa}_{G_{V\setminus\{t\}}}(v) = \mathrm{pa}_G(v) = \mathrm{pa}(v) \text{ for all } v \neq t.$$

Hence, (2.5) is equal to (2.6) if

$$\mathrm{de}(v) = \mathrm{de}_{G_{V\setminus\{t\}}}(v) \cup \{t\}.$$

The case that $t \notin \mathrm{de}(v)$ is still considered. We have to add t to (2.6). Let $t \notin \mathrm{de}(v)$. In other words, $v \notin \mathrm{pa}(t)$. We know from (2.6) that

$$v \perp\!\!\!\perp (V \setminus [\mathrm{pa}(v) \cup \mathrm{de}(v)]) \setminus \{t\} | \mathrm{pa}(v).$$

If we can argue that

$$v \perp\!\!\!\perp t | \mathrm{pa}(v) \cup (V \setminus [\mathrm{pa}(v) \cup \mathrm{de}(v)]) \setminus \{t\} \tag{2.7}$$

and

$$\mathrm{pa}(v) \cup (V \setminus [\mathrm{pa}(v) \cup \mathrm{de}(v)]) \setminus \{t\} = V \setminus [\mathrm{de}(v) \cup \{t\}]$$

the property of conditional independence gives Eq. (2.5), and the proof is done. However we know from Eq. (2.4) that

$$t \perp\!\!\!\perp V \setminus [\mathrm{pa}(v) \cup \{t\}] | \mathrm{pa}(t)$$

that implies (2.7) by properties of conditional independence.

□

2.3 Causal Model

In many research fields, such as biology, sociology, and economics, the problem of learning the dependency structure using certain measurements from observational data is a major challenge. In order to start causal modeling, we need to understand a causal structure. This structure entails a probability model.[9]. In this chapter, we will introduce a fundamental framework of causal effect inference.[5]. The definitions adopted from [4] which are originally provided by [9, 5]. This chapter begins with an interesting experiment. Consider an example of an experiment to investigate vitamin supplements' effect on health. To implement an observational study, we randomly select n experiment participants and observe each participants:

$$X = \text{"Vitamin supplement"}$$
$$Y = \text{"Health outcome"}$$

Then, we have two mathematical models for the joint distribution of (X,Y). Firstly, X causes Y. Secondly, Y causes X. Since both cases yield that X and Y may be arbitrarily dependent, these are Markov equivalents. However, those models say intuitively something very different. Now, we carry out a similar but different experiment. In this case, we randomly divide the participants into a treatment and a control group. Therefore, the treatment group takes the prescribed amount of supplements, and the control group takes no vitamin supplements. From both experiments, it is expected that this intervention changes the distribution of the system compared to the behavior only with the observational outcome. In detail, we will introduce the intervention's main framework.
One interesting starting point for investigating causal relations is that the correlation does not directly imply a causal relation. For example, consider variables X and Y again, and assume that there are correlations between X and Y. Then, the correlation in an observational study may arise from the three cases shown in Figure 2.3. Consider a observed random vector $X = (X_1, \ldots, X_m)$, and a directed acyclic graph (DAG) $G = (V,E)$ with $V = \{1, \ldots, m\}$. The function f denotes the joint density of X with respect to a product measure. The Factorization according to G is as follows :

$$f(x) = \prod_{v \in V} f(x_v | x_{\mathrm{pa}(v)}).$$

In order to formalize the intuitive understanding of how G captures causal relations, we have to figure out a model for the joint distribution of X when we intervene on a subvector X_A of the vector X.

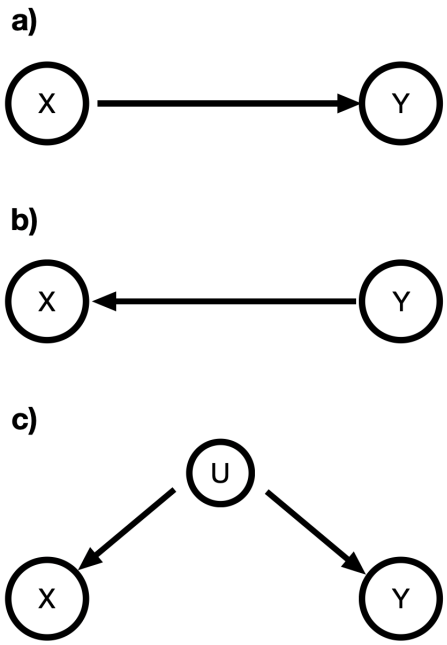


Figure 2.3: a) X has a causal effect on Y. b) Y has an effect on X. c) X and Y are both affected by an unobserved/latent variable U

Definition 2.3.1 (Interventional Distribution)**.** A full family of interventional distribution is a family

$$(P^X_{A,x^*_A})_{A\subseteq V, x^*_A \in \mathbb{R}^d}$$

where each P^X_{A,x^*_A} is a joint distribution for the considered random vector X such that $\{X_A = x^*_A\}$ with probability 1 when $X \sim P^X_{A,x^*_A}$.

The distribution P^X_{A,x^*_A} can be interpreted as the distribution P^X_{A,x^*_A} is the distribution of X under a intervention $\mathrm{do}(X_A = x^*_A)$. If $A = \emptyset$, $P^X_A = P^X_\emptyset = P^X$ is usual joint distribution P^X which is the observational distribution. The following notation can be used alternatively.

$$P(X \in \cdot \ ; \mathrm{do}(X_A = x^*_A)) = P^X_{A,X^*_A}(\cdot)$$

Remark. Interventions that fix values are also called perfect interventions.

Causal analysis in graphical models starts with the understanding that all causal effects are identifiable whenever the model satisfies the causal Markov property. That is, the graph is DAG, and all the error terms are independent. Models which do not satisfy the property, for example, a model with correlated errors, allow identification only under certain conditions. These conditions can only be determined from the structure of the graph. [5] In the following, we will introduce the definition of causal Markov property and its relation to causal analysis.

Definition 2.3.2 (Causal Markov Property)**.** A full family of interventional distributions for random vector X satisfies the causal Markov property for the DAG $G = (V,E)$ if for all $A \subseteq V, x^*_A \in \mathbf{R}^d$:

i) $P(x \in \cdot \ ; \mathrm{do}(X_A = x_A^*))$ factorizes according to G, and thus satisfies the local/ global M.P for G

i) $P(x_v \in \cdot \quad | X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}; \mathrm{do}(X_A = x_A^*)) = P(x \in \cdot \quad | X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)})$ for all $v \notin A$ and $x_u = x_u^*$ for all $u \in \mathrm{pa}(v) \cap A$.

Remark. i) Taking fixed values by intervention does not induce new dependencies

ii) A stochastic transition transforms $X_{\mathrm{pa}(v)}$ into X_v. Interventions don't change the stochastic transition mechanism for variables not intervened upon.

Proposition 2 (Truncated Factorization)**.** A full family of intervention distribution satisfies the causal Markov property if and only if the joint density of $P(X \in \cdot \quad ; \mathrm{do}(X_A = x_A^*))$ factors as

$$f(x; \mathrm{do}(X_A = x_A^*)) = \prod_{v \notin A} f(x_v | x_{\mathrm{pa}(v)}) \prod_{v \in A} \mathbb{1}_{\{x_v = x_v^*\}} \tag{2.8}$$

We give here a example of the truncated factorization. Consider a graph in Figure 2.4. The observational distribution of the graph factors as

$$f(x) = f(x_1) f(x_2|x_1) f(x_3|x_1, x_2).$$

However, interventional distribution under $\mathrm{do}(X_2 = x_2^*)$ factors as

$$f(x; \mathrm{do}(X_2 = x_2^*)) = f(x_1) f(x_3|x_1, x_2 = x_2^*) \text{ if } X_2 = x_2^*$$

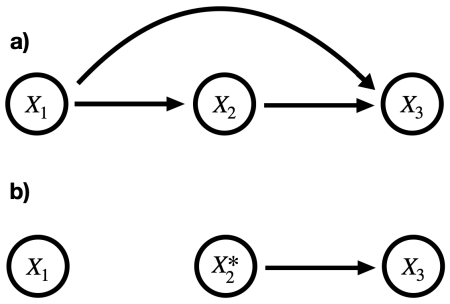


Figure 2.4: a) Observational distribution. b) Interventional distribution under $\mathrm{do}(X_2 = x_2*)$

The graph intervened upon is called mutilated DAG. We define the mutilated DAG.

Definition 2.3.3. Let $G = (V, E)$ be a DAG, and let $A \subseteq V$. The subgraph $G_{\mathrm{do}(A)} = (V, E_{\mathrm{do}(A)})$ with edge set $E_{\mathrm{do}(A)} = E \setminus \{w \to v : w \in V, v \in A\}$ is called the mutilated DAG (representing an intervention on A)

Figure 2.4 shows an example of a mutilated graph. The upper graph is a DAG, and the lower graph is the DAG under $\mathrm{do}(X_2 = x_2^*)$. Now, we give a practical example of applying the intervention. The experiment is called a sanity check experiment.

Example 2.3.1 (Sanity check)**.** In the experiment, the participants will be asked if they are smoking or not. Furthermore, it is also observed how yellow the teeth of the participants are. S denotes observation of smoking, and Y denotes observation of yellow teeth. Intuitively, we know that smoking causes yellow teeth. That is, the smoking behavior of a participant affects their yellow teeth. There is a dependency between S and Y. The observational distribution of this dependency factors as

$$f(Y,S) = f(Y|S)f(S).$$

Moreover, the distributions of mutilated DAGs under $\text{do}(S = s)$ and $\text{do}(Y = y)$ are

$$f(Y;\text{do}(S = s)) = f(Y|s)$$
$$f(S;\text{do}(Y = y)) = f(S).$$

Suppose we intervene on the variable S smoking status and set it to 0 or 1. In that case, this intervention changes the distribution of the system compared to the behavior only with the observational outcome. We construct intervention distributions from a structure causal model (SCM). We obtain the distributions by modifying the SCM and considering the new entailed distribution. Generally, intervention distributions and observation distributions are different. [9]

2.4 Linear Structural Equation Model

Definition 2.4.1. Let X be a random vector. Structure equation models assume X solves a system of the equation given by a DAG G:

$$X_v = g_v(X_{\text{pa}(v)}, \varepsilon_v) \quad v = 1,\dots,m. \tag{2.9}$$

Since G is a DAG, the equation system admits a unique solution: the system is triangular when considering a topological order.

Satisfying the local M.P. for a graph G yields many properties. The following proposition shows that an assumption for error terms $\varepsilon_1,\dots,\varepsilon_m$ implies the local M.P. for a graph G.

Proposition 3. If $\varepsilon_1,\dots,\varepsilon_m$ are independent then the joint distribution of X, the solution to the Eq.(2.9), satisfies the local M.P. for G

Proof. In order to prove this proposition, we use mathematical induction on m.

Base $m = 1$ X is trivially the solution for the Eq. (2.9)

Step $m \to m+1$ Let t be a terminal vertex. Then $X_{V\setminus\{t\}}$ solves the Eq.(2.9) with the equation for $v = t$ dropped. By induction assumption, $X_{V\setminus\{t\}}$ satisfies local M.P. for $G_{V\setminus\{t\}}$. Since $X_{V\setminus\{t\}} = h(\varepsilon_v, v \neq t)\varepsilon_t \perp\!\!\!\perp \varepsilon_t$, the conditional distribution of X_t given $X_{V\setminus\{t\}} = x_{V\setminus\{t\}}$ is the distribution of $g_t(x_{\text{pa}(t)}, \varepsilon_t)$. Hence,

$$t \perp\!\!\!\perp V \setminus [\text{pa}(t) \cup \{t\}]|\text{pa}(t).$$

To conclude that $X_{V\setminus\{t\}}$ satisfies local M.P. for G (not only $G_{V\setminus\{t\}}$) we may argue as in proof of Lemma 2.2.1

□

Remark. There exist functions g_v and independent random variables $\varepsilon_1, \ldots, \varepsilon_m$ such that the distribution of the random vector X equals P^X where X solves the Eq.(2.9) for every distribution which satisfies the local M.P. for G.

These structural equations can be differently viewed for a causal interpretation as making an assignment:

$$X_v := g_v(X_{\mathrm{pa}(v)}, \varepsilon_v) \quad v = 1, \ldots, m. \tag{2.10}$$

The interventional model for X under the intervention $\mathrm{do}(X_A = x_A^*)$ is obtained by replacing the equations for $w \in A$ by

$$X_w := x_w^* \quad , w \in A.$$

$X(\mathrm{do}(X_A = x_A^*))$ denotes the new solution for the Eq.(2.10) under the intervention $\mathrm{do}(X_A = x_A^*)$. The new solution has the new distribution which consists of distribution of $(\varepsilon_v)_{v \notin A}, (g_v)_{v \notin A}, x_A^*$.

Example 2.4.1. Consider the DAG again in Figure 2.4. Given parametrizing functions $g = (g_1, g_2, g_3)$ and distributions $Q = (Q_1, Q_2, Q_3)$, where $\varepsilon_i \sim Q_i$ for $i = 1, 2, 3$, the structural equations with observational data are

$$\begin{aligned} X_1 &= g_1(\varepsilon_1) \\ X_2 &= g_2(X_1, \varepsilon_2) \\ X_3 &= g_3(X_1, X_2, \varepsilon_3) \end{aligned}$$

where $\varepsilon_1, \ldots, \varepsilon_4$ are independent. Intervening on the variable X_2, we have the graph G as in Figure 2.4 b). The equations for the interventional data are

$$\begin{aligned} X_1 &= g_1(\varepsilon_1) \\ X_2 &= x_2^* \\ X_3 &= g_3(X_1, X_2, \varepsilon_3) \end{aligned}$$

Definition 2.4.2. Let $X = (X_1, \ldots, X_d)$ be a random vector. A structural causal model (SCM) or structural equation model (SEM) $\mathfrak{C} = (S, Q)$ for X consists of a collection of d structural assignments

$$X_j := f_j(\mathrm{pa}(j), \varepsilon_j), \quad j = 1, \ldots, d, \tag{2.11}$$

where $\mathrm{pa}(j) \subset \{X_1, \ldots, X_d\} \setminus \{X_j\}$ are called parents of X_j, and a distribution Q over the noise variables $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d) \sim Q$. Hereby, we require Q to be a product distribution, that is, $\varepsilon_1, \ldots, \varepsilon_d$ are independent.

Definition 2.4.3. A linear structural equation model (LSEM) for X is an SEM of the form

$$X_j = \sum_{i \in \mathrm{pa}(j)} \beta_{ji} X_i + \beta_{0j} + \varepsilon_j, \quad i, j = 1, 2, \ldots, d \tag{2.12}$$

where $\beta_{jk} \in \mathbb{R}$ for all $j \in \{1, \ldots, d\}$ and all $k \in \mathrm{pa}(j)$ and $\varepsilon_1, \ldots, \varepsilon_d$ are independent random variables with mean zero.

We define the matrix $B = (\beta_{jk}^d)_{j,k=1}$ such that $\beta_{jk} = 0$ for all pairs $(j,k) \in \{1,\dots,d\}^2$ where $k \notin \mathrm{pa}(j)$, and the random vector $\varepsilon = (\varepsilon_1,\dots,\varepsilon)$, we obtain equivalent model in the compact form

$$X = BX + \varepsilon.$$

The dependence structure in this model is entirely encoded by the matrix B. Since $\mathscr{E} = \{(k,j) \in \{1,\dots,d\}^2 : \beta_{jk} \neq 0\}$, the edge set $\mathscr{E}$ of the corresponding DAG $\mathscr{G}$ relies on the model (2.12) through the matrix B. The value $\beta_{jk} = 0$ implies the nonexistence of a causal relationship of X_k and X_j, while the value $\beta_{jk} \neq 0$ reflects the existence of the relationship. In the following, we will prove that the matrix $(\mathrm{Id} - B)$ is invertible so that we can express X in a simple formulation

$$X = (\mathrm{Id} - B)^{-1}\varepsilon. \tag{2.13}$$

Suppose σ is a topological ordering for $\mathscr{G}$ and S denotes the permutation matrix $(\mathbb{1}_{\{\sigma(i)=j\}})_{i,j=1}^d$. This yields that the (i,j)-th entry of the matrix $S^T BS$ is

$$[S^T BS]_{ij} = \sum_{\mu,\nu=1}^{d} \mathbb{1}_{\{i=\sigma(\mu)\}} B_{\mu\nu} \mathbb{1}_{\{j=\sigma(\nu)\}} = B_{\sigma^{-1}(i)\sigma^{-1}(j)}.$$

We know that for $i \leq j$ $\sigma^{-1}(i) \notin \mathrm{de}(\sigma^{-1}(j))$ and therefore $B_{\sigma^{-1}(i)\sigma^{-1}(j)} = 0$. This yields that the acyclicity of the graph depicting the model (2.12) concludes that the matrix B is a permutation similar to a strictly lower triangular matrix. This fact reflects that $S^T(\mathrm{Id} - B)S = (\mathrm{Id} - S^T BS)$ is invertible and thus $(\mathrm{Id} - B)$ is also invertible. Hence, we have the alternative representation of X (2.13). This formulation shows that $\mathbb{E}[X] = 0$ and

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[XX^T] = (\mathrm{Id} - B)^{-1}\mathbb{E}[\varepsilon\varepsilon^T](\mathrm{Id} - B)^{-T}. \tag{2.14}$$

By the assumption in Definition 2.4.2 the matrix $\mathbb{E}[\varepsilon\varepsilon^T]$ is determined by variance of ε_i

$$\mathbb{E}[\varepsilon\varepsilon^T] = \begin{pmatrix} \mathbb{E}[\varepsilon_1\varepsilon_1] & \cdots & \mathbb{E}[\varepsilon_1\varepsilon_d] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\varepsilon_d\varepsilon_1] & \cdots & \mathbb{E}[\varepsilon_d\varepsilon_d] \end{pmatrix} = \begin{pmatrix} \mathbb{V}\mathrm{ar}[\varepsilon_1] & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \mathbb{V}\mathrm{ar}[\varepsilon_d] \end{pmatrix}$$

Theorem 2.4.1. Suppose that an LSEM (2.12) for a random vector X is given. Let σ be a topological ordering for X and let $i,j \in \{1,\dots,d\}$. Then

$$\begin{aligned} \mathscr{C}(j \to i) =& \left[(\mathrm{Id} - B)^{-1}\right]_{ij} \\ =& \delta_{ij} + \beta_{ij} + \sum_{\substack{k_1=1 \\ \sigma(1)>\sigma(k_1)>\sigma(j)}}^{d} \beta_{i,k_1}\beta_{k_1,j} + \sum_{\substack{k_1,k_2=1 \\ \sigma(1)>\sigma(k_1)>\sigma(k_2)>\sigma(j)}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\beta_{k_2,j} \\ &+ \cdots + \sum_{\substack{k_1,k_2,\dots,k_{\sigma(j)-\sigma(i)}=1 \\ \sigma(1)>\sigma(k_1)>\sigma(k_2)>\cdots>\sigma\left(k_{\sigma(j)-\sigma(i)-1}\right)>\sigma(j)}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdots\beta_{k_{\sigma(j)-\sigma(i)-1},j}. \end{aligned} \tag{2.15}$$

Proof. The proof is provided in [4]. In this proof, we will be concerned many times with sums of the form

$$\sum_{(k_1,\dots,k_l)\in B} \beta_{i,k_1}\beta_{k_1,k_2}\beta_{k_2,k_3}\cdots\beta_{k_{l-1},k_l}\alpha_{k_l},$$

where the α_{k_l} denote numbers or random variables and B is a specific index set contained is $\{1,\dots,d\}^l$. Due to the acyclic structure of the LSEM, $\beta_{k_j,k_{j+1}} = 0$ if $\sigma(k_j) \leq \sigma(k_{j+1})$ and hence

$$\begin{aligned}\sum_{(k_1,\dots,k_l)\in B} &\beta_{i,k_1}\beta_{k_1,k_2}\beta_{k_2,k_3}\cdots\beta_{k_{l-1},k_l}\alpha_{k_l} \\ &= \sum_{\substack{(k_1,\dots,k_l)\in B \\ \sigma(i)>\sigma(k_1)>\cdots>\sigma(k_l)}} \beta_{i,k_1}\beta_{k_1,k_2}\beta_{k_2,k_3}\cdots\beta_{k_{l-1},k_l}\alpha_{k_l}\end{aligned} \tag{2.16}$$

Especially, if $d \leq l$, then we have for all $i, j \in \{1,\dots,d\}$

$$\sum_{k_1,\dots,k_l=1}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\beta_{k_2,k_3}\cdots\beta_{k_{l-1},k_l}\alpha_{k_l} = 0$$

and we conclude that $B^l = 0$ for all $d \leq l$. Thus,

$$(\mathrm{Id}+B+B^2+\cdots+B^{d-1})(\mathrm{Id}-B) = \mathrm{Id}-B^d = Id.$$

This yields

$$(\mathrm{Id}-B)^{-1} = \sum_{k=0}^{d-1} B^k$$

and

$$\begin{aligned}[(\mathrm{Id}-B)^{-1}]_{ij} =& \delta_{ij} + \beta_{ij} + \sum_{k_1=1}^{d} \beta_{i,k_1}\beta_{k_1,j} + \sum_{k_1,k_2=1}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\beta_{k_2,j} \\ &+\cdots+ \sum_{k_1,k_2,\dots,k_{d-1}=1}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdots\beta_{k_{d-1},j}.\end{aligned} \tag{2.17}$$

The right hand side in Eq.(2.17) is equal to the last expression in Eq.(2.15) due to Eq.(2.16).
Now we fix $i, j \in \{1,\dots,d\}$. If $i = j$, then due to (2.16) every term except δ_{ij} on the right hand side of (2.17) vanishes. Since $\mathscr{C}(i \to i) = 1$ for all i, we have equality of the diagonal entries. Furthermore, suppose that $i \neq j$ and consider linear SEM obtained from the original linear SEM in Eq.(2.12) after we

intervene on the variable $X_j := x$. In the following, we show that

$$\begin{aligned} X_i =& \varepsilon_i + \beta_{ij}x + \sum_{\substack{k_1=1 \\ k_1\neq j}}^{d} \beta_{i,k_1}\varepsilon_{k_1} + \sum_{k_1=1}^{d} \beta_{i,k_1}\beta_{k_1,j}x + \sum_{\substack{k_1,k_2=1 \\ k_1,k_2\neq j}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\varepsilon_{k_2} + \\ &+ \sum_{k_1,k_2=1}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\beta_{k_2,j}x \\ &+ \cdots + \sum_{\substack{k_1,k_2,\ldots,k_{l-1}=1 \\ k_1,k_2,\ldots,k_{l-1}\neq j}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdot\beta_{k_{l-2},k_{l-1}}\varepsilon_{k_{l-1}} + \\ &+ \sum_{k_1,k_2,\ldots,k_{l-1}=1}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdot\beta_{k_{l-2},k_{l-1}}x + \sum_{\substack{k_1,k_2,\ldots,k_l=1 \\ k_1,k_2,\ldots,k_l\neq j}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdot\beta_{k_{l-1},k_l}X_{k_l} \end{aligned} \tag{2.18}$$

We continue the proof by mathematical induction.

$\underline{l=1}$ Trivially, since we set $X_j = x$,

$$X_i = \varepsilon_i + \sum_{k=1}^{d} \beta_{ik}X_k = \varepsilon_i + \beta_{ij}x + \sum_{\substack{k=1 \\ k\neq j}}^{d} \beta_{ik}X_k.$$

$\underline{l \to l+1}$ Suppose (2.18) holds for some $l < d$. Then

$$\begin{aligned} &\sum_{\substack{k_1,k_2,\ldots,k_l=1 \\ k_1,k_2,\ldots,k_l\neq j}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdots\beta_{k_{l-1}k_l}X_{k_l} \\ &\quad= \sum_{\substack{k_1,k_2,\ldots,k_l=1 \\ k_1,k_2,\ldots,k_l\neq j}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdots\beta_{k_{l-1}k_l}\left[\varepsilon_{k_l} + \sum_{k_{l+1}=1}^{d} X_{k_{l+1}}\right] \\ &\quad= \sum_{\substack{k_1,k_2,\ldots,k_l=1 \\ k_1,k_2,\ldots,k_l\neq j}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdots\beta_{k_{l-1}k_l}\varepsilon_{k_l} + \sum_{k_1,k_2,\ldots,k_l=1}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdots\beta_{k_{l-1}k_l}\beta_{k_l,j}x \\ &\qquad+ \sum_{\substack{k_1,k_2,\ldots,k_{l+1}=1 \\ k_1,k_2,\ldots,k_{l+1}\neq j}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdots\beta_{k_{l-1}k_l}\beta_{k_l,k_{l+1}}\varepsilon_{k_{l+1}} \end{aligned}$$

Plugging this into (2.18) gives exactly (2.18) with $l+1$ in place of l.

Now, we consider the Eq.(2.18) for the case $l = d$. Then the last term

$$\sum_{\substack{k_1,k_2,\ldots,k_d=1 \\ k_1,k_2,\ldots,k_d\neq j}}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdots\beta_{k_{d-1}k_d}X_{k_d} = 0.$$

Now, we take a expectation with respect to the interventional distribution, then

$$\mathbb{E}[X_i; \mathrm{do}(X_j = x)] = \beta_{ij}x + \sum_{k_1=1}^{d} \beta_{i,k_1}\beta_{k_1,j}x + \cdots + \sum_{k_1,k_2,\ldots,k_d=1}^{d} \beta_{i,k_1}\beta_{k_1,k_2}\cdots\beta_{k_d,j}x. \tag{2.19}$$

By differentiating with respect to the value x we can conclude that $\mathscr{C}(i \to j)$ is equal to the expression in Eq.(2.17) for $i \neq j$ □

2.5 Causal Effects

Many statistical studies are mainly aimed at predicting the effects of interventions. In the previous chapter, we learned that interventions change the joint distribution of a causal model. In this section, we introduce the concept of the total causal effect between two variables in a graph and the identifiability of a graphical structure from a joint distribution.

Definition 2.5.1. Let $G = (V, E)$ be a DAG with $V = \{1, \ldots, m\}$ and $X = (X_1, \ldots, X_m)$ be a random vector with factorizing density

$$f(x) = \prod_{v \in V} f(x_v | x_{\mathrm{pa}(v)}).$$

Assume that all conditional distributions are uniquely determined as

$$f(x_A | x_B) = \frac{f(x_A, x_B)}{f(x_B)}$$

and are positive. Assuming for the DAG G the full family of the interventional distributions $P(X \in \cdot\ ; \mathrm{do}(X_A = x_A^*))$ satisfies the causal Markov property with $A \subseteq V$, $x_A^* \in \mathbb{R}^A$, we have densities such that

$$\begin{aligned} f(x; \mathrm{do}(X_A = x_A^*)) &= \prod_{v \notin A} f(x_v | x_{\mathrm{pa}(v)} \prod_{v \in A} \mathbb{1}_{\{x_v = x_v^*\}} \\ &= \frac{f(x)}{\prod_{v \in A} f(x_v | x_{\mathrm{pa}(v)})} \prod_{v \in A} \mathbb{1}_{\{x_v = x_v^*\}}. \end{aligned}$$

Let $T, R \subseteq V$. The causal effect of X_T on X_R is the map

$$X_T \mapsto P(X_R \in \cdot\ ; \mathrm{do}(X_T = x_T)), \quad x_T \in \mathbb{R}^T$$

Alternatively, we write the definition of the (total) causal effect of a variable on another variable

$$\mathscr{C}(i \to j) := \frac{d}{dx} \mathbb{E}[X_i; \mathrm{do}(X_i = x_i^*)],$$

where $i, j \in V$.

Note that the notations $P(\cdot)$, and $\mathbb{E}[\cdot]$ denote calculating probabilities and taking expectation with respect to the distribution of an original SEM $\mathfrak{C}$. The notations $P(\cdot; \mathrm{do}(X_k = x_k^*)$, and $\mathbb{E}[\cdot; \mathrm{do}(X_k = x_k^*)]$ will be used to express calculating probabilities and taking expectation, modifying a SEM $\tilde{\mathfrak{C}}$ obtained from the original SEM $\mathfrak{C}$ by intervening on the variable X_k and setting this to x_k^*.

