ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

MASTER THESIS

LITERATURE REVIEW OF CAUSAL INFERENCE WITH THE USE OF STRUCTURAL CAUSAL MODELS (SCM)

Anastasios Dionysopoulos

Supervisor:

Dr. Fotios Siannis
Dr. Loukia Meligotsidou
Dr. Samis Trevezas

Athens 2020

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

Abstract

Department of Mathematics

Master In Statistical And Operational Research

Literature Review of Causal Inference with the use of Structural Causal Models (SCM)

by Anastasios Dionysopoulos

In this dissertation, we are tying to model Causality inference so as to identify and extract her from empirical data. In statistics, when we find two random variables to be dependent, doesn't mean that they have also causal relation,Cause-Effect. As a result, if we want to model Causality between random variables, we need to model the direction of this dependency, from the Causes to Effects, except from how depended or correlated are these variables. In order to do that we use the Directed Acyclic Graphs, DAGs. As a result, in the First,Second and Third Chapter we illustrate how capable the DAGs are as tool to store Probabilistic dependence-independence Knowledge. Also, we illustrate two basic assumptions: the Markov and the Faithfulness. These two play a significant role in that procedure. At the Forth chapter, we propose the Structural Causal Models, SCMs, as a way to model Causal information. The SCMs can induce distribution Functions and compatible DAGs to that distribution at the same time. In statistics, we use Distribution Functions as a data generation process. In Causality inference we use the SCMs the same way with the exception that these can give us much more information about the data than classical Distribution Functions. The main reason that we use the SCMs as a modelling tool is their additional ability to produce information about randomized-trial or their ability to induce Intervention distributions. One way to identify that one variable has causal influence in the outcome of an other variable is by keeping all the factors that influence the outcome variable static except the one we are interested in. This is very difficult in practice. However, the SCMs give us the solution. In the Fifth Chapter of this dissertation, assuming a known SCM which generate the data, we give a brief illustration about how we can compute the Causal influence of a variable in a system based on the randomize trial or precisely, the knowledge of intervention distributions. Finally, in the Sixth Chapter we illustrate algorithms which extract the correct SCM from empirical data, and in the last Seventh chapter we compare these algorithms under their ability to predict a correct SCM.

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Περίληψη

Τμήμα Μαθηματικών

Μεταπτυχιακό στην Στατιστική και Επιχειρησιακή Έρευνα

Βιβλιογραφική Ανασκόπηση της αιτιώδους συμπερασματολογίας με την Χρήση των Δομικών Αιτιώδων Μοντέλων

