National Institute of Science Education and Research

School of Mathematical Sciences

M499: Project Report

Identifiability of bivariate causal structures

Student
Tanikella Padma Ragaleena
4th year Integrated M.Sc.
School of Mathematical Sciences
NISER Bhubaneswar (HBNI)
tp.ragaleena@niser.ac.in

Supervisor
Dr. Shyamal Krishna De
Reader-F
School of Mathematical Sciences
NISER Bhubaneswar (HBNI)
sde@niser.ac.in

Spring 2020

Acknowledgements

I would like to express my sincere gratitude to my supervisor *Dr. Shyamal Krishna De* for his patience, motivation, and immense knowledge.

Contents

1 Why study causation? 2
1.1 Weekly exercise and cholesterol levels . . . 2
1.2 Chocolates and Nobel Prizes . . . 3
1.3 Randomized Control Trials . . . 4

2 Causal Models and Assumptions 6
2.1 Causal learning and causal reasoning . . . 6
2.2 Causal Models . . . 7
2.3 S.E.Ms vs causal B.Ns . . . 8
2.3.1 Our primary focus . . . 8
2.4 Assumptions for causal inference . . . 8
2.4.1 Independence of cause and mechanism . . . 8
2.4.2 DAG assumption . . . 10
2.4.3 Jointly independent noise variables . . . 10

3 Learning cause and effect from data 12
3.1 Relation between BNs and SEMs . . . 12
3.1.1 Bayesian Networks represent joint distribution . . . 12
3.2 Structural Identifiability . . . 15
3.2.1 Assumptions that provide identifiability . . . 16
3.3 Identifiability results . . . 16
3.3.1 Additive Noise Models . . . 17
3.4 Data Analysis . . . 19
3.4.1 Information geometric causal inference (IGCI) based models and Post Non Linear Models (PNL) . . . 20

4 LiNGAM 21
4.1 Identifiability of LiNGAM . . . 21
4.1.1 Hilbert Space of Random Variables . . . 25
4.2 Estimating the LiNGAM model . . . 26
4.2.1 Basic setup . . . 27
4.2.2 Likelihood of LiNGAM . . . 28
4.2.3 Independent component analysis (ICA) . . . 30
4.2.4 ICA model . . . 30
4.2.5 ICA-LiNGAM algorithm . . . 32
4.3 Data Analysis . . . 33

Bibliography 38

Chapter 1

Why study causation?

The primary references for the content of this chapter are [Pea13], [H$^+$08], and [Pet15]. The Figure 1.3 has been taken from [Mes12]. The other figures have been taken from [JPJ16] and [Pet15].

The importance of learning causality can be best understood when one goes through some examples where the traditional ideas of statistics cannot be applied to arrive at an answer. A few illustrations have been given in this chapter to signify the importance of developing causal inference along with other statistical methods.

1.1 Weekly exercise and cholesterol levels

Assume a hypothetical study that was conducted to measure weekly exercise and cholesterol in various age groups. This data was plotted in two ways. One approach was to plot the data without segregating participants by age, while the other approach was to plot the data after being segregated by age.

The inference drawn from 1.1 is that an increase in exercise leads to an increase in cholesterol levels and hence patients with the problem of high cholesterol should be recommended not to exercise. On the contrary, 1.2 concludes that though the cholesterol levels increase with age, the amount of cholesterol we have at a particular age can be minimized with exercise. So given a patient with high cholesterol, should we or should we not recommend exercise or in other words should we believe in 1.1 or 1.2 ?

To answer this question we need to know the "story" behind the data generation i.e. how the data was generated. Assume that biological evidence suggests that the total cholesterol in a person increases naturally (due to some biological process) with age regardless of the amount of exercise he/she does. In such a scenario, it is clear that we need to believe in the segregated plot and recommend exercise to patients suffering from high cholesterol. Here, the inference we have drawn about the data we observed is not based on the numerical data alone but on the data generation process i.e. what **caused** the data to make it look the way it is. This problem couldn't haven been answered without this **non-statistical** piece of evidence.

Remark 1.1. Note that there can exist evidence which supports analysis of unsegregated data over segregated data.

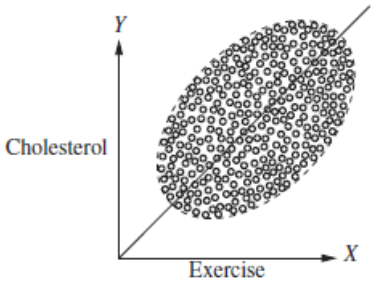

Figure 1.1: Exercise-cholesterol data without age segregation [JPJ16].

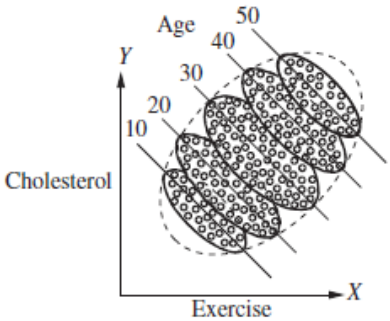

Figure 1.2: Exercise-cholesterol data with age segregation [JPJ16]

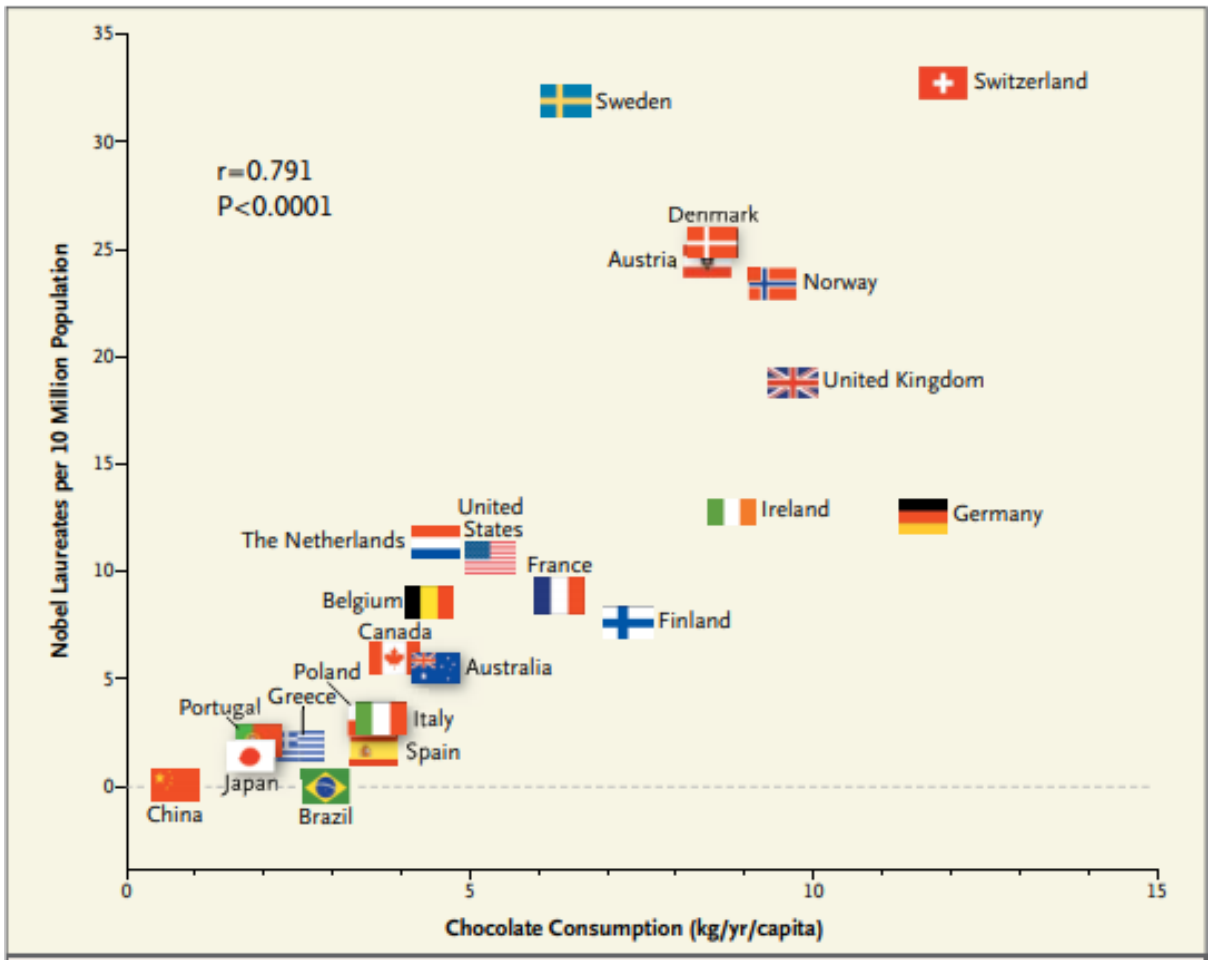

Figure 1.3: Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population [Mes12, Figure 1].

1.2 Chocolates and Nobel Prizes

Messerli (add reference) conducted a study in which a country's per capita chocolate consumption in kg per year was plotted against the number of noble prize winners per ten million people. A high correlation ($r = 0.791$)is observed between the two variables as shown in 1.3.

Figure 1.4: Claim: Eating chocolate produces Nobel prize winners.

Figure 1.5: Claim: Nobel prize winners are more likely to eat lots of chocolates.

Figure 1.6: Causal interpretations [Pet15]

Two different **causal** interpretations have been made out of the observed correlation by two different sources as shown in 1.6. (add ref) Given the correlation in 1.3 how do we arrive at the right answer? The possible interpretations of the data are as follows:

1. An increase in chocolate consumption causes an increase in number of Nobel laureates in the country.

2. Increase in Nobel laureates in a country causes an increase in chocolate consumption

3. The plot describes a correlation between variables which are **not** cause-effect pairs.

It is obvious to most of us that the correlation shown above cannot be a cause-effect relationship. Most of us might have made an intelligent guess that the observed association might be due to the existence of latent variables. But the question is whether our hypothesis can be proven mathematically.

Even though the example above gives us an intuition of whether or not the variables are causally related, there are many real life scenarios where it is not obvious to make an intelligent guess about whether or not the observed correlation implies a causal relationship. For instance, a strong correlation between exercise and cholesterol levels did not imply that an increase in exercise leads to an increase in cholesterol. Here the non-existence of causal relationship was neither obvious nor intuitive.

1.3 Randomized Control Trials

Randomized controlled trials (RCT) are studies which involve the use of a control group (hence the word "controlled") which acts as the reference group in the study. The term "randomized" implies that subjects under study are randomly assigned either to the control group or to the treatment group. The design of a RCT study is such that it constitutes the primary tool for identifying causal relationships. But such experiments are in many cases unethical, too expensive or technically impossible. Hence this lead to the development of causal discovery methods to infer causal relationships from uncontrolled data. (add ref)

An example will help us better understand the limitations of randomised control trials. It is well known that RCTs can be used to test for the significance of a new drug against a placebo or a standard treatment because, here we can **intervene** on the control group. Now consider the scenario where we have been asked to determine the cause and effect relationship between the variables "waiting time between eruptions" and the "duration of the eruption" for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA i.e. we want to determine which one of the two hypotheses is true:

1. The waiting time between the two consecutive eruptions e_1 and e_2 determines the eruption time for e_2 i.e. how long it erupts

2. How long a given geyser erupts determines how long it will take for the next eruption to occur.

Clearly, it is impossible to intervene on the eruption time or the time duration between eruptions to perform a RCT. Here, we can only be provided with observational data. So the natural question is as to whether there are methods to interpret causal relations from such purely observational or uncontrolled data.