Example 2.5.1. Here, we give two examples with specific derived quantities.

a) X_T is binary, then

$$\mathbb{E}[X_R;\ \mathrm{do}(X_T = 1)] - \mathbb{E}[X_R;\ \mathrm{do}(X_T = 0)]$$

b) The structural equation is linear. Consider $\mathbb{E}[X_R;\ \mathrm{do}(X_T = x_T)] = \beta_0 + \beta x_T$, then

$$\frac{\partial}{\partial x_T}\mathbb{E}[X_R;\ \mathrm{do}(X_T = x_T)] = \beta$$

If a variable holds no parents, there exists a straightforward way to find the causal effect of the variable on other variables in a DAG g with intervention. Let $v \in V$ be a variable of the DAG G such that v has no parents, i.e., $\mathrm{pa}(v) = \emptyset$. Then, the causal effect in this case is

$$\begin{aligned} f(x_{V\setminus\{w\}}; \mathrm{do}(X_w = x_w^*)) &= \prod_{v \neq w} f(x_v | x_{\mathrm{pa}(v)})\Big|_{x_w = x_w^*} \\ &= \frac{f(x_w^*, x_{V\setminus\{w\}})}{f(x_w^*)} = f(x_{V\setminus\{w\}} | x_w^*) \end{aligned}$$

This result conclude that the causal effect for $\mathrm{do}(x_w = x_w^*)$ is determined by usual probabilistic conditioning $f(x_{V\setminus\{w\}}|x_w^*)$. Now, we consider the case of a variable equipped with parents in a DAG g. The following theorem shows how one computes the causal effect of a variable with parents on other variables in g.

Theorem 2.5.1. Let $t \in V$, and let $R \subseteq V \setminus [\{t\} \cup \mathrm{pa}(t)]$. Then,

$$f(x_R; \mathrm{do}(X_t = x_t^*)) = \int f(x_R | x_t^*, x_{\mathrm{pa}(t)}) f(x_{\mathrm{pa}(t)}) d\mu_{\mathrm{pa}(t)}(x_{\mathrm{pa}(t)})$$

Proof. Let X be a random vector such that $t_t = x_t^*$. Then, we have

$$\begin{aligned} f(X; \mathrm{do}(X_t = x_t^*) &= \frac{f(X)}{f(x_t | x_{\mathrm{pa}(t)})} = \frac{f(X)}{f(x_t, x_{\mathrm{pa}(t)})} f(x_{\mathrm{pa}(t)}) \\ &= f(x_{V\setminus[\{t\}\cup\mathrm{pa}(t)]} | x_t, x_{\mathrm{pa}(t)}) f(x_{\mathrm{pa}(t)}) \end{aligned} \tag{2.20}$$

Now, we compute the marginal density $f(x_R; \mathrm{do}(X_t = x_t^*))$. Let $S = V \setminus [\{t\} \cup \mathrm{pa}(t) \cup R]$. Then

$$\begin{aligned} f(x_R; \mathrm{do}(X_t = x_t^*)) &= \int f(x_R, x_S, x_{\mathrm{pa}(t)}; \mathrm{do}(X_t = x_t^*)) d\mu_{S\cup\mathrm{pa}(t)}(x_S, x_{\mathrm{pa}(t)}) \\ &= \int\int f(x_R, x_S | x_t^*, x_{\mathrm{pa}(t)}) f(x_{\mathrm{pa}(t)}) d\mu_S(x_S) d\mu_{\mathrm{pa}(t)}(x_{\mathrm{pa}(t)}) \\ &= \int f(x_R | x_t^*, x_{\mathrm{pa}(t)}) f(x_{\mathrm{pa}(t)}) d\mu_{\mathrm{pa}(t)}(x_{\mathrm{pa}(t)}) \end{aligned}$$

where the second equation follows from the above Eq.(2.20). □

Remark. The result of this theorem shows that the causal effect of X_t on X_R is uniquely calculated by the marginal distribution of $(X_R, X_t, X_{\mathrm{pa}(t)})$.

In the previous chapter, we introduced the definition of SCM in which the effect X is determined by the cause $\mathrm{pa}(X)$ using functions g_x. In a bivariate case, X is computed from a cause Y by using a function g. However, the joint distribution $P_{X,Y}$ of two variables does not exactly tell us if an SCM induces it from X causes Y, or Y causes X. That is, one may want to figure out that the structure is identifiable from the joint distribution. The following proposition shows a model which is not identifiable from the joint distribution.[5]

Proposition 4. For every joint distribution $P_{X,Y}$ of two real-valued variables, there is an SCM

$$Y = g_Y(X, N_Y), \quad X \perp\!\!\!\perp N_Y,$$

where f_Y is a measurable function and N_Y is a real-valued noise variable.

Proof. Define the conditional cumulative distribution function

$$F_{Y|x}(y) := P(Y \leq y | X = x).$$

Then define

$$f_Y(x, n_y) := F^{-1}_{Y|x}(n_y),$$

where $F^{-1}_{Y|x}(n_y) = \inf\{x \in \mathbb{R} : F_{Y|x} \geq n_Y\}$. Then, let N_Y be uniformly distributed on $[0,1]$ and independent of X. □

The result of this proposition can be used for both cases, X to Y and Y to X. In other words, every joint distribution $P_{X,Y}$ admits SCMs for both directions. Now, we consider a more general case. Suppose we observe the variables X_T, X_R, and X_C for a set $C \subseteq V \setminus [T \cup R]$. In this case, we wonder whether the causal effect of X_T on X_R is identifiable. In other words, one may wonder if the distribution $P(X_R \in \cdot; \mathrm{do}(X_T = x_T^*)$ uniquely determined by the marginal distribution of X_T, X_R, and X_C. We introduce two criteria and theorems showing that the causal effects are identifiable under these criteria.

Definition 2.5.2. Let $r, v \in V$ and $r \neq t$. A set $C \subseteq V \setminus \{r, t\}$ satisfies the back-door criterion with respect to the ordered pair (t, r) if

i) $C \cap \mathrm{de}_G(t) = \emptyset$

ii) C blocks all back-door paths from t to r, that is, there is no path from t to r that starts with an edge of the form $t \leftarrow w$ and that d-connects t and r given C.

Theorem 2.5.2. If C satisfies the back-door criterion with respect to (t, r)*then*

$$f(x_r; \mathrm{do}(X_t = x_t^*)) = \int f(x_r | x_t x_C) f(x_C) d\mu_C(x_C). \tag{2.21}$$

If $C = \emptyset$, then $f(x_r; \mathrm{do}(x_t = x_t^*)) = f(x_r | x_t^*)$.

Example 2.5.2.

Remark. The back-door criterion is sufficient but not necessary for the covariate adjustment formula in Eq.(2.21). The necessary condition is provided in the following theorem [10].

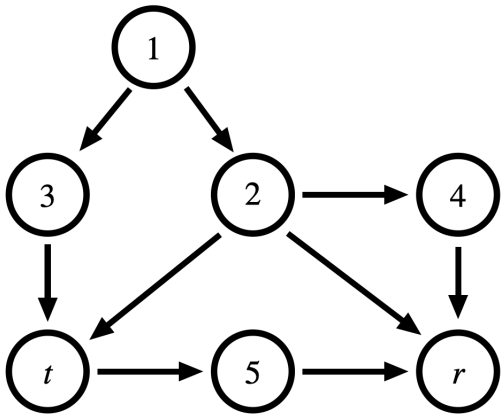


Figure 2.5: a) Observational distribution. b) Interventional distribution under do($X_2 = x_2*$)

Theorem 2.5.3 (Adjustment criterion)**.** the covariate adjustment formula in Eq.(2.21) holds for all factorizing distribution f if and only if

i) for all $v \in \mathrm{de}(t) \cap \mathrm{an}(r), \;\; v \neq t: \quad C \cap \mathrm{de}(v) = \emptyset$.

ii) every path from t to r that is d-connecting given C is a directed path from t to r.

We will prove this theorem later since we need to introduce a new definition before proving it. Here, we introduce the second criterion to identify causal effects.

Definition 2.5.3. Let $r,t \in V, \;\; r \neq t$.A set $C \subseteq V \setminus \{r,t\}$ satisfies the front-door criterion with respect to the (ordered) pair (t,r) if

i) every directed path from t to r contains a node in C.

ii) there is no unblocked back-door path from t to C, that is, no back door path from t to C is d-connecting given $\emptyset$.

iii) there does not exist a back-door path from C to r that is d-connecting given $\{t\}$.

Theorem 2.5.4. If C satisfies the front-door criterion with respect to (t,r) then

$$f(x_r; \mathrm{do}(X_t = x_t^*)) = \int f(x_C|x_t^*) \left[\int f(x_r|x_t,x_c) f(x_t) d\mu_t(x_t) \right] d\mu_C(x_C).$$

Remark. Front-door criterion shows that the identification of causal effects is achieved through another formula from covariate adjustment in the Eq.(2.21).

Now, we prove the Theorem 2.5.3. We need a new definition which is intervention graphs and variables. The necessity of the definitions arises from expressing intervention distribution in a different way to use in the proof of the Theorem 2.5.3. In Fig.2.6, there is an example of an intervention graph in which the blue node indicates an intervention variable. Generally, an intervention graph and variable are defined as

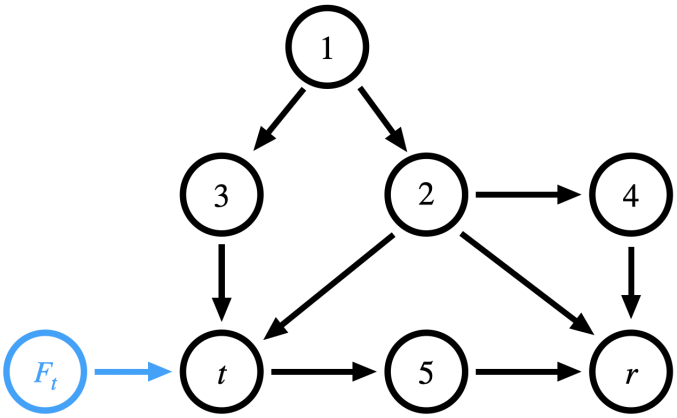


Figure 2.6: A example of intervention graph

follows. Suppose that we take an intervention on the variables $(X_t)_{t\in A}$. Let $t \in V$ be a variable in a graph G in Fig.2.6. A intervention variable F_t intervenes on the variable t. The variable F_t can take values in $\mathbb{R}\cup\{\emptyset\}$, where $F_t = \emptyset$ states that there is no intervention.

Definition 2.5.4 (Intervention graph)**.** The intervention graph is an augmented DAG with vertex set V $\cup\{F_v : v \in A\}$ and edge set $E\cup\{F_v \to v : v \in A\}$. Furthermore, the new conditional densities for $v \in A$:

$$f'(x_v|\mathrm{pa}(v), F_v = f_v) = \begin{cases} f(x_v|x_{\mathrm{pa}(v)}) & \text{if } f_v = \emptyset \\ \mathbb{1}_{\{x_v = x_v^*\}} & \text{if } f_v = x_v^* \in \mathbb{R}. \end{cases}$$

Marginal distributions of F_V are arbitrary and positive.

In the above set up, f denotes the joint density of $X = (X_1, \ldots, X_m)$. Moreover, f' denotes the joint density of the random vector X and intervention variables $(F_v : v \in A)$, which is determined by conditional distributions:

$$f'(x, f_A) = \prod_{v\notin A} f(x_v|x_{\mathrm{pa}(v)}) \prod_{v\in A} f'(x_v|x_{pa(v)}, f_v)$$

Interventional densities can also be expressed by the function f'

$$\begin{aligned} f(x) &= f'(x|F_v = \emptyset, v \in A) \\ f(x; \mathrm{do}(X_A = x_A^*)) &= f'(x|F_A = x_A^*) \\ f(x; \mathrm{do}(X_B = x_B^*)) &= f'(x|F_B = x_B^*, F_v = \emptyset, v \in A \setminus B) \text{ for } B \subseteq A. \end{aligned}$$

Recall the theorem and prove it,

Theorem 2.5.5. If C satisfies the back-door criterion with respect to (t, r)*then*

$$f(x_r; \mathrm{do}(X_t = x_t^*)) = \int f(x_r|x_t^*, x_C) f(x_C) d\mu_C(x_C).$$

If $C = \emptyset$, then $f(x_r; \mathrm{do}(x_t = x_t^*)) = f(x_r|x_t^*)$.

Proof. Let $r \in V$ in the graph G. We use the intervention graph for $A = \{t\}$, then we have

$$\begin{aligned} f(x_r; \mathrm{do}(X_t = x_t^*)) &= f'(x_r | F_t = x_t^*) \\ &= \int f'(x_r, x_C | F_t = x_t^*) d\mu_C(x_C) \\ &= \int f'(x_r | x_C, F_t = x_t^*) f'(x_C | F_t = x_t^*) d\mu_C(x_C) \\ &= \int f'(x_r | x_C, x_t^*, F_t = x_t^*) f'(x_C | F_t = x_t^*) d\mu_C(x_C) \end{aligned}$$

The proof is completed by showing that

$$f'(x_r | x_C, x_t^*, F_t = x_t^*) = f(x_r | x_C, x_t^*) \tag{2.22}$$

$$f'(x_C | F_t = x_t^*) = f(x_C). \tag{2.23}$$

Firstly, we show the Eq.(2.22). The sufficient condition for the Eq.(2.22) is $r \perp\!\!\!\perp F_t | \{t\} \cup C$. That is, F_t is d-separated from r given $C \cup \{t\}$. C satisfies the back-door criterion. In other words, C blocks all back-door paths from t to r. Hence, there is no path on the form $F_t \to t \leftarrow \ldots \leftarrow r$ d-connects given $C \cup \{t\}$. Moreover, every path of the form $F_t \to t \to \ldots$ is blocked by t. The condition $r \perp\!\!\!\perp F_t | \{t\} \cup C$ yields that

$$\begin{aligned} f'(x_r | x_C, x_t^*, F_t = x_t^*) &= f'(x_r | x_C, x_t^*, F_t = \emptyset) \\ &= f(x_r | x_C, x_t^*). \end{aligned}$$

The second equation in Eq.(2.23) holds if $C \perp\!\!\!\perp F_t$ in intervention graph, since

$$f'(x_C | F_t = x_t^*) = f'(x_C | F_t = \emptyset) = f(x_C).$$

It remains to argue that in the intervention graph $C \perp\!\!\!\perp F_t$ holds. However, $F_t \perp\!\!\!\perp C$ follows from the local M.P. for the intervention graph. Indeed, F_t has no parents and C does not contain any descendants of F_t as the assumption $C \cup \mathrm{de}_G(t) = \emptyset$ □

Example 2.5.3. Let $X = (X_1, \ldots, X_d)$ be a random vector, and follows the LSEM

$$X = BX + \varepsilon \Leftrightarrow X = (I - B)^{-1} \varepsilon$$

where $B = (B_{ij})_{i,j=1,\ldots d}$. Suppose that we contemplate an intervention on the variable X_t = x_t^*. Then,

$$X = \tilde{B}X + \tilde{\varepsilon}$$

where

$$\tilde{B}_{vw} = \begin{cases} \beta_{vw} & \text{if } v \neq t \\ 0 & \text{if } v = t, \end{cases}$$

and

$$\tilde{\varepsilon}_{vw} = \begin{cases} \varepsilon_{vw} & \text{if } v \neq t \\ x_t^* & \text{if } v = t \end{cases}$$

Hence,

$$X(\mathrm{do}(X_t = x_t^*)) = (I - \tilde{B})^{-1}\tilde{\varepsilon}.$$

The total effects in the linear SCM can be computed by using the above expression. Let $r \in V$ and $r \neq t$. Then the total effect of X_t on X_r is equal to

$$\begin{aligned}
\mathbb{E}[X_r; \mathrm{do}(X_t = x_t^*)] &= \mathbb{E}[\sum_{w=1}^{m} ((I - \tilde{B})^{-1})_{rw}\tilde{\varepsilon}_w] \\
&= \mathbb{E}[((I - \tilde{B})^{-1})_{rt}x_t^* + \sum_{w \neq t}^{m} ((I - \tilde{B})^{-1})_{rw}\varepsilon_w] \\
&= \mathbb{E}[((I - \tilde{B})^{-1})_{rt}x_t^*] \\
&= \mathbb{E}[((I - B)^{-1})_{rt}x_t^*]
\end{aligned}$$

We can compute the total causal effect of this case by differentiating the above result

$$[(\mathrm{Id} - B)^{-1}]_{rt} = \sum_{\text{paths from } t \text{ to } r} \text{prodect of } \beta_{vw} \text{ for } v \to w \text{ on the path.}$$

The exact result of the right-hand side is explained and displayed in Theorem 2.4.1. Let $C \subseteq V$ and satisfies the back-door criterion. Then the adjustment formula in Eq.(2.21) holds. This formula in the case of a linear SCM is

$$\begin{aligned}
\mathbb{E}[X_r; \mathrm{do}(X_t = x_t^*)] &= \mathbb{E}_{X_C}[\mathbb{E}[X_r | X_t = x_t^*, X_C]] \\
&= \mathbb{E}_{X_C}[\alpha_t * x_t^* + \sum_{w \in C} \alpha_w X_w] = const. + \alpha_t x_t^*
\end{aligned}$$

In other words, if (2.21) holds, then total effect of X_t on X_r is the coefficient for X_t in the conditional expectation $\mathbb{E}[X_r | X_t, X_C]$. Estimating the coefficients is achieved by linear regression of X_r on (X_t, X_C)

Example 2.5.4 (Harmfulness of mother's smoking during pregnancy to babies [7])**.** Now we consider a specific example in the real world. A study is carried out to investigate the harmfulness of a mother's smoking during pregnancy to her baby. The study records the baby's birth weight and the number of cigarettes per day the mom smoked in the first trimester. The scientists argued that there is a causal effect of smoking on the birth weight of babies, and both factors are negatively correlated, while cigarette companies say that smoking does not harm babies. The companies viewed that unobserved factors can lead to heavier smoking. To clarify this hidden relation, the scientists designed a clever idea to show that smoking has a direct effect by using the tax rate on cigarettes. The graphical model which is considered is displayed in Fig.2.7. Assume that the structural causal model is linear

$$\begin{aligned}
X_1 &= \beta_{01} + \varepsilon_1 \\
X_2 &= \beta_{02} + \beta_{21}X_1 + \beta_{24}X_4 + \varepsilon_2 \\
X_3 &= \beta_{03} + \beta_{32}X_2 + \beta_{34}X_4 + \varepsilon_3 \\
X_4 &= \beta_{04} + \varepsilon_4
\end{aligned}$$

where the errors $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ are independent and β_{32} is the total effect of smoking on baby's weight.

The result of the above example will be very similar to our work to figure out the causal effect of 2-dimensional and 3-dimensional cases of the LSEM equipped with normally distributed independent error terms. Therefore, we will use this idea again and formulate it formally in the later chapter.

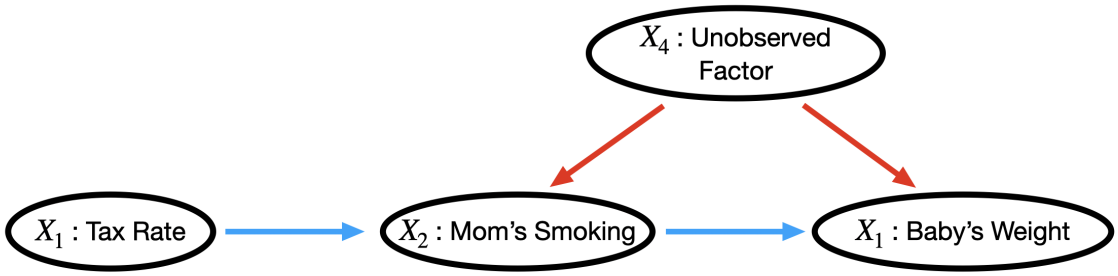


Figure 2.7: The graph designed by scientists who argue that smoking during pregnancy harms the baby and leads to low birth weight.

3 Linear Regression

As mentioned in Example 2.5.4, the strength of causal coefficients of a linear SEM can be easily estimated by the linear regression method. We will estimate the parameters for both 2,3-dimensional cases by using linear regression. Therefore, this chapter introduces the main framework of linear regression to help understand the thesis more comprehensively. The definitions and explanations about linear regression methods in this chapter are initially provided by [11].

3.1 introduce

The main assumption of a linear regression model is that the regression function $\mathbb{E}[Y|X]$ is linear in the inputs $X_1,\ldots,X_d$. Consider a linear SEM $\mathfrak{C}$ for a graph G, then a variable X_v in graph G has the form of

$$X_v = \sum_{w\in \mathrm{pa}(v)} \beta_{vw} X_w + \varepsilon_v.$$

This model also satisfies exactly the assumption of a linear regression model, that is, the variable X_v and the parents of this variable have a linear relation. Hence, we can apply the linear methods for regression. Moreover, we assume that the error terms are normally distributed as done in the paper [2]. In the following parts of this chapter, we will introduce how one can estimate the parameter of the linear regression model with normally distributed error terms.

3.2 Linear Regression Models and Least Squares

Assume that we have an input vector $X^T = (X_1, X_2, \ldots, X_d)$, and want to predict a real-valued output Y. The linear regression model has the following form

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j. \tag{3.1}$$

The linear model assumes that the regression function $\mathbb{E}[Y|X]$ is linear. The coefficients β_j for $j \in \{1,\ldots,d\}$ are unknown parameters, and the variables X_j are in the form

i) quantitative inputs

ii) transformation of input such as log, square-root, etc.

iii) polynomial representation for example $X_2 = X_1^2$

iv) numeric or "dummy" coding of the levels of inputs.

v) interaction between variables, for example $X_1 = X_2 \cdot X_3$.

Even though the source of the X_j is the log transformation of a variable or square of variables, the model is linear in parameters. Typically, a set of training data $(x_1, y_1) \dots (x_N, y_N)$ is gathered, and the parameters β are estimated by the training set. $x_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ is a vector of feature measurements for i-th element of the set.

Definition 3.2.1. Let $(x_1, y_1) \dots (x_N, y_N)$ be a set of data. Consider the model in Eq.(3.1), then the quantity $RSS(\beta)$ is called residual sum of squares and defined as

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \end{aligned} \tag{3.2}$$

Remark. Alternatively, the residual sum of squares can be written in the matrix form

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta). \tag{3.3}$$

where $\mathbf{X}$ denote the $N \times (d+1)$ matrix with each row an input vector, and $\mathbf{y}$ is N vector of outputs.

The most simplest and widely used estimation method is least method, where we choose the coefficients $\beta = (\beta_0, \beta_1, \dots, \beta_d)$ to minimize $RSS(\beta)$ in Eq.(3.2). The Figure 3.1 shows the geometry of least-squares fitting in the $\mathbb{R}^{d+1}$-dimensional space occupied by the pairs (X, Y). Now, we pose the question. How do we minimize the $RSS(\beta)$ with respect to β? The Equation (3.3) is a quadratic function in the $d+1$ parameters. Taking the derivative with respect to β, we obtain

$$\begin{aligned} \frac{\partial RSS}{\partial \beta} &= -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \\ \frac{\partial^2 RSS}{\partial \beta \partial \beta^T} &= 2\mathbf{X}^T\mathbf{X} \end{aligned} \tag{3.4}$$

We assume that the matrix $\mathbf{X}$ has full column rank. This yield that $\mathbf{X}^T\mathbf{X}$ is positive definite. Now, we equate the lower equation in (3.4) to 0 :

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X})\beta = 0.$$

This equation can be solved uniquely by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.5}$$

Hence, the fitted value at the training inputs $\mathbf{X}$ are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.6}$$

Now, suppose that the error term ε is normally distributed as

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

where σ is a positive value. Then, the likelihood function of y is equal to

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{2\sigma^2}(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j)^2\right)$$

Alternatively, one can write it down as

$$L = (2\pi\sigma^2)^{-N/2} \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j)^2\right)$$

One can see from the last term that the parameter $\beta_0, \ldots, \beta_d$, which maximize the likelihood L are the same as the ones that minimize the residual sum of square *RSS*. Thus, the maximum likelihood estimator of the parameter β is the same as (3.5). Now, we estimate the variance σ of the error term ε. Substituting the maximum likelihood estimates $\hat{\beta}$ into L, we obtain

$$L = (2\pi\sigma^2)^{-N/2} \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \sum_{j=1}^{d} x_{ij}\hat{\beta}_j)^2\right).$$

Taking the logarithm, we obtain the log-likelihood function $l = \log L$, which makes differentiation easier:

$$l \propto -\frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \hat{\beta}_0 - \sum_{j=1}^{d} x_{ij}\hat{\beta}_j$. Taking the derivative with respect to σ^2, we obtain

$$\frac{d}{d\sigma^2} l = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Setting to zero and solving for σ^2 yields:

$$\hat{\sigma}^2_{mle} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

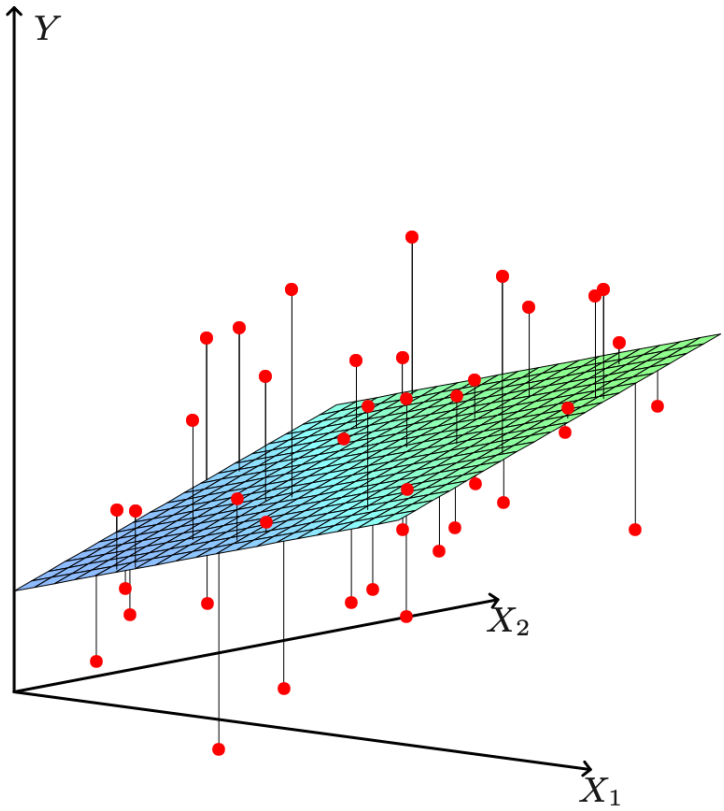


Figure 3.1: Linear least squares fitting with $X \in \mathbb{R}^2$. The main focus is to find the linear function $f(X)$, which minimizes the residual sum of squares. [11]

4 Universial Inference

We will use the *split ratio test* (SRT) to understand confidence intervals of causal effects. In this section, we introduce the main framework of universal inference, which is firstly introduced by [6] in order to help to understand the following sections in the thesis. First of all, the mathematical background which we work on is introduced, closely following the thesis [4]. The sample space $\mathbb{X} \neq \emptyset$ equipped with a certain σ-field $\mathscr{F}$ and the parameter space Θ as well as the parametric statistical model $\mathscr{P} = \{P_\Theta : \Theta \in \Theta\}$ composed of probability measures $P_\Theta : \mathscr{F} \to [0,1]$. Suppose that an i.i.d. sample $Y_1, \ldots, Y_n \sim P_{\Theta_*}$ is given where $\Theta_* \in \Theta$ denotes the true value of the unknown parameter. In addition, it is assumed that the model $\mathscr{P}$ is dominated by some σ-finite measure $\mu : \mathscr{F} \to [0,\infty]$, i.e., for every event $A \in F$ we know that $\mu(A) = 0$ implies $P_\Theta(A) = 0$ for all $\theta \in \Theta$. The measure μ will be either a counting measure or the d-dimenstional Lebesgue measure, depending on the sample space $\mathbb{X}$. If $\mathbb{X}$ is a counting set, the measure μ is also a counting measure. If $\mathbb{X}$ is equal to $\mathbb{R}^d$ or a subset of $\mathbb{R}^d$, the measure is the d-dimensional Lebesgue measure. In the case of the d-dimensional Lebesgue measure, there exists for every $\theta \in \Theta$ a μ-density $p_\theta := \frac{dP}{d\mu}$ for P_θ, that is, a function $p_\theta : \mathbb{X} \to [0,\infty)$ such that for every $A \in \mathscr{F}$

$$P_\theta(A) = \int_A p_\theta d\mu$$

by the following well-known theorem.