Αναστάσιος Διονυσόπουλος

Σκοπός αυτής της διπλωματικής είναι η μοντελοποίηση της αιτιώδους συμπερασματολογίας, καθώς και η μελέτη της σε εμπειρικά δεδομένα. Στην στατιστική, όταν βρίσκουμε δυο μεταβλητές εξαρτημένες δεν συνεπάγεται ότι η εξάρτηση αυτή είναι αιτιακή, δηλαδή ότι η μια είναι η αιτία και η άλλη το αποτέλεσμα. Δηλαδή, για την μοντελοποίηση της αιτιώδους συμπερασματολογίας εκτός απο μέτρα συσχέτισης χρειαζόμαστε και μέτρα που να δείχνουν την κατεύθυνση της πληροφορίας. Για αυτό τον σκοπό χρησιμοποιούμε τους κατευθυνόμενουν άκυκλους γράφους (ΚΑΓ). Στα πρώτα τρία κεφάλαια, γίνεται μελέτη των ΚΑΓ ως προς την ικανότητα τους να αποθηκεύουν πληροφορία που σχετίζεται με την εξάρτηση και ανεξαρτησία τυχαίων μεταβητών. Επίσης, διατυπόνονται και δύο βασικές υποθέσεις 1) η Μαρκοβιανή και 2) η πιστότητα που παίζουν καθοριστικό ρόλο σε αυτη την διαδικασία. Στην συνέχεια, στο τέταρτο κεφάλαιο παρουσιάζονται τα Αιτιώδη Δομικα Μοντέλα (ΑΔΜ), τα οποία είναι ο τρόπος που χρησιμοποιήσαμε στην διπλωματική αυτή για την μοντελοποίηση την αιτιακής συμπερασματολογίας. Τα ΑΔΜ εχουν την ικανότητα να μοντελοποιούν κατανομές πιθανότητας αλλα και συμβατους γράφους με την κατανομή την ίδια στιγμή. Όπως η κατανομές στην στατιστική θεωρούμε οτι δημιουργούν τα δεδόμενα, έτσι και τα ΑΔΜ στην αιτιώδη συμπερασματολογία εχουν τον ίδιο ρόλο με την διαφορά οτι παρέχουν περισσότερες πληροφοριές. Ο λόγος όμως που τα κάνει ιδιαίτερα προσιτό μέσο για την μελέτη της αιτιώδους συμπερασματολογίας είναι οτι έχουν την δυνατότητα να μοντελοποιούν και κατανομές που προκύπτουν απο τυχαιοποιημένες δοκιμές ή αλλιώς κατανομές που προέκυψαν μετά απο επέμβαση στο σύστημα. Ένας τρόπος για να εξαταστεί αν μια μεταβλητή εχει αιτιώδη επίδραση σε μια άλλη είναι, κρατώντας όλες τις παραμέτρους του συστήματος σταθερές, να μεταβάλεις μόνο την ζητούμενη και να δείς τις αλλαγές που επιφέρει αυτή στο σύστημα. Αυτο όμως ειναι αρκετά δύσκολο να γίνει στην πραγματικότητα. Παρόλα αυτά τα ΑΔΜ δίνουν την λύση σε αυτό το πρόβλημα. Στο πέμπτο κεφάλαιο με την υπόθεση οτι το ΑΔΜ που δημιουργεί τα δεδομένα ειναι γνωστό, παρουσιάζεται ενδελεχώς ο τρόπος υπολογισμού της αιτιώδους επίδρασης μιας μεταβλητής βασισμένος στις τυχαιοποιημένες δοκιμές-κατανομές που προκύποτουν μετά απο επέμβαση. Τέλος στο έκτο κεφάλαιο παρουσιάζονται αλγόρρυθμοι εξαγωγής των ΑΔΜ απο εμπειρικά δεδομένα και στο έβδομο και τελευταίο κεφάλαιο γίνεται η σύγκρισή τους.

Contents

Chapter 1

Graph Notation

1. Graph Terminology

A graph consists of vertices (or nodes),the set of all nodes in the graph is symbolized by V, and the connecting links between them are known as edges,the set of all edges in the graph is symbolized by $\mathcal{E}$.

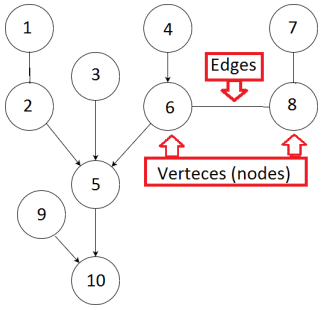


Figure 1.1: A graph example

A graph can be characterized as directed (Figure 1.2(a)),contains only directed edges ($i \rightarrow j$). Similarly, an undirected graph (Figure 1.2(b)), contains only undirected edges ($i - j$), and partially oriented graph like (Figure 1.1), contains both directed and undirected edges.

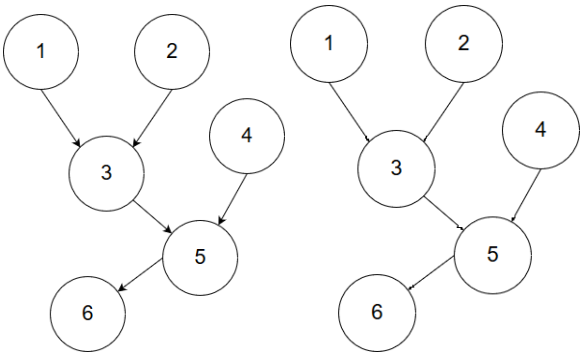


Figure 1.2: (a) A directed graph example. (b) An undirected graph example.