The following questions arise from the examples given above:

1. Not all inferences can be drawn from the data alone. In order to arrive at sound conclusions, it is important to understand the data generation process. Are there statistical methods that take into consideration the data generation process to arrive at a conclusion?

2. Correlation may or may not imply causation. Then how do we mathematically prove the existence (or non-existence) of a cause-effect relationship between variables.

3. What is the alternative procedure or method if RCT cannot be employed to conclude the cause and effect relationship?

These problems mentioned above are addressed by causal inference (and not traditional statistics) where our aim is to model causal relationships found in data that we obtain by taking into consideration the data generation process. Causal inference is a paradigm that gives us the unprecedented advantage of understanding systems under various **interventions**. For eg. it lets us answer the question "What would be the eruption time if we **hypothetically** intervene on the waiting time between consecutive eruptions?".

Objective:

The focus of the report is on determining the causal structure of the system under consideration using purely observational data. The details of the assumptions we wish to follow is given in the next chapter. We will mainly focus on the two variable case, though some of the proofs are for the general n variable case.

Chapter 2

Causal Models and Assumptions

The content of this chapter has been taken from [P+17], [Pea13], [Pet12], [Nea03], [Pea] and [dW].

As mentioned in the previous chapter, causal inference deals with various kinds of problems related to cause and effect relationships like "What would be the outcome if I (hypothetically) intervene on a system?", "What would have been the outcome had I used a different approach?" , "Which variables are causally effecting a given variable?" etc.

2.1 Causal learning and causal reasoning

Understanding the relationship between causal inference and causal learning is made easy if one understand the relation between probability theory and statistical reasoning. Probability theory is the branch of mathematics which allows us to reason about various outcomes that we observe by the end of a random experiment, provided that we are given a mathematical structure *beforehand*. Structural learning on the other hand deals with the inverse problem of identifying the properties of the *unknown* underlying mathematical structure given the outcomes of the random experiment.

Similarly, causal reasoning helps to explain the outcomes, observations and interventions of a random experiment given the underlying mathematical structure (or the **causal model**). Conversely, causal learning is its inverse problem which pertains to predicting the underlying unknown model given the data from various observational and interventional studies. Causal leaning subsumes statistical learning which makes the former harder than the latter. The complexity of statistical learning stems from the fact that we are trying to solve an inverse problem based on empirical data alone. A finite set of observations never has all the information about the underlying mathematical structure. This makes the problems of statistical learning **ill-posed**. Along with the ill-posedness of statistical learning, causal learning has an additional layer of ill-posedness as we are usually unable to determine the causal structure of a system even after being provided with complete information about the observational distribution. The ideas discussed in this section can be summarized using the following diagram

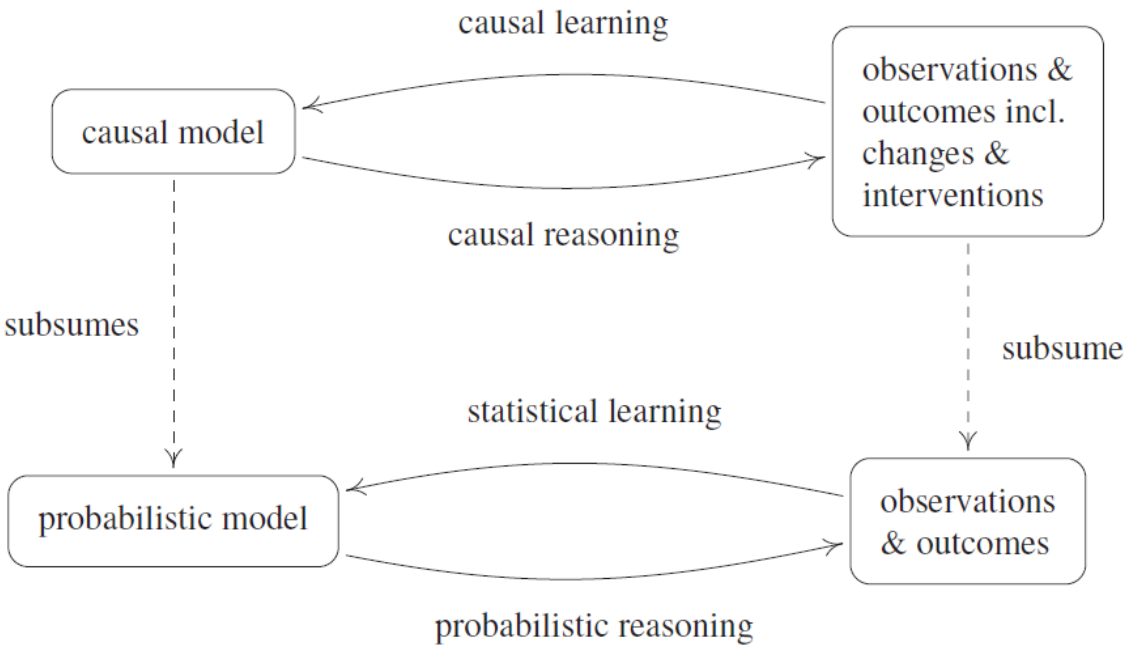

Figure 2.1: Various terminologies used in this report [P+17, Figure 1.1].

2.2 Causal Models

There are two ways of giving a causal structure to a system we want to study. One choice is to define a structural equation model (SEM) while the other to to define a causal Bayesian network. While both the models mentioned above are probabilistic graphical models which are used to model causality, there is a fundamental difference in how each model assumes the way nature's laws work. In order to understand the philosophy behind each model, it is important that we know how each models is defined.

Definition 2.1. Structural Equation Models (S.E.M)
A structure causal model is a system of equations defined as follows:

$$X_i = f_i(PA_i, \epsilon_i) \text{ such that } \epsilon_i \perp PA_i \text{ for all } i \in \{1, 2, \ldots, n\} \tag{2.1}$$

Where $X = \{X_1, X_2, \ldots, X_n\}$ represents a set of random variables. PA_i is called the parent set of X_i which is defined as the subset of X which directly determines the value of X_i. ϵ_i represents noise or disturbance which is independent of the variables in the parent set.2.1 is called a **causal model** if it describes the process generating the data.

Remark 2.1. X_i and ϵ_i for all $i \in \{1, 2 \ldots, n\}$ are random variables.

The graphical model or the **causal graph** associated with the given SEM can be constructed by considering the set random variables X as set containing node and drawing directed edges from parents of X_i to X_i for all i.

Definition 2.2. Markov Condition
Let $\mathcal{P}$ be the joint distribution defined on set of variables $X = \{X_1, X_2, \ldots X_n\}$. Let $\mathcal{G} = (X, E)$ be a directed acyclic graph(DAG) with nodes from set X and edges from set E (E is a subset of $X \times X$.). Then the pair $(\mathcal{G}, \mathcal{P})$ is said to satisfy the Markov

condition if for each variable X_i in X, $\{X_i\}$ is conditionally independent of set of all its non-descendants (ND_i) given the set of all its parents (PA_i). This is denoted as:

$$X_i \perp ND_i \mid PA_i \tag{2.2}$$

Definition 2.3. Bayesian Network (B.N)
A pair $(\mathcal{G}, \mathcal{P})$ is called a Bayesian network if it satisfies the Markov condition.

Definition 2.4. Causal Bayesian Network
A causal Bayesian network is a Bayesian network which allows us to answer interventional queries. The detailed definition can be found in Causality by Judea Pearl.

2.3 S.E.Ms vs causal B.Ns

As it was mentioned briefly earlier, the two graphical models have a fundamental difference in their assumptions about how nature works. SEMs rely on the assumption that nature's laws are deterministic while error surfaces due to our ignorance of underlying boundary conditions. On the contrary, causal B.N.s assume all relationships given in definition to be inherently stochastic.

Along with a fundamental difference in the philosophy behind the models, there is also a difference in the kind of question these models are generally utilized to answer. A SEM is constructed if the objective is to identify the factors determining a given value. On the other hand, a causal B.N. is constructed when one is interested to obtain interventional probabilities.

2.3.1 Our primary focus

Though a brief introduction to causal Bayesian networks has been given, it is worth noting that understanding SEMs is the primary aim of this report. This report will mainly focus on scenarios when cause and effect (2 variable case) can be distinguished by looking at the joint distribution alone.

2.4 Assumptions for causal inference

Now that we have a basic idea about what an SEM is, it is good to pause for a while so that we can better understand the assumptions that we have seen so far.

2.4.1 Independence of cause and mechanism

This principle of independence between cause and mechanism is better explained with the help of an example. Consider a hypothetical position where have to determine the causal structure of the altitude(A)-temperature(T) relationship.

The first idea we would get is to try and identify the **effect of interventions**. Consider a hypothetical world where we are able to intervene on the altitude values (A) i.e. we can alter the value of A so as to observe how T changes. Suppose that this lead to a drop in temperature of the place. Keeping this result in mind, we apply a second intervention where we change the value of T to determine its effect on A. If we observe

no change in altitude values after the second intervention, then we can conclude that altitude determines the temperature of the place.

However, how is it that the description of such an intervention is considered reasonable given that it is not always possible to intervene on real life systems. This intervention is reasonable in spite of the impossibility to intervene at all times because we assume that changing the altitude does not change the physical mechanism which generates the temperature output for a given place. To understand the physical process mentioned just now, suppose for simplicity that temperature of a place is determined by an exhaustive set containing A and 2 other terms (K, L). So nature takes as input the instantaneous values of A, K, L, undergoes a natural process (like a function taking inputs) and finally gives the output which is the instantaneous temperature of the place. So our assumption states that even if we intervene on A (which determines temperature), the natural process (which takes inputs to give out temperature) generating the temperature remains the same. But note that intervening on A changes the value T if A causes T

Remark 2.2. If altitude determines temperature and not vice-versa is true, then the corresponding SEM would look as follows:

$$A := \epsilon_1$$
$$T := f(A, \epsilon_2)$$

By the definition of causal graph given earlier, the graph corresponding to the SEM would be:

$$A \longrightarrow T$$

We can formalize our intuition in the form of two postulates given below. If the statement $A \longrightarrow T$ is true, then

1. it is **possible to intervene** on A (i.e. change the distribution of A in the SEM) without changing the natural process f that generates T given the value of A (f(t|a)).

2. Cause and the mechanism leading to the effect are **autonomous or invariant mechanisms** in the world

The principle of cause and mechanism is for the case of two variables (one cause and the other effect). A more general version of this principle for n variables, which is mentioned below, is called as the *principle of independent mechanisms.*

Definition 2.5. Principle of independent mechanisms
The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.
In the probabilistic case, this means that the conditional distribution of each variable given its causes does not inform or influence the other conditional distributions. In case we have only two variables, this reduces to an independence between the cause distribution and the mechanism producing they effect distribution.

2.4.2 DAG assumption

A crucial assumption made through out the report is the acyclicity of the graphs corresponding to SEMs we come across. This is because the assumption of acyclicity gives us an opportunity to simplify our system using the following property that all DAGs satisfy.