Theorem 4.0.1 (Radon-Nikodym theorem)**.** Let ($\mathbb{X}$,Σ) be a measurable space on which two σ-finite measures are defined, μ and ν. If $\nu \ll \mu$ (that is, if ν is absolutely continuous with respect to μ), then there exists a Σ-measurable function $f : \mathbb{X} \to [0,\infty)$, such that for any measurable set $A \subseteq \mathbb{X}$,

$$\nu_\theta(A) = \int_A p_\theta d\mu$$

In order to avoid having measurability issues, we will consider the parameter space as a measurable space $(\Theta, \mathscr{A})$ and the function

$$\Theta \times \mathbb{X} \to [0,\infty), (\theta, x) \mapsto p_\theta(x)$$

is $\mathscr{A} \otimes \mathscr{F}/\mathscr{B}([0,\infty))$-measurable, where the product σ-field on the product space $\Theta \times \mathbb{X}$ is denoted by $\mathscr{A} \otimes \mathscr{F}$ and the Borel σ-field on $[0,\infty)$ is by $\mathscr{B}([0,\infty))$. As mentioned above, we will use the SRT method. Therefore we will not work with the whole data sample $Y_1, \ldots, Y_n$, but split the data set into two groups and estimate each parameter by using both groups separately. Each index set of both groups is denoted by D_0, D_1, which are subsets of the index set $1, \ldots, n$. In addition we have the properties such that $D_0 \cap D_1 = \emptyset$ and $D_0 \cup D_1 = \{1, \ldots, n\}$. In the following, we will evaluate the likelihood based on D_0 and compute the maximum likelihood estimator under $\mathbf{H}_0$ or $\mathbf{H}_0$ calculated from D_0. To avoid mathematical ambiguity, we introduce here exactly both objects. The likelihood function evaluated based on D_0 is denoted by

$$\mathfrak{L}^{(0)}(\theta) := \prod_{i \in D_0} p_\theta(Y_i)$$

Furthermore, we denote the maximal likelihood estimator under $\mathbf{H}_0$, which is computed on D_0 by

$$\hat{\theta}_0 := \arg\max_{\theta \in \Theta_0} \mathfrak{L}^{(0)}(\theta)$$

Considering the above setup, we define the following objects.

Definition 4.0.1. Suppose that we have the above setup

(*i*) The split likelihood ratio statistic is defined as

$$T_n(\theta) := \frac{\mathfrak{L}^{(0)}(\hat{\theta}_1)}{\mathfrak{L}^{(0)}(\theta)}.$$

(*ii*) Let $\alpha \in (0,1)$. The universal confidence set of level 1 - α is defined as

$$C_n := C_n(\alpha) := \left\{ \theta \in \Theta : T_n(\theta) \leq \frac{1}{\alpha} \right\}$$

It is proven that the universal confidence set is valid finite sample confidence set for an unknown parameter at level 1 - α [6]. In the following, we introduce the simple proof, which is also given by [4]. The proof is quite simple and only requires a simple tool from probability theory. First of all, we give the simple tool from probability theory which is known as *Markov's inequality*.

Theorem 4.0.2 (Markov's Inequality)**.** Let X be a real-valued random variable. Them for every $\beta > 0$

$$P(|X| \geq \beta) \leq \frac{1}{\beta} \mathbb{E}(|X|)$$

Proof. This immediately follows from the inequality

$$\mathbb{E}[|X|] \geq \mathbb{E}[|X| \mathbb{1}_{\{|X| \geq \beta\}}] \geq \beta P(|X| \geq \beta)$$

□

This theorem will be used to prove the following theorem.

Theorem 4.0.3. Let $\alpha \in (0,1)$ and consider the set-up from above. Then $C_n(\alpha)$ is a finite sample valid $(1-\alpha)$-confidence set for θ_*, that is,

$$P_{\theta_*}(\theta_* \in C_n(\alpha)) \geq 1 - \alpha.$$

Proof. Let $\psi \in \Theta$ be fixed, let $\int \ldots dy_{D_0}$ denote integration with respect to all variables contained in $D_=$ and let $A := \{p_{\theta^* > 0}\}^k$. Then,

$$\begin{aligned} \mathbb{E}_{\theta^*}\left[\frac{\mathfrak{L}^{(0)}(\psi)}{\mathfrak{L}^{(0)}(\theta^*)}\right] &= \mathbb{E}_{\theta^*}\left[\frac{\prod_{i \in D_0} p_\psi(Y_i)}{\prod_{i \in D_0} p_{\theta^*}(Y_i)}\right] \\ &= \int_A \frac{\prod_{i \in D_0} p_\psi(y_i)}{\prod_{i \in D_0} p_{\theta^*}(y_i)} \prod_{i \in D_0} p_{\theta^*}(y_i) dy_{D_0} \\ = \int_A \prod_{i \in D_0} p_\psi(y_i) dy_{D_0} \leq \prod_{i \in D_0} \left[\int_{\mathbb{X}} p_\psi(y_i) dy_i\right] &= 1 \end{aligned} \tag{4.1}$$

Since $\hat{\theta}_1$ is a function only of the samples contained in D_1, and since $Y_{D_=}$ and Y_{D_1} are independent,

$$\begin{aligned}\mathbb{E}_{\theta^*}[T_n(\theta^*)] &= \mathbb{E}_{\theta^*}\left[\mathbb{E}_{\theta^*}[T_n(\theta^*)|Y_{D_1}]\right] \\ &= \int_{\mathbb{X}^{n-k}} \underbrace{\mathbb{E}_{\theta^*}\left[\frac{\mathfrak{L}^{(0)}(T(y_{D_1})}{\mathfrak{L}^{(0)}(\theta^*)}\right]}_{\leq 1 \text{ by } (??)} d(P_{\theta^*}^{\otimes(n-k)}(y_{D_1}) \leq 1.\end{aligned}$$

Finally, Markov's inequality gives us that

$$P_{\theta^*}(\theta^* \notin C_n) = P_{\theta^*}(T_n(\theta^*) > \frac{1}{\alpha}) \leq \alpha \mathbb{E}_{\theta^*}[T_n(\theta^*)] \leq \alpha.$$

□

In many cases, a certain part of the information encoded by the true parameter is taken into account to construct confidence sets. For example, suppose that we have a parameter $\theta = (\theta_1, \ldots, \theta_n)$, and we only want to construct confidence set of θ_i whereas the other components $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n$ are so-called *nuisance parameter.* More formally, we construct confidence sets of some function $\psi = g(\theta)$ where $g : \Theta \to \Psi$. In this case, we can consider set

$$B_n := \{\psi \in \Psi | C_n \cap g^{-1}(g(\theta)) \neq \emptyset\}$$

where $g^{-1} = \{\theta \in \Theta | g(\theta) = \psi\}$, and it has the following property

$$P_\theta(g(\theta) \in B_n) = P_\theta(C_n \cap g^{-1}(g(\theta)) \neq \emptyset) \geq P_\theta(\theta \in C_n) \geq 1 - \alpha$$

Thus, this yield that B_n is a $(1-\alpha)$-confidence set for $\psi = g(\theta)$. We can write down this set B_n in a different way by using the so-called profile likelihood. We define the profile likelihood as follows

Definition 4.0.2. Consider the situation from above. If it is well-defined, the function

$$\mathfrak{L}_\dagger^{(0)} : \Psi \to \mathbb{R}, \quad \psi \mapsto \max_{\theta \in \Theta : g(\theta) = \psi} \mathfrak{L}^{(0)}(\theta) \tag{4.2}$$

is called the profile likelihood function.

Proposition 5. Consider the situation from above and suppose the profile likelihood function is well-defined. Then

$$B_n = \left\{ \psi \in \Psi : \frac{\mathfrak{L}^{(0)}(\hat{\theta}_1)}{\mathfrak{L}^{(0)}(\psi)} \leq \frac{1}{\alpha} \right\}$$

Proof. Let $\psi \in \Psi$. Then $C_n \cap g^{-1}(\psi) \neq \emptyset$ is equivalent to the fact that there exists some $\theta \in \Theta$ such that $g(\theta) = \psi$ and $\mathfrak{L}^{(0)}(\hat{\theta}_1)/\mathfrak{L}^{(0)}(\theta) \leq 1/\alpha$. The latter is equivalent to $\mathfrak{L}^{(0)}(\hat{\theta}_1)/\mathfrak{L}^{(0)}(\psi) \leq 1/\alpha$ □

Now, suppose that we carry out a hypothesis testing with a testing problem of the following form

$$\mathbf{H}_0 : \theta \in \Theta_0 \quad \text{versus} \quad \mathbf{H}_1 : \theta \in \Theta_1$$

where $\Theta_1, \Theta_0 \subset \Theta$

Definition 4.0.3. Let $\alpha \in (0,1)$. The split likelihood ratio test at level α is defined by the rule

$$\text{Reject } \mathbf{H}_0 \text{ if } U_n > \frac{1}{\alpha}, \text{ where } U_n = \frac{\mathfrak{L}^{(0)}(\hat{\theta}_1)}{\mathfrak{L}^{(0)}(\hat{\theta}_0)}$$

where $\hat{\theta}_1$ is a estimator estimated by using D_1 and $\hat{\theta}_0$ is a estimator calculated by using D_0. The following theorem is the fundamental result that will be applied in the thesis.

Theorem 4.0.4. The split likelihood ratio test controls the type I error at lever α, that is,

$$\sup_{\theta_0 \in \Theta_0} P_{\theta_0}\left(U_n > \frac{1}{\alpha}\right) \leq \alpha.$$

Proof. We reject $\mathbf{H}_0$ if and only if $C_n(\alpha) \cap \Theta_0 = \emptyset$. The type I error of this test is simply

$$\sup_{\theta_0 \in \Theta_0} P_{\theta_0}(C_n(\alpha) \cap \Theta_0 = \emptyset) \leq \sup_{\theta_0 \in \Theta_0} (\theta_0 \notin C_n(\alpha)) \leq \alpha$$

□

To simplify the calculation, we use a logarithmic version of U_n

$$\xi_n := l^{(0)}(\hat{\theta}_1) - l^{(0)}(\hat{\theta}_0) \tag{4.3}$$

where $l^{(\theta)} := \log \mathfrak{L}^{(0)}(\theta)$. Equivalently to the previous test, one rejects $\mathbf{H}_0$ if $\log(U_n) < -\log(\alpha)$.

5 A confidence set for the causal effect in Linear Structural Equation Models (LSEM) by using interventional data set via split likelihood ratio test

In this chapter, we focus on constructing a confidence set of valid hypothesis tests. We will repeat explaining the concept of the linear structural equation model briefly and show how one can apply this model to investigate the causal effect and construct confidence set for this in two- and three-dimensional cases. Additionally, the boundaries and maximal width of the confidence set will be calculated analytically in a two-dimensional case. In [2] and [4], a observational data set in the form of a sample of independent copies of a random vector $Y = (Y_1, \ldots, Y_d)$ is considered, where each component has zero mean, without loss of generality. A further assumption is that the distribution of X underlies a dependence structure given by the following linear structural equation model

$$X_j = \sum_{k \in \mathrm{pa}(j)} \beta_{ji} X_i + \varepsilon_j, \quad j = 1, \ldots, d.$$

Furthermore, it is assumed that the errors are homoscedastic, i.e., for an unknown variance parameter $\sigma^2 \in (0, \infty)$, the variance of each error is

$$Var[\varepsilon_1] = \cdots = Var[\varepsilon_d] = \sigma^2.$$

However, in this thesis, we work with a non-homoscedastic setup. Hence, we consider the following model

$$X_j = \sum_{k \in \mathrm{pa}(j)} \beta_{ji} X_i + \varepsilon_j, \quad j = 1, \ldots, d.$$
$$Var[\varepsilon_j] = \sigma_j^2 \in (0, \infty), \quad j = 1, \ldots, d$$

Once we take a intervention on a variable X_t under $\mathrm{do}(X_t = x_t^*)$, we obtain a modified LSEM $\tilde{\mathscr{C}}$ from the original LSEM $\mathscr{C}$ in the form of

$$X_j = \sum_{k \in \mathrm{pa}(j)} \tilde{B}_{ji} X_i + \tilde{\varepsilon}_j, \quad j = 1, \ldots, d.$$

where

$$\tilde{B}_{vw} = \begin{cases} \beta_{vw} & \text{if } v \neq t \\ 0 & \text{if } v = t, \end{cases}$$

and

$$\tilde{\varepsilon}_{vw} = \begin{cases} \varepsilon_{vw} & \text{if } v \neq t \\ x_t^* & \text{if } v = t \end{cases}$$

Example 5.0.1. Consider the LSEMs

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= \beta_{21} X_1 + \varepsilon_2 \\ X_3 &= \beta_{31} X_1 + \varepsilon_3 \end{aligned}$$

where $\varepsilon_1, \varepsilon_2$ are standard normally distributed. Then we have the model in the matrix form as

$$X = (\mathrm{Id} - B)^{-1}\varepsilon, \quad B = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & 0 & 0 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

Intervening on the variable $\mathrm{do}(X_2 = x_2^*)$, we obtain modified LSEMs

$$X(\mathrm{do}(X_2 = x_2^*)) = (\mathrm{Id} - \tilde{B})^{-1}\tilde{\varepsilon}$$

where

$$\tilde{B} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \beta_{31} & 0 & 0 \end{pmatrix}, \quad \tilde{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ x_2^* \\ \varepsilon_3 \end{pmatrix}.$$

Figure 5.1 illustrates the graphical structure of this example.

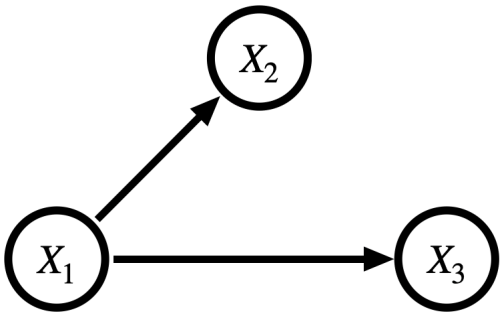


Figure 5.1: The graphical structure of Example 5.0.1

5.1 Dataset

The thesis focuses mainly on investigating causal effects by using interventional data. Hence, data sets that are used in the thesis contain observations of random vectors $(X^{(j)})_{j\in\{1,\dots,n\}}$ where there are causal

effects among variables which are elements of the vectors. The i-th element of the j-th random vector is denoted by $X_i^{(j)}, i \in \{1,\dots,d\}, j \in \{1,\dots,n\}$ where n is the sample size of the data set. Moreover, there is one additional element in a vector for the interventional setup. This additional variable is called the intervention indicator and is denoted by F_j. If F_j takes a value $i \in \{1,\dots,d\}$, then the j-th vector is intervened on the variable $X_i^{(j)}$, that is, $X_i^{(j)}$ take a fix value $x_i^{*(j)}$. If F_j takes the value 0, then the j-th vector $X^{(j)}$ is not intervened. That is, it is observational. In other words, the data set can be divided into disjoint subsets according to the level of F. The first subset N_0 is an observational subset of the data set, in which all vectors have the property $F = 0$. The subsets N_i, $i \in \{1,\dots,d\}$ are interventional subsets, where all vectors have the property $F = i$. N denotes the entire data set. Thus, the data set factors as

$$N = \cup_{i=0}^{n} N_i$$

Furthermore, we define index sets D to indicate indices of vectors in the subsets. The index sets are defined as

$D^{F=i}$: Index set for the vectors in N_i.

$D^{F\neq i}$: Index set for the vectors in $\cup_{k\neq i} N_k$.

n_i: Sample size of the subset N_i.

Equivalently,

$$N_i = (X^{(j)})_{j \in D^{F=i}}$$

where

$$X^{(j)} = \begin{pmatrix} X_1^{(j)} \\ X_2^{(j)} \\ \vdots \\ X_{d-1}^{(j)} \\ X_d^{(j)} \\ F_j \end{pmatrix}, \text{ for } j \in \{1,2,\dots,n\}$$

is a random vector in the data set. In the following, we introduce notations for the split data set. In order to conduct the split likelihood ratio test introduced in Chapter 4, the data set needs to be split. The first split data set is denoted by $N^{(0)}$, and $N^{(1)}$ denotes the second one. Again, the $N^{(0)}, N^{(1)}$ factors as

$$N^{(0)} = \cup_{i=0}^{n} N_i^{(0)}$$
$$N^{(1)} = \cup_{i=0}^{n} N_i^{(1)},$$

depending on the intervention indicator F value in a random vector. Moreover, we introduce notations of index sets and sample sizes of the index set

$D_0^{F=i}$: Index set for the vectors in $N_i^{(0)}$.

$D_0^{F\neq i}$: Index set for the vectors in $\cup_{k\neq i} N_k^{(0)}$.

$D_1^{F=i}$: Index set for the vectors in $N_i^{(1)}$.

$D_1^{F\neq i}$: Index set for the vectors in $\cup_{k\neq i} N_k^{(1)}$.

$n_{0,i}$: Sample size of the subset $N_i^{(0)}$.

$n_{1,i}$: Sample size of the subset $N_i^{(1)}$.

5.2 Estimating Parameters in Linear Structural Equation Models

In the following, we will show how one can estimate parameters explicitly in a two-dimensional case without using directly the method introduced in Chapter 3. The *linear structural equation model* (LSEM) is considered for 2 dimension. There are two different directions of dependency

(M2.1)	(M2.2)
$X_1 := \beta_{01} + \varepsilon_1$	$X_1 := \beta_{01} + \beta_{12}X_2 + \varepsilon_1$
$X_2 := \beta_{02} + \beta_{21}X_1 + \varepsilon_2$	$X_2 := \beta_{02} + \varepsilon_2$.

There is additionally the case where there is no dependency between two variables. This case is equivalent to either $\beta_{21} = 0$ or $\beta_{12} = 0$ from (M2.1) or (M2.2). We assume that X_1 causes X_2. That is, without loss of generality, we only consider (M2.1) where the data follows the LSEMs

$$\begin{aligned} X_1 &= \beta_{01} + \varepsilon_1, \\ X_2 &= \beta_{02} + \beta_{21}X_1 + \varepsilon_2 \end{aligned} \tag{5.1}$$

where ε_1, ε_2 are normally distributed independent error terms which have zero mean and variance of σ_1 and σ_2, respectively. The parameter β_{21} is an unknown parameter that represents a direct causal effect between two variables. We assume in the thesis that the errors are not necessary to be homoscedastic,i.e., each variance of $\varepsilon_i, i \in \{1,2\}$ is not required to be the same. From Proposition 1, we know that the joint distribution of the observational data factors as

$$f(X_1, X_2) = f(X_2|X_1)f(X_1) \tag{5.2}$$

Since X_2 is the sum of X_1 and independent normally distributed error term ε_2, X_2 given that X_1 has taken on the value x is clearly normally distributed such that

$$X_2|X_1 = x_1 \sim \mathcal{N}(\beta_{02} + \beta_{21}x_1, \sigma_2) \tag{5.3}$$

Now, we consider the two systems of LSEMs in which the interventions on a variable are taken. The model intervening on X_1 has the form of

$$\begin{aligned} X_1 &= x_1^*, \\ X_2 &= \beta_{02} + \beta_{21} \cdot x_1^* + \varepsilon_2. \end{aligned} \tag{5.4}$$

The interventional distribution under $\mathrm{do}(X_1 = x_1^*)$ factors according to Eq. (2.8) as

$$f(X; \mathrm{do}(X_1 = x_1^*)) = \prod_{v\neq 1} f(X_v|X_{\mathrm{pa}(v)}) \prod_{v=1} \mathbb{1}_{\{X_v = x_v^*\}} = f(X_2|X_1 = x_1^*) \tag{5.5}$$

The model under intervention $\text{do}(X_2 = x_2^*)$ is in the form of

$$\begin{aligned} X_1 &= \beta_{01} + \varepsilon_1 \\ X_2 &= x_2^* \end{aligned} \tag{5.6}$$

The interventional distribution under $\text{do}(X_1 = x_1^*)$ factors as

$$f(X; \text{do}(X_2 = x_2^*)) = f(X_1 | X_2 = x_2^*) = f(X_1) \tag{5.7}$$

Typically, we work with independent identical distributed set of observations $(x_1^{(1)}, x_2^{(1)}, F_1) \dots (x_1^{(n)}, x_2^{(n)}, F_n)$. Note that the joint densities in (5.2) (5.5) (5.7) consist of two following density functions of observations

$$f(X_2^{(i)} = x_2^{(i)} | X_1^{(i)} = x_1^{(i)}) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{1}{2\sigma_2^2}(x_2^{(i)} - (\beta_{01} + \beta 21 x_1^{(i)}))^2\right) \tag{5.8}$$

$$f(X_1^{(i)} = x_1^{(i)}) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{1}{2\sigma_1^2}(x_1^{(i)} - \beta_{02})^2\right). \tag{5.9}$$

Now, we figure out the likelihood function of the data set that contains both observational and interventional data. Assuming that the data set contains n samples which are i.i.d, then the likelihood function is

$$L_n(\beta, \sigma) = \prod_{i=1}^{n} f(x_1^{(i)}, x_2^{(i)}, F_i | \beta, \sigma) \tag{5.10}$$

where the function f is the joint distribution of $X_1^{(i)}$ and $X_2^{(i)}$ and $\beta = (\beta_{01}, \beta_{01}, \beta_{21})$, $\sigma = (\sigma_1, \sigma_2)$ are the parameter vectors of X_1 and X_2. As mentioned before, our data set consists of three subsets. Therefore one can split the likelihood function separately, depending on the value of F_i, which is the interventional indicator. The likelihood function in (5.10) factors as

$$L_n(\beta, \sigma) = \prod_{i=1}^{n} f(x_1^{(i)}, x_2^{(i)} | \beta, \sigma)^{\mathbb{1}(F_i=0)} f(x_1^{(i)}, x_2^{(i)} | \beta, \sigma)^{\mathbb{1}(F_i=1)} f(x_1^{(i)}, x_2^{(i)} | \beta, \sigma)^{\mathbb{1}(F_i=2)} \tag{5.11}$$

The density function of observational model can factors as in Eq.(5.2) and the density functions of both interventional model are equal to the Eq.(5.5) and Eq.(5.7). Thus, the first term in the Eq.(5.11) is equal to

$$f(x_1^{(i)}, x_2^{(i)} | \beta, \sigma)^{\mathbb{1}(F_i=0)} = (f(x_2^{(i)} | x_1^{(i)}, \beta, \sigma) f(x_1^{(i)} | \beta, \sigma))^{\mathbb{1}(F_i=0)}$$

Furthermore, the second and third terms are

$$\begin{aligned} f(x_1^{(i)}, x_2^{(i)} | \beta, \sigma)^{\mathbb{1}(F_i=1)} &= f(x_2^{(i)} | x_1^{(i)}, \beta, \sigma)^{\mathbb{1}(F_i=1)} \\ f(x_1^{(i)}, x_2^{(i)} | \beta, \sigma)^{\mathbb{1}(F_i=0)} &= f(x_1^{(i)} | \beta, \sigma)^{\mathbb{1}(F_i=2)} \end{aligned}$$

Substituting the above results in Eq.(5.11), we obtain the likelihood function of our data set

$$
\begin{aligned}
L_n(\beta,\sigma) &= \prod_{i=1}^{n} (f(x_2^{(i)}|x_1^{(i)},\beta,\sigma) f(x_1^{(i)}|\beta,\sigma))^{\mathbb{1}(F_i=0)} f(x_2^{(i)}|x_1^{(i)},\beta,\sigma)^{\mathbb{1}(F_i=1)} f(x_1^{(i)}|\beta,\sigma)^{\mathbb{1}(F_i=2)} \\
&= \prod_{i=1}^{n} f(x_2^{(i)}|x_1^{(i)},\beta_{02},\beta_{21},\sigma_2)^{\mathbb{1}(F_i=0,1)} \cdot f(x_1^{(i)}|\beta_{01},\sigma_1)^{\mathbb{1}(F_i=0,2)}. \\
&= \prod_{i=1}^{n} f(x_2^{(i)}|x_1^{(i)},\beta_{02},\beta_{21},\sigma_2)^{\mathbb{1}(F_i\neq 2)} \cdot f(x_1^{(i)}|\beta_{01},\sigma_1)^{\mathbb{1}(F_i\neq 1)}. \qquad (5.12)
\end{aligned}
$$

Now we determine exactly the likelihood function of the LSEM with the interventional and observational data where the variables are normally distributed. We substitute those terms in Eq.(3) and obtain the likelihood function of the model

$$
\begin{aligned}
L_n(\beta,\sigma) = \prod_{i=1}^{n} &\left(\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp - \frac{1}{2\sigma_1^2}(X_1^{(i)} - \beta_{01})^2\right)\right)^{\mathbb{1}(F_i\neq 1)} \\
&\cdot \left(\left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp - \frac{1}{2\sigma_2^2}(X_2^{(i)} - \beta_{02} - \beta_{21}\cdot x_1^{*(i)})^2\right)\right)^{\mathbb{1}(F_i\neq 2)}.
\end{aligned}
$$

In order to estimate parameters β and σ, the maximum likelihood method is used due to the simplicity of our likelihood function. The maximum likelihood method estimates the parameters which maximize the likelihood function. In order to simplify the calculation, the log-likelihood function is used.

$$
\begin{aligned}
l_n(\beta,\sigma) = \sum_{i=1}^{n} &\mathbb{1}(F_i\neq 1)\cdot\left(-\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_1^{(i)} - \beta_{01})^2\right) \\
&+ \mathbb{1}(F_i\neq 2)\cdot\left(-\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_2^{(i)} - \beta_{02} - \beta_{21}\cdot x_1^{(i)})^2\right). \qquad (5.13)
\end{aligned}
$$

The following lemma shows how the parameters β,σ are estimated.