Definition 1..1 (Graph). *A graph $\mathcal{G} = (V, \mathcal{E})$ consists of (finetely many) nodes or vertices V and edges $\mathcal{E} \subseteq V^2$ with $(v, v) \notin \mathcal{E}$for any $v \in V$.*

The members of $\mathcal{E}$ are ordered pairs of vertices. For example, in figure 1.3 depicts two graphs:
$\mathcal{G}_1 = (V_1, \mathcal{E}_1)$ which $V_1 = \{1, 2\}$ and $\mathcal{E}_1 = \{(1, 2)\}$
$\mathcal{G}_2 = (V_2, \mathcal{E}_2)$ which $V_2 = \{1, 2\}$ and $\mathcal{E}_2 = \{(1, 2), (2, 1)\}$

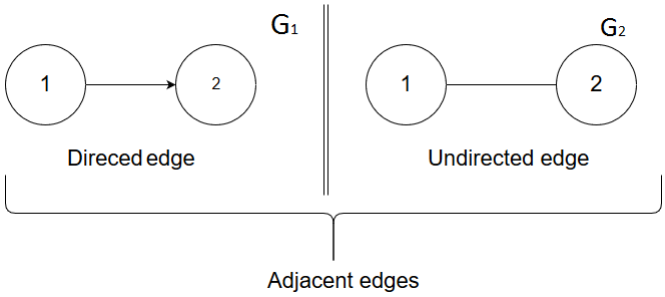


Figure 1.3: adjacent,un-directed edges example

Let $\mathcal{G} = (V, \mathcal{E})$ be a graph with $V := (1, ..., d)$. Then:

- Two nodes i and j are adjacent if either $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$.See, both graphs in Figure 1.3.
- We say that there is an undirected edge between two adjacent nodes i and j if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$.See, right graph in Figure 1.3.
- An edge between two adjacent is directed if is not undirected .We then write $i \to j$for $(i, j) \in \mathcal{E}$. See, left graph in Figure 1.3.

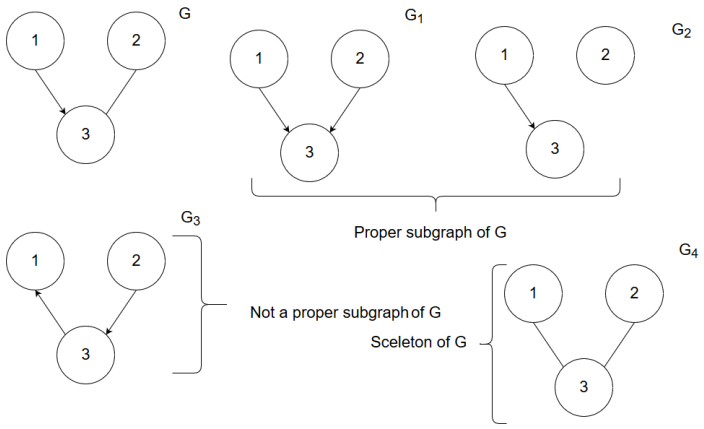


Figure 1.4: Proper - Sub-graph skeleton example

- A graph $\mathcal{G}_1 = (V_1, \mathcal{E}_1)$ is called subgraph of $\mathcal{G}$ if $V_1 = V$ and $\mathcal{E}_1 \subseteq \mathcal{E}$. If additionally $\mathcal{E}_1 \subset \mathcal{E}$ then $\mathcal{G}_1$ is a proper subgraph of $\mathcal{G}$. See, Figure 1.4
- Skeleton of $\mathcal{G}$ does not take the directions of the edges into account. It's the graph $(V, \tilde{\mathcal{E}})$ with $\{(i,j),(j,i)\} \in \tilde{\mathcal{E}}$ if (i,j)$\in \mathcal{E}$ or (j,i)$\in \mathcal{E}$. See, Figure 1.4

We can make free use of the terminology of kinship (e.g., parents, children, descendants, ancestors) to denote various relationships in a graph. More precisely a node i is called a parent of j if $(i,j) \in \mathcal{E}$ and $(j,i) \notin \mathcal{E}$ and the node i is called child of j if $(i,j) \notin \mathcal{E}$ and $(j,i) \in \mathcal{E}$. The set of parents of j is denoted by $PA_j^{\mathcal{G}}$ and the set of its children by $CH_j^{\mathcal{G}}$ in graph $\mathcal{G}$. For example in graph $\mathcal{G}$ of Figure 1.1 we have: $PA_5^{\mathcal{G}} = \{2,3,6\}$, $CH_5^{\mathcal{G}} = \{10\}$. We denote all the descendants of i by $DE_i^{\mathcal{G}}$, excluding i[1]and all non-descendant ,excluding i, by $ND_i^{\mathcal{G}}$.
In fiure 1.1 $DE_2^{\mathcal{G}} = \{5,10\}$ and $ND_{10}^{\mathcal{G}} = \{5,2,3,6,4,9\}$.