Proposition 2.1 (Existence of a "source" node)**.** *Every $\mathcal{G}$ which is a DAG has at least one node with no incoming edges.*

Proof. On the contrary, let us assume that every node of a given DAG has at least one incoming edge into every node of the graph. Pick any vertex V. Hypothesis claims that there exists V_{-1} such that $V_{-1} \longrightarrow V$ is a directed edge of the graph.Continuing this process generates a sequence of nodes enumerated as $V_{-2}, V_{-3} \dots V_{-k} \dots$. Due to the finite number of nodes, there exists V_{-j} which we come across for the first time such that $V_{-j} = V$. Hence we have:

$$\cdots \longrightarrow V_{-j} = V \longrightarrow V_{-(j-1)} \longrightarrow \cdots \longrightarrow V_{-2} \longrightarrow V_{-1} \longrightarrow V$$

Above sequence produces a cycle which leads to a contradiction. □

Remark 2.3. The nodes of a DAG with no incoming edges are called "source nodes" due to the reason explained below.

Example 2.1. Consider the following SEM with mutually independent noise distributions ϵ_i:

$$\begin{aligned} X_1 &:= \epsilon_1 \text{ where } \epsilon_1 \sim \mathcal{N}(0,1) \\ X_2 &:= 5X_1 - \epsilon_2 \text{ where } \epsilon_2 \sim \mathcal{N}(0,1) \\ X_3 &:= 2X_1 + \epsilon_3 \text{ where } \epsilon_3 \sim \mathcal{N}(1,1) \end{aligned}$$

The corresponding graph would be:

$$X_3 \longleftarrow X_1 \longrightarrow X_2 \tag{2.3}$$

(2.3) is a DAG with the source node X_1. This node acts as a source in the sense that we can start by sampling a point from the distribution ϵ_1 to obtain the realization for X_1. This realized value can be used to find the realization of the random variables X_2 and X_3 after sampling points from their respective noises. Repeating this procedure several times will give a large sample of (X_1, X_2, X_3) values from which we can procure a joint density for the variables X_1, X_2, X_3

Hence the existence of a source node along with the independence of noise variables allows us to compute the joint distribution of graph nodes given the distribution of noise variables. △

2.4.3 Jointly independent noise variables

This is another assumption we make throughout the report. One advantage of this choice can be seen in Example 2.1 where we could sample values from distributions ϵ_i independently of one another due to the assumption of mutual independence of noise.

The other motivation behind the choice of mutually independent noise terms is linked to the principle of independence between cause and mechanism. To understand this

relation, the following interpretation is useful:
For a given equation $X = f(Y; \epsilon)$ from an SEM, sampling a point from ϵ makes the equation mentioned a deterministic one. Call this $X = f^{\epsilon}(Y)$. Therefore the realization of a noise ϵ can be considered as the process of choosing one of the many possible *states* f^{ϵ}.

Now, assume that there exists a directed edge from the node X_j to X_k. Then the equations corresponding to X_j and X_k look as follows:

$$X_k = f(PA_k) + \epsilon_1 \text{ where } X_j \notin PA_k$$
$$X_j = g(PA_j) + \epsilon_2 \text{ where } X_k \in PA_j$$

Let us assume, for instance that ϵ_1 and ϵ_2 are dependent on each other in such a way that identifying the state at node X_k determines the state at the other node. This would imply that knowing the state f^s at node X_k determines the state g^t at node X_j when X_k causes X_j. This violates the independence of cause and mechanism as it implies in this present example that change in distribution of X_k should be independent of the mechanism g that generates the output X_j.

To summarize, the assumption of mutually independent noise variables is made in order to avoid the situations where dependence of noise can violate the principle of independence between cause and mechanism.

Chapter 3

Learning cause and effect from data

The following chapter is based on [P+17], [Pet12], [Pea13], [Nea03], [EW08], [JSK17], [H+08], [JS10] and [JPS11].

Prior to understanding how cause and effect variables are identified, we need to familiarize ourselves with some terminology and concepts. This is the purpose of the first section of this chapter. The second section provides an outline of the state-of-the art techniques available to distinguish cause from effect in the two variable case.

3.1 Relation between BNs and SEMs

As pointed out in the previous chapter, every SEM can be associated with a causal graph where a directed edge is drawn from vertex X to vertex Y if the value of the latter is determined by the former. In the two variable case, the only possible causal graphs are:

$$X \longrightarrow Y \iff X \text{ causes } Y$$

$$X \cdot \quad \cdot Y \iff X \perp Y$$

Though the concept of visualizing how the graph looks like is not very important in the two variable case, which is our primary concern, it is clear that as the number of variables increases the complexity of the graph is bound to increase. Even though the graphical representation doesn't come into picture very often in our report, it is important to understand the relation SEMs have with graphical models. This section tries to provide a brief introduction in this direction. Few theorems have been mentioned without proofs.

3.1.1 Bayesian Networks represent joint distribution

We have already been introduced to the definition of a B.N in the previous chapter. Bayesian networks are a class of multivariate statistical models with a wide range of applications in science and technology. One important application of these networks, as pointed out earlier, is in causal learning where the causal relations are encoded in the structure or topology of the network. Bayesian networks are important because they provide us with an efficient representation of the joint distribution when the number of variables is very large to handle. In addition to that, it is an important tool used to do Bayesian inference on a large number of variables. In this section, we focus on understanding why B.Ns are considered as representatives of joint distributions under large

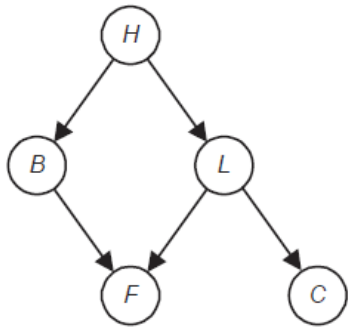

Figure 3.1: DAG illustrating Markov condition [Nea03]

instances. This is made possible since B.Ns exploit the Markov condition to represent large instances effectively. Before going further into the topic, it is good to go have a refresh on Markov condition using an example.

Example 3.1. Consider the DAG given in the figure (add ref). The following conditional independence assertions can be made with respect to the given DAG using Markov condition.

Node	PA	Conditional independence
C	{ L }	C ⊥ { H,B,F } | { L }
B	{ H }	B ⊥ { L,C } | { H }
F	{ B,L }	F ⊥ { H,C } | { B,L }
L	{ H }	L ⊥ { B } | { H }

△

Each node X of the graph can be associated with a conditional probability distribution (CPD) which is the conditional density $P(X|PA)$ where PA denotes the parents of X. If the node X has no parents, then the associated density is just the unconditional density $P(X)$.

Example 3.2. The CPDs associated with the nodes F, C, B, L, H of DAG in 3.1 are as follows:

$$F \to P(f|b,l) \quad C \to P(c|l) \quad B \to P(b|h) \quad L \to P(l|h) \quad H \to P(h)$$

△

Remark 3.1. All the graphs we consider are DAGs unless otherwise stated.

Theorem 3.1. *If $(\mathcal{G}, \mathcal{P})$ satisfies the Markov condition, then $\mathcal{P}$ is equal to the product of its conditional distributions of all nodes given values of their parents, whenever these conditional distributions exist.*

Proof. We prove for the case when $\mathcal{P}$ is discrete. Since $\mathcal{G}$ is a DAG, there exists a topological ordering of its nodes which we denote by $X_1, X_2, X_3, \ldots X_n$. We prove the theorem by induction on n.

- The statement is true for the case of $n = 1$ because X_1 is a root node due to topological ordering.
- The following is the induction hypothesis:

$$\mathcal{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathcal{P}(X_1 = x_1 | PA_1 = pa_1) \ldots \mathcal{P}(X_n = x_n | PA_n = pa_n)$$

- We want to prove for the case of $n + 1$ i.e. we want to prove that

$$\mathcal{P}(X_1 = x_1, \ldots, X_{n+1} = x_{n+1}) = \mathcal{P}(X_1 = x_1 | PA_1 = pa_1) \ldots \mathcal{P}(X_{n+1} = x_{n+1} | PA_{n+1} = pa_{n+1})$$

  1. **Case 1**: $\mathcal{P}(x_1, x_2, \ldots, x_n) = 0$
     then, $\mathcal{P}(x_1, x_2, \ldots, x_{n+1}) = P(x_{n+1} | x_1, x_2, \ldots, x_n) \times \mathcal{P}(x_1, x_2, \ldots, x_n) = 0$
     Also, by induction hypothesis, there exists a k such that $\mathcal{P}(x_k | pa_k) = 0$.
     Hence, the result holds for $n + 1$
  2. **Case 2**: $\mathcal{P}(x_1, x_2, \ldots, x_n) \neq 0$
     Then $\dfrac{\mathcal{P}(x_1, x_2, \ldots, x_{n+1})}{\mathcal{P}(x_1, x_2, \ldots, x_n)} = \mathcal{P}(x_{n+1} | x_1, x_2, \ldots, x_n)$
     Using induction hypothesis, topological ordering of nodes and the observation above can prove the desired result.

□

Remark 3.2. 1. Observe that the our proof depends on the assumption that $\mathcal{P}(pa_i) \neq 0$ for $i \in \{1, 2, \ldots, n\}$

2. The theorem 3.1 is an important one because it reduces the trouble of determining a huge number of probabilities to that of determining relatively few.

3. The theorem states that if we start with a joint distribution satisfying the Markov condition with respect to some DAG, then that joint distribution can be decomposed as the product of conditional distributions.

Theorem 3.2. *Let a DAG G be given in which each node is a random variable, and let a discrete conditional probability distribution of each node given values of its parents in $\mathcal{G}$ be specified. Then the product of these conditional distributions yields a joint probability distribution $\mathcal{P}$ of the variables, and $(\mathcal{G}, \mathcal{P})$ satisfies the Markov condition*

Proof. Order the nodes according to an ancestral ordering. Let $X_1, X_2 \ldots, X_n$ be the resultant ordering. Now define a function $\mathcal{P}(x_1, x_2 \ldots, x_n)$ as the product of all specified conditional densities $\mathcal{P}(x_i | pa_i)$. So,

$$\mathcal{P}(X_1 = x_1, \ldots, X_n = x_n) := \mathcal{P}(X_1 = x_1 | PA_1 = pa_1) \ldots \mathcal{P}(X_n = x_n | PA_n = pa_n) \quad (3.1)$$

Claim - 1: $\mathcal{P}(x_1 \ldots x_n)$ is a well defined joint density function.

$$\sum_{\forall x_1} \cdots \sum_{\forall x_n} \mathcal{P}(x_1, \ldots, x_n) = \quad (3.2)$$

$$\sum_{\forall x_1} \left[\sum_{\forall x_2} \left[\sum_{\forall x_3} \cdots \sum_{\forall x_n} [\mathcal{P}(x_n | pa_n)] \, \mathcal{P}(x_{n-1} | pa_{n-1}) \ldots \right] \mathcal{P}(x_1 | pa_1) \right] \mathcal{P}(x_1 | pa_1) \quad (3.3)$$

The claim can be proved by observing that $\sum_{\forall x_i} \mathcal{P}(x_1|pa_i) = 1 \forall i$.
Claim - 2: Markov condition is satisfied.
Order the nodes so that all non-descendants of X_k proceed X_k in the ordering. For such an ordering, we have $ND_k = \{X_1 \dots X_{k-1}\}$ and $D_k = \{X_{k+1} \dots X_n\}$ for the node X_k. (here, ND_k represents the set of non-decedents of X_k while D_k represents the set of descendants of X_k
We wish to prove the following:

$$\mathcal{P}(x_k|nd_k, pa_k) = \mathcal{P}(x_k|pa_k) \forall k \in \{1, 2 \dots, n\}$$

Since $PA_k \subset ND_k$, it is enough to prove that

$$\mathcal{P}(x_k|nd_k) = \mathcal{P}(x_k|pa_k) \forall k \in \{1, 2 \dots, n\}$$

Now, we have

$$\mathcal{P}(X_k = x_{k0}|ND_k = nd_{k0}) = \frac{\mathcal{P}(x_{k0}, n\bar{d}_{k0})}{\mathcal{P}(n\bar{d}_{k0})} \text{ where } n\bar{d}_{k0} = (x_{10}, x_{20} \dots, x_{k-1,0})$$

Careful simplification of above equation proves the statement. During simplification we use the fact that product of conditionals associated with a subset of the nodes which form a sub-graph of $\mathcal{G}$ also give a valid density. □

Remark 3.3.
1. Notice that the theorem requires the given conditional distributions to be discrete. It mostly holds for the continuous case but not always.
2. This theorem states is the opposite of the previous theorem. It states that if we start with conditional distributions, then the product of these conditionals is a joint pdf satisfying the Markov condition.