Lemma 5.2.1. Consider the LSEMs in (5.1), (5.4), (5.6). Let $x = (x_1^{(1)},x_2^{(1)})\dots(x_1^{(d)},x_2^{(d)})$ be a set of observations which follow one of the LSEMs above. Then, the log-likelihood function is

$$
\begin{aligned}
l_n(\beta,\sigma) = \sum_{i=1}^{n} &\mathbb{1}(F_i\neq 1)\cdot\left(-\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_1^{(i)} - \beta_{01})^2\right) \\
&+ \mathbb{1}(F_i\neq 2)\cdot\left(-\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_2^{(i)} - \beta_{02} - \beta_{21}\cdot x_1^{(i)})^2\right).
\end{aligned}
$$

The maximal likelihood estimators for β are

$$
\begin{aligned}
\arg\max_{\beta_{01}} l_n(\beta,\sigma) &= \bar{x}_1^{F_i\neq 1} \\
\arg\max_{\beta_{02}} l_n(\beta,\sigma) &= \bar{x}_2^{F\neq 2} - \hat{\beta}_{21}\left(\frac{n_0+n_2}{n_0+n_1}\right)\bar{x}_1^{F\neq 2} \\
\arg\max_{\beta_{21}} l_n(\beta,\sigma) &= \frac{\sum_{i\in D^{F_i\neq 2}}(\bar{x}_2^{F\neq 2} - x_2^{(i)})x_1^{(i)}}{\sum_{i\in D^{F_i\neq 2}}\left(\left(\frac{n_0+n_2}{n_0+n_1}\bar{x}_1^{F\neq 2}\right)^2 - (x_1^{(i)})^2\right)}.
\end{aligned}
$$

where n_0 is the number of samples with observational data, and n_1 is the number of samples with interventional data under $\text{do}(X_i = x_1)$ and

$$\bar{x}_i^{F\neq j} = \frac{1}{n-n_j}\sum_{k\in D^{F\neq j}} x_i^{(k)}, \quad i,j=1,2.$$

Proof. Taking the derivative with respect to β_{01}, we obtain the equation

$$\frac{\partial l_n(\beta_{01},\sigma)}{\partial \beta_{01}} = \sum_{i=1}^{n} \mathbb{1}(F_i \neq 1)\cdot\left(\frac{x_1^{(i)}-\beta_{01}}{\sigma_1^2}\right)$$

Substituting the maximizer $\hat{\beta}_{01}$ in the above equation, we equate this to 0. This equation is solved by

$$\hat{\beta}_{01} = \frac{1}{n_0+n_2}\sum_{i\in D^{F_i\neq 1}} x_1^{(i)}.$$

The derivative of the log-likelihood function with respect to β_{02} is

$$\frac{\partial l_n(\beta_{02},\beta_{21},\sigma)}{\partial \beta_{02}} = \sum_{i=1}^{n} \mathbb{1}(F_i \neq 2)\left(\frac{x_2^{(i)}-\beta_{02}-\beta_{21}\cdot x_1^{(i)}}{\sigma_2^2}\right).$$

Equating this derivative to zero gives the maximizer $\hat{\beta}_{02}$

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{1}(F_i \neq 2)\cdot\left(\frac{x_2^{(i)}-\hat{\beta}_{02}-\hat{\beta}_{21}\cdot x_1^{(i)}}{\sigma_2^2}\right) &= 0\\
\sum_{i\in D^{F_i\neq 2}} \hat{\beta}_{02} &= \sum_{i\in D^{F_i\neq 2}} (x_2^{(i)}-\hat{\beta}_{21}x_1^{(i)})\\
\hat{\beta}_{02} &= \frac{1}{n_0+n_1}\sum_{i\in D^{F_i\neq 2}} (x_2^{(i)}-\hat{\beta}_{21}x_1^{*(i)})\\
\hat{\beta}_{02} &= \frac{1}{n_0+n_1}\sum_{i\in D^{F_i\neq 2}} x_2^{(i)} - \hat{\beta}_{21}\frac{1}{n_0+n_1}\sum_{i\in D^{F_i\neq 2}} x_1^{(i)}\\
\hat{\beta}_{02} &= \bar{x}_2^{F\neq 2} - \hat{\beta}_{21}\left(\frac{n_0+n_2}{n_0+n_1}\right)\bar{x}_1^{F\neq 2}
\end{aligned}$$

The number of entire samples is equal to $n = n_0+n_1+n_2$. Now, we differentiate the log-likelihood function with respect to β_{21} and equate to 0

$$\frac{\partial l_n(\hat{\beta}_{02},\hat{\beta}_{21},\sigma)}{\partial \beta_{21}} = \sum_{i=1}^{n} \mathbb{1}(F_i \neq 2)\cdot\left(\frac{x_2^{(i)}-\hat{\beta}_{02}-\hat{\beta}_{21}\cdot x_1^{(i)}}{\sigma_2^2}\right)\cdot x_1^{(i)} = 0$$

which is solved by

$$\begin{aligned}
\sum_{i\in D^{F_i\neq 2}}\left(\frac{x_2^{(i)}x_1^{(i)}-\hat\beta_{02}x_1^{(i)}-\hat\beta_{21}(x_1^{(i)})^2}{\sigma_2^2}\right)&=0\\
0&=\sum_{i\in D^{F_i\neq 2}}\left(x_2^{(i)}x_1^{(i)}-\left(\bar x_2^{F\neq 2}-\hat\beta_{21}\left(\frac{n_0+n_2}{n_0+n_1}\right)\bar x_1^{F\neq 2}\right)x_1^{(i)}-\hat\beta_{21}(x_1^{(i)})^2\right)\\
0&=\sum_{i\in D^{F_i\neq 2}}x_2^{(i)}x_1^{(i)}-\bar x_2^{F\neq 2}x_1^{(i)}+\left(\frac{n_0+n_2}{n_0+n_1}\right)\hat\beta_{21}\bar x_1^{F\neq 2}x_1^{(i)}-\hat\beta_{21}(x_1^{(i)})^2\\
0&=\sum_{i\in D^{F_i\neq 2}}(x_2^{(i)}-\bar x_2^{F\neq 2})x_1^{(i)}+\hat\beta_{21}(\sum_{i\in D^{F_i\neq 2}}-(x_1^{(i)})^2+\left(\frac{(n_0+n_2)^2}{n_0+n_1}\right)(\bar x_1^{F\neq 2})^2)\\
0&=\sum_{i\in D^{F_i\neq 2}}(x_2^{(i)}-\bar x_2^{F\neq 2})x_1^{(i)}+\hat\beta_{21}(\sum_{i\in D^{F_i\neq 2}}(\left(\frac{n_0+n_2}{n_0+n_1}\bar x_1^{F\neq 2}\right)^2-(x_1^{(i)})^2))\\
\hat\beta_{21}&=\frac{\sum_{i\in D^{F_i\neq 2}}(\bar x_2^{F\neq 2}-x_2^{(i)})x_1^{(i)}}{\sum_{i\in D^{F_i\neq 2}}(\left(\frac{n_0+n_2}{n_0+n_1}\bar x_1^{F\neq 2}\right)^2-(x_1^{(i)})^2)}
\end{aligned}$$

□

Proposition 6. Consider the log-likelihood function (5.13). Then,

$$\begin{aligned}
\arg\max_{\sigma_1>0} l_n(\beta_{02},\beta_{21},\sigma_1,\sigma_2)&=\frac{\sum_{i\in D^{F_i\neq 1}}(x_1^{(i)}-\beta_{01})^2}{n_0+n_2}\\
\arg\max_{\sigma_2>0} l_n(\beta_{02},\beta_{21},\sigma_1,\sigma_2)&=\frac{\sum_{i\in D^{F_i\neq 2}}(x_2^{(i)}-\beta_{02}-\beta_{21}x_1^{*(i)})^2}{n_0+n_1}.
\end{aligned}$$

Proof. We take the derivative of the log-likelihood function l_n with respect to σ_1

$$\begin{aligned}
\frac{\partial l_n(\beta_{02},\beta_{21},\sigma_1,\sigma_2)}{\partial\sigma_1}&=\sum_{i\in D^{F_i\neq 1}}\frac{\partial}{\partial\sigma_1}\Big(-\frac{1}{2}\log(2\pi\sigma_1^2)-\frac{1}{2\sigma_1^2}(x_1^{(i)}-\beta_{01})^2\Big)\\
&=\sum_{i\in D^{F_i\neq 1}}-\frac{1}{\sigma_1}+\frac{1}{\sigma_1^3}(x_1^{(i)}-\beta_{01})^2\\
&=-\frac{n_0+n_2}{\sigma_1}+\sum_{i\in D^{F_i\neq 1}}\frac{1}{\sigma_1^3}(x_1^{(i)}-\beta_{01})^2.
\end{aligned}$$

Equating this to zero, we obtain the maximizer $\hat\sigma_1$

$$\begin{aligned}
0&=-\frac{n_0+n_2}{\hat\sigma_1}+\sum_{i\in D^{F_i\neq 1}}\frac{1}{\hat\sigma_1{}^3}(x_1^{(i)}-\beta_{01})^2\\
\hat\sigma_1&=\frac{\sum_{i\in D^{F_i\neq 1}}(x_1^{(i)}-\beta_{01})^2}{n_0+n_2}.
\end{aligned}$$

Analogously, we can compute $\hat{\sigma}_2$

$$\hat{\sigma}_2 = \frac{\sum_{i \in D^{F_i \neq 2}} (x_2^{(i)} - \beta_{02} - \beta_{21} x_1^{*(i)})^2}{n_0 + n_1}.$$

□

Remark. Estimating the parameters β_{02} ,β_{21}, and σ_2 is equivalent to estimating the parameter of a linear regression model $Y = \beta_{02} + \beta_{21} X + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_2)$. In our case, Y and X are corresponding to $X_2^{(i)}$ and $X_1^{(i)}$ for $i \in D^{F \neq 2}$. Estimating the parameter β_{01}, σ_1 is equivalent to estimating parameters of the model $X_1^{(i)} = \varepsilon$ for $i \in D^{F \neq 1}$ where $\varepsilon \sim \mathcal{N}(\beta_{01}, \sigma_2)$. Thus we can rewrite the previous results of the estimators in the matrix form as follows:

$$\hat{\beta}_{01} = (X_{0,F \neq 1}^T X_{0,F \neq 1}) X_{0,F \neq 1} X_{1,F \neq 1}$$

where $X_{0,F \neq 1}^T = (1, 1, \ldots, 1, 1)$ which has the length of $n_0 + n_2$ and and

$$X_{1,D^{F \neq 1}}^T = \begin{pmatrix} x_1^{(i_1)} & \ldots & x_1^{(i_{n_0+n_2-1})} & x_1^{(i_{n_0+n_2})} \end{pmatrix} \text{ for } i_k \in D^{F \neq 1},$$

$$\hat{\beta}_2 := \begin{pmatrix} \beta_{02} \\ \beta_{21} \end{pmatrix} = (X_{01,F \neq 2}^T X_{01,F \neq 2}) X_{01,F \neq 2} X_{(2,F \neq 2)}$$

where

$$X_{01,D^{F \neq 2}}^T = \begin{pmatrix} 1 & \ldots & 1 & 1 \\ x_1^{(i_1)} & \ldots & x_1^{(i_{n_0+n_1-1})} & x_1^{(i_{n_0+n_1})} \end{pmatrix} \text{ for } i_k \in D^{F \neq 2}$$

and

$$X_{2,D^{F \neq 2}}^T = \begin{pmatrix} x_2^{(i_1)} & \ldots & x_2^{(i_{n_0+n_1-1})} & x_2^{(i_{n_0+n_1})} \end{pmatrix} \text{ for } i_k \in D^{F \neq 2}$$

5.3 Construction of a Confidence Set in Two-dimensional Case

In the following, we assume, without loss of generality, that $\beta_{01} = \beta_{02} = 0$. Solving a non-zero case is achieved by centralizing or standardizing the variables. Consider the linear structural equation model (2.12) and a data set that consists of both observational and interventional data. Then we have three models.

i) Model with observational data

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= \beta_{21} X_1 + \varepsilon_2 \end{aligned}$$

ii) Model with interventional data under $\text{do}(X_1 = x_1^*)$

$$\begin{aligned} X_1 &= x_1^* \\ X_2 &= \beta_{21} x_1^* + \varepsilon_2. \end{aligned}$$

iii) Model with interventional data under $\text{do}(X_2 = x_2^*)$

$$X_1 = \varepsilon_1$$
$$X_2 = x_2^*$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i), i = 1,2$. Consider the following observations of random vectors $x^{(1)}, \ldots, x^{(n)}$, where

$$x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ F_i \end{pmatrix},$$

and i is a element of the index set $D = \{1,2,\ldots n-1,n\}$ which consists of three groups, that is, $D = D^{F=0} \cup D^{F=1} \cup D^{F=2}$ and $D^{F=0} \cap D^{F=1} \cap D^{F=2} = \emptyset$. The samples $x_1^{(i)}, x_2^{(i)}, i \in D^{F=0}$ follow the first model (i). Analogously, the samples $x_1^{(i)}, x_2^{(i)}, i \in D^{F=1}$ follow the second model (ii) and the samples for $i \in D^{F=2}$ follow the model (iii). The empirical covariance matrix is defined as $\Sigma^{F=k} = \frac{1}{n_k} \sum_{F_i=k} X^{(i)} X^{(i)^T}$ for $k = 0,1,2$ and $\Sigma^{F \neq k} = \frac{1}{n-n_k} \sum_{F_i=k} X^{(i)} X^{(i)^T}$ for $k = 0,1,2$. The joint distribution and log-likelihood function of each model are

i) Model with observational data

$$\begin{aligned} f(x_1^{(i)}, x_2^{(i)} | \beta_{21}, \sigma_1, \sigma_2) &= f(x_2^{(i)} | x_1^{(i)}, \beta_{21}, \sigma_2) \cdot f(x_1^{(i)} | \sigma_1) \\ &= \frac{1}{\sqrt{2\pi\sigma_2^2}} \cdot e^{-\frac{1}{2\sigma_2^2}(x_2^{(i)} - \beta_{21} \cdot x_1^{(i)})^2} \frac{1}{\sqrt{2\pi\sigma_1^2}} \cdot e^{-\frac{1}{2\sigma_1^2}(x_1^{(i)})^2} \\ l_{n_0}(\beta_{21}, \sigma_1, \sigma_2 | x^{(i)}, i \in D^{F=0}) &= \sum_{i \in D^{F=0}} \Big(-\frac{1}{2}\log(2\pi\sigma_2) - \frac{1}{2\sigma_2}(x_2^{(i)} - \beta_{21} x_1^{(i)})^2 \\ &\qquad - \frac{1}{2}\log(2\pi\sigma_1) - \frac{1}{2\sigma_1}(x_1^{(i)})^2 \Big) \end{aligned}$$

ii) Model with interventional data under $\text{do}(X_1^{(i)} = x_1^{(i)})$

$$\begin{aligned} f(x_1^{(i)}, x_2^{(i)} | \beta_{21}, \sigma_1, \sigma_2) &= f(x_2^{(i)} | x_1^{(i)}, \beta_{21}, \sigma_2) \\ &= \frac{1}{\sqrt{2\pi\sigma_2^2}} \cdot e^{-\frac{1}{2\sigma_2^2}(x_2^{(i)} - \beta_{21} \cdot x_1^{(i)})^2} \\ l_{n_1}(\beta_{21}, \sigma_1, \sigma_2 | x^{(i)}, i \in D^{F=1}) &= \sum_{i \in D^{F=1}} -\frac{1}{2}\log(2\pi\sigma_2) - \frac{1}{2\sigma_2}(x_2^{(i)} - \beta_{21} x_1^{(i)})^2 \end{aligned}$$

iii) Model with interventional data under $\text{do}(X_2^{(i)} = x_2^{(i)})$

$$\begin{aligned} f(x_1^{(i)}, x_2^{(i)} | \beta_{21}, \sigma_1, \sigma_2) &= f(x_1^{(i)} | \sigma_1) \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \cdot e^{-\frac{1}{2\sigma_1^2}(x_1^{(i)})^2} \\ l_{n_2}(\beta_{21}, \sigma_1, \sigma_2 | x^{(i)}, i \in D^{F=2}) &= \sum_{i \in D^{F=2}} -\frac{1}{2}\log(2\pi\sigma_1) - \frac{1}{2\sigma_1}(x_1^{(i)})^2 \end{aligned}$$

To calculate the profile likelihood function evaluated by the data set with D_0, we assume that the parameter of the causal effect β_{21} takes a fixed value Ψ. The log-likelihood function of this case is

$$l^{(0)}(\Psi,\sigma_1,\sigma_2) = \frac{n_{0,0}+n_{0,1}}{2}\log(2\pi\sigma_2) - \frac{1}{2\sigma_2}\sum_{i\in D_0^{F_i\neq 2}}(x_2^{(i)}-\Psi x_1^{(i)})^2 + \tilde{l}^{(0)}(\sigma_1) \tag{5.14}$$

where $\tilde{l}^{(0)}(\sigma_1)$ denotes a part of likelihood function which does not depend on Ψ:

$$\tilde{l}^{(0)}(\sigma_1) = \frac{n_{0,0}+n_{0,2}}{2}\log(2\pi\sigma_1) - \frac{1}{2\sigma_1}\sum_{i\in D_0^{F_i\neq 1}}(x_1^{(i)})^2$$

In order to determine the profile likelihood function, we first assumed that $\mathscr{C}(1\to 2)$ takes a fixed value Ψ and under this constraint maximize with respect to the parameters σ_1 and σ_2.

Lemma 5.3.1. Consider the profile likelihood in (5.14). Then, the maximal likelihood estimators calculated by the data set D_0 are

$$\begin{aligned}
&\arg\max_{\sigma_1^2>0} l_n(\Psi\sigma_1,\sigma_2|x^{(1)},\ldots,x^{(n)}) = \hat{\Sigma}_{0,11}^{F\neq 1}\\
&\arg\max_{\sigma_2^2>0} l_n(\Psi,\sigma_1,\sigma_2|x^{(1)},\ldots,x^{(n)}) = \hat{\Sigma}_{0,11}^{F\neq 2}\Psi^2 - 2\hat{\Sigma}_{0,12}^{F\neq 2}\Psi + \hat{\Sigma}_{0,22}^{F\neq 2}
\end{aligned}$$

where the matrix $\hat{\Sigma}_0^{F\neq i}$ is the empirical covariance matrix based on the data with D_0 such that

$$\hat{\Sigma}_0^{F\neq i} = \begin{pmatrix} \hat{\Sigma}_{0,11}^{F\neq i} & \hat{\Sigma}_{0,12}^{F\neq i} \\ \hat{\Sigma}_{0,21}^{F\neq i} & \hat{\Sigma}_{0,22}^{F\neq i} \end{pmatrix} = \frac{1}{n_{0,0}+n_{0,l}}\sum_{k\in D_0^{F\neq i}} x^{(k)}x^{(k)^T}, \quad l,i=1,2, \text{ and } l\neq i$$

Proof. From the Proposition (6), we know that

$$\begin{aligned}
\hat{\sigma_1}^2 &:= \arg\max_{\sigma_1^2>0} l_n(\beta_{02},\beta_{21},\sigma_1,\sigma_2) = \frac{\sum_{F_i\neq 1}(x_1^{(i)}-\beta_{01})^2}{n_0+n_2}\\
\hat{\sigma_2}^2 &:= \arg\max_{\sigma_2^2>0} l_n(\beta_{02},\beta_{21},\sigma_1,\sigma_2) = \frac{\sum_{F_i\neq 2}(x_2^{(i)}-\beta_{02}-\beta_{21}x_1^{(i)})^2}{n_0+n_1}.
\end{aligned}$$

Since we assumed that the β_{01} and β_{02} are zero and we set β_{21} to Ψ, we obtain

$$\begin{aligned}
\arg\max_{\sigma_1^2>0} l_n(\Psi,\sigma_1,\sigma_2) &= \frac{\sum_{i\in D_0^{F\neq 1}}(x_1^{(i)})^2}{n_{0,0}+n_{0,2}}\\
&= \hat{\Sigma}_{0,11}^{F\neq 1}
\end{aligned}$$

and

$$\begin{aligned}\arg\max_{\sigma_2^2>0} l_n(\Psi,\sigma_1,\sigma_2) &= \frac{\sum_{i\in D_0^{F\neq 2}}(x_2^{(i)}-\Psi x_1^{*(i)})^2}{n_{0,0}+n_{0,1}}\\ &= \frac{\sum_{i\in D_0^{F\neq 1}}(x_2^{(i)})^2+\Psi^2(x_1^{(i)})^2-2\psi x_1^{(i)}x_2^{(i)}}{n_{0,0}+n_{0,1}}\\ &= \hat{\Sigma}_{0,11}^{F\neq 2}\Psi^2-2\hat{\Sigma}_{0,12}^{F\neq 2}\Psi+\hat{\Sigma}_{0,22}^{F\neq 2}\end{aligned}$$

□

Case 1: $\Psi\neq 0$

We determined logarithmic version of the profile likelihood function $l_\dagger^{(0)}(\Psi)$ based on the data set $D_0=D_0^{F=0}\cup D_0^{F=1}\cup D_0^{F=2}$ from above result. Substituting the maximal likelihood estimators of remaining parameters σ_1,σ_2 in Eq.(5.14), we obtain

$$\begin{aligned}l_\dagger^{(0)}(\Psi) :=& \max_{\sigma_1,\sigma_2} l^{(0)}(\psi,\sigma_1,\sigma_2) = l^{(0)}(\psi,\hat{\sigma}_1,\hat{\sigma}_2)\\ =& -\frac{n_{0,0}+n_{0,1}}{2}\log(2\pi\hat{\sigma}_2)-\frac{1}{2\hat{\sigma}_2}\sum_{i\in D_0^{F\neq 2}}(x_2^{(i)}-\Psi x_1^{(i)})^2+\tilde{l}^{(0)}(\hat{\sigma}_1)\\ =& -\frac{n_{0,0}+n_{0,1}}{2}\log(2\pi(\hat{\Sigma}_{0,11}^{F\neq 2}\Psi^2-2\hat{\Sigma}_{0,12}^{F\neq 2}\Psi+\hat{\Sigma}_{0,22}^{F\neq 2}))\\ &-\frac{n_{0,0}+n_{0,1}}{2\sum_{i\in D_0^{F\neq 2}}(x_2^{(i)}-\Psi x_1^{(i)})^2}\sum_{i\in D_0^{F\neq 2}}(x_2^{(i)}-\Psi x_1^{(i)})^2+\tilde{l}^{(0)}(\hat{\sigma}_1)\\ =& -\frac{n_{0,0}+n_{0,1}}{2}\log(2\pi(\hat{\Sigma}_{0,11}^{F\neq 2}\Psi^2-2\hat{\Sigma}_{0,12}^{F\neq 2}\Psi+\hat{\Sigma}_{0,22}^{F\neq 2}))-\frac{n_{0,0}+n_{0,1}}{2}+\tilde{l}^{(0)}(\hat{\sigma}_1)\end{aligned}\tag{5.15}$$

where $\Sigma_0^{F\neq 2}$ and $\Sigma_0^{F\neq 1}$ are the empirical covariance matrices calculated based on the data set with the index set D_0.

Case 2: $\Psi=0$

Suppose that $\mathscr{C}(1\to 2)=\Psi=0$. In (M2.2), the direct causal effect $\mathscr{C}(1\to 2)$ is always zero. Therefore, (M2.1) with $\Psi=0$ is a special case of (M2.2). Since all parameter involved in (M2.2) are independent on the constraint $\mathscr{C}(1\to 2)$, we obtain the profile likelihood function by maximizing the likelihood function of (M2.2). The log-likelihood of (M2.2) is

$$\begin{aligned}l_{2\to 1}^{(0)}(\beta_{12},\sigma_1,\sigma_2) =& \sum_{i\in D_0}\mathbb{1}(F_i\neq 2)\cdot\Big(-\frac{1}{2}\log(2\pi\sigma_2^2)-\frac{1}{2\sigma_2^2}(x_2^{(i)})^2\Big)\\ &+\mathbb{1}(F_i\neq 1)\cdot\Big(-\frac{1}{2}\log(2\pi\sigma_1^2)-\frac{1}{2\sigma_1^2}(x_1^{(i)}-\beta_{12}\cdot x_2^{(i)})^2\Big).\end{aligned}\tag{5.16}$$

The maximum likelihood estimators of this model are

$$\hat{\beta}_{12}=\arg\max_{\beta_{12}}\sum_{i\in D_0^{F\neq 1}}(x_1^{(i)}-\beta_{12}\cdot x_2^{(i)})^2=\frac{\Sigma_{0,12}^{F\neq 1}}{\Sigma_{0,22}^{F\neq 1}},$$

$$\hat{\sigma}_2^2 = \arg\max_{\sigma_2^2} \sum_{i \in D_0^{F\neq 2}} \left(-\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_2^{(i)})^2 \right) = \Sigma_{0,22}^{F\neq 2},$$

and

$$\hat{\sigma}_1^2 = \arg\max_{\sigma_1^2} \sum_{i \in D_0^{F\neq 1}} \left(-\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_1^{(i)} - \beta_{12}\cdot x_2^{(i)})^2 \right) = \Sigma_{0,11}^{F\neq 1} - \frac{(\Sigma_{0,12}^{F\neq 1})^2}{\Sigma_{0,22}^{F\neq 1}}$$

where $\Sigma_0^{F\neq i}$ is empirical covariance matrix based on the data with the index set $D_0^{F\neq i}$ for $i = 1,2$. Therefore, the profile likelihood in this case is

$$\begin{aligned} l_\dagger^{(0)}(\Psi) &:= \max_{\beta_{21},\sigma_1,\sigma_2} l_{n_0,2\to 1}^{(0)}(\beta_{12},\sigma_1,\sigma_2) \\ &= -\frac{n_{0,0}+n_{0,1}}{2}\log(2\pi\Sigma_{0,22}^{F\neq 2}) - \frac{n_{0,0}+n_{0,1}}{2} \\ &\quad - \frac{n_{0,0}+n_{0,2}}{2}\log(2\pi(\Sigma_{0,11}^{F\neq 1} - \frac{(\Sigma_{0,12}^{F\neq 1})^2}{\Sigma_{0,22}^{F\neq 1}})) - \frac{n_{0,0}+n_{0,2}}{2}. \end{aligned} \tag{5.17}$$

Now, we determine the boundaries of the confidence interval of the split likelihood function by using the profile likelihood function in Equation (5.15). The logarithmic version of the split likelihood ratio ξ_n from the Equation (4.3) is therefore

$$\begin{aligned} \xi_n &= l^{(0)}(\hat{\theta}_1) - l_\dagger^{(0)}(\Psi) \\ &= l^{(0)}(\hat{\theta}_1) + \frac{n_{0,0}+n_{0,1}}{2}\log(2\pi(\hat{\Sigma}_{0,11}^{F\neq 2}\Psi^2 - 2\hat{\Sigma}_{0,12}^{F\neq 2}\Psi + \hat{\Sigma}_{0,22}^{F\neq 2})) + \frac{n_{0,0}+n_{0,1}}{2} - \tilde{l}_n(\hat{\sigma}_1) \end{aligned}$$

where the likelihood function $l^{(0)}(\hat{\theta}_1)$ is the likelihood function evaluated by the data with the index setD_0, given the maximum likelihood estimator calculated by the data with the index set D_1

$$\hat{\theta}_1 = \begin{pmatrix} \hat{\beta}_{1,12} \\ \hat{\sigma}_{1,1} \\ \hat{\sigma}_{1,2} \end{pmatrix}.$$

According to Theorem 4.0.4, the set

$$C_n := \left\{ \Psi \in \mathbf{R} : \xi_n \leq \log(\frac{1}{\alpha}) \right\}$$

is $1-\alpha$ confidence interval set for $\mathscr{C}(1\to 2)$. Rearranging the inequality $\xi_n \leq \log(\frac{1}{\alpha})$ in the case $\Psi \neq 0$, we get

$$\hat{\Sigma}_{0,11}^{F\neq 2}\Psi^2 - 2\hat{\Sigma}_{0,12}^{F\neq 2}\Psi + \hat{\Sigma}_{0,22}^{F\neq 2} - B \leq 0 \tag{5.18}$$

where

$$B = \frac{1}{2\pi}\exp\left(\frac{2}{n_{0,0}+n_{0,1}}(\tilde{l}(\hat{\sigma}_1) - l^0(\hat{\theta}_1) - \frac{n_{0,0}+n_{0,1}}{2} + \log(\frac{1}{\alpha}))\right).$$

The left-hand side is a function of Ψ and a parabola opening to the top. The solution of this inequality is all Ψ in between the two zeros of this parabola if they exist. Using a well-known formula for the roots of a general quadratic polynomial gives that the inequality (5.18) is equivalent to

$$\Psi \in \begin{cases} [L(\hat{\Sigma}_0^{F\neq 2}), U(\hat{\Sigma}_0^{F\neq 2})] & \text{if} \quad G \geq 0 \\ \emptyset & \text{if} \quad G < 0 \end{cases}$$

where

$$G = (\hat{\Sigma}_{0,12}^{F\neq 2})^2 - \hat{\Sigma}_{0,11}^{F\neq 2}(\hat{\Sigma}_{0,22}^{F\neq 2} - B)$$

and

$$L(\hat{\Sigma}_0^{F\neq 2}) = \frac{(\hat{\Sigma}_{0,12}^{F\neq 2})^2 - \sqrt{G}}{\hat{\Sigma}_{0,11}^{F\neq 2}},$$
$$U(\hat{\Sigma}_0^{F\neq 2}) = \frac{(\hat{\Sigma}_{0,12}^{F\neq 2})^2 + \sqrt{G}}{\hat{\Sigma}_{0,11}^{F\neq 2}}.$$

Rearranging the inequality $\xi_n > log(1/\alpha)$ in the case $\Psi = 0$ delivers

$$\begin{aligned} H =& l^{(0)}(\hat{\theta}_1) + \frac{n_{0,0}+n_{0,1}}{2}\log(2\pi\Sigma_{0,22}^{F\neq 2}) + \frac{n_{0,0}+n_{0,1}}{2} \\ &+ \frac{n_{0,0}+n_{0,2}}{2}\log(2\pi(\Sigma_{0,11}^{F\neq 1} - \frac{(\Sigma_{0,12}^{F\neq 1})^2}{\Sigma_{0,22}^{F\neq 1}})) + \frac{n_{0,0}+n_{0,2}}{2} \geq 0. \end{aligned}$$

Totally, we rewrite the description of a $1-\alpha$ confidence set for $\mathscr{C}(1\to 2)$ in detail as follows

$$B_n = \begin{cases} ([L(\hat{\Sigma}_0^{F\neq 2}), U(\hat{\Sigma}_0^{F\neq 2})] \cap (\mathbb{R}\setminus\{0\})) \cup \{0\}, & \text{if} \quad G \geq 0 \text{ and } H \leq 0, \\ \{0\}, & \text{if} \quad G < 0 \text{ and } H \leq 0, \\ L(\hat{\Sigma}_0^{F\neq 2}), U(\hat{\Sigma}_0^{F\neq 2})] \cap (\mathbb{R}\setminus\{0\}), & \text{if} \quad G \geq 0 \text{ and } H > 0, \\ \emptyset, & \text{if} \quad G < 0 \text{ and } H > 0. \end{cases}$$

Now, we write down the test problem we consider in detail.