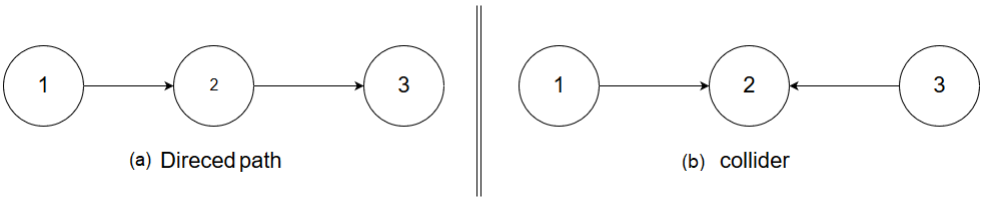


Figure 1.5: directed path collision example

Furthermore, we may define the following:

- A path in $\mathcal{G}$ is a sequence of (at least two) distinct vertices $i_1, ..., i_n$ such that there is an edge between i_k and $i_{k+1} \forall k \in \{1, ..., n\}$. For example in Figure 1.1 two possible paths are the $(9, 10, 5, 2, 1)$ and the $(7, 8, 6, 4)$.
- If $i_k \to i_{k+1} \forall k$ we speak of a directed path from i_1 toi_n and call the i_n a descendant of i_1 see Figure 1.5(a).

[1] In this dissertation we assume that a node i is neither a descendant nor a non-descendant of itself.

- Three nodes are called an immorality or v-structure if one node is a child of two others that themselves are not adjacent .

In figure 1.1 we can find the immoralities or v-structures : $2 \rightarrow 5 \leftarrow 6$ or $9 \rightarrow 10 \leftarrow 5$ etc , but if we have an arrow like $2 \rightarrow 6$

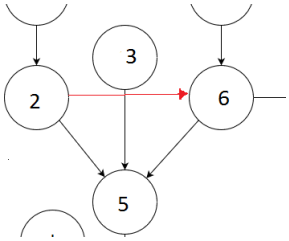


Figure 1.6: An alternative of figure 2.1

we lose the first immorality $2 \rightarrow 5 \leftarrow 6$.

- A graph $\mathcal{G}$ is fully connected if all pairs of nodes are adjacent. See, Figure 1.7.

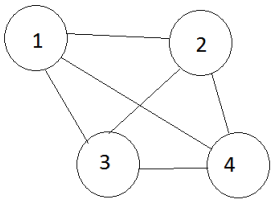


Figure 1.7: Fully connected graph.

- A graph $\mathcal{G}$ is called directed if all its edges are directed.
- A graph $\mathcal{G}$ is called a partially directed acyclic graph PDAG if there is no directed cycle, i.e. if there is no pair (j, k) with directed paths from j to k and from k to j.
- $\mathcal{G}$ is called directed and acyclic graph DAG if it is a PDAG and all the edges are directed.

In figure 1.8 there is an example of DAG left graph PDAG in right graph and cyclic directed graph in the middle graph.

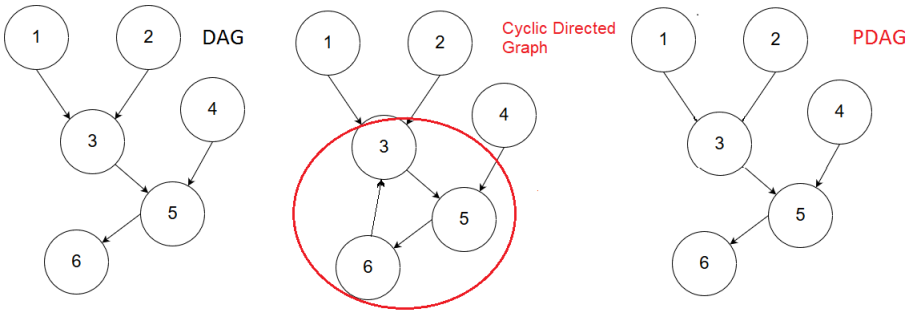


Figure 1.8: DAG Partially-DAG and Cyclic Directed graph example

If a graph $\mathcal{G}$ is a DAG then easily proved that the graph has at least one node with no incoming edges.