It is important to note that there exist B.Ns whose nodes are all continuous random variables (continuous Bayesian Networks) and those whose nodes consist of both continuous and discrete random variables (hybrid random variables).

3.2 Structural Identifiability

Causal learning, as mentioned earlier is an ill-posed problem at two levels. One of the issues is that looking at a joint density is not always sufficient to identify the causal structure i.e. to obtain the causal relations. But, there are some instances where the joint density $P_{x,y}$ is sufficient to tell cause from effect. This is when we say that the "structure is identifiable" from the joint distribution. The following theorem explains why causal structure is not always identifiable using the the joint density alone.

Theorem 3.3. *For every joint distribution $P_{X,Y}$ of two real valued variables X, Y, there is* at least one SEM *in each of the possible causal directions, i.e. there exist measurable functions f, g and noise variables N_1, N_2 such that:*

- $Y = f(X, N_1)$; $X \perp N_1$
- $X = f(Y, N_2)$; $Y \perp N_2$

Proof. We wish to construct an SEM for the graph $X \longrightarrow Y$ using the density $P_{X,Y}$. For that we use inverse of cumulative distribution function ($\mathcal{F}^{-1}_{Y|x}$) as shown below.

$$\text{Define } Y = f(x, n_1) := \mathcal{F}^{-1}_{Y|x}(n_1) \text{ such that } \mathcal{F}_{Y|x}(y) = \mathcal{P}(Y \leq y | X = x)$$

Here we assume N_1 to be normally distributed on $[0, 1]$ and independent of X. This gives an SEM supporting the mechanism $X \longrightarrow Y$
The same steps can be followed to obtain an SEM supporting $Y \longrightarrow X$. □

Remark 3.4. This theorem states that there exists at least one SEM that represents the joint distribution at hand but it does not claim that the SEM given above describes the causal generative process all the time. There can be multiple SEMs which describe the same joint density but not all explain the cause-effect mechanism involved.

3.2.1 Assumptions that provide identifiability

Now, our next step is to identify the conditions or assumptions under which the two variable causal structure can be recovered from the joint density. Note that these assumptions are made in addition to the SEM assumptions that we already saw. One possible approach, that we employ here, is to restrict the class of functions f and/or to restrict the class of noise distributions. It is worth mentioning that independence of noise renders causal directions identifiable only after restricting the function class. After stating these facts, it is important to notice that theorem 3.3 has no restrictions placed on its noise and function class. Therefore, SEMs in both the directions could be produced.

3.3 Identifiability results

Identifiability results are the theorems that prove that the causal direction is identifiable under a certain set of assumptions or conditions. Current section provides examples of some of such identifiability results. The state-of-the-art models for causal discovery in the bivariate case are:

- Additive Noise Models (ANMs)
- Post Non-Linear Models (PNL models)
- Information Geometric Causal Inference models (IGCI models)

Another important class of models constitutes the class of **Linear Non-Gaussian Additive Noise Models (LiNGAM)**. This is so because LiNGAM is commonly applied to data that is observed to be continuous valued. This is not necessarily because the linear model describes the process well, but because these models are well understood and easy to work with. However, LiNGAM identifies the causal structure if the data we are working with fits a LiNGAM model reasonable well. This section provides an overview of ANMs, PNLs and IGCI models. The next chapter will be devoted for LiNGAM.

3.3.1 Additive Noise Models

Additive noise models were proposed to deal with non-linear relationships. The model assumed that effect can be expressed as a functional model of the cause X and additive noise N such that the cause and additive noise are independent of each other. Mathematically, this represents the following model:

$$Y = f(X) + N \ ; \ X \perp N$$

The model is learnt by performing regression in both directions and testing the independence between the assumed cause and noise for each direction. Decision rule choose the direction with "less dependence" as the true causal direction. Given below is a step by step for model estimation.

Model Estimation

1. Test whether the variables X and Y are statistically independent.
2. If X and Y are not independent, then test whether the model $y = f(x) + n$ is consistent with the data by doing a non-linear regression of y on x. This gives us the estimate $\hat{f}$ of f. These estimates can be used to calculate the corresponding residuals $\hat{n} = y - \hat{f(x)}$. Test for independence between residuals($\hat{n}$) and x. The model can be accepted if the test supports independence between variables. If this is not the case, then the same procedure is applied to the reverse model i.e. to $x = g(y) + n'$
3. Above steps result in one of the given possible scenarios:
   - $x \perp y \implies$ we infer that a causal relation between X and Y does not exist.
   - $x \not\perp y$ and both directional models are accepted $\implies$ either model may be correct but we cannot infer it from the data.
   - $x \not\perp y$ and we are able to accept exactly one direction (and reject the other) $\implies$ the accepted model gives the the correct causal direction.
   - $x \not\perp y$ and neither direction is consistent with data $\implies$ the generating mechanism is more complex and can't be described using this model.

Remark 3.5. The above estimation method can be generalized to the case where N number of variables are involved. The optimality of the model is not claimed here.

Do ANMs admit the true causal direction?

The earlier estimation procedure assumed the true causal direction to be the unique direction for which an ANM exists. However it is natural to wonder if our data would ever admit an ANM in the wrong causal direction. The question being posed is elaborated below:
Suppose that $Y \longrightarrow X$ is the correct causal structure with respect to some non-ANM model for a given phenomenon. Then would the distribution corresponding to this causal model exhibit an additive noise model in the wrong direction i.e. from $X \longrightarrow Y$?

This question was addressed by DJanzing and BSteudel in their paper titled "Justifying additive-noise-model based causal discovery via algorithmic information theory" (add

ref). Using algorithmic information theory and kolmogorov complexity, it has been shown that the above mentioned situation is a rare scenario which requires $p(y)$ and $p(x|y)$ to be tuned in a specific way (See the result below). Fortunately this specific relation between the probabilities does not occur in our scenario due to our assumption of independence between cause and mechanism. The following untypical relation (3.4) has to be followed to obtain an ANM in the wrong direction. The relation is stated without proof because my intention is only to give the reader an understanding of how unlikely it is for ANMs to occur in the wrong causal direction as mentioned earlier.

$$\frac{\partial^2}{\partial y^2} logp(y) = -\frac{\partial^2}{\partial y^2} logp(x|y) - \frac{1}{f'(x)} \frac{\partial^2}{\partial x \partial y} logp(x|y) \tag{3.4}$$

Discrete Additive noise models

The additive models we have discussed about so far were for continuous valued random variables. Surprisingly, additive noise models can be defined on random variables which take values in a ring (like $\mathbb{Z}$, $\mathbb{Z}/n\mathbb{Z}$). ANMs defined on random variables with countable (discrete) support are called discrete additive noise models (DANMs).

DANMs over $\mathbb{Z}$ are defined in case where variables under consideration are taking integer values. For example, in the case of a variable recording the number of kids a family has. On the other hand, DANMs defined on $\mathbb{Z}/n\mathbb{Z}$ have a inherent cyclic structure like the variable "Day of the week". This is a cyclic variable with a cycle of length n. The models of the former type are called **integer models** while the latter are called **cyclic models**.

- **Integer Models:** Suppose that X and Y are random variables taking values in $\mathbb{Z}$. An additive noise model is said to exist from X to Y if there exists $f : \mathbb{Z} \longrightarrow \mathbb{Z}$ and a discrete noise N taking values in $\mathbb{Z}$ such that the joint pdf admits the ANM:

$$Y = f(X) + N \ ; \ N \perp X \text{ and } n(0) \geq n(j) \forall j$$

- **Cyclic models:** Suppose that X and Y are random variables taking values in a periodic fashion. Given a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, a function $f : \Omega \longrightarrow \mathbb{Z}/n\mathbb{Z}$ is called an **m-cyclic** random variable if $X^{-1}(k) \in \mathcal{F} \ \forall k \in \mathbb{Z}/n\mathbb{Z}$.
  Suppose that X is a m- cyclic random variable while Y is a - cyclic random variable. We say that there exists an ANM from X to Y if there is a function $f : \mathbb{Z}/m\mathbb{Z} \longrightarrow \mathbb{Z}/\mathbb{Z}$ and a -noise N such that

$$Y = f(X) + N \ ; \ N \perp X \text{ and } n(0) \geq n(j) \ \forall j$$

We earlier saw that it is very unlikely for the ANM to be reversible in the continuous case. As expected, it has been proved by JPeters, DJanzing and BScholkopf that invertibility of above mentioned causal models is also very unlikely. One can refer to their paper titled "Causal Inference on Discrete Data using Additive Noise Models" for further information. As done previously, the direction in which the causal model is obtained, is considered as the true causal direction.

3.4 Data Analysis

We will now try to identify the cause and effect relationship in a real data set by fitting an additive noise model to the data. For this we use the "Old Faithful Geyser Data" available on R. The data involves the variables "waiting time between distributions in minutes (waiting)" and "duration of a given eruption in minutes (eruption)" for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We have come across this example in the first chapter as well. There, we observed that identifying cause and effect relationship in this particular case is not very obvious. Hence we now try to apply an ANM in order to obtain the causal structure.

```
> library(MASS)
> data("faithful")
> head(faithful)
  eruptions  waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

A bunch of packages needed to be installed before using the command that fits the data with an ANM.

```
> if (!requireNamespace("BiocManager", quietly = TRUE))
   + install.packages("BiocManager")

> BiocManager::install("graph")

> if (!requireNamespace("BiocManager", quietly = TRUE))
   + install.packages("BiocManager")

> BiocManager::install("RBGL")

> install.packages("pcalg")
> library(pcalg)

> install.packages("CompareCausalNetworks")
> library(CompareCausalNetworks)

> install.packages("kernlab")
> library(kernlab)
```

After downoading the packages above, use the getParents() to fit an ANM model.