H_0 : The data follows a LSEM and the causal effect $\mathscr{C}(1\to 2) = \Psi$ versus

H_1: H_0 does not hold.

In order to construct a test with a valid finite-sample bound on the type I error for this problem, we calculate the split likelihood ratio test statistic. Again, it is necessary to distinguish two cases. In the case $\Psi \neq 0$, the logarithmic version of this test statistic is determined by

$$\xi_n = l^0(\hat{\theta}_1) + \frac{n_{0,0}+n_{0,1}}{2}\log(2\pi(\hat{\Sigma}_{0,11}^{F\neq 2}\Psi^2 - 2\hat{\Sigma}_{0,12}^{F\neq 2}\Psi + \hat{\Sigma}_{0,22}^{F\neq 2})) + \frac{n_{0,0}+n_{0,1}}{2} - \tilde{l}(\hat{\sigma}_1)$$

whereas in the case $\Psi = 0$,

$$\begin{aligned} \xi_n =& l^{(0)}(\hat{\theta}_1) + \frac{n_{0,0}+n_{0,1}}{2}\log(2\pi\Sigma_{0,22}^{F\neq 2}) + \frac{n_{0,0}+n_{0,1}}{2} \\ &+ \frac{n_{0,0}+n_{0,2}}{2}\log(2\pi(\Sigma_{0,11}^{F\neq 1} - \frac{(\Sigma_{0,12}^{F\neq 1})^2}{\Sigma_{0,22}^{F\neq 1}})) + \frac{n_{0,0}+n_{0,2}}{2}. \end{aligned}$$

The split likelihood ratio test rejects H_0 if and only if $\xi_n > \log(1/\alpha)$.

5.4 Construction of a Confidence Set in Three-dimensional Case

Here, the three-dimensional case is considered. This case consists of a more complex structure than the two-dimensional case. There exist six directions of dependency

(M3.1)	(M3.2)	(M3.3)
$X_1 := \varepsilon_1$	$X_1 := \varepsilon_1$	$X_1 := \beta_{12}X_2 + \varepsilon_1$
$X_2 := \beta_{21}X_1 + \varepsilon_2$	$X_2 := \beta_{21}X_1 + \beta_{23}X_3 + \varepsilon_2$	$X_2 := \varepsilon_2$
$X_3 := \beta_{31}X_1 + \beta_{32}X_2 + \varepsilon_3$	$X_3 := \beta_{31}X_1 + \varepsilon_3$	$X_3 := \beta_{31}X_1 + \beta_{32}X_2 + \varepsilon_3$
(M3.4)	(M3.5)	(M3.6)
$X_1 := \beta_{13}X_3 + \varepsilon_1$	$X_1 := \beta_{12}X_2 + \beta_{13}X_3 + \varepsilon_1$	$X_1 := \beta_{12}X_2 + \beta_{13}X_3 + \varepsilon_1$
$X_2 := \beta_{21}X_1 + \beta_{23}X_3 + \varepsilon_2$	$X_2 := \varepsilon_2$	$X_2 := \beta_{23}X_3 + \varepsilon_2$
$X_3 := \varepsilon_3$	$X_3 := \beta_{32}X_2 + \varepsilon_3$	$X_3 := \varepsilon_3$.

In all models, it is also allowed that the dependency takes the value 0. Without loss of generality, we observe the first model (M3.1). Fig.(5.2). displays the graph of this model. Calculations of the other models are derived by reindexing variables and errors appropriately according to the LSEMs of each model.

i) LSEM of Observational data

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= \beta_{21}X_1 + \varepsilon_2 \\ X_3 &= \beta_{31}X_1 + \beta_{32}X_2 + \varepsilon_3. \end{aligned} \tag{5.19}$$

ii) LSEM of interventional data under $\mathrm{do}(X_1 = x_1^*)$

$$\begin{aligned} X_1 &= x_1^* \\ X_2 &= \beta_{21}x_1^* + \varepsilon_2 \\ X_3 &= \beta_{31}X_1 + \beta_{32}X_2 + \varepsilon_3. \end{aligned} \tag{5.20}$$

iii) LSEM of interventional data under $\mathrm{do}(X_2 = x_2^*)$

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= x_2^* \\ X_3 &= \beta_{31}X_1 + \beta_{32}x_2^* + \varepsilon_3. \end{aligned} \tag{5.21}$$

iv) LSEM of interventional data under $\mathrm{do}(X_3 = x_3^*)$

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= \beta_{21}X_1 + \varepsilon_2 \\ X_3 &= x_3^*. \end{aligned} \tag{5.22}$$

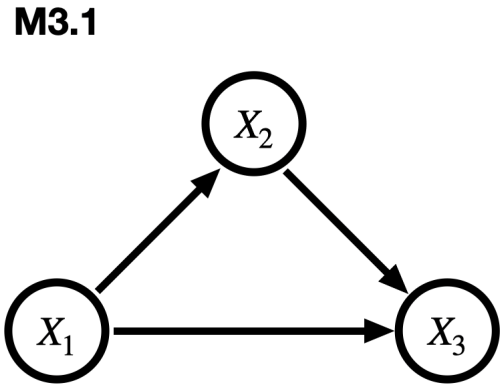


Figure 5.2: M3.1

As implemented in the two-dimensional case, we determine the density factorization. By Proposition 1, The density of the observational data which follow the LSEM in (5.20) factors as

$$f(X_1,X_2,X_3) = \sum_{i=1}^{3} f(X_i|X_{\mathrm{pa}(i)}) = f(X_1)f(X_2|X_1)f(X_3|X_1,X_2)$$

The interventional densities under $\mathrm{do}(X_i = x_i)$ for $i \in \{1,2,3\}$ factor as

$$f(X_1,X_2,X_3;\mathrm{do}(X_i = x_i)) = \begin{cases} f(X_2|X_1 = x_1)f(X_3|X_1 = x_1,X_2) & \text{for} \quad i = 1 \\ f(X_1)f(X_3|X_1,X_2 = x_2) & \text{for} \quad i = 2 \\ f(X_1)f(X_2|X_1) & \text{for} \quad i = 3 \end{cases}$$

Again, we consider the observations of random vector $x = ((x_1^{(1)},x_2^{(1)},x_3^{(1)},F_1)\dots(x_1^{(n)},x_2^{(n)},x_3^{(n)},F_n))$, where F_i take one of the values $0,1,2,3$. Let D be a index set $\{1,2,\dots,n\}$. Depending on the value of $F_i = j$, the index i of $x^{(i)} = (x_1^{(i)},x_2^{(i)},x_3^{(i)},F_i)$ is contained in the subset D_j. Hence, the index set D consists of 4 groups such that $D = D_0 \cap D_1 \cap D_2 \cap D_3$. In the following, we show how the likelihood function of our model factors. It follows exactly the same factorization rules as the two-dimensional

case:

$$
\begin{aligned}
L(\beta,\sigma) &= \prod_{i=1}^{n} f(x_1^{(i)},x_2^{(i)},x_3^{(i)}|\beta,\sigma) \\
&= \prod_{i=1}^{n} f(x_1^{(i)},x_2^{(i)},x_3^{(i)}|\beta,\sigma)^{\mathbb{1}(F_i=0)} f(x_1^{(i)},x_2^{(i)},x_3^{(i)}|\beta,\sigma)^{\mathbb{1}(F_i=1)} \\
&\quad \cdot f(x_1^{(i)},x_2^{(i)},x_3^{(i)}|\beta,\sigma)^{\mathbb{1}(F_i=2)} f(x_1^{(i)},x_2^{(i)},x_3^{(i)}|\beta,\sigma)^{\mathbb{1}(F_i=3)} \\
&= \prod_{i=1}^{n} \left(f(x_1^{(i)}) f(x_2^{(i)}|x_1^{(i)}) f(X_3|x_1^{(i)},x_2^{(i)}) \right)^{\mathbb{1}(F_i=0)} \left(f(x_2^{(i)}|x_1^{(i)}) f(x_3^{(i)}|x_1^{(i)},x_2^{(i)}) \right)^{\mathbb{1}(F_i=1)} \\
&\quad \cdot \left(f(x_1^{(i)}) f(x_3^{(i)}|x_1^{(i)},x_2^{(i)}) \right)^{\mathbb{1}(F_i=2)} \left(f(x_1^{(i)}) f(x_2^{(i)}|x_1^{(i)}) \right)^{\mathbb{1}(F_i=3)} \\
&= \prod_{i=1}^{n} f(x_1^{(i)}|\sigma_1)^{\mathbb{1}(F_i=0,2,3)} f(x_2^{(i)}|x_1^{(i)},\beta_{21},\sigma_2)^{\mathbb{1}(F_i=0,1,3)} \\
&\quad \cdot f(x_3^{(i)}|x_1^{(i)},x_2^{(i)},\beta_{31},\beta_{32},\sigma_3)^{\mathbb{1}(F_i=0,1,2)} \\
&= \prod_{i=1}^{n} f(x_1^{(i)}|\sigma_1)^{\mathbb{1}(F_i\neq 1)} f(x_2^{(i)}|x_1^{(i)},\beta_{21},\sigma_2)^{\mathbb{1}(F_i\neq 2)} \\
&\quad \cdot f(x_3^{(i)}|x_1^{(i)},x_2^{(i)},\beta_{31},\beta_{32},\sigma_3)^{\mathbb{1}(F_i\neq 3)}
\end{aligned} \tag{5.23}
$$

Since we assume that all error are normally distributed, we have the terms that arise in (5.23) as

$$
f(x_1^{(i)}|\sigma_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{1}{2\sigma_1^2} (x_1^{(i)})^2 \right)
$$

$$
f(x_2^{(i)}|x_1^{(i)},\beta_{21},\sigma_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{1}{2\sigma_2^2} (x_2^{(i)} - \beta_{21} x_1^{(i)})^2 \right)
$$

$$
f(x_3^{(i)}|x_1^{(i)},x_1^{(i)},\beta_{31},\beta_{32},\sigma_1) = \frac{1}{\sqrt{2\pi\sigma_3^2}} \exp\left(\frac{1}{2\sigma_3^2} (x_3^{(i)} - \beta_{31} x_1^{(i)} - \beta_{32} x_2^{(i)})^2 \right).
$$

Substituting the terms in (5.23), we obtain the likelihood function of the entire data

$$
\begin{aligned}
L_n(\beta,\sigma) = \prod_{i=1}^{n} &\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{1}{2\sigma_1^2} (x_1^{(i)})^2 \right) \right)^{\mathbb{1}(F_i\neq 1)} \\
&\cdot \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{1}{2\sigma_2^2} (x_2^{(i)} - \beta_{21} x_1^{(i)})^2 \right) \right)^{\mathbb{1}(F_i\neq 2)} \\
&\cdot \left(\frac{1}{\sqrt{2\pi\sigma_3^2}} \exp\left(\frac{1}{2\sigma_3^2} (x_3^{(i)} - \beta_{31} x_1^{(i)} - \beta_{32} x_2^{(i)})^2 \right) \right)^{\mathbb{1}(F_i\neq 3)}.
\end{aligned}
$$

Taking logarithm to the likelihood function, we obtain the log-likelihood function

$$\begin{aligned} l_n(\beta,\sigma) = \sum_{i=1}^{n} & \mathbb{1}(F_i \neq 1)\Big(-\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_1^{(i)})^2 \Big) \\ & + \mathbb{1}(F_i \neq 2)\Big(-\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_2^{(i)} - \beta_{21}x_1^{(i)})^2 \Big) \\ & + \mathbb{1}(F_i \neq 3)\Big(-\frac{1}{2}\log(2\pi\sigma_3^2) - \frac{1}{2\sigma_3^2}(x_3^{(i)} - \beta_{31}x_1^{(i)} - \beta_{32}x_2^{(i)})^2 \Big). \end{aligned} \tag{5.24}$$

Taking the derivative with respect to σ_i for $i = 1,2,3$ similarly to the two-dimensional case, the maximal likelihood estimators for the parameters are determined. The maximal likelihood estimators are in the following

$$\begin{aligned} \hat{\sigma}_1 &:= \frac{1}{n_0 + n_2 + n_3} \sum_{F_i \neq 1} (x_1^{(i)}) \\ \hat{\sigma}_2 &:= \frac{1}{n_0 + n_1 + n_3} \sum_{F_i \neq 2} (x_2^{(i)} - x_1^{(i)}\beta_{21}) \\ \hat{\sigma}_3 &:= \frac{1}{n_0 + n_1 + n_2} \sum_{F_i \neq 3} (x_3^{(i)} - x_1^{(i)}\beta_{31} - x_2^{(i)}\beta_{32}) \end{aligned} \tag{5.25}$$

where n_0 is the number of samples of the observational data, and n_j is the number of samples of the interventional data under $\mathrm{do}(X_j^{(i)} = x_j^{(i)})$. Using this result, we calculate the profile likelihood functions for the total causal effects $\mathscr{C}(1 \to 2), \mathscr{C}(1 \to 3), \mathscr{C}(2 \to 3)$. The total causal effect of X_1 on X_2 and X_2 on X_3 is directly the parameters β_{21} and β_{32}, since

$$\begin{aligned} \mathscr{C}(1 \to 2) &= \frac{d}{dx_1}\mathbb{E}[X_2; (X_1 = x_1)] \\ &= \frac{d}{dx_1}\beta_{21}x_1 = \beta_{21} \end{aligned}$$

and

$$\begin{aligned} \mathscr{C}(2 \to 3) &= \frac{d}{dx_2}\mathbb{E}[X_3; (X_2 = x_2)] \\ &= \frac{d}{dx_2}(\beta_{31}\,\mathbb{E}[X_1] + \beta_{32}x_2) = \beta_{32}. \end{aligned}$$

For the total causal effect $\mathscr{C}(1 \to 3)$, estimating the parameters is more complicated since the fixed value Ψ is a function of the parameter. The total causal effect of X_1 on X_3 is

$$\begin{aligned} \mathscr{C}(1 \to 3) &= \frac{d}{dx_1}\mathbb{E}[X_3; (X_1 = x_1)] \\ &= \frac{d}{dx_1}(\beta_{31}x_1 + \beta_{32}\,\mathbb{E}[X_2; (X_1 = x_1)]) \\ &= \frac{d}{dx_1}(\beta_{31}x_1 + \beta_{32}\beta_{21}x_1) = \beta_{31} + \beta_{32}\beta_{21}. \end{aligned}$$

Assuming that the causal effect $\mathscr{C}(1 \to 2)$ or $\mathscr{C}(2 \to 3)$ takes a fixed value Ψ such that $\beta_{21} = \Psi$ or $\beta_{32} = \Psi$, the profile likelihood function is determined by maximizing the likelihood function $l_n^{(0)}$ based on the split data set with the index set D_0 with respect to the remaining parameters such as σ_1, σ_2, and β_{31}. The maximal likelihood estimator of σ_1, σ_2, and σ_3 are listed in (5.25). In order to apply this calculation for every model, we define three cases. Each case represents the calculation of the profile likelihood function for each causal effect arising in the model (M3.1). Table 5.1 shows which causal effect of the given models from (M3.1) to (M3.6) belongs to one of the three cases. For the other models, the calculations can analogously be applied according to Table 5.1 by setting the corresponding values to the indices i, j, k in the following LSEMs. All our three-dimensional models have the form of

$$\begin{aligned} X_i &:= \varepsilon_i \\ X_j &:= \beta_{ji} X_i + \varepsilon_j \\ X_k &:= \beta_{ki} X_i + \beta_{kj} X_j + \varepsilon_k \end{aligned}$$

for $i, j, k \in \{1,2,3\}$ and $i \neq j \neq k$. For this model, the three total causal effects $\mathscr{C}(i \to j)$, $\mathscr{C}(j \to k)$, and $\mathscr{C}(i \to k)$ can be considered to proceed a valid hypothetical test. For example, if $i = 1, j = 2, k = 3$, we have the mode (M3.1). If $i = 2, j = 1, k = 3$, we have the model (M3.3). Thus, the following calculations of the cases are valid for all possible three-dimensional models from (M3.1) to (M3.6) as listed in Tabel 5.1. In the following, the calculation of profile likelihood functions of the three cases is given.

Model	Case I	Case II	Case III
M3.1	$\mathscr{C}(1 \to 2)$	$\mathscr{C}(2 \to 3)$	$\mathscr{C}(1 \to 3)$
M3.2	$\mathscr{C}(1 \to 3)$	$\mathscr{C}(3 \to 2)$	$\mathscr{C}(1 \to 2)$
M3.3	$\mathscr{C}(2 \to 1)$	$\mathscr{C}(1 \to 3)$	$\mathscr{C}(2 \to 3)$
M3.4	$\mathscr{C}(3 \to 1)$	$\mathscr{C}(1 \to 2)$	$\mathscr{C}(3 \to 2)$
M3.5	$\mathscr{C}(2 \to 3)$	$\mathscr{C}(3 \to 1)$	$\mathscr{C}(2 \to 1)$
M3.6	$\mathscr{C}(3 \to 2)$	$\mathscr{C}(2 \to 1)$	$\mathscr{C}(3 \to 1)$

Table 5.1: Sorting which causal effect of the models belongs to one of the three cases.

Case I : First, we consider the case $\beta_{21} = \Psi$, then the maximal likelihood estimators of β_{31}, β_{32} are calculated by differentiating the last term of the log-likelihood function in (5.24) with respect to β_{31}, and β_{32}. Applying the maximal likelihood method introduced in Chapter 3, we estimate β_{31}, β_{32} as follows

$$\hat{\beta}_3 := \begin{pmatrix} \hat{\beta}_{31} \\ \hat{\beta}_{32} \end{pmatrix} = (X^T_{12,D_0^{F\neq 3}} X_{12,D_0^{F\neq 3}})^{-1} X^T_{12,D_0^{F\neq 3}} X_{3,D_0^{F\neq 3}}. \tag{5.26}$$

where

$$X_{jl,D_0^{F\neq m}} = \begin{pmatrix} x_j^{(i_1)} & \dots & x_j^{(i_{n_{0,0}+n_{0,j}+n_{0,l}-1})} & x_j^{(i_{n_{0,0}+n_{0,j}+n_{0,l}})} \\ x_l^{(i_1)} & \dots & x_l^{(i_{n_{0,0}+n_{0,j}+n_{0,l}-1})} & x_l^{(i_{n_{0,0}+n_{0,j}+n_{0,l}})} \end{pmatrix} \text{ for } i_k \in D_0^{F\neq m} \text{ and } j,l,m \in \{1,2,3\}, j \neq l \neq m$$

and

$$X_{j,D_0^{F\neq m}} = \begin{pmatrix} x_j^{(i_1)} & \dots & x_j^{(i_{n_{0,0}+n_{0,j}+n_{0,l}-1})} & x_j^{(i_{n_{0,0}+n_{0,j}+n_{0,l}})} \end{pmatrix} \text{ for } i_k \in D_0^{F\neq m} \text{ and } j,l,m \in \{1,2,3\}, j \neq l \neq m.$$

Substituting all above ML estimators and the fixed value Ψ for β_{21} in (5.24), we obtain the profile likelihood function

$$\begin{aligned} l_\dagger^{(0)}(\Psi) := l(\hat\beta, \hat\sigma, \Psi) = & \sum_{i\in D_0^{F_i\neq 1}} \left(-\frac{1}{2}\log(2\pi\hat\sigma_1^2) - \frac{1}{2\hat\sigma_1^2}(x_1^{(i)})^2 \right) \\ & + \sum_{i\in D_0^{F_i\neq 2}} \left(-\frac{1}{2}\log(2\pi\hat\sigma_2^2) - \frac{1}{2\hat\sigma_2^2}(x_2^{(i)} - \hat\beta_{21}x_1^{(i)})^2 \right) \\ & + \sum_{i\in D_0^{F_i\neq 3}} \left(-\frac{1}{2}\log(2\pi\hat\sigma_3^2) - \frac{1}{2\hat\sigma_3^2}(x_3^{(i)} - \hat\beta_{31}x_1^{(i)} - \hat\beta_{32}x_2^{(i)})^2 \right) \\ = & -\frac{n_{0,0}+n_{0,2}+n_{0,3}}{2}\log\left(2\pi\hat\Sigma_{0,11}^{F\neq 1}\right) - \frac{n_{0,0}+n_{0,2}+n_{0,3}}{2} \\ & - \frac{n_{0,0}+n_{0,1}+n_{0,3}}{2}\log\left(2\pi(\Psi^2\hat\Sigma_{0,11}^{F\neq 2} - 2\Psi\hat\Sigma_{0,12}^{F\neq 2} + \hat\Sigma_{0,22}^{F\neq 2})\right) - \frac{n_{0,0}+n_{0,1}+n_{0,3}}{2} \\ & - \frac{n_{0,0}+n_{0,1}+n_{0,2}}{2}\log\Big(2\pi(\hat\Sigma_{0,33}^{F\neq 3} + \hat\Sigma_{0,22}^{F\neq 3}\hat\beta_{32}^2 + \hat\Sigma_{0,11}^{F\neq 3}\hat\beta_{31}^2 \\ & \qquad + 2\hat\Sigma_{0,12}^{F\neq 3}\hat\beta_{31}\hat\beta_{32} - 2\hat\Sigma_{0,23}^{F\neq 3}\hat\beta_{32} - 2\hat\Sigma_{0,13}^{F\neq 3}\hat\beta_{31})\Big) - \frac{n_{0,0}+n_{0,1}+n_{0,2}}{2} \end{aligned}$$

where $\hat\beta_{32}$ and $\hat\beta_{31}$ are explicitly given in (5.26) and

$$\hat\Sigma_0^{F\neq i} = \begin{pmatrix} \hat\Sigma_{0,11}^{F\neq i} & \hat\Sigma_{0,12}^{F\neq i} & \hat\Sigma_{0,13}^{F\neq i} \\ \hat\Sigma_{0,21}^{F\neq i} & \hat\Sigma_{0,22}^{F\neq i} & \hat\Sigma_{0,23}^{F\neq i} \\ \hat\Sigma_{0,31}^{F\neq i} & \hat\Sigma_{0,32}^{F\neq i} & \hat\Sigma_{0,33}^{F\neq i} \end{pmatrix} = \frac{1}{n_{0,0}+n_{0,j}+n_{0,k}} \sum_{l\in D_0^{F\neq i}} (X^{(l)})^T X^{(l)}, \quad i,j,k \in \{1,2,3\} \text{ and } i\neq j\neq k$$

is the empirical covariance matrix.

Case II : Now, we consider the second case where the total causal effect $\mathscr{C}(2\to 3)$ takes a fixed value Ψ. The variances of the error terms can be estimated as in (5.25). β_{21} and β_{31} are estimated by using the maximal likelihood method in Chapter 3 for both following linear models

$$\begin{aligned} X_2 &= \beta_{21}X_1 + \varepsilon_2 \\ X_3 - \Psi X_2 &= \beta_{31}X_1 + \varepsilon_3. \end{aligned}$$

The estimated parameters $\hat\beta_{21}$ and $\hat\beta_{31}$ are given by

$$\hat\beta_{21} = (X_{1,D_0^{F\neq 2}}^T X_{1,D_0^{F\neq 2}})^{-1} X_{1,D_0^{F\neq 2}}^T X_{2,D_0^{F\neq 2}},$$

and

$$\hat\beta_{31}(\Psi) := \hat\beta_{31} = (X_{1,D_0^{F\neq 3}}^T X_{1,D_0^{F\neq 3}})^{-1} X_{1,D_0^{F\neq 3}}^T X_{3-2\Psi,D_0^{F\neq 3}}$$

where

$$X_{j-\Psi l, D_0^{F\neq m}} = \left(\begin{array}{cccc} x_j^{(i_1)} - \Psi x_l^{(i_1)} & \cdots & x_j^{(i_{n_{0,0}+n_{0,j}+n_{0,l}-1})} - \Psi x_l^{(i_{n_{0,0}+n_{0,j}+n_{0,l}-1})} & x_j^{(i_{n_{0,0}+n_{0,j}+n_{0,l}})} - \Psi x_l^{(i_{n_{0,0}+n_{0,j}+n_{0,l}})} \end{array} \right)$$
$$\text{for } i_k \in D_0^{F\neq m} \text{ and } j,l,m \in \{1,2,3\}, j \neq l \neq m.$$

Substituting $\hat{\beta}_{21}, \hat{\beta}_{31}, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3$ and Ψ into Eq.(5.24), the profile likelihood function is obtained as follows

$$\begin{aligned} l_\dagger(\Psi) = &- \frac{n_{0,0}+n_{0,2}+n_{0,3}}{2} \log\left(2\pi\hat{\Sigma}_{0,11}^{F\neq 1}\right) - \frac{n_{0,0}+n_{0,2}+n_{0,3}}{2} \\ &- \frac{n_{0,0}+n_{0,1}+n_{0,3}}{2} \log\left(2\pi(\hat{\beta}_{21}^2\hat{\Sigma}_{0,11}^{F\neq 2} - 2\hat{\beta}_{21}\hat{\Sigma}_{0,12}^{F\neq 2} + \hat{\Sigma}_{0,22}^{F\neq 2})\right) - \frac{n_{0,0}+n_{0,1}+n_{0,3}}{2} \\ &- \frac{n_{0,0}+n_{0,1}+n_{0,2}}{2} \log\Big(2\pi(\hat{\Sigma}_{0,33}^{F\neq 3} + \hat{\Sigma}_{0,22}^{F\neq 3}\Psi^2 + \hat{\Sigma}_{0,11}^{F\neq 3}\hat{\beta}_{31}(\Psi)^2 \\ &\qquad + 2\hat{\Sigma}_{0,12}^{F\neq 3}\Psi\hat{\beta}_{31}(\Psi) - 2\hat{\Sigma}_{0,23}^{F\neq 3}\Psi - 2\hat{\Sigma}_{0,13}^{F\neq 3}\hat{\beta}_{31}(\Psi))\Big) - \frac{n_{0,0}+n_{0,1}+n_{0,2}}{2}. \end{aligned}$$

Case III : The last case is the total causal effect of X_1 on X_3. Assuming that the total causal effect $\mathscr{C}(1\to 3)$ takes a fixed value Ψ, that is, $\mathscr{C}(1\to 3) = \beta_{31} + \beta_{32}\beta_{21} = \Psi$, we have to solve the following optimization problem in order to estimate the parameters

$$\begin{aligned} \max_{\beta,\sigma} \quad & l^{(0)}(\beta,\sigma) \\ \text{s.t.} \quad & \beta_{31} + \beta_{32}\beta_{21} = \Psi \end{aligned} \tag{5.27}$$

where

$$\beta = \left(\begin{array}{c} \beta_{21} \\ \beta_{31} \\ \beta_{32} \end{array} \right)$$

and

$$\sigma = \left(\begin{array}{c} \sigma_1 \\ \sigma_2 \end{array} \right)$$

and the function $l^{(0)}(\beta,\sigma)$ is the log-likelihood function based on the data set with D_0 given in (5.24). To solve the optimization problem, we take a derivative with respect to each parameter. Solving the constraint function for β_{31}, we have $\beta_{31} = \Psi - \beta_{21}\beta_{32}$. Thus, we have the objective function $l^{(0)}(\beta,\sigma)$ as

$$\begin{aligned} l^{(0)}(\beta,\sigma) = &\sum_{i\in D_0^{F_i\neq 1}} \left(-\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_1^{(i)})^2 \right) \\ &+ \sum_{i\in D_0^{F_i\neq 2}} \left(-\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_2^{(i)} - \beta_{21}x_1^{(i)})^2 \right) \\ &+ \sum_{i\in D_0^{F_i\neq 3}} \left(-\frac{1}{2}\log(2\pi\sigma_3^2) - \frac{1}{2\sigma_3^2}(x_3^{(i)} - (\Psi - \beta_{21}\beta_{32})x_1^{(i)} - \beta_{32}x_2^{(i)})^2 \right). \end{aligned}$$