Proposition 1..1. If $\mathcal{G}$ is a DAG, then $\mathcal{G}$ has a node with no incoming edges.

***Proof* 1.** We suppose that $\mathcal{G}$ is a DAG and every node of $\mathcal{G}$ has at least one incoming edge. We pick any node $v \in V$, and follow edges backward from v, and repeat this procedure until we visit a node, say w, twice. Let C denote the sequence of nodes encountered between successive visits to w. C exists because $|V| < \infty$. As a result C is a cycle but $\mathcal{G}$ is DAG and conclude to contradiction. So there must be a node with no incoming edges.

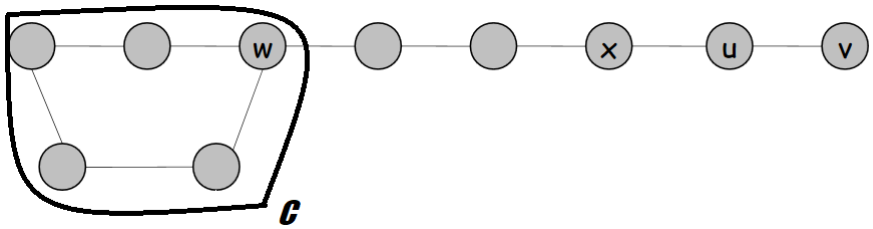


Figure 1.9

We continue with the definition of d-separation generally in DAGs this graphical tool is very important for the graphical representation of Probabilistic knowledge.

Definition 1..2 (d-sepereation). In a DAG $\mathcal{G}$, a path between node i_1 and node i_n is blocked by set of nodes S (with neither i_1 nor i_n in S) whenever there is a node i_k such that one of the following two posibilities holds:

1. $i_k \in S$ and $i_{k-1} \to i_k \to i_{k+1}$ or $i_k \in S$ and $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \to i_{k+1}$
2. $i_{k-1} \to i_k \leftarrow i_{k+1}$and neither i_k nor any of its descendants is in S.

We say that two disjoint subsets of vertices A and B are d-separated by a third subset S (disjoint too), if every path between nodes in A and B is blocked by S.

To understand when two nodes i.e. $i_1,\ i_2$ are d-separated by a set of node S, we follow the algorithm bellow.

Step 1: Find all the paths between nodes i_1, i_2.

Step 2: In every path, check all the paths between possible triplets of nodes, if is blocked (not active)or not blocked (active) by S.

(a): If all the triplets of variables are active, then the path is active by S.

(b): If there is a triplet which is not active, then the path is not active by S.

Step 3: If all the paths between i_1 and i_2 are not active by S then i_1, i_2 is d-separated by S. But if there is a path which is active, then is not d-separated by S.

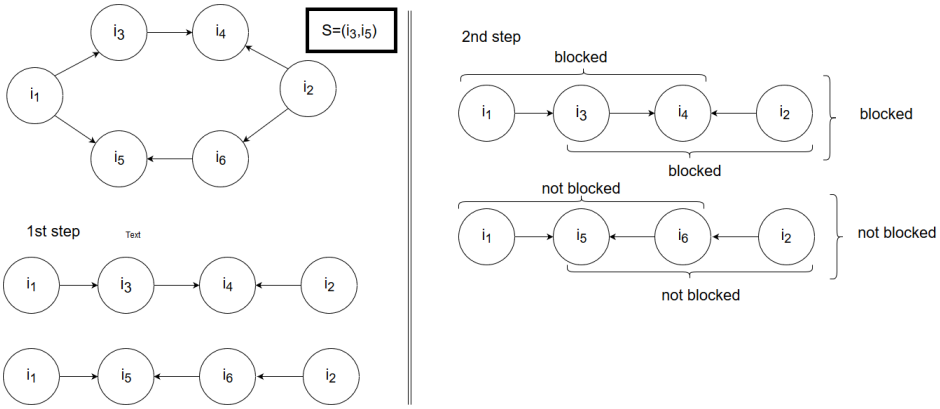


Figure 1.10: d-separation algorithm

Let‘s turn algorithm into practice with the example below:

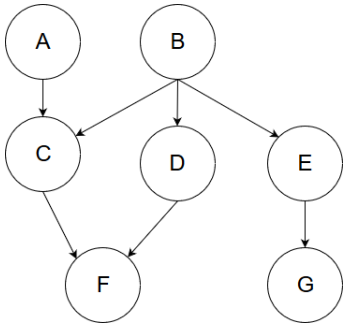


Figure 1.11

Example 1..2. (a) Given the graph of Figure 1.11 assume that we are interested in whether the nodes A and B are d-separated from $S := \{D\}$. We start with the paths. In that case, we have two paths $A \to C \leftarrow B$ and $A \to C \to F \leftarrow D \leftarrow B$. The first path $A \to C \leftarrow B$ is a v-structure triplet and since $C \notin S$ or any descendant of C, in our case the F, then the path become blocked by S, so the path is not active. In the second path $A \to C \to F \leftarrow D \leftarrow B$, we start with the triplet A, C, F which isn't block ,so the triplet is active. Then we continue with C, F, D which is not active

by D, so the path is not active. Thus, A,B are d-separated by $S = \{D\}$.
(b) Assume that we are interested in whether the nodes A and B are d-separated by $S := \{C\}$. We start with the paths and we have the same paths as in the first case $A \rightarrow C \leftarrow B$ and $A \rightarrow C \rightarrow F \leftarrow D \leftarrow B$. In the first path $A \rightarrow C \leftarrow B$, this triplet of variables is not blocked by C. Thus the path is active. In the second path $A \rightarrow C \rightarrow F \leftarrow D \leftarrow B$, triplet A, C, F is blocked by C, so the path is not active. Thus, A, B is not d-separated by $S = \{C\}$.
(c) assume that we are interested in whether the nodes A and B are d-separated by $S = \{C, D\}$. We start with the same paths as in the first case $A \rightarrow C \leftarrow B$ and $A \rightarrow C \rightarrow F \leftarrow D \leftarrow B$. In first path $A \rightarrow C \leftarrow B$, the triplet of variables is not blocked by C. Thus, A, B is not d-separated by $S := \{C, D\}$.
(d) Assume that we are interested in whether the nodes A and B are d-separated by $S := \{F\}$. We start with the same paths, as in the first case $A \rightarrow C \leftarrow B$ and $A \rightarrow C \rightarrow F \leftarrow D \leftarrow B$. In the first path $A \rightarrow C \leftarrow B$, the triplet of variables is not blocked by F, since F is descendant of C and $F \in S$. Thus, A, B is not d-separated by $S := \{F\}$.
(e) Assume that we are interested in whether F,G are d-separated by $S_1 := \{C\}, S_2 := \{C, D\}, S_3 := \{D, E\}, S_4 := \{B\}$. Then for the S_1 we have: F,G is not d-separated by $\{C\}$,as the path $F \leftarrow D \leftarrow B \rightarrow E \rightarrow G$ is not blocked by C. For the S_2: F,G is d-separated by $\{C, D\}$, as all the paths are blocked. For the S_3: F,G is d-separated by $\{D, E\}$, as all the paths are blocked. Finally for the set S_4: F,G is d-separated by $\{B\}$, as all the paths are blocked.

Definition 1..3 (Topological-Causal ordering). *Given a DAG $\mathcal{G}$, we say that the $\pi \in S_p$ that is a bijective mapping*

$$\pi : \{1, ..., p\} \rightarrow \{1, ..., p\}$$

is a topological-causal ordering of the variables if it is satisfies

$$\pi(i) < \pi(j) \text{ if } j \in \boldsymbol{DE}_i^{\mathcal{G}}$$

Because of the acyclic structure of the DAG, there is always a topological ordering. But this order does not have to be unique.

Proposition 1..3. For each DAG $\mathcal{G}$ there is a topological ordering.