```
> getParents(faithful,method = "bivariateANM")
          eruptions waiting
eruptions     FALSE    TRUE
waiting       FALSE   FALSE
```

The output obtained is the adjacency matrix of the correponding causal graph. Here, we have a causal graph from "eruption" to "waiting". **Conclusion:** We can conclude that the eruption time of a given geyser determines the amount of time it takes for the next geyser to erupt i.e. eruption time determines the waiting time for the consecutive eruption. Therefore the causal structure is :

$$\text{eruption} \longrightarrow \text{waiting}$$

The other two classes of models are briefly explained below.

3.4.1 Information geometric causal inference (IGCI) based models and Post Non Linear Models (PNL)

A breif introduction to these models has been provided considering their importance in causal discovery.

Post Non-Linear Models

This model, just like the ANM, assumes the effect variable to vary with the cause in a non-linear fashion along with some additive internal noise. The difference lies in the assumption of an external non-linear distortion in the case of PNL models.

$$Y = f(g(x) + N) \ ; \ N \perp X \tag{3.5}$$

The assumptions involded in a PNL can be summarized as follows:

1. effects are non-linear transformations of causes with some "inner" additive noise.
2. This has to be followed by an external distortion.

IGCI models

These models are based on the hypothesis that if "X causes Y", then the marginal distribution $P(x)$ and conditional distribution $P(y|x)$ are "independent" in a particular way. This model gives an information-theoretic view of additive noise models and defines independence using "orthogonality" in information space. Hence the words "geometric and information theory " in the name of the model. The advantage this model provides is that it can infer causal directions even when the noise is not additive.

Chapter 4

LiNGAM

The main sources for this chapter are [H+08], [P+17], [Pet08], [Zit13], [Wol18], [Sti11], [AHO01], [Slo18], [Shi14], [HZSH10], and [SHHK06].

As already pointed out, "Linear Non-Gaussian Additive Models (LiNGAM)" are an important class of causal models due to the inherent ease of working with these well understood statistical models. This chapter consists of two sections. The first section deals with the invertibility of (i.e. existance of causal structures in both directions.) LiNGAM models where we find out the explicit conditions under which the causal model is identifiable. The second section discusses various methods used to estimate the LiNGAM model that fits the data.

Definition 4.1. LiNGAM As the name suggests, this model assumes the effect variable to vary linearly with the cause variable upto an additive non-Gaussian noise term.

$$Y = \alpha X + N \text{ where } N \perp X$$

Multivariate LiNGAM is defined as follows:

$$Y = \sum_{i=1}^{n} \phi_i X_i + N \ ; \ N \perp (X_1, X_2, \ldots, X_n)$$

4.1 Identifiability of LiNGAM

The following is the major theorem of this section that we wish to prove. The theorem along with its implications is first mentioned to emphasize the significance of this theorem.

Theorem 4.1. *Assume that $P_{x,y}$ admits the following linear model for the continuous random variables X, N_Y*

$$Y = \alpha X + N_Y \ ; \ N_Y \perp X$$

Then there exists a model in the opposite direction i.e. there exists $\beta \in \mathbb{R}$ and a continuous random variable N_X such that

$$X = \beta X + N_X \ ; \ N_X \perp Y$$

$$\boxed{\textit{if and only if}}$$

X, N_Y (and hence Y, N_X) are Gaussian random variables.

Before proving this theorem, it is good to appreciate how important the above theorem is. The theorem states that LiNGAM models are identifiable as long as the noise and X are not Gaussian i.e. as long as we are not following the assumptions associated with **linear regression**. Therefore, linear regression's inability to explain cause and effect relationships can be explained by the invertibility of the regression model that is mentioned in the above theorem.

Another important message the theorem conveys is that causal direction can be rendered identifiable if the noise distribution is non-gaussian. This gives a valuable property for non-gaussian noise which is generally considered undesirable. It is worth pointing out that, although the proof are given for the bivariate case, the proofs can easily be extended to multivariate case. A couple of results need to be proved before proving the main theorem. Definition of the characteristic function is also provided below which will be used in the process of proving the lemma.

Definition 4.2. The characteristic function of a random variable X is given by

$$\psi_X(t) = E[\exp(itX)]$$

It has the property that $\psi_{X+Y} = \psi_X \cdot \psi Y$ when $X \perp Y$.
If $\tilde{X} = (X_1, X_2, \ldots X_n)$, then the multivariate characteristic function is given by:

$$\psi_{\tilde{X}}(t_1, t_2, \ldots, t_n) = E\left[exp\left(i\sum_{i=1}^{n} t_k X_k\right)\right] \tag{4.1}$$

Here, $\psi_{\tilde{X}}(t_1, \ldots, t_n) = \prod_{i=1}^{n} \psi_{X_i}(t_i)$ for all $(t_1, \ldots, t_n) \in \mathbb{R}^n$.

Lemma 4.1. *Suppose that X and N are independent variables such that N is non-deterministic. Then*

$$N \not\perp (X + N) \tag{4.2}$$

Proof.

- **Case - 1:** $Var(X) < \infty$ and $Var(N) < \infty$
  $cov(N, X + N) = cov(N, X) + cov(N, N) = cov(N, N) \neq 0$ since N is non-deterministic. Therefore $N \not\perp (X + N)$

- **Case - 2:** At least one of $Var(X), Var(N)$ is not finite. We use proof by contradiction. Assume, on the contrary that $N \perp (X + N)$ .We first show the following using the fact that $X \perp N$.

  $$\psi_{N,X}(u + v, v) = \psi_{N,X+N}(u, v) = \psi_N(u + v) \cdot \psi_X(v) \ \forall \ (u, v) \in \mathbb{R}^2$$

  Now, the assumption that $N \perp (X + N)$ is used to show to show that

  $$\psi_{N,X+N}(u, v) = \psi_N(u) \cdot \psi_X(v) \cdot \psi_N(v) \ \forall \ (u, v) \in \mathbb{R}^2$$

  Since $\psi_X(0) \neq 0$ and ψ_X is continuous for any random variable X, we can say that there exists open interval $V = (-r, r) \subset \mathbb{R}$ such that $|\psi_X(v)| > 0 \ \forall \ v \in V$. So for all $v \in V$, we have

  $$\psi_N(u + v) = \psi_N(u) \cdot \psi_N(v) \tag{4.3}$$

for a given v choose $n_v \in \mathbb{R}$ such that $\left|\frac{v}{n_v}\right| \leq r$. Now, using (4.3) repeatedly, we get

$$\begin{aligned}\psi_N(u+v) &= \psi_N\left(u + (n-1)\frac{v}{n_v} + \frac{v}{n_v}\right) \\ &= \psi_N\left(u + (n-1)\frac{v}{n_v}\right) \cdot \psi_N\left(\frac{v}{n_v}\right) \\ &= \psi_N\left(u + (n-2)\frac{v}{n_v}\right) \cdot \left[\psi_N\left(\frac{v}{n_v}\right)\right]^2\end{aligned}$$

Repeating the same procedure again and again, we finally get:

$$\psi_N(u+v) = \psi_N\left(u + (n-2)\frac{v}{n_v}\right) \cdot \left[\psi_N\left(\frac{v}{n_v}\right)\right]^2$$

$$\vdots$$

$$= \psi_N(u) \cdot \left[\psi_N(\frac{v}{n_v})\right]^{n_v} = \psi_N(u) \cdot \psi_N(v) \tag{4.4}$$

Hence, $\psi_N(u+v) = \psi_N(u) \cdot \psi_N(v)$ for all $u, v \in \mathbb{R}$. This implies that $\psi_N(u) = z^u$ for some $z \in \mathbb{C}$ $\{c \in \mathbb{C} | Im(c) = 0; Re(c) < 0\}$ Now, let $z = \exp(a + ib)$. Then we can deduce that $a = 0$ since $||\psi_N||_\infty \leq 1$. It follows that

$$\psi_N(u) = \exp(ib \cdot u)$$

Due to uniqueness of characteristic function, we have $P(N = 1) = 1$. This gives a contradiction.

□

Theorem 4.2. ***[Darmois-Skitovich]*** *Let $X_1, X_2 \ldots, X_n$ be independent, non-degenerate random variables. The the two linear combinations*

$$l_1 = a_1X_1 + \cdots + a_nX_n \ ; \ a_i \neq 0 \text{ for all } i$$
$$l_2 = b_1X_1 + \cdots + b_nX_n \ ; \ b_i \neq 0 \text{ for all } i$$

are independent, then each X_i is normally distributed.

The proof of above theorem uses the following theorem:

Theorem 4.3. *Let $f_1, f_2, \ldots, f_n$ be characteristic functions which satisfy $\prod_{i=1}^n f_i^{\alpha_i}(t) = f(t)$ for some $\alpha_i > 0$ and $\forall t$ in a neighbourhood of 0. Here, f is a characteristic function of the normal distribution.*
Then, every f_i will also be a characteristic function of the normal distribution.

Proof. We now prove the theorem 4.2 as follows. Without loss of generality, we can assume $a_i = 1$ for all i. Let ψ_i be the characteristic function for X_i. Then using (4.4) proved in the earlier theorem, we get

$$\prod_{i=1}^n \psi_i(u + b_iv) = \prod_{i=1}^n \psi_i(u) \prod_{i=1}^n \psi_i(b_iv) \tag{4.5}$$

Claim : none of the ψ_i vanish on the real line.
We use proof by contradiction. On the contrary, assume that one of the ψ_i vanishes on $\mathbb{R}$. So, there is a root u_0 of some ψ_j such that u_0 is a root of $\prod_{i=1}^{n} \psi_i$ as well such that u_0 is with smallest possible absolute value. So,

$$\forall\ v \in \mathbb{R}\ \prod_{i=1}^{n} \psi_i(u_0 + b_i v) = 0 \tag{4.6}$$

now, choose $v \in \mathbb{R}$ such that $|b_i v| < |\frac{u_0}{2}|\ \forall i$. Writing $u_0 = \frac{u_0}{2} + \frac{u_0}{2}$ in (4.6), we get

$$\prod_{i=1}^{n} \psi_i(u_0/2) \prod_{i=1}^{n} \psi_i(u_0/2 + b_i v) = 0 \tag{4.7}$$

$$\psi_j(u_0) = 0 \implies \psi_j(u_0 + b_j v) = \psi_j(u_0) \cdot \psi_j(b_j v) = 0 = \psi_j(u_0/2) \cdot \psi_j(u_0/2 + b_j v)$$

Hence $\frac{u_0}{2}$ or $\frac{u_0}{2} + b_j v$ is a root of $\prod_{i=1}^{n} \psi_i$ which is a contradiction since both have absolute values less than u_0. Hence the claim is true.
Due to the claim we proved, we can apply logarithm on both sides of (4.5). Now, assume that $\phi_i(x) = \log(\psi_i(x))$. Application of logarithm gives

$$\sum_{i=1}^{n} \phi_i(u + b_i v) = \sum_{i=1}^{n} \phi_i(u) + \sum_{i=1}^{n} \phi_i(b_i v) := A(u) + B(v) \tag{4.8}$$

Multiply both sides of (4.8) by $(x - u)$ and integrate over u. This gives

$$\int_0^x \sum_{i=1}^{n} \phi_i(u_0 + b_i v)(x - u)\mathrm{d}u = \left[\int_0^x A(u)(x - u)\mathrm{d}u\right] + B(v)\frac{x^2}{2} := C(x) + B(v)\frac{x^2}{2}$$

Apply the change of variable $t := u + b_i v \implies \mathrm{d}t = \mathrm{d}u$. This gives

$$\sum_{i=0}^{n} \int_0^{x+b_i v} \phi_i(t)(x - t + b_i v)\mathrm{d}t = C(x) + B(v)\frac{x^2}{2} + B_2(v) + B_3(v)$$

Differentiating both sides twice and setting v to 0 gives(we use Leibniz rule here):

$$\sum_{i=0}^{n} \phi_i(x) b_i^2 = B"(0)\frac{x^2}{2} + B_2"(0) + B_3"(0) = R(x)$$

Using the relation between ϕ and ψ we get the following, where $R(x)$ is a polynomial of degree 2 over $\mathbb{C}$

$$\prod_{i=1}^{n} \psi_i(x)^{b_i^2} = \exp R(x)$$

Using $\psi_i(0) = 1$, $\psi_i(-x) = \overline{\psi_i(v)}$, we can conclude that $\exp R(x)$ denotes the characteristic function of a normal distribution. Using theorem 4.3 now proves that each X_i is normally distributed. $\square$

4.1.1 Hilbert Space of Random Variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A real vector space $\mathcal{L}_2$ space of square integrable random variables (R.Vs) is defined below:

$$\mathcal{L}_2(\Omega, \mathcal{F}, P) = \{X | X \text{ is real valued R.V. and } E[X^2] < \infty\}$$

This space is actually a space of equivalence classes of random variables where X, Y are equivalent if $X - E(X) = Y - E(Y)$ almost surely. $\mathcal{L}_2$ with respect to the operation $(X, Y) = cov(X, Y)$ forms a Hilbert space. Hence we can define the projection of a vector $v \in \mathcal{L}_2$ onto a subspace $U = span(u_1, u_2, \ldots, u_k)$ $(P_U(v))$ of $\mathcal{L}_2$ as follows:

$$P_U(v) = \sum_{i=1}^{k} \frac{(v, u_i)}{(u_i, u_i)} u_i \tag{4.9}$$

The following is the proof of the main theorem mentioned at the beginning of the chapter. The theorem has been restated for the sake of reader's convenience.

Theorem 4.4. *Assume that $P_{x,y}$ admits the following linear model for the continuous random variables X, N_Y*

$$Y = \alpha X + N_Y \ ; \ N_Y \perp X$$

Then there exists a model in the opposite direction i.e. there exists $\beta \in \mathbb{R}$ and a continuous random variable N_X such that

$$X = \beta X + N_X \ ; \ N_X \perp Y$$

$$\boxed{\textit{if and only if}}$$

$$X, N_Y \textit{ (and hence } Y, N_X\textit{) are Gaussian random variables.}$$

Proof.

- **Given:** given $\phi \in \mathbb{R}$ such that $Y = \phi X + \epsilon$ where $\epsilon \perp X$ and ϵ, X are normally distributed.
  To prove: there exists noise $\tilde{\epsilon} \perp Y$ such that $X = \tilde{\phi} Y + \tilde{\epsilon}$
  Let $P_U(X)$ be the projection of X onto the subspace $U = span(Y)$ given by

$$P_U(X) = \frac{(X, Y)}{(Y, Y)} Y$$

  Then define $\tilde{\epsilon}$ as follows:

$$\tilde{\epsilon} = X - P_U(X) \implies X = \frac{(X, Y)}{(Y, Y)} Y + \tilde{\epsilon}$$

  Above $\tilde{\epsilon}$ is the required noise variable with $\tilde{\phi} = \dfrac{(X, Y)}{(Y, Y)}$ since $\tilde{\epsilon} \perp Y$

$$cov(\tilde{\epsilon}, Y) = cov(X, Y) - \frac{(X, Y)}{(Y, Y)} cov(Y, Y) = 0$$

  Hence proved.

- **Given:** X,Y are continuous R.Vs such that $Y = \phi X + \epsilon$ is true such that $\phi \neq 0$ is in $\mathbb{R}$ and $\epsilon \perp X$. Also, this process can be reversed in the sense that there exists $\psi \in \mathbb{R}$ such that $Y \perp \tilde{\epsilon}$
  to prove: $X, Y, \epsilon, \tilde{\epsilon}$ are all Gaussian R.Vs
  $\tilde{\epsilon} = X - \psi Y = X - \psi(\phi X + \epsilon) = (1 - \psi\phi)X - \psi\epsilon$ with $\tilde{\epsilon} \perp Y$. The following cases are possible.

  1. $1 - \psi\phi \neq 0$ and $\psi \neq 0$
     Taking $l_1 = Y = \phi X + \epsilon$ and $l_2 = \tilde{\epsilon} = (1 - \psi\phi)X - \psi\epsilon$ and applying theorem 4.2 proves the claim in this case.
  2. $1 - \psi\phi \neq 0$ and $\psi = 0$
     We have $Y = \phi X + \epsilon$ and $\tilde{\epsilon} = (1 - \phi\psi)X$ such that $Y \perp \tilde{\epsilon}$. We use the following property of probability theory now:
     If U and V are independent R.Vs, then for all measurable functions f, g, we have $f(U) \perp g(U)$.
     Let Y be the variable U and $\tilde{\epsilon}$ be the variable V in the result just stated. Take f to be the identity map and take $g(x) = \dfrac{x}{1 - \phi\psi}\phi$. Notice that $g(x)$ is well defined. Using this result, and applying the functions f and g, we get $\phi X + \epsilon \perp \phi X$. But this contradicts lemma 4.1. Therefore, this case is not possible.
  3. $1 - \psi\phi = 0$ and $\psi \neq 0$
     This case be proved in a way similar to the previous case.
  4. $1 - \psi\phi = 0$ and $\psi = 0$
     then $\tilde{\epsilon} = 0$ which is a contradiction since it has to be Gaussian.

□

Theorem 4.5. *Generalized version Let $X_1, X_2 \ldots, X_n$ and Y be random variables for which $Y = \sum_{i=1}^{n} \phi_i X_i + \epsilon$ such that $\epsilon \perp (X_1, X_2, \ldots, X_n)$ and $\phi_i \neq 0$. Then we can reverse the process i.e. there exist $\psi \in \mathbb{R} for i \in \{1, 2, \ldots, n\}$ and a noise $\tilde{\epsilon}$ such that $X_1 = \sum_{i=1}^{n} \psi_i X_i + \psi Y + \tilde{\epsilon}$*

$$\boxed{\textit{if and only if}}$$

$$X_1, X_2, \ldots, X_n, Y, \epsilon, \tilde{\epsilon} \textit{ are Gaussian random variables.}$$

Proof. The proof is similar to the proof in the proof for two variable case. □

4.2 Estimating the LiNGAM model

The previous section successfully proved that linear causal models with **non-gaussian additive noise** are identifiable. Our task now is to estimate the causal model that describes a given phenomenon. One might wonder about the role of using linear regression here to estimate the causal model. The answer is that it can be used provided the causal ordering is known beforehand. Unfortunately, this is not the case in general. Hence, there are two things that need to be estimated from the observed data alone - one is the causal order of the variables involved and the other is to estimate the corresponding SEM that fits the data. Though most of the report focuses on the two variable case, here we state a few methods or results for the general n variable case.

4.2.1 Basic setup

The basic setup for identifying causal structures is explained in this subsection. Like it was previously assumed, causal relations of the observed variables are assumed to correspond to a DAG i.e. the causal graph does not involve directed cycles or feedback loops. The noise or exogenous variables are assumed to be independent of one another. This forces the model not to have any latent or unobserved confounding variables that causally influence more than one variable. Our primary focus is on continuous variables. We assume without loss of generality that each variable x_i has zero mean.

A causal ordering of the nodes of DAG is an ordering of its nodes in such a way that no later variable (in terms of ordering) has a directed path to an earlier variable in the DAG. The causal ordering of the ith variable is denoted by $k(i)$. For example, consider the simple graph $X_3 \longrightarrow X_1 \longrightarrow X_2$. Here we have $k(3) = 1; k(1) = 2; k(2) = 3$.

To summarize, the following is our linear acyclic SEM with no latent confounders and mutually independent noise variables.

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_j \tag{4.10}$$

Here, b_{ij} is called the connection strengths from x_j to x_i. Each e_i has zero mean and non-zero variance such that $e_i \perp e_j$ for all $i \neq j$. Writing (4.10) in matrix form gives the following.

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \tag{4.11}$$

Here, bold font letters in lower case represent vectors while bold font letters in upper case represent matrices. $\mathbf{B}$ is called the **connection strength matrix**. Observe that the non-zero entries of $\mathbf{B}$ match with that of the transpose of the adjacency matrix $\mathbf{A}$ for the causal graph. The following is an important observation that will be used later.

Lemma 4.2. *It is possible to perform simultaneous, equal row and column permutations on the connection strength matrix* $\boldsymbol{B}$ *so that it becomes strictly lower triangular.*