By maximizing this function, the optimization problem is solved. Taking the derivative with respect to σ_1, σ_2, and σ_3 and equating the results to zero, then we have

$$\begin{aligned}
\sigma_1^2 &= \frac{1}{n_{0,0}+n_{0,2}+n_{0,3}} \sum_{i \in D_0^{F_i \neq 1}} (x_1^{(i)})^2 \\
\sigma_2^2 &= \frac{1}{n_{0,0}+n_{0,1}+n_{0,3}} \sum_{i \in D_0^{F_i \neq 2}} (x_2^{(i)} - x_1^{(i)}\beta_{21})^2 \\
\sigma_3^2 &= \frac{1}{n_{0,0}+n_{0,1}+n_{0,2}} \sum_{i \in D_0^{F_i \neq 3}} (x_3^{(i)} - x_1^{(i)}(\Psi - \beta_{21}\beta_{32}) - x_2^{(i)}\beta_{32})^2
\end{aligned} \tag{5.28}$$

Now we take derivative of $l_n(\beta, \sigma)$ with respect to β_{21} and β_{32}:

$$\begin{aligned}
\frac{\partial}{\partial \beta_{32}} l^{(0)}(\beta, \sigma) =& \frac{\partial}{\partial \beta_{32}} \sum_{i \in D_0^{F_i \neq 3}} \left(-\frac{1}{2}\log(2\pi\sigma_3^2) - \frac{1}{2\sigma_3^2}(x_3^{(i)} - (\Psi - \beta_{21}\beta_{32})x_1^{(i)} - \beta_{32}x_2^{(i)})^2 \right) \\
=& -\frac{1}{\sigma_3^2} \sum_{i \in D_0^{F_i \neq 3}} \left(x_3^{(i)} - (\Psi - \beta_{21}\beta_{32})x_1^{(i)} - \beta_{32}x_2^{(i)} \right) \left(\beta_{21}x_1^{(i)} - x_2^{(i)} \right).
\end{aligned} \tag{5.29}$$

and

$$\begin{aligned}
\frac{\partial}{\partial \beta_{21}} l^{(0)}(\beta, \sigma) =& \frac{\partial}{\partial \beta_{32}} \sum_{i \in D_0^{F_i \neq 2}} \left(-\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_2^{(i)} - \beta_{21}x_1^{(i)})^2 \right) \\
&+ \sum_{i \in D_0^{F_i \neq 3}} \left(-\frac{1}{2}\log(2\pi\sigma_3^2) - \frac{1}{2\sigma_3^2}(x_3^{(i)} - (\Psi - \beta_{21}\beta_{32})x_1^{(i)} - \beta_{32}x_2^{(i)})^2 \right) \\
=& -\frac{1}{\sigma_2^2} \sum_{i \in D_0^{F_i \neq 2}} \left(x_2^{(i)} - \beta_{21}x_1^{(i)} \right) x_1^{(i)} \\
&- \frac{1}{\sigma_3^2} \sum_{i \in D_0^{F_i \neq 3}} \left(x_3^{(i)} - (\Psi - \beta_{21}\beta_{32})x_1^{(i)} - \beta_{32}x_2^{(i)} \right) \beta_{32}x_1^{(i)}.
\end{aligned} \tag{5.30}$$

We equate the results in (5.29) and (5.30) to 0 as

$$\sum_{i \in D_0^{F_i \neq 3}} \left(x_3^{(i)} - (\Psi - \beta_{21}\beta_{32})x_1^{(i)} - \beta_{32}x_2^{(i)} \right) \left(\beta_{21}x_1^{(i)} - x_2^{(i)} \right) = 0 \tag{5.31}$$

$$\sigma_3^2 \sum_{i \in D_0^{F_i \neq 2}} \left(x_2^{(i)} - \beta_{21}x_1^{(i)} \right) x_1^{(i)} + \sigma_2^2 \sum_{i \in D_0^{F_i \neq 3}} \left(x_3^{(i)} - (\Psi - \beta_{21}\beta_{32})x_1^{(i)} - \beta_{32}x_2^{(i)} \right) \beta_{32}x_1^{(i)} = 0. \tag{5.32}$$

σ_1 is directly determined by the equation in (5.28). The remaining parameters $\sigma_2, \sigma_3, \beta_{32}$ and β_{21} are calculated by solving the system of equations (5.28),(5.31),(5.32). Substituting σ_1 and σ_2

from (5.28) in Eq.(5.32), we obtain

$$\frac{1}{n_{0,0}+n_{0,1}+n_{0,2}} \sum_{i\in D_0^{F_i\neq 3}} (x_3^{(i)} - x_1^{(i)}(\Psi-\beta_{21}\beta_{32}) - x_2^{(i)}\beta_{32})^2 \sum_{i\in D_0^{F_i\neq 2}} \left(x_2^{(i)} - \beta_{21}x_1^{(i)}\right)x_1^{(i)}$$
$$+ \frac{1}{n_{0,0}+n_{0,1}+n_{0,3}} \sum_{i\in D_0^{F_i\neq 2}} (x_2^{(i)} - \beta_{21}x_1^{(i)})^2 \sum_{i\in D_0^{F_i\neq 3}} \left(x_3^{(i)} - (\Psi-\beta_{21}\beta_{32})x_1^{(i)} - \beta_{32}x_2^{(i)}\right)\beta_{32}x_1^{(i)} = 0 \tag{5.33}$$

Rearrange (5.31) for β_{32},we obtain

$$0 = \sum_{i\in D_0^{F_i\neq 3}} \left(x_3^{(i)} - (\Psi-\beta_{21}\beta_{32})x_1^{(i)} - \beta_{32}x_2^{(i)}\right)\left(\beta_{21}x_1^{(i)} - x_2^{(i)}\right)$$
$$0 = \beta_{21}\hat{\Sigma}_{13}^{F\neq 3} - \hat{\Sigma}_{23}^{F\neq 3} - \Psi(\beta_{21}\hat{\Sigma}_{11}^{F\neq 3} - \hat{\Sigma}_{12}^{F\neq 3}) + \beta_{32}(\beta_{21}^2\hat{\Sigma}_{11}^{F\neq 3} - 2\beta_{21}\hat{\Sigma}_{12}^{F\neq 3} + \hat{\Sigma}_{22}^{F\neq 3})$$
$$\beta_{32}^*(\beta_{21}) := \beta_{32} = \frac{\Psi(\beta_{21}\hat{\Sigma}_{11}^{F\neq 3} - \hat{\Sigma}_{12}^{F\neq 3}) - \beta_{21}\hat{\Sigma}_{13}^{F\neq 3} + \hat{\Sigma}_{23}^{F\neq 3}}{\beta_{21}^2\hat{\Sigma}_{11}^{F\neq 3} - 2\beta_{21}\hat{\Sigma}_{12}^{F\neq 3} + \hat{\Sigma}_{22}^{F\neq 3}} \tag{5.34}$$

Substituting β_{32} from (5.34) in Eq.(5.33), we obtain a higher degree polynomial equation

$$P(\beta_{21}) = \frac{1}{n_{0,0}+n_{0,1}+n_{0,2}} \sum_{i\in D_0^{F_i\neq 3}} (x_3^{(i)} - x_1^{(i)}(\Psi-\beta_{21}\beta_{32}^*(\beta_{21})) - x_2^{(i)}\beta_{32}^*(\beta_{21}))^2 \sum_{i\in D_0^{F_i\neq 2}} \left(x_2^{(i)} - \beta_{21}x_1^{(i)}\right)x_1^{(i)}$$
$$+ \frac{1}{n_{0,0}+n_{0,1}+n_{0,3}} \sum_{i\in D_0^{F_i\neq 2}} (x_2^{(i)} - \beta_{21}x_1^{(i)})^2 \sum_{i\in D_0^{F_i\neq 3}} \left(x_3^{(i)} - (\Psi-\beta_{21}\beta_{32}^*(\beta_{21}))x_1^{(i)} - \beta_{32}^*(\beta_{21})x_2^{(i)}\right)\beta_{32}^*(\beta_{21})x_1^{(i}$$

β_{21}^* denotes the solution of the polynomial equation $P(\beta_{21}) = 0$. Solving this equation, that is, finding the roots of the polynomial, is achieved by using a numerical method(e.g. **POLYROOT** in R). Once β_{21}^* is obtained by solving the polynomial equation $P(\beta_{21}) = 0$, the other parameters such as β_{32}, σ_2 and σ_3 can be easily found by substituting the solution β_{21}^* in (5.34),(5.28).

Previously, we assumed $\Psi \neq 0$ and calculated the profile likelihood function. Now, we assume $\Psi = 0$ and compute the profile likelihood functions of each case. For the possible causal effects in the model (M3.1), it is straightforward to see

$$\mathscr{C}(1\to 2) = \begin{cases} \beta_{21} & \text{in} \quad (M3.1) \\ \beta_{23}\beta_{31}+\beta_{21} & \text{in} \quad (M3.2) \\ 0 & \text{in} \quad (M3.3) \\ \beta_{21} & \text{in} \quad (M3.4) \\ 0 & \text{in} \quad (M3.5) \\ 0 & \text{in} \quad (M3.6) \end{cases} \tag{5.35}$$

and

$$\mathscr{C}(2\to 3) = \begin{cases} \beta_{32} & \text{in} \quad (M3.1) \\ 0 & \text{in} \quad (M3.2) \\ \beta_{31}\beta_{12}+\beta_{32} & \text{in} \quad (M3.3) \\ 0 & \text{in} \quad (M3.4) \\ \beta_{32} & \text{in} \quad (M3.5) \\ 0 & \text{in} \quad (M3.6) \end{cases} \tag{5.36}$$

and

$$\mathscr{C}(1\to 3)=\begin{cases} \beta_{32}\beta_{21}+\beta_{31} & \text{in} \quad (M3.1)\\ \beta_{31} & \text{in} \quad (M3.2)\\ \beta_{31} & \text{in} \quad (M3.3)\\ 0 & \text{in} \quad (M3.4)\\ 0 & \text{in} \quad (M3.5)\\ 0 & \text{in} \quad (M3.6). \end{cases} \tag{5.37}$$

Case I : Suppose that $\mathscr{C}(1\to 2)=0$. According to (5.35) and Table 5.1, the data came from one of the models (M3.1), (M3.3). However, (M3.1) is a special case of (M3.3). Thus, the profile likelihood function for $\mathscr{C}(1\to 2)=\Psi=0$ is the maximum of the likelihood function of the model (M3.3), since all parameter involved in (M3.3) are independent of the constraint $\mathscr{C}(1\to 2)$. Hence, the profile likelihood function is

$$\begin{aligned} l_{\dagger}^{(0)}(\Psi) =& -\frac{n_{0,0}+n_{0,1}+n_{0,3}}{2}\log\left(2\pi\hat{\Sigma}_{0,22}^{F\neq 2}\right)-\frac{n_{0,0}+n_{0,1}+n_{0,3}}{2}\\ &-\frac{n_{0,0}+n_{0,2}+n_{0,3}}{2}\log\left(2\pi(\hat{\beta}_{12}^2\hat{\Sigma}_{0,22}^{F\neq 1}-2\hat{\beta}_{12}\hat{\Sigma}_{0,12}^{F\neq 1}+\hat{\Sigma}_{0,11}^{F\neq 1})\right)-\frac{n_{0,0}+n_{0,2}+n_{0,3}}{2}\\ &-\frac{n_{0,0}+n_{0,1}+n_{0,2}}{2}\log\Big(2\pi(\hat{\Sigma}_{0,33}^{F\neq 3}+\hat{\Sigma}_{0,22}^{F\neq 3}\hat{\beta}_{32}^2+\hat{\Sigma}_{0,11}^{F\neq 3}\hat{\beta}_{31}^2\\ &\qquad +2\hat{\Sigma}_{0,12}^{F\neq 3}\hat{\beta}_{31}\hat{\beta}_{32}-2\hat{\Sigma}_{0,23}^{F\neq 3}\hat{\beta}_{32}-2\hat{\Sigma}_{0,13}^{F\neq 3}\hat{\beta}_{31})\Big)-\frac{n_{0,0}+n_{0,1}+n_{0,2}}{2} \end{aligned}$$

where

$$\begin{pmatrix}\hat{\beta}_{31}\\ \hat{\beta}_{32}\end{pmatrix}=(X_{12,D_0^{F\neq 3}}^T X_{12,D_0^{F\neq 3}})^{-1}X_{12,D_0^{F\neq 3}}^T X_{3,D_0^{F\neq 3}}$$

and

$$\begin{pmatrix}\hat{\beta}_{12}\end{pmatrix}=(X_{2,D_0^{F\neq 1}}^T X_{2,D_0^{F\neq 1}})^{-1}X_{2,D_0^{F\neq 1}}^T X_{1,D_0^{F\neq 1}}.$$

Case II : Now, we assume $\mathscr{C}(2\to 3)=0$. According to (5.36) and Table 5.1, the data came from one of the models (M3.1), (M3.2). Analogously in Case I, (M3.1) is a special case of (M3.2). Therefore the profile likelihood function is given as

$$\begin{aligned} l_{\dagger}^{(0)}(\Psi) =& -\frac{n_{0,0}+n_{0,2}+n_{0,3}}{2}\log\left(2\pi\hat{\Sigma}_{0,11}^{F\neq 1}\right)-\frac{n_{0,0}+n_{0,2}+n_{0,3}}{2}\\ &-\frac{n_{0,0}+n_{0,1}+n_{0,2}}{2}\log\left(2\pi(\hat{\beta}_{31}^2\hat{\Sigma}_{0,11}^{F\neq 3}-2\hat{\beta}_{31}\hat{\Sigma}_{0,13}^{F\neq 3}+\hat{\Sigma}_{0,33}^{F\neq 3})\right)-\frac{n_{0,0}+n_{0,1}+n_{0,2}}{2}\\ &-\frac{n_{0,0}+n_{0,1}+n_{0,3}}{2}\log\Big(2\pi(\hat{\Sigma}_{0,22}^{F\neq 2}+\hat{\Sigma}_{0,33}^{F\neq 2}\hat{\beta}_{23}^2+\hat{\Sigma}_{0,11}^{F\neq 2}\hat{\beta}_{21}^2\\ &\qquad +2\hat{\Sigma}_{0,13}^{F\neq 2}\hat{\beta}_{21}\hat{\beta}_{23}-2\hat{\Sigma}_{0,23}^{F\neq 2}\hat{\beta}_{23}-2\hat{\Sigma}_{0,12}^{F\neq 2}\hat{\beta}_{21})\Big)-\frac{n_{0,0}+n_{0,1}+n_{0,3}}{2} \end{aligned}$$

where

$$\begin{pmatrix}\hat{\beta}_{21}\\ \hat{\beta}_{23}\end{pmatrix}=(X_{13,D_0^{F\neq 2}}^T X_{13,D_0^{F\neq 2}})^{-1}X_{13,D_0^{F\neq 2}}^T X_{2,D_0^{F\neq 2}}$$

and

$$\begin{pmatrix} \hat{\beta}_{31} \end{pmatrix} = (X^T_{1,D_0^{F\neq 3}} X_{1,D_0^{F\neq 3}})^{-1} X^T_{1,D_0^{F\neq 3}} X_{3,D_0^{F\neq 3}}.$$

Case III : Assuming $\mathscr{C}(1 \to 3) = 0$, according to (5.36) and Table 5.1 the data came from the model (M3.1), (M3.4), (M3.5) or (M3.6). Since the likelihood functions of the models do not depend on β_{32}, β_{21}, and β_{31}, the profile likelihood function for $\mathscr{C}(1 \to 3) = \Psi$ is the maximum of the likelihood functions of the models (M3.4), (M3.5) or (M3.6). The maximum of likelihood functions of the models are

$$\begin{aligned} l^{(0)}_{M3.4} = &- \frac{n_{0.0} + n_{0,1} + n_{0,2}}{2} \log\left(2\pi \hat{\Sigma}^{F\neq 3}_{0,33}\right) - \frac{n_{0.0} + n_{0,1} + n_{0,2}}{2} \\ &- \frac{n_{0.0} + n_{0,2} + n_{0,3}}{2} \log\left(2\pi (\hat{\beta}^2_{13} \hat{\Sigma}^{F\neq 1}_{0,33} - 2\hat{\beta}_{13} \hat{\Sigma}^{F\neq 1}_{0,13} + \hat{\Sigma}^{F\neq 1}_{0,11})\right) - \frac{n_{0.0} + n_{0,2} + n_{0,3}}{2} \\ &- \frac{n_{0.0} + n_{0,1} + n_{0,3}}{2} \log\Big(2\pi (\hat{\Sigma}^{F\neq 2}_{0,22} + \hat{\Sigma}^{F\neq 2}_{0,33} \hat{\beta}^2_{23} + \hat{\Sigma}^{F\neq 2}_{0,11} \hat{\beta}^2_{21} \\ &\qquad + 2\hat{\Sigma}^{F\neq 2}_{0,13} \hat{\beta}_{21} \hat{\beta}_{23} - 2\hat{\Sigma}^{F\neq 2}_{0,23} \hat{\beta}_{23} - 2\hat{\Sigma}^{F\neq 2}_{0,12} \hat{\beta}_{21})\Big) - \frac{n_{0.0} + n_{0,1} + n_{0,3}}{2} \end{aligned}$$

where

$$\begin{pmatrix} \hat{\beta}_{21} \\ \hat{\beta}_{23} \end{pmatrix} = (X^T_{13,D_0^{F\neq 2}} X_{13,D_0^{F\neq 2}})^{-1} X^T_{13,D_0^{F\neq 2}} X_{2,D_0^{F\neq 2}}$$

and

$$\begin{pmatrix} \hat{\beta}_{13} \end{pmatrix} = (X^T_{3,D_0^{F\neq 1}} X_{3,D_0^{F\neq 1}})^{-1} X^T_{3,D_0^{F\neq 1}} X_{1,D_0^{F\neq 1}}$$

$$\begin{aligned} l^{(0)}_{M3.5} = &- \frac{n_{0.0} + n_{0,1} + n_{0,3}}{2} \log\left(2\pi \hat{\Sigma}^{F\neq 2}_{0,22}\right) - \frac{n_{0.0} + n_{0,1} + n_{0,3}}{2} \\ &- \frac{n_{0.0} + n_{0,1} + n_{0,2}}{2} \log\left(2\pi (\hat{\beta}^2_{32} \hat{\Sigma}^{F\neq 3}_{0,22} - 2\hat{\beta}_{32} \hat{\Sigma}^{F\neq 3}_{0,23} + \hat{\Sigma}^{F\neq 3}_{0,33})\right) - \frac{n_{0.0} + n_{0,1} + n_{0,2}}{2} \\ &- \frac{n_{0.0} + n_{0,2} + n_{0,3}}{2} \log\Big(2\pi (\hat{\Sigma}^{F\neq 1}_{0,11} + \hat{\Sigma}^{F\neq 1}_{0,33} \hat{\beta}^2_{13} + \hat{\Sigma}^{F\neq 1}_{0,22} \hat{\beta}^2_{12} \\ &\qquad + 2\hat{\Sigma}^{F\neq 1}_{0,23} \hat{\beta}_{12} \hat{\beta}_{13} - 2\hat{\Sigma}^{F\neq 1}_{0,13} \hat{\beta}_{13} - 2\hat{\Sigma}^{F\neq 1}_{0,12} \hat{\beta}_{12})\Big) - \frac{n_{0.0} + n_{0,2} + n_{0,3}}{2} \end{aligned}$$

where

$$\begin{pmatrix} \hat{\beta}_{12} \\ \hat{\beta}_{13} \end{pmatrix} = (X^T_{23,D_0^{F\neq 1}} X_{23,D_0^{F\neq 1}})^{-1} X^T_{23,D_0^{F\neq 1}} X_{1,D_0^{F\neq 1}}$$

and

$$\begin{pmatrix} \hat{\beta}_{32} \end{pmatrix} = (X^T_{2,D_0^{F\neq 3}} X_{2,D_0^{F\neq 3}})^{-1} X^T_{2,D_0^{F\neq 3}} X_{3,D_0^{F\neq 3}}$$

$$
\begin{aligned}
l^{(0)}_{M3.6} = &-\frac{n_{0.0}+n_{0,1}+n_{0,2}}{2}\log\left(2\pi\hat{\Sigma}^{F\neq 3}_{0,33}\right)-\frac{n_{0.0}+n_{0,1}+n_{0,2}}{2}\\
&-\frac{n_{0.0}+n_{0,1}+n_{0,3}}{2}\log\left(2\pi(\hat{\beta}^2_{23}\hat{\Sigma}^{F\neq 2}_{0,33}-2\hat{\beta}_{23}\hat{\Sigma}^{F\neq 2}_{0,23}+\hat{\Sigma}^{F\neq 2}_{0,22})\right)-\frac{n_{0.0}+n_{0,1}+n_{0,3}}{2}\\
&-\frac{n_{0.0}+n_{0,2}+n_{0,3}}{2}\log\Big(2\pi(\hat{\Sigma}^{F\neq 1}_{0,11}+\hat{\Sigma}^{F\neq 1}_{0,33}\hat{\beta}^2_{13}+\hat{\Sigma}^{F\neq 1}_{0,22}\hat{\beta}^2_{12}\\
&\qquad +2\hat{\Sigma}^{F\neq 1}_{0,23}\hat{\beta}_{12}\hat{\beta}_{13}-2\hat{\Sigma}^{F\neq 1}_{0,13}\hat{\beta}_{13}-2\hat{\Sigma}^{F\neq 1}_{0,12}\hat{\beta}_{12})\Big)-\frac{n_{0.0}+n_{0,2}+n_{0,3}}{2}
\end{aligned}
$$

where

$$
\begin{pmatrix}\hat{\beta}_{12}\\ \hat{\beta}_{13}\end{pmatrix}=(X^T_{23,D_0^{F\neq 1}}X_{23,D_0^{F\neq 1}})^{-1}X^T_{23,D_0^{F\neq 1}}X_{1,D_0^{F\neq 1}}
$$

and

$$
\begin{pmatrix}\hat{\beta}_{23}\end{pmatrix}=(X^T_{3,D_0^{F\neq 2}}X_{3,D_0^{F\neq 2}})^{-1}X^T_{3,D_0^{F\neq 2}}X_{2,D_0^{F\neq 2}}.
$$

Thus, the profile likelihood functions for $\mathscr{C}(1\to 3)=0$ is

$$
l^{(0)}_{\dagger}(\Psi)=\max\{l^{(0)}_{M3.4},l^{(0)}_{M3.5},l^{(0)}_{M3.6}\}.
$$

Now suppose $\hat{\Sigma}^{F\neq j}_1$ is the maximal likelihood estimator of the covariance matrix computed based on $x_i, i\in D_1^{F\neq j}$, that is,

$$
\hat{\Sigma}^{F\neq j}_1:=\begin{pmatrix}\hat{\Sigma}^{F\neq j}_{1,11} & \hat{\Sigma}^{F\neq j}_{1,12} & \hat{\Sigma}^{F\neq j}_{1,13}\\ \hat{\Sigma}^{F\neq j}_{1,21} & \hat{\Sigma}^{F\neq j}_{1,22} & \hat{\Sigma}^{F\neq j}_{1,23}\\ \hat{\Sigma}^{F\neq j}_{1,31} & \hat{\Sigma}^{F\neq j}_{1,32} & \hat{\Sigma}^{F\neq j}_{1,33}\end{pmatrix}=\frac{1}{n_{1,0}+n_{1,k}+n_{1,m}}\sum_{i\in D_1^{F\neq j}}x_ix_i^T\quad\text{for}\quad j,k,m\in\{1,2,3\}.
$$

Moreover, the maximal likelihood estimator of $\beta_{21},\beta_{32},\beta_{31}$ based $x_i, i\in D_1^{F\neq j}$ are

$$
\begin{pmatrix}\hat{\beta}^{(1)}_{31}\\ \hat{\beta}^{(1)}_{32}\end{pmatrix}:=(X^T_{12,D_1^{F\neq 3}}X_{12,D_1^{F\neq 3}})^{-1}X^T_{12,D_1^{F\neq 3}}X_{3,D_1^{F\neq 3}}
$$

and

$$
\begin{pmatrix}\hat{\beta}^{(1)}_{23}\end{pmatrix}:=(X^T_{1,D_1^{F\neq 2}}X_{1,D_1^{F\neq 2}})^{-1}X^T_{1,D_1^{F\neq 2}}X_{2,D_1^{F\neq 2}}.
$$

The maximal likelihood estimator of σ_1,σ_2, and σ_3 are

$$
\begin{aligned}
\hat{\sigma^2}^{(1)}_1&=\hat{\Sigma}^{F\neq 1}_{1,11}\\
\hat{\sigma^2}^{(1)}_2&=\hat{\Sigma}^{F\neq 2}_{1,11}(\hat{\beta}^{(1)}_{21})^2-2\hat{\Sigma}^{F\neq 2}_{1,11}(\hat{\beta}^{(1)}_{21})+\hat{\Sigma}^{F\neq 2}_{1,22}\\
\hat{\sigma^2}^{(1)}_3&=\hat{\Sigma}^{F\neq 3}_{1,33}+\hat{\Sigma}^{F\neq 3}_{1,22}(\hat{\beta}^{(1)}_{32})^2+\hat{\Sigma}^{F\neq 3}_{1,11}(\hat{\beta}^{(1)}_{31})^2+2\hat{\Sigma}^{F\neq 3}_{1,12}\hat{\beta}^{(1)}_{32}\hat{\beta}^{(1)}_{31}-2\hat{\Sigma}^{F\neq 3}_{1,23}\hat{\beta}^{(1)}_{32}-2\hat{\Sigma}^{F\neq 3}_{1,13}\hat{\beta}^{(1)}_{31}.
\end{aligned}
$$

As in the 2-dimensional case, $l^{(0)}(\Sigma)$ denote log likelihood function of the samples $x_i \in D_0$, given that the covariance matrix of each sample is Σ. The log likelihood function given $\hat{\Sigma}_1$ is

$$\begin{aligned} l^{(0)}(\hat{\Sigma}_1) = &-\frac{n_{0,0}+n_{0,2}+n_{0,3}}{2}\log\left(2\pi\hat{\sigma^2}_1^{(1)}\right) - \frac{n_{0,0}+n_{0,2}+n_{0,3}}{2\hat{\sigma^1}_2^{(1)}}\hat{\Sigma}_{0,11}^{F\neq 1} \\ &-\frac{n_{0,0}+n_{0,1}+n_{0,3}}{2}\log\left(2\pi\hat{\sigma^2}_2^{(1)}\right) - \frac{n_{0,0}+n_{0,1}+n_{0,3}}{2\hat{\sigma^2}_2^{(1)}}((\hat{\beta}_{21}^{(1)})^2\hat{\Sigma}_{0,11}^{F\neq 2} - 2\hat{\beta}_{21}^{(1)}\hat{\Sigma}_{0,12}^{F\neq 2} + \hat{\Sigma}_{0,22}^{F\neq 2}) \\ &-\frac{n_{0,0}+n_{0,1}+n_{0,2}}{2}\log\left(2\pi\hat{\sigma^2}_3^{(1)}\right) - \frac{n_{0,0}+n_{0,1}+n_{0,2}}{2\hat{\sigma^2}_3^{(1)}}(\hat{\Sigma}_{0,33}^{F\neq 3} + \hat{\Sigma}_{0,22}^{F\neq 3}(\hat{\beta}_{32}^{(1)})^2 + \hat{\Sigma}_{0,11}^{F\neq 3}(\hat{\beta}_{31}^{(1)})^2 \\ &+ 2\hat{\Sigma}_{0,12}^{F\neq 3}\hat{\beta}_{32}^{(1)}\hat{\beta}_{31}^{(1)} - 2\hat{\Sigma}_{0,23}^{F\neq 3}\hat{\beta}_{32}^{(1)} - 2\hat{\Sigma}_{0,13}^{F\neq 3}\hat{\beta}_{31}^{(1)}) \end{aligned}$$