Proof. Since we assumed our causal graph to be a DAG, we can give its nodes a topological ordering. Using this topological ordering of nodes, adjacency matrix of a DAG can be made strictly upper triangular. Hence, the topologically ordered nodes can make $\mathbf{B}$ strictly lower triangular. The reason for equal row and column transformations can be understood with the help of an example which is given below. □

Example 4.1. Consider the following SEM which corresponds to an acyclic DAG:

$$\begin{aligned} x_1 &= 2x_3 + e_1 \\ x_2 &= 4x_3 + 2x_1 + e_2 \\ x_3 &= e_3 \end{aligned}$$

The 3×3 connection matrix $\mathbf{B}$ which we want to convert to a lower triangular matrix is:

$$\left[\begin{array}{c|ccc} & x_1 & x_2 & x_3 \\ \hline x_1 & 0 & 0 & 1 \\ x_2 & 2 & 0 & 4 \\ x_3 & 0 & 0 & 0 \end{array}\right] \tag{4.12}$$

The variables x_1, x_2, x_3 are written to keep track of how the vector $\mathbf{x}$ changes in $\mathbf{x} = \mathbf{Bx} + \mathbf{e}$. The first column keeps track of how the vector x on the LHS changes while the first row tracks how the elements of $\mathbf{x}$ on RHS need to be permuted to keep the equalities of the SEM intact. Now we apply row and permutations to (4.12) in order to make it strictly lower triangular.

$$\left[\begin{array}{c|ccc} & x_1 & x_2 & x_3 \\ \hline x_1 & 0 & 0 & 1 \\ x_2 & 2 & 0 & 4 \\ x_3 & 0 & 0 & 0 \end{array}\right] \xrightarrow[C_3 \Leftrightarrow C_1]{} \left[\begin{array}{c|ccc} & x_3 & x_2 & x_1 \\ \hline x_1 & 1 & 0 & 0 \\ x_2 & 4 & 0 & 2 \\ x_3 & 0 & 0 & 0 \end{array}\right] \xrightarrow[C_2 \Leftrightarrow C_3]{} \left[\begin{array}{c|ccc} & x_3 & x_1 & x_2 \\ \hline x_1 & 1 & 0 & 0 \\ x_2 & 4 & 2 & 0 \\ x_3 & 0 & 0 & 0 \end{array}\right] \tag{4.13}$$

Observe that this makes The matrix B strictly lower triangular but the sequence of $x_i's$ in first column and first row (and hence in (4.11)) is not the same. Hence applying the same column permutations gives the desired matrix as shown below.

$$\left[\begin{array}{c|ccc} & x_3 & x_1 & x_2 \\ \hline x_1 & 1 & 0 & 0 \\ x_2 & 4 & 2 & 0 \\ x_3 & 0 & 0 & 0 \end{array}\right] \xrightarrow[R_3 \Leftrightarrow R_1]{} \left[\begin{array}{c|ccc} & x_3 & x_2 & x_1 \\ \hline x_3 & 0 & 0 & 0 \\ x_2 & 4 & 2 & 0 \\ x_1 & 1 & 0 & 0 \end{array}\right] \xrightarrow[R_2 \Leftrightarrow R_3]{} \left[\begin{array}{c|ccc} & x_3 & x_1 & x_2 \\ \hline x_3 & 1 & 0 & 0 \\ x_1 & 1 & 0 & 0 \\ x_2 & 4 & 2 & 0 \end{array}\right] \tag{4.14}$$

Observe that X_3, X_1, X_2 is a causal ordering as well as a topological ordering. △

Remark 4.1. Now, note that any matrix $\mathbf{B}$ can be permuted to become strictly lower triangular according to the causal ordering $k(i)$. The causal ordering need not always be unique. I believe that number of causal orders possible depends on the number of source nodes, since they are the potential points where we can start our causal enumeration.

Objective:

Our objective is to estimate the connection strength matrix $\mathbf{B}$ using the observed data alone. We assume that the data is randomly sampled from a linear acyclic SEM with no latent confounding variables.

4.2.2 Likelihood of LiNGAM

First approach one would want to consider is the likelihood approach. Therefore, we try to obtain the expression for the likelihood of the LiNGAM model. We use the following result from probability to get the likelihood function.

Theorem 4.6. *Let x and $\boldsymbol{y}$ be n-dimensional random vectors related by an invertible linear transformation $\boldsymbol{A}$ i.e. $\boldsymbol{y} = \boldsymbol{Ax}$. Then the density of $\boldsymbol{y}$ in terms of the density for $\boldsymbol{x}$ is given by:*

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = \frac{1}{|\det \boldsymbol{A}|} \cdot f_{\boldsymbol{x}}(\boldsymbol{A}^{-1}\boldsymbol{y}) \tag{4.15}$$

In order to use this theorem , we need to modify (4.11) in the following way

$$\mathbf{x} = \mathbf{Bx} + \mathbf{e} \implies \mathbf{(I - B)x} = \mathbf{e} \implies \mathbf{x} = \mathbf{Ae} \tag{4.16}$$

where $A = (I - B)^{-1}$.Also, suppose that $A^{-1} = W$ Now applying theorem Theorem 4.6 to (4.16), we get the following

$$f_{\tilde{x}}(\tilde{x}) = \frac{1}{|\det \mathbf{A}|} \cdot f_{\tilde{e}}(\mathbf{Wx})$$

Denote the ith row of Wx by $w_i^T x$. Using this notation, we get

$$f_{\tilde{x}}(\tilde{x}) = \frac{1}{\det A} \cdot \prod_{i=1}^{n} f_{e_i}(w_i^T x)$$

If the vectors observed are $x_1, x_2, \ldots, x_t$, then the likelihood is given as

$$L(W;x_1, x_2, \ldots, x_t) = \left[\frac{1}{\det A} \prod_{j=1}^{t} \prod_{i=1}^{n} f_{e_i}(w_i^T x_j) \right] - t \log (\det A)$$

Observe that $A = (I - B)^{-1}$ and $I - B$ is a lower triangular matrix with ones on its diagonal. Since the determinant of a lower triangular matrix is the product of its diagonal elements, we can say that $\det(I - B) = 1$. Hence $\det A = \dfrac{1}{\det(I - B)} = 1$. Therefore, the likelihood function is:

$$\log L(W;x_1, x_2, \ldots, x_t) = \sum_{j=1}^{t} \sum_{i=1}^{n} f_{e_i}(w_i^T x_j)$$

Let the jth row of $\mathbf{B}$ be given by b_j while the vector $x_i = (x_{i0}, x_{i1}, \ldots, x_{in})$. Using $w_i^T x_j = x_{ij} - b_j^T x_i$ and using the normalized pdf $\tilde{f}_{e_i}$ for each e_i with variance σ_i^2 , we get:

$$\log L(W;x_1, x_2, \ldots, x_t) = \sum_{j=1}^{t} \sum_{i=1}^{n} \tilde{f}_{e_i} \left(\frac{x_{ij} - b_j^T x_i}{\sigma_j} \right) + t \sum_{j=1}^{n} \log(\sigma_j) \tag{4.17}$$

Once we have obtained the log likelihood of the LiNGAM model as in (4.17), our next step should be to estimate the connection strength matrix $\mathbf{B}$ which maximizes the log likelihood function over all possible causal orderings. This approach has the following two problems due to which it is generally avoided:

- The number of possible causal orderings for nodes (which need to be identified here) increase quickly when large number of variables are involved. Hence, this procedure will be **computationally costly**.
- In order to apply the method we proposed, we need to *estimate* the densities $\tilde{f}_{e_i}$ for all $1 \leq i \leq n$. These estimates are again used to estimate the likelihood function. This means that we use the data twice, which is not desirable.

Methods discussed later in the section explain in brief the methods proposed by Shimizu et. al. to estimate $\mathbf{B}$ that do not require the investigation of all possible causal orderings or the estimation of densities. Before stating this method, it is important to have a basic understanding of "independent component analysis".

4.2.3 Independent component analysis (ICA)

The problems that ICA deals with can be better understood with the help of a motivating example which has been given below. This example serves as a motivation for us to mathematically define the problem we wish to address. Then, we try to identify the conditions under which this model can be estimated.

Motivation

Assume that there are three people locked up in a room. The three people talk simultaneously and all the voices/sound signals in the room are being recorded by three microphones kept at three different places. The voice recorded in each microphone will depend on its distance from each of the three person present in the room. Hence, each microphone records a unique superimposition of the three sound signals. Let these (observed) recordings/signals be given by $x_1(t), x_2(t), x_3(t)$. Now assuming that the signals being emitted by the three people (three independent sources) is given by $s_1(t), s_2(t), s_3(t)$, we can conclude the following:

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \\ x_3(t) &= a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t) \end{aligned}$$

Here, we wish to retrieve the original speech signals $s_1(t), s_2(t), s_3(t)$ using the observed signals $x_1(t), x_2(t), x_3(t)$. Note that the parameters $a'_{ij}s$ as well as the source signals $s_i(t)$ are unknown to us. Hence, the aim of the section is to estimate $s_i(t)$ for $1 \leq i \leq 3$ and $a_{i,j}$ for all i, j using the observed signals $(x_i(t))$ alone.

4.2.4 ICA model

As observed earlier, ICA is used to estimate the parameters a_{ij} and the independent components $s_i(t)$ using the information we have from the observed signals $x_i(t)$. We rigorously define ICA in the following manner

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n \text{ for all } i \in \{1, 2, \ldots, n\} \tag{4.18}$$

Notice that the source variables and observed variables in the above definition do not involve the time variable. Instead each s_i and x_i is simply a random variable such that all s_i are mutually statistically independent variables. The **independent components(IC)** s_i are called latent variables because they cannot be directly observed. All we observe here are the random variables x_i, and we must estimate the mixing coefficients a_{ij} and the ICs s_i using the x_i. (4.18) is called the **basic ICA model**. This model is said to be a **generative model**, which means that it describes how the observed data are generated by a process of mixing the component s_j. The matrix version of the ICA model is given below:

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{n} \mathbf{a}_i s_i \tag{4.19}$$

In (4.19), $\mathbf{A}$ is an $n \times n$ matrix consisting of all parameters a_{ij} while $\mathbf{x}$= $(x_1, x_2 \ldots, x_n)$ and $\mathbf{s} = (s_1, s_2 \ldots, s_n)$ denote the vectors consisting of observed and latent variables

respectively. Another definition of the ICA model using column vectors $\mathbf{a}_i$ of $\mathbf{A}$ is also given in (4.18).

Restrictions we assume for the ICA model

1. The independent components are assumed to be statistically independent.

2. The independent components must have a non-gaussian distribution. It is important to note that in the basic model, we do not assume that we know the form of the non gaussian distributions that the independent components follow. Instead we only assume that the distribution is not gaussian.

3. We assume for the sake of simplicity that the unknown mixing matrix is an invertible square matrix.

Under the above mentioned assumptions, ICA model is identifiable i.e. the mixing matrix and the independent components can be estimated upto some indeterminacies. These indeterminacies in the final solution have been explained below.

Ambiguities of ICA

ICA is capable of estimating the matrix $\mathbf{A}$ but it has the limitation of being able to estimate it only upto a permutation, scaling and sign indeterminacy. These ambiguities have been elaborated below:

1. We cannot determine the variances of the independent components because $\mathbf{x}$ can be written as $\mathbf{x} = \sum_i (\frac{1}{\alpha_i} a_i)(s_i \alpha_i)$ which means that the estimate we obtain for $\mathbf{A}$ is correct upto left multiplication by a diagonal matrix. This is because each column of $\mathbf{A}$ is being multiplied by a different scalar. Observe that this also explains the sign indeterminacy of ICA.

2. We cannot determine the order of the independent components. Since both $\mathbf{s}$ and $\mathbf{A}$ are unknown, the order of the terms in (4.19) can be changed the way we want i.e. we can call any one of the ICs as the first IC. In other words, we can estimate $\mathbf{A}$ only upto multiplication by a permutation matrix $\mathbf{P}$ to the left (which permutes the rows).

Suppose that $\mathbf{A}$ is the true solution for the ICA model we are given to solve. Then the estimate we obtain ($\mathbf{A}_{ICA}$) will be related to $\mathbf{A}$ as follows:

$$\mathbf{A}_{ICA} = \mathbf{PDA} \tag{4.20}$$

where $\mathbf{P}$ is a permutation matrix while $\mathbf{D}$ is a diagonal matrix. There a number of different approaches to estimate the independent components. These methods are not discussed here.

4.2.5 ICA-LiNGAM algorithm

ICA-LiNGAM algorithm is an estimation algorithm that utilizes independent component analysis in order to estimate the connection strength matrix $\mathbf{B}$. We now convert the problem of LiNGAM identification into a problem of ICA estimation as follows:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \implies \mathbf{x} = \mathbf{A}\mathbf{e} \text{ where } \mathbf{A} = (\mathbf{I}\text{-}\mathbf{B})^{-1} \tag{4.21}$$

Here, $\mathbf{x}$ denotes the vector of observed values while $\mathbf{e}$ denotes the vector of non-Gaussian noise variables. Hence the problem of estimating $\mathbf{A}$ has now become the problem of ICA model estimation.

Instead of assuming that the observed and source/error variables to be random variables, consider the case where the observed and error/source variables are random vectors $\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_n$ and $\mathbf{e}_1, \mathbf{e}_2 \ldots \mathbf{e}_m$ respectively. The problem of LiNGAM identification now becomes

$$\mathbf{X} = \mathbf{A}\mathbf{E} \text{ where } \mathbf{X}_{m\times n} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \text{ and } \mathbf{E}_{m\times n} = [\mathbf{e}_1, \ldots, \mathbf{e}_m]^T \tag{4.22}$$

Note that $\mathbf{A}$ is called the mixing matrix whose inverse is denoted by $\mathbf{W}$. The following are the steps of ICA-LiNGAM algorithm

1. Given an $m \times n$ data matrix $\mathbf{X}$, where each column contains one sample vector $\mathbf{x}$, first subtract the mean from each row of $\mathbf{X}$, then apply an ICA algorithm to obtain a decomposition $\mathbf{X} = \mathbf{A}\mathbf{E}$ where $\mathbf{E}$ has the same size as $\mathbf{X}$ and contains in its rows the independent components. From here on, we will exclusively work with $\mathbf{W} = \mathbf{A}^{-1}$.

2. Find the one and only permutation of rows of $\mathbf{W}$ which yields a matrix $\mathbf{W}^{est}$ without any zeros on the main diagonal. In practice, small estimation errors will cause all elements of $\mathbf{W}$ to be non-zero, and hence the permutation is sought which minimizes $\sum_i (1/|\mathbf{W}^{est}_{ii}|)$.

3. Divide each row of $\mathbf{W}^{est}$ by its corresponding diagonal element, to yield a new matrix $\mathbf{W}^*$ with all ones on the diagonal.

4. Compute an estimate $\mathbf{B}^*$ of $\mathbf{B}$ using $\mathbf{B}^* = \mathbf{I} - \mathbf{W}^*$.

5. Finally, to find a causal order, find the permutation matrix $\mathbf{P}$ (applied equally to both rows and columns) of $\mathbf{B}^*$ which yields a matrix $\mathbf{B}^\# = \mathbf{P}\mathbf{B}^*\ \mathbf{P}^T$ which is as close as possible to strictly lower triangular. This can be measured for instance using $\sum_{i\leq j} [\mathbf{B}^\#_{ij}]^2$

First step involves the use of ICA to obtain an estimate for $\mathbf{A}$. Any standard ICA algorithm found in literature can be applied here. This gives us the estimate $\mathbf{A}$ which as pointed out earlier has permutation,scaling and sign indeterminacies which have to be resolved.

The second step uses the property of a DAG. The estimated matrix $\mathbf{W}$ has rows arranged in a random order. Hence the observed vectors are not necessarily in correspondence with the error vectors (We are not trying to give the nodes a causal ordering, we only wish to have the right correspondence). This issue is resolved by using the fact that there exists a unique permutation of rows of $\mathbf{W}$ that would give a matrix with no zeroes on the diagonal. This is because $\mathbf{W} = \mathbf{I} - \mathbf{B}$, where there is a permutation of rows of $\mathbf{B}$

that can make the matrix **B** strictly lower triangular. This happens because we assumed the graph to be a DAG. Hence **W** turns into be a lower triangular matrix with 1 as its diagonal entries. The uniqueness of the permutation however needs a proof which is not mentioned here.

The result mentioned above would occur in real life provided we were able to estimate $\mathbf{W}^{est}$ exactly. However, this is not the case in general. Since every permutation of the rows of **W**, except for one , contains at least one zero, finding the matrix which minimizes $\sum_i(1/|\mathbf{W}_{ii}^{est}|)$ would give us the matrix with non-zero diagonal entries. As mentioned earlier, this is the matrix with the right correspondence between source and observed vectors.Again, this need not necessarily be the causal ordering.

The third step is to divide the elements of each row with its corresponding diagonal element because we know that the true value of **W** is **I - B**. This is due to the fact that all the diagonal entries for **B** are zero when the matrix **B** is made strictly lower triangular due to its causal ordering. Hence the true estimate for **W** must contain all 1s on its diagonal. This solves the scaling indeterminacy associated with ICA estimation.

The final step intends to solve the permutation indeterminacy. The indeterminacy is solved by identifying the causal ordering. If $\mathbf{B}^*$ is an exact estimate of **B**, then the causal ordering can be obtained by permuting the rows and columns of $\mathbf{B}^*$ such that it becomes strictly lower triangular. However, the estimate we obtain is almost never exact. Hence the causal order is identified by permuting the rows and columns of $\mathbf{B}^*$ until it minimizes the sum of upper triangular elements.

Remark 4.2.

1. In the case of second step, we have to identify the unique permutation of rows that minimizes $\sum_i(1/|\mathbf{W}_{ii}^{est}|)$. This can be done by by performing an exhaustive search as long as the dimension of the data is low. High dimensional data require specialized methods to identify the required row permutations.

2. This method exhibits the drawback that ICA algorithms may not converge to a correct solution in a finite number of steps if the initially guessed state is badly chosen.

4.3 Data Analysis

We will again use the Old Faithful Geyser data that we came across in the last chapter. This time, we use the LiNGAM model to determine the underlying causal structure. A plot of the data after fitting the least squares regression line is provided. It can be observed that the data is clustered in two different regions which indicates that the noise is likely to be bi-modal. We calculate the coefficient of determination which gives us enough evidence to believe that the variables involved have a linear relationship.

```
> attach(faithful)
> h = lm(waiting~eruptions)
> summary(h)$r.squared
[1] 0.8114608
```

Now, we give enough evidence to claim that the noise is infact a mixture distribution made up of two gaussian distributions. Observe that the first cluster is found for the values of the x-variable (eruptions) less than 3, while the other one is found for x-values greater than 3. Hence we subset the data as follows:

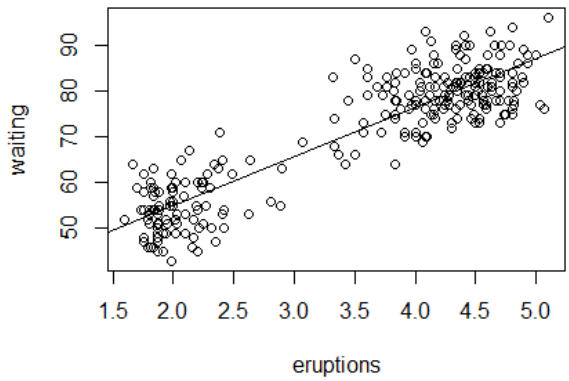

Figure 4.1: Clustered old faithful geyser data

```
> erup = faithful[,1]
> a = h$residuals[erup<3]
> length(h$residuals[erup<3])
[1] 97
> ks.test(h$residuals[erup<3], rnorm(97,mean(a), sd(a)))
Two-sample Kolmogorov-Smirnov test