Now consider hypothesis testing for Case I in the model (M3.1). the problem we focus on is

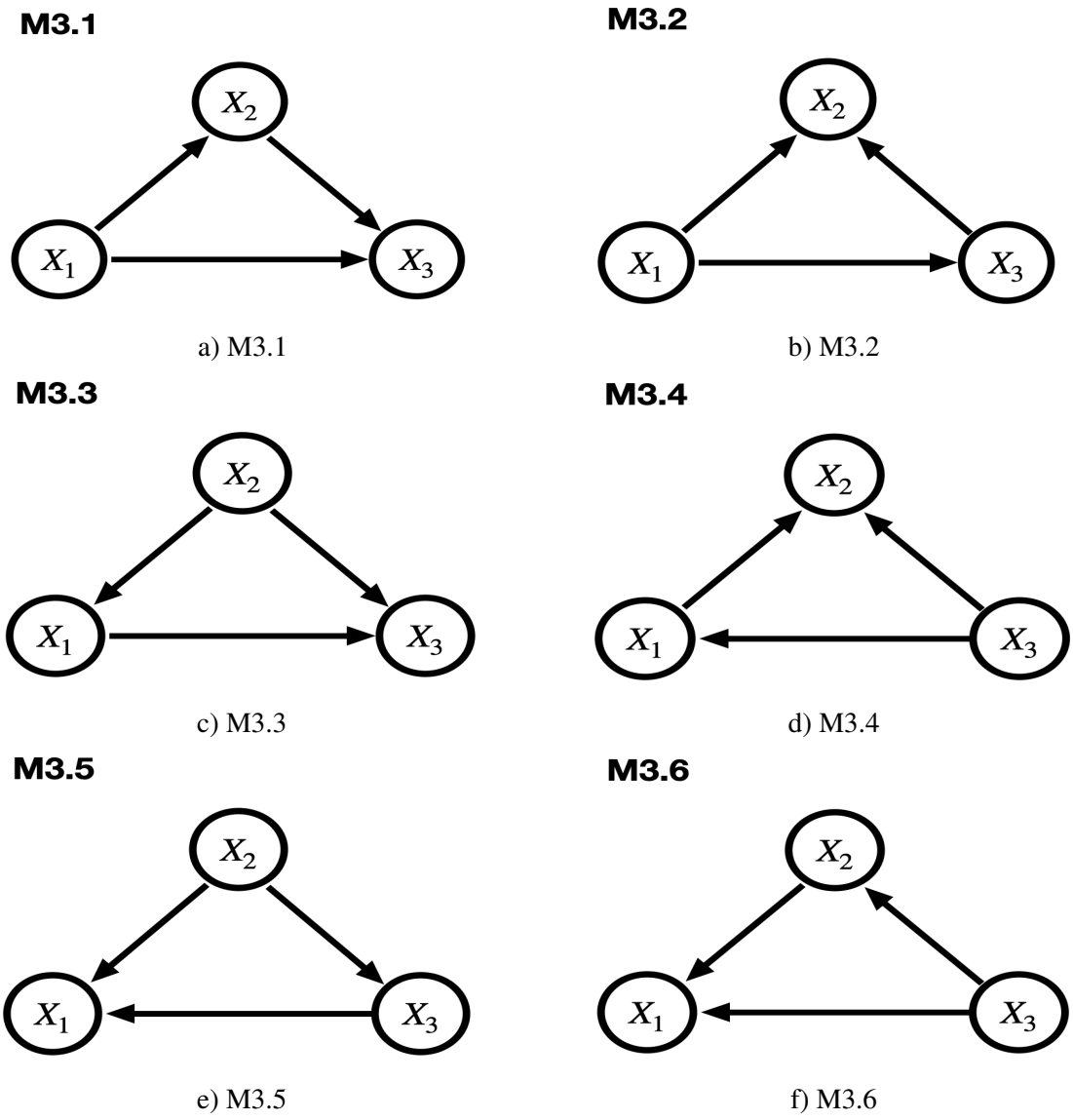
H_0 : The data follows a linear structural equation model and $\mathscr{C}(1 \to 2)$ vs.

H_1 : H_0 does not hold.

For all values of Ψ which satisfy the inequality $l^{(0)}(\hat{\Sigma}_1) - l_\dagger^{(0)}(\Psi) \leq \log(\frac{1}{\alpha})$, H_0 is accepted. However, it is not simple to solve the inequality $l^{(0)}(\hat{\Sigma}_1) - l_\dagger^{(0)}(\Psi) \leq \log(\frac{1}{\alpha})$ in order to compute analytically a confidence set for each case. Thus, we use the strategy for computing the confidence set, which is to guess a reasonable interval $[l,u]$ such that for a reasonable step size ε, all values in the range

$$\hat{\Psi} = l, l+\varepsilon, \ldots u-\varepsilon, u$$

are included in the confidence set, that is, if the inequality $l^{(0)}(\hat{\Sigma}_1) - l_\dagger^{(0)}(\hat{\Psi}) \leq \log(\frac{1}{\alpha})$ hold true. Moreover, we add the value zero if zero is accepted.



a) M3.1 b) M3.2

c) M3.3 d) M3.4

e) M3.5 f) M3.6

Figure 5.3: Graphs of the models from (M3.1) to (M3.6)

6 Simulations

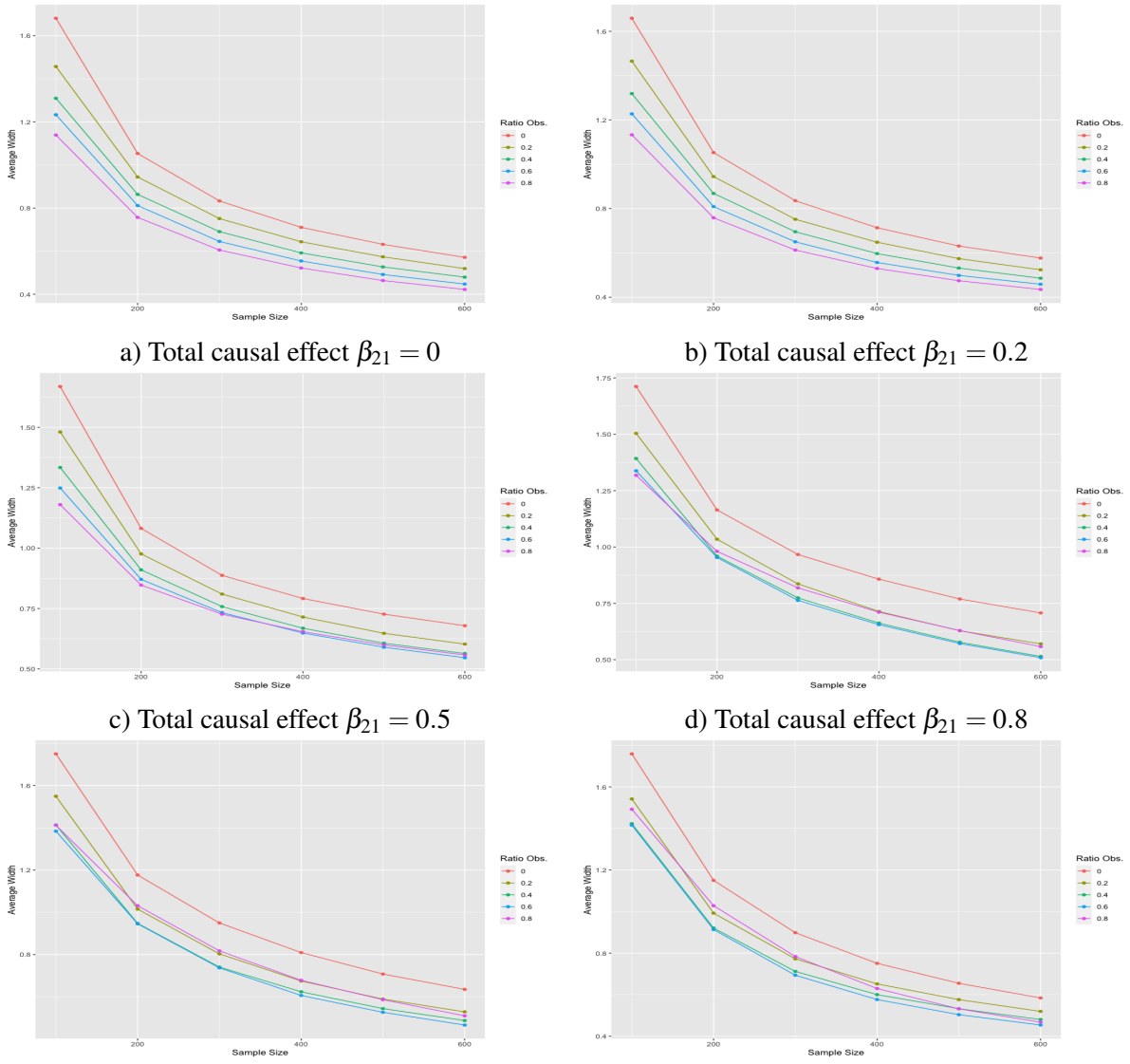
In this chapter, we investigate the results of the simulation study. We focus on the model's performance in a non-homoscedastic environment and compare the result with the results from [2]. The simulation is carried out both in two- and three-dimensional versions. In order to implement all simulations in this chapter, the programming language R is used.

6.1 Two-Dimensional Case

We analyze the performance of the bivariate case first. Our first interest is whether the algorithm introduced in the thesis performs well in the non-homoscedastic environment. If one can access only the observational data, the non-homoscedastic model is not identifiable [3]. Thus, we will show that the method using interventional data also works with the different variances of both error terms ε_1 and ε_2. Our second interest is analyzing the confidence set and its maximal width, changing the ratio of the observational data in the whole data set, that is, n_0/n. From this analysis, we want to determine if the ratio of either the observational or interventional data impacts the maximal width of the confidence interval of a valid split likelihood ratio test and its empirical coverage.
We generated pseudo-random numbers following the model $(1 \to 2)$ with standard normal errors to compare the result from [2]. The errors with different values of the variances are also considered later. We select values of $\beta_{21} \in \{0, 0.2, 0.5, 0.8, 1.0, 1.2\}$ and sample sizes $n \in \{100, 200, 300, 400, 500, 600\}$and ratios of observational data set $n_0/n \in \{0, 0.2, 0.4, 0.6, 0.8\}$ for the simulation. We generate both interventional data set under $\mathrm{do}(X_i = x_i)$ for $i = 1, 2$ with same size, that is, $n_1 = n_2$. Ten thousand independent data sets are simulated, and the confidence set is constructed for $\alpha = 0.05$. The empirical coverage probabilities for all sample sizes and values of β_{21} and all ratios of the observational data set are reported in Table 6.1. All cases achieve the desired coverage frequency of 0.95. As we know, the split likelihood ratio test is a very conservative method. From the result of the experiment in this thesis, it can also be demonstrated that the split likelihood ratio test is a conservative method since we obtained the values of 1.00 overall for the empirical coverage probabilities. We also checked the coverage probabilities of the calculated confidence set for other settings in terms of other parameters, such as different values of standard error of error terms or β_{21}. In every case, we achieved an overall coverage probability of 1. For example, the constructed confidence set has a very narrow average width if the variance of the error term ε_1 is relatively high than ε_2. Even in this case, the coverage probability is at least 0.9993.
Fig.6.1 shows the average width of the smallest interval containing the constructed confidence set against the sample size. Note that the confidence set is usually an interval. However, a confidence set can also contain a disconnected element zero. The split likelihood ratio test is a conservative method that yields a vast confidence interval [2]. For the values $\beta_{21} \in \{0, 0.2\}$, the data containing observational data in

$\mathscr{C}(1 \to 2)$	$n \setminus$ Ratio	0	0.2	0.4	0.6	0.8
$\beta_{21} = 0$	100	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00
	300	1.00	1.00	1.00	1.00	1.00
	400	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00
	600	1.00	1.00	1.00	1.00	1.00
$\beta_{21} = 0.2$	100	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00
	300	1.00	1.00	1.00	1.00	1.00
	400	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00
	600	1.00	1.00	1.00	1.00	1.00
$\beta_{21} = 0.4$	100	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00
	300	1.00	1.00	1.00	1.00	1.00
	400	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00
	600	1.00	1.00	1.00	1.00	1.00
$\beta_{21} = 0.5$	100	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00
	300	1.00	1.00	1.00	1.00	1.00
	400	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00
	600	1.00	1.00	1.00	1.00	1.00
$\beta_{21} = 0.8$	100	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00
	300	1.00	1.00	1.00	1.00	1.00
	400	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00
	600	1.00	1.00	1.00	1.00	1.00
$\beta_{21} = 1.0$	100	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00
	300	1.00	1.00	1.00	1.00	1.00
	400	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00
	600	1.00	1.00	1.00	1.00	1.00

Table 6.1: Empirical coverage of 95%-confidence intervals for the total causal effect of X_1 and X_2,selecting standard normal errors (10000 independent data sets).



a) Total causal effect $\beta_{21} = 0$

b) Total causal effect $\beta_{21} = 0.2$

c) Total causal effect $\beta_{21} = 0.5$

d) Total causal effect $\beta_{21} = 0.8$

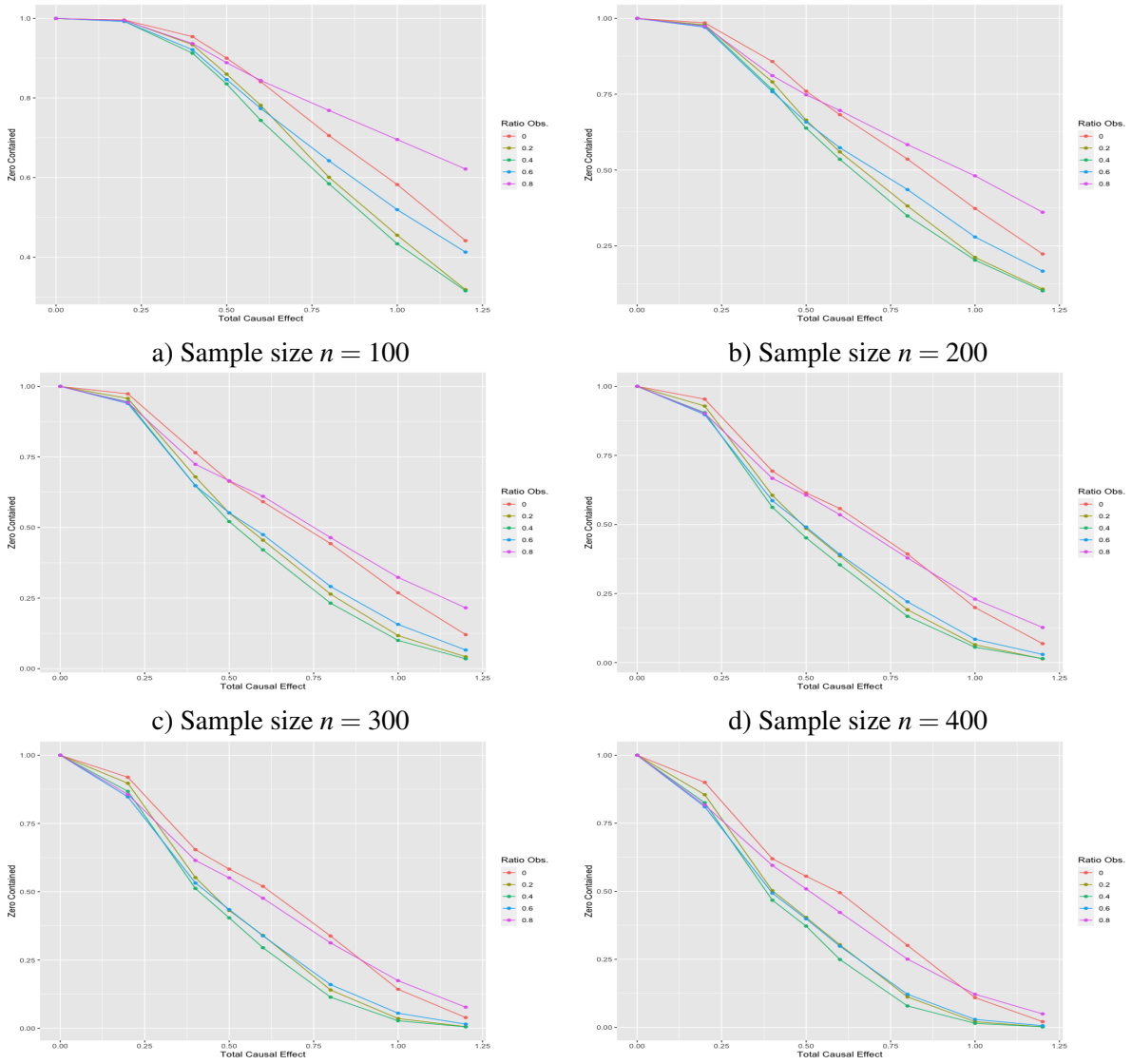
e) Total causal effect $\beta_{21} = 1.0$

f) Total causal effect $\beta_{21} = 1.2$

Figure 6.1: Average maximum width of 95%-confidence intervals for the causal effect of X_1 on X_2. $\sigma_1 = 1$, $\sigma_1 = 1$ (10000 replications).

the ratio of 0.8 yields the best result. In other words, the confidence set constructed by the data has the smallest average width. For the values $\beta_{21} \in \{0.5, 0.8\}$, the average width calculated by the data with the ratio of 0.8 becomes wider than the width computed based on the data with the ratio 0.6, increasing the sample size of the data set. For the values $\beta_{21} \in \{1, 1.2\}$, the average width of the data with the ratio of 0.6 returns least conservative result. Moreover, for some sample sizes, the average width of the confidence set of the data containing observational data in a ratio of 0.6 is even wider than the data containing observational data in a ratio of 0.2. This result changes dramatically if the variances of each error term are not equal. We will discuss it later in this chapter. In compare to the method implemented

by [2], for the value $\beta_{21} = 0.5$ the method in this paper provides more conservative result for all ratios of observational data.



a) Sample size $n = 100$

b) Sample size $n = 200$

c) Sample size $n = 300$

d) Sample size $n = 400$

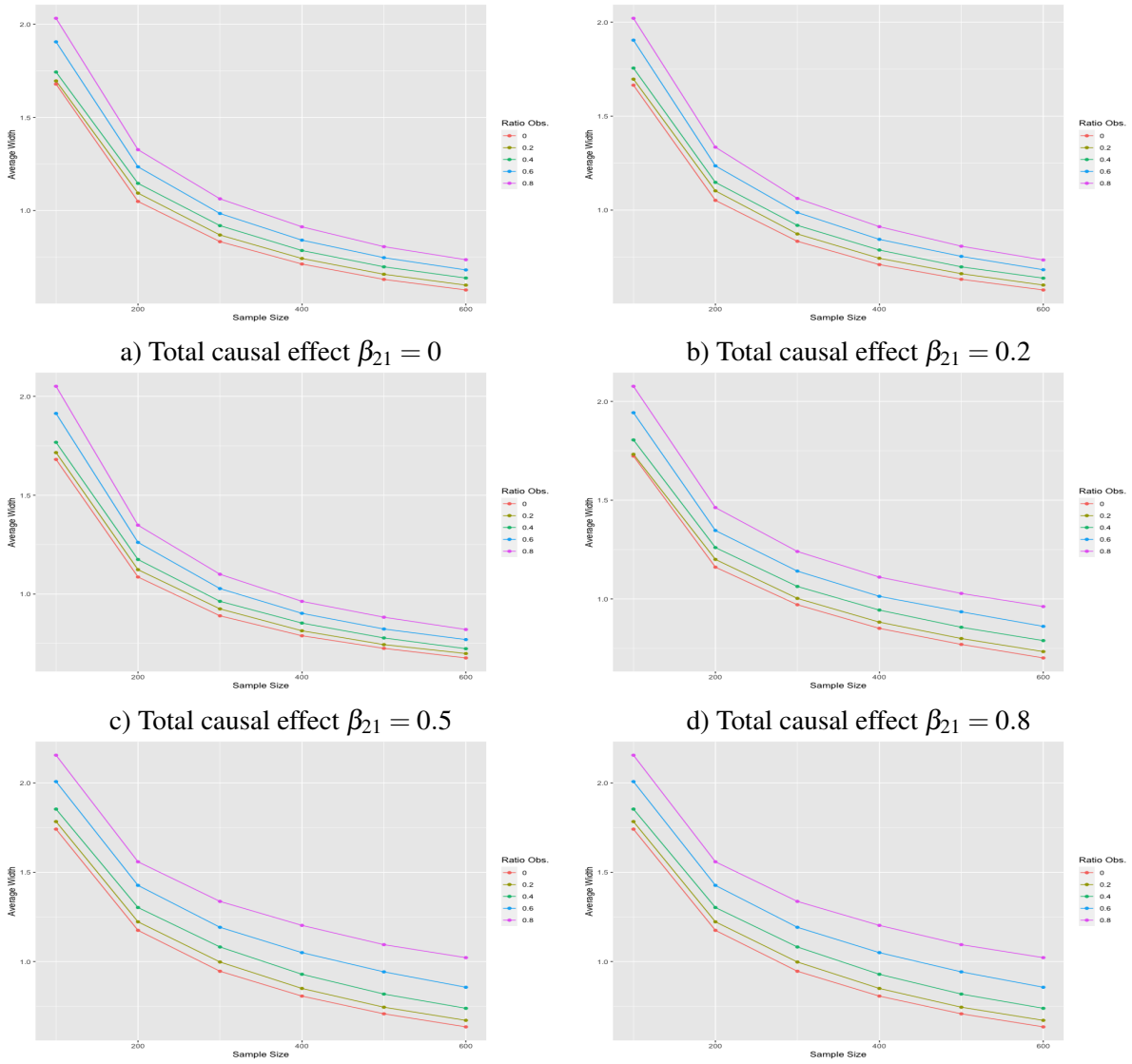
e) Sample size $n = 500$

f) Sample size $n = 600$

Figure 6.2: Percentage of times zero contained in the 95%-confidence intervals for the causal effect of X_1 on X_2 for different sample sizes. $\sigma_1 = 1$, $\sigma_2 = 1$ (10000 replications).

Fig. 6.2 shows the percentage of times zero contained against total causal effects from 0 to 1.2 for ratios of the observational data set n_0/n and sample sizes n. All cases with different ratios of observational data exclude the possibility of no causal effect with increasing sample size. The data set containing observational data in the ratio of 0.4 indicates the lowest percentage of times zero contained. Compared to the result from [2], the percentage for the total causal effect of 0.5 is higher. Almost 50 percent of replications contain zero in the calculated confidence set. This yield wider average width of the

calculated confidence set, which is observed in Fig.6.1. For all sample sizes, the confidence set of the data without observational data contains mostly the case of $\Psi = 0$ for the total causal effects around zero. In contrast, for total causal effects $\beta_{21} > 0.6$, the confidence set of the data with the ratio of 0.8 contains most often zero in itself. For sample sizes $n \geq 300$, the percentages of the data with ratios of $0, 0.8$ have a similar downward trend, and the percentages of the data with ratios of $0.2, 0.4$ and 0.6 decrease also in a similar trend. The confidence set of the data containing observational data in the ratio of 0.4 contains the value zero least often.



a) Total causal effect $\beta_{21} = 0$

b) Total causal effect $\beta_{21} = 0.2$

c) Total causal effect $\beta_{21} = 0.5$

d) Total causal effect $\beta_{21} = 0.8$

e) Total causal effect $\beta_{21} = 1.0$

f) Total causal effect $\beta_{21} = 1.2$

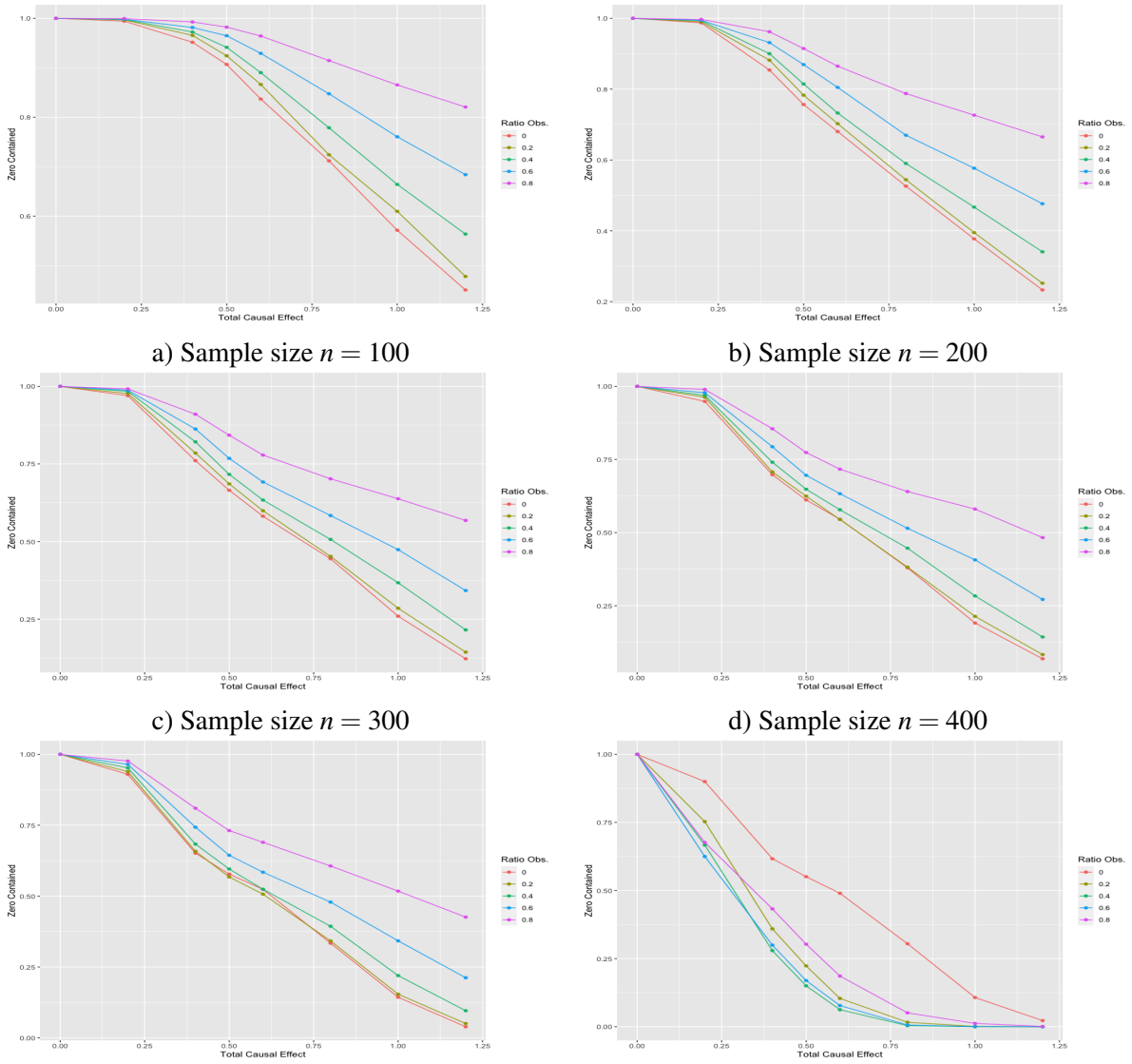
Figure 6.3: Average maximum width of 95%-confidence intervals for the causal effect of X_1 on X_2. $\sigma_1 = 0.5$, $\sigma_2 = 1$ (10000 replications).