data:  h$residuals[erup < 3] and rnorm(97, mean(a), sd(a))
D = 0.10309, p-value = 0.6812
alternative hypothesis: two-sided
```

From a p-value of 0.6812, we can conclude that we do not have enough evidence to reject the null hypothesis (the noise is normally distributed) at a significance level of 5% . We apply the same procedure for the second cluster.

```
> b = h$residuals[erup > 3]
> length(h$residuals[erup>3])
[1] 175
> ks.test(h$residuals[erup>3], rnorm(175,mean(b), sd(b)))
Two-sample Kolmogorov-Smirnov test

data:  h$residuals[erup > 3] and rnorm(175, mean(b), sd(b))
D = 0.085714, p-value = 0.5412
alternative hypothesis: two-sided
```

Here also, we do not have enough evidence to reject the null hypothesis. Hence the model has an inherent linearity along with a noise that follows a mixed distribution made out of two gaussian distributions. Hence this is an ideal situation to apply the LiNGAM model in order to identify the causal direction.

```
getParents(faithful,method = "LINGAM")
          eruptions waiting
eruptions         0       1
waiting           0       0
```

We again get the same result as we obtained earlier in the previous chapter. Yet again we conclude that eruption time of a given geyser determines the amount of time it takes for the next geyser to erupt. Therefore the causal structure is :

$$\text{eruption} \longrightarrow \text{waiting}$$

Bibliography

[AHO01] Juha Karhunen Aapo Hyvarinen and Erkki Oja. *Independent Component Analysis*. Wiley series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. A Jhon Wiley and Sons publication, New York, 2001.

[dW] Mathijs de Weerdt. Dags and topological ordering. Lecture slides for the course Algoritmiek at TUDelft. `https://ocw.tudelft.nl/wp-content/uploads/Algoritmiek_DAGs_and_Topological_Ordering.pdf`. Last visited on 3 June 2020.

[EW08] Bryon Ellis and Wing Hung Wong. Learning causal bayesian network structures from experimental data. *Journal of the American Statistical Association*, 103(482):778–789, 2008.

[H+08] Patrik O. Hoyer et al. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems 21*, 2008. `https://papers.nips.cc/paper/3548-nonlinear-causal-discovery-with-additive-noise-models.pdf`.

[HZSH10] Aapo Hyvarinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731, 2010. `http://jmlr.org/papers/v11/hyvarinen10a.html`. Last visited on 9 June 2020.

[JPJ16] Madelyn Glymour Judea Pearl and Nicholas P. Jewell. *Causal inference in statistics*. A Jhon Wiley and Sons publication, United Kingdom, 2016.

[JPS11] Dominik Janzing Jonas Peters and Bernhard Scholkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436 – 2450, 2011.

[JS10] Dominik Janzing and Bastian Steudel. Justifying additive noise model-based causal discovery via algorithmic information theory. *Open Systems and Information Dynamics*, 17(2):189–212, 2010.

[JSK17] Satoshi Oyama Jing Song and Masahito Kurihara. Tell cause from effect: models and evaluation. *International Journal of Data Science and Analytics*, 2017.

[Mes12] Franz H. Messerli. Chocolate consumption, cognitive function, and nobel laureates. *The New England Journal of Medicine*, pages 1562–1564, 2012.

[Nea03] Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall series in Artificial Intelligence. Prentice Hall, 2003.

[P+17] Jonas Peters et al. *Elements of causal inference*. The Adaptive Computation and Machine Learning series. The MIT Press, 2017.

[Pea] Judea Pearl. Causal analysis in theory and practice. Reply given for a query raised by Jim Grace. http://causality.cs.ucla.edu/blog/index.php/2012/12/07/on-structural-equations-versus-causal-bayes-networks/. Last visited on 3 June 2020.

[Pea13] Judea Pearl. *Causality*. Cambridge University Press, New York, 2 nd edition, 2013.

[Pet08] Jonas Peters. Asymmetries of time series under inverting their direction. Diploma thesis submitted at Ruprecht-Karls-Universitat Heidelberg., 2008. http://web.math.ku.dk/~peters/jonas_files/diplomaJonasPeters.pdf. Last visited on 9 June 2020.

[Pet12] Jonas Martin Peters. Restricted structural equation models for causal inference. Doctoral thesis submitted at ETH Zurich., 2012. https://www.research-collection.ethz.ch/handle/20.500.11850/60302. Last visited on 3 June 2020.

[Pet15] Jonas Peters. Causality. Lecture notes from ETH Zurich., Spring 2015. http://web.math.ku.dk/~peters/jonas_files/scriptChapter1-4.pdf. Last visited on 3 June 2020.

[SHHK06] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvarinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7, 2006. http://www.jmlr.org/papers/v7/shimizu06a.html. Last visited on 9 June 2020.

[Shi14] Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 2014. http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/papers/Shimizu13BHMK.pdf. Last visited on 9 June 2020.

[Slo18] George M. Slota. Dags, trees. Lecture notes for the course CSCI 4260/MATH 4150 Graph Theory used at Rensselaer Polytechnic Institute, Spring 2018. https://www.cs.rpi.edu/~slotag/classes/SP18/slides/lec06.pdf. Last visited on 9 June 2020.

[Sti11] Robert Stine. Hilbert spaces. Lecture notes for the course Statistics 910: Time Series Analysis used at University of Pennsylvania, Spring 2011. http://www-stat.wharton.upenn.edu/~stine/stat910/lectures/16_hilbert.pdf. Last visited on 9 June 2020.

[Wol18] Robert L. Wolpert. Probability and measure theory. Lecture notes for the course STA 711: Probability and Measure Theory used at Duke University, Fall 2018. http://www2.stat.duke.edu/courses/Fall18/sta711/lec/wk-05.pdf. Last visited on 9 June 2020.

[Zit13] Gordan Zitkovic. Characteristic functions. Lecture notes for the course Theory of probability 1 used at University of Texas at Austin, Fall 2013. https://web.ma.utexas.edu/users/gordanz/notes/characteristic.pdf. Last visited on 9 June 2020.