Now, we analyze the performance of the method in a non-homoscedastic environment. We generate

samples of two random vectors X_1 and X_2 again. The random variables follow the model (M2.1), that is, X_1 causes X_2. The error term ε_2 has again the value of $\sigma_2 = 1$ for the standard error. However, we now select the error terms ε_1 such that its standard error equals 0.5. We fix again confidence level $1 - \alpha = 0.95$. As expected that the coverage probability exceeds the level, we skip to add the table of the coverage probabilities for this case of $\sigma_1 = 0.5$ and $\sigma_2 = 1$ since here we also have the exact same result as the simulation with standard normal errors. As mentioned above, the coverage probabilities are 1 for all ratios, sample sizes, and strengths of causal effects. However, we observed in this experiment remarkable differences in average widths and percentages of times zero contained from the previous experiment.

As seen in Fig.6.1, the data with a higher ratio of observational data yields a narrower width tendentially. However, if we set the variance of ε_1 to a lower value, as selected in this experiment, decreasing the ratio of observational data yields a more conservative result. Comparing Fig.6.1 and Fig.6.3 the average widths without observational data in both experiments are almost identical. Meanwhile, the average width of confidence set for $\sigma_1 = 1$ decreases and the average width of confidence set for $\sigma_1 = 0.5$ increases, if data contain more observational data. Thus we can demonstrate from the figures in Fig. 6.3 that consisting of increased ratio of observational data results in large confidence set for the value $\sigma_1 = 0.5, \sigma_2 = 1$. Fig.6.4 shows percentage of times zero contained of the simulation for the variances of $\sigma_1 = 0.5, \sigma_2 = 1$. The result of this experiment is very different from the result of the experiment with standard normal errors. The percentage of times zero contained increases if the data set contains interventional data in a bigger ratio of observational data. Every cases of data sets has a percentage of 1 for the value $\beta_{21} = 0$. Moreover, the confidence sets constructed based on the data without observational data have the lowest percentage among the rest.

The next part of this chapter analyzes the simulation result for error terms of random variables with standard errors of $\sigma_1 = 1.5$, $\sigma_2 = 1$. Again, Fig.6.5 depicts the average width of calculated confidence sets against sample sizes. The result of this simulation shows a similar downward trend of average widths as the experiment with standard normal errors. For the low values $\beta_{21} = 0$ and $\beta_{21} = 0.2$, the average widths become narrow, when the ratio of observational data go up to. Moreover, we see in Fig.6.5 that for the value $\beta_{21} = 0.5$ the data set containing observational data in the ratio of 0.8 results wider average widths than the data with the ratio of 0.6 for bigger sample sizes such as $n > 200$. However, for the values $\beta_{21} \in \{0.8, 1.0, 1.2\}$ the relative small sample sizes such as $n = 100, 200$ provide conservative confidence sets for the data with the ratio of 0.8. Except data containing observational data in the ratio of 0.8, the average widths for all total causal effects $\beta_{21} \in \{0, 0.2, 0.5, 0.8, 1, 1.2\}$ move in a similar declining trends. In addition, the data containing observational data return a better result in terms of the average width of computed confidence sets compared to the data without observational data. Fig.6.6 shows the percentage of times zero contained against total causal effects, selecting $\sigma_1 = 1.5, \sigma_2 = 1$ for the variances of the error terms. As we have seen from Fig. 6.2, the confidence set computed based on the data set containing observational data in the ratio of 0.4 contains least often zero case in itself. The percentages in this experiment have a similar downward trend as the previous experiment with the standard normal errors. However, as we can see in the figures, the differences between the percentages from each ratio are larger than in the previous experiment with the standard normal errors. From the results of the experiments in this section, we see that contribution of interventional data in the average width of a confidence interval and the percentage of times zero contained is highly dependent on the



a) Sample size $n = 100$

b) Sample size $n = 200$

c) Sample size $n = 300$

d) Sample size $n = 400$

e) Sample size $n = 500$

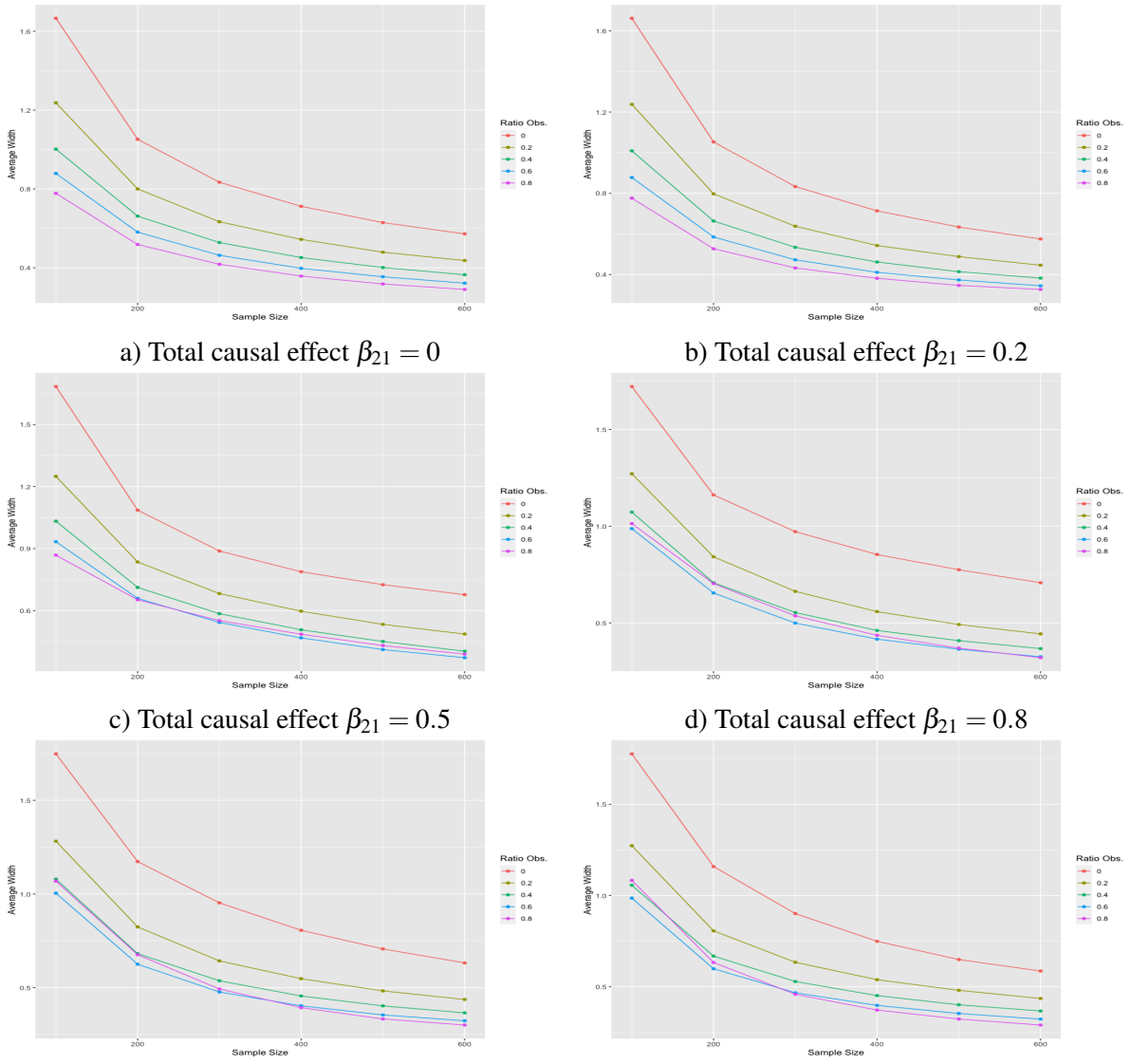
f) Sample size $n = 600$

Figure 6.4: Percentage of times zero contained in the 95%-confidence intervals for the causal effect of X_1 on X_2 for different sample sizes. $\sigma_1 = 0.5$, $\sigma_1 = 1$ (10000 replications).

strength of direct causal effects among variables and variances of the errors.

6.2 Three-Dimensional Case

The analysis proceeds in this section similarly. We will generate samples of three random variables which follow the mode (M3.1). As the 2-dimensional case, our main interest is to learn the effect of interventional data and how it influences the confidence set of a valid hypothetical test. Thus we will compare the average maximal width of the confidence set of data sets containing interventional data



a) Total causal effect $\beta_{21} = 0$

b) Total causal effect $\beta_{21} = 0.2$

c) Total causal effect $\beta_{21} = 0.5$

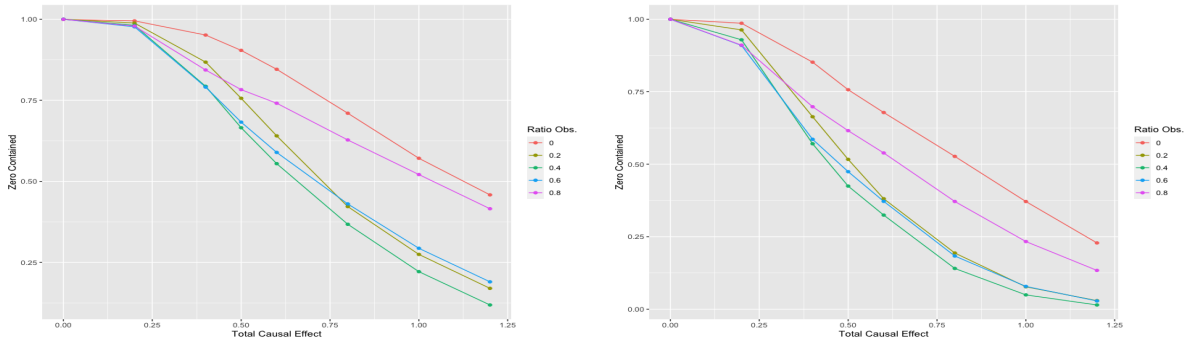
d) Total causal effect $\beta_{21} = 0.8$

e) Total causal effect $\beta_{21} = 1.0$

f) Total causal effect $\beta_{21} = 1.2$

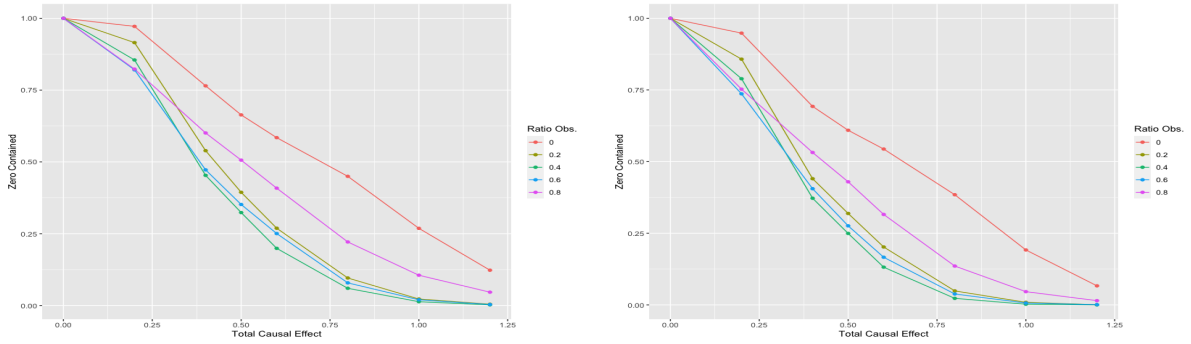
Figure 6.5: Average maximum width of 95%-confidence intervals for the causal effect of X_1 on X_2.$\sigma_1 = 1.5$, $\sigma_1 = 1$(10000 replications).

in different ratios. Again, we fix confidence level $1 - \alpha = 0.95$. Possible values Ψ of all total causal effects, that is, $\mathscr{C}(1 \to 2), \mathscr{C}(2 \to 3)$ and $\mathscr{C}(1 \to 3)$ are tested in the interval $[0,1]$ with step size 0.1. Due to the high computational expanse, we simulated 1000 independent data sets with longer step sizes than 2-dimensional cases. We used the R-package **Rsolnp** to solve the optimization problem in Case III.



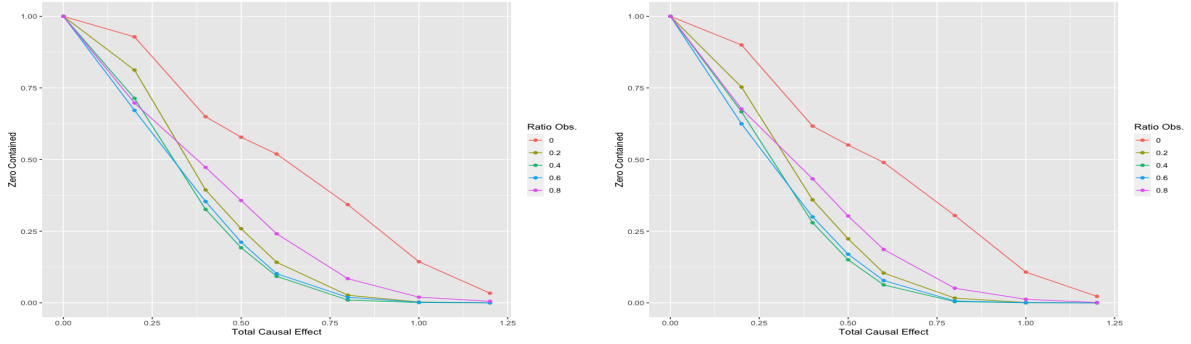
a) Sample size $n = 100$

b) Sample size $n = 200$



c) Sample size $n = 300$

d) Sample size $n = 400$



e) Sample size $n = 500$

f) Sample size $n = 600$

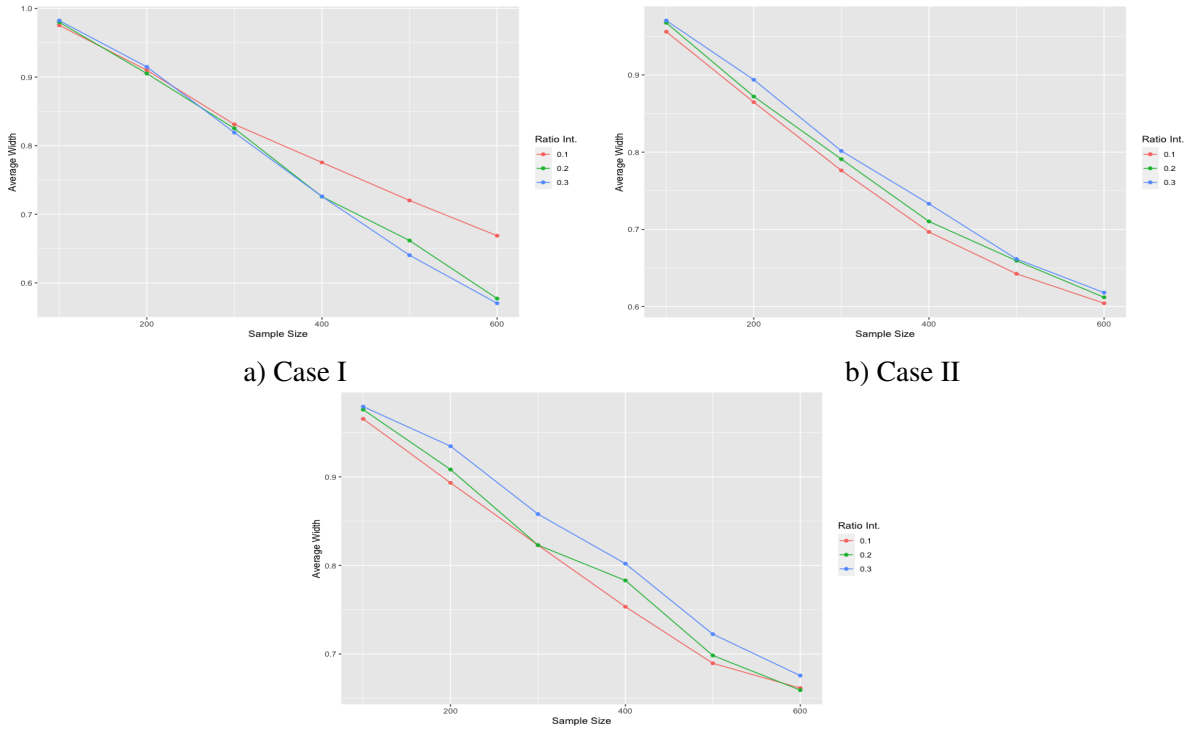
Figure 6.6: Percentage of times zero contained in the 95%-confidence intervals for the causal effect of X_1 on X_2 for different sample sizes. $\sigma_1 = 1.5$, $\sigma_1 = 1$ (10000 replications).

We consider the model (M3.1) and compute the confidence set of Cases I, II, and III

$$\begin{aligned} X_1 &:= \varepsilon_1 \\ X_2 &:= \beta_{21} X_1 + \varepsilon_2 \\ X_3 &:= \beta_{32} X_2 + \beta_{31} X_1 + \varepsilon_3 \end{aligned}$$

where $\varepsilon_1, \varepsilon_2$, and ε_3 are standard normal errors. We choose $\beta_{21} = 0.5, \beta_{32} = 0.5, \beta_{31} = 0.25$ in order to assum $\mathscr{C}(1 \to 2) = \mathscr{C}(2 \to 3) = \mathscr{C}(1 \to 3) = 0.5$. We calculated the average widths from $n \in$

$\{100,200,300,400,500,600\}$ independently simulated data sets containing interventional data in the ratios of 0.1, 0.2, 0.3. Figure 6.7 shows the average width of confidence sets calculated by simulated data sets against sample size for the three ratios of interventional data in the data set.
The coverage probabilities in all cases are 1. In three dimensional case, we can therefore demonstrate that the method yields a very conservative result. In Case II and Case III, the confidence sets based on the data set containing interventional data in a ratio of 0.1 have the smallest width among the three ratios. In Case I, the average width of the confidence set based on the data set containing interventional data in the ratio of 0.1 is distinctly wider than other data sets in a different ratio.



a) Case I

b) Case II

c) Case III

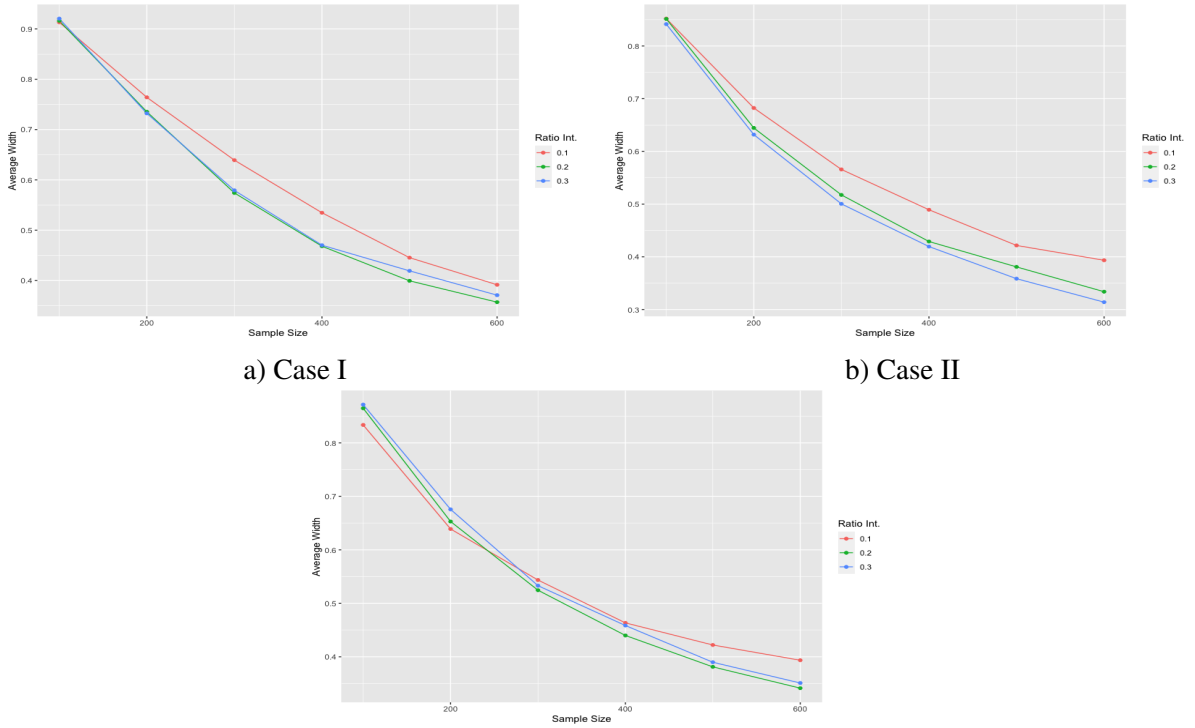
Figure 6.7: Average maximum width of 95%-confidence intervals for the causal effects. $\sigma_1 = 1$, $\sigma_2 = 1$,$\sigma_3 = 1$(1000 replications).

Secondly, We consider the model (M3.1) and compute confidence set of Case I, II, and III

$$\begin{aligned} X_1 &:= \varepsilon_1 \\ X_2 &:= \beta_{21} X_1 + \varepsilon_2 \\ X_3 &:= \beta_{32} X_2 + \beta_{31} X_1 + \varepsilon_3 \end{aligned}$$

where the errors are $\varepsilon_1 \sim \mathcal{N}(0, 0.5^2), \varepsilon_2 \sim \mathcal{N}(0, 0.7^2)$, $\varepsilon_3 \sim \mathcal{N}(0, 1^2)$. For non-homoscedastic errors, our method yields a very conservative result, as the coverage probabilities have the value 1. In Figure 6.8, the average maximal width of each Case from I to III against sample size is depicted for the three ratios of interventional data in the data set. In contrast with the previous experiment with standard normal

errors, the average width of the confidence interval based on the data set containing interventional data in the ratio of 0.1 is the widest in both Case I and Case II. In Case III, the average width of the data set with the ratio of 0.1 is narrowest for a small sample size such as $n = 100$ or 200. However, for the sample sizes $n > 300$, this average width is the widest among the three ratios.



a) Case I

b) Case II

c) Case III

Figure 6.8: Average maximum width of 95%-confidence intervals for the causal effects. $\sigma_1 = 1$, $\sigma_2 = 1$,$\sigma_3 = 1.5$(1000 replications).

7 Summary and conclusion

The main focus of this thesis is to calculate a confidence set of a valid hypothesis for total causal effects among two variables in linear structural equation models and evaluate the result of the method.
In Chapter 2, the basic mathematical backgrounds of the graphical model are introduced. A directed graph represents the causal structure underlying a model. Moreover, we consider models represented by a directed acyclic graph (DAG). At the beginning of the chapter, the graphical structure is introduced. We continue by discussing the causal model and linear structure equation. At the end of the chapter, the definition of total causal effect and identifiability of the structure of a causal model is provided.
In Chapter 3, we discuss the main idea of the linear regression model. In order to estimate the parameters which arise in distributions of the interesting variables, we apply the linear regression method. Maximizing a likelihood function in related parameters allows us to estimate the parameters. The maximal likelihood method is introduced. In the linear regression model, it is assumed that the variables are associated with each other by linear relations, and error terms are normally distributed. In our causal analysis, we also assumed that the variables in our data set follow linear equation structure models. Under conditions in some of our cases, the problems we have are equivalent to the problem of a linear regression model. Hence, the method from this chapter is applied in Chapter 5 to estimate the parameter in LSEMs. This led to calculating and determining a confidence interval for the causal effect between two variables.
Chapter 4 begins by discussing the universal inference concepts, especially the split likelihood ratio test. The solutions for a confidence interval are based on the theory of universal inference[6] and likelihood ratio tests of order constraints[12]. To determine the confidence set, we fix a significant level of $\alpha \in (0,1)$ and suppose that a total causal effect between two variables takes a fixed value. Then, we derive confidence set from a hypothetical test's acceptance region for the total causal effect.
Chapter 5 begins by applying models and methods introduced in the previous chapter. Firstly, we provide the LSEMs in d-dimensional cases by equipping them with the interventional data in the data set. The interventions to observations change the structure of LSEMs. After that, we move to specific cases. To estimate a valid interval of causal effects, we need to calculate maximal likelihood estimators of parameters in LSEMs. Maximizing the log-likelihood functions of the random vectors with respect to the parameter, we obtain the maximal likelihood estimators. The calculation is simplified by accessing the interventional data since it provides access to conditional density functions. This yields that parameter estimation is simply achieved by solving a linear regression problem. In the three-dimensional case, we classify the parameter and profile likelihood function calculation into 3 cases. In the first two cases, confidence intervals are computed similarly to the two-dimensional case. However, in Case III, we use the numerical method to solve the problem we have due to the complexity of the problem.
Our main interest is to evaluate how the interventional data in generated sample impact the result of maximal average width and how often zero cases are obtained in the confidence interval. In Chapter 6, we carry out simulation experiments and show the result of simulation experiments in two- and

three-dimensional cases. Our approach yields overly conservative results with a coverage probability of the true parameter. The probabilities for all cases are equal to 1. In our experiment, we select that the error term ε_2 is standard normally distributed and fix this value for ε_2. After that, we change the value of standard error σ_1 of ε_1 and check the change of results. From the results, we demonstrate that for a low variance of the error term ε_1 such as $\sigma_1 = 0.5$, the results become more conservative, increasing the ratio of observational data. However, for increased variance such as $\sigma_1 = 1.5$, the average maximal width of the confidence interval is wider if we increase the ratio of the interventional data set. For the three-dimensional case, we carried out two simulation experiments for the values $\sigma_1 = \sigma_2 = \sigma_3 = 1$ and $\sigma_1 = 0.5, \sigma_2 = 0.7, \sigma_3 = 1$. The analysis of the coverage probability reports that the method in the three-dimensional case also yields a very conservative result. From the result of both two- and three-dimensional cases, we demonstrate that the influences of the amount of interventional data in a data set highly depend on the parameter of linear structural equation models and the variances of the error terms. Depending on the parameters in the models, more interventional data in the data set yields a more conservative result in some cases. In contrast, containing more interventional data results in a less conservative confidence set in other cases, as shown in Chapter 6.

Furthermore, the study in this thesis can be extended to higher dimensional cases. In other words, the calculation and modeling could be generalized to models of any finite dimension. Due to the simplicity of the calculation of confidence set for a split likelihood ratio test, we focused on determining confidence intervals in terms of the split likelihood ratio test. One could also study how the other methods introduced in [2] are applied to interventional data. One could select error distributions that are not normal and noise functions that are not linear.

List of Figures

List of Tables

Bibliography

[1] S. Wright. “Correlation and causation”. In: *Journal of agricultural research* 20.7 (1921), pp. 557–585.

[2] D. Strieder, T. Freidling, S. Haffner, and M. Drton. “Confidence in Causal Discovery with Linear Causal Models”. In: (2021). DOI: 10.48550/ARXIV.2106.05694. URL: https://arxiv.org/abs/2106.05694.

[3] J. Peters and P. Bühlmann. “Identifiability of Gaussian structural equation models with equal error variances”. In: *Biometrika* 101.1 (Nov. 2013), pp. 219–228. DOI: 10.1093/biomet/ast043. URL: https://doi.org/10.1093%2Fbiomet%2Fast043.

[4] S. Haffner. “Two likelihood-ratio based approaches for interval estimation of causal effects in linear structural equation models”. Masterarbeit. Garching b. München: Technische Universität München, Apr. 2021.

[5] J. Pearl. *Causality*. 2nd ed. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511803161.

[6] L. Wasserman, A. Ramdas, and S. Balakrishnan. “Universal inference”. In: *Proceedings of the National Academy of Sciences* 117.29 (July 2020), pp. 16880–16890. DOI: 10.1073/pnas.1922664117. URL: https://doi.org/10.1073%2Fpnas.1922664117.

[7] M. Drton. *Graphical Models in Statistics*. Munich, Bayern, Germany: Technical University of Munich, 2021.

[8] W. Cook. *Combinatorial optimization*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 1998. ISBN: 9780471558941. URL: http://books.google.com/books?id=jFDvAAAAMAAJ.

[9] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: The MIT Press, 2017.

[10] I. Shpitser, T. VanderWeele, and J. M. Robins. *On the Validity of Covariate Adjustment for Estimating Causal Effects*. 2012. DOI: 10.48550/ARXIV.1203.3515. URL: https://arxiv.org/abs/1203.3515.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[12] M. Silvapulle and P. Sen. *Constrained statistical inference: Inequality, order and shape restrictions*. English. 1st ed. United States of America: John Wiley Sons, 2005. ISBN: 0471208272. DOI: 10.1002/9781118165614.