***Proof* 2.** (by induction on n)
Base case: If n=1 (only one node), we have only one topological ordering.
Hypothesis: If $\mathcal{G}$ is a DAG of size $\leq$n, then $\mathcal{G}$ has topological ordering.
Step: Given DAG $\mathcal{G}$ with n+1 nodes, we can find a node v with no incoming edges (proposition 1.1). Then the graph $\mathcal{G}\backslash\{v\}$ is a DAG, as we cannot create cycles by deleting v. By inductive hypothesis, the $\mathcal{G}\backslash\{v\}$ has a topological ordering.Now we can create topological ordering of $\mathcal{G}$ by placing v first and then append topological ordering of $\mathcal{G}\backslash\{v\}$.This is valid since v has no incoming edges.

Example 1..4. Figure 1.12 depicts two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$. $\mathcal{G}_1$ has only one causal ordering π, while the second graph $\mathcal{G}_2$ has more than one. We will show two of them π_1, π_2 :

In $\mathcal{G}_1$ $\pi : 3 \mapsto 1, 1 \mapsto 2, 2 \mapsto 3, 4 \mapsto 4$.

In $\mathcal{G}_2$ π_1: $1 \mapsto 1, 7 \mapsto 2, 2 \mapsto 3, 3 \mapsto 4, 5 \mapsto 5, 4 \mapsto 6, 6 \mapsto 7$

In $\mathcal{G}_2$ π_2: $7 \mapsto 1, 1 \mapsto 2, 2 \mapsto 3, 3 \mapsto 4, 5 \mapsto 5, 4 \mapsto 6, 6 \mapsto 7$

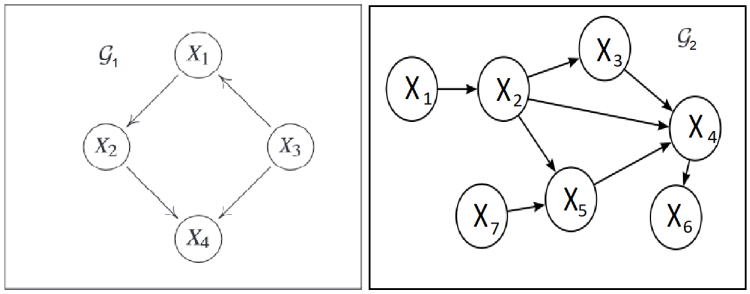


Figure 1.12: $\mathcal{G}_1$ (left) and $\mathcal{G}_2$ (right)

Part I

Graphical Representation of Dependency knowledge

Chapter 2

Introduction to Bayesian Networks

As Pearl refers in his book (Pearl, 2009), p.13, and in paper, (Verma & Pearl, 1990a) p.2 .
“ The role of DAG in probabilistic and statistical modelling is threefold:

1. to provide convenient means of expressing substantive assumption;
2. to facilitate economical representation of joint probability function ;
3. to facilitate efficient inferences from observations. ”

In this chapter we will illustrate the second item.

Let $X_1, ..., X_n$ be n dichotomous variables characterized with joint density function $P(X_1 = x_1, ..., X_n = x_n)$. To store the $P(X_1 = x_1, ..., X_n = x_n)$ explicitly would require a table 2^n entries. For illustration, consider the following example, alteration of (Pearl, 1988):

Example 0..1. Let that we want model the following story under a probabilistic approach: In my city burglary and earthquakes are not uncommon and both can cause the alarm of my house. In case of alarm, two neighbours John and Mary may call.
In this domain problem the dichotomous variables are :

- Burglary (B) with domain $D_B = \{yes, no\}$
- Earthquake (E) » $D_E = \{yes, no\}$
- Alarm (A) » $D_A = \{yes, no\}$
- John Calls (J) » $D_J = \{yes, no\}$
- Mary Calls (M) » $D_M = \{yes, no\}$

To solve any problem under a probabilistic approach we need the knowledge of density function:

$$P(B, E, A, J, M)$$

For example:

B	E	A	J	M	Prob	B	E	A	J	M	Prob
y	y	y	y	y	.00001	n	y	y	y	y	.0002
y	y	y	y	n	.000025	n	y	y	y	n	.0004
y	y	y	n	y	.000025	n	y	y	n	y	.0004
y	y	y	n	n	.00000	n	y	y	n	n	.0002
y	y	n	y	y	.00001	n	y	n	y	y	.0002
y	y	n	y	n	.000015	n	y	n	y	n	.0002
y	y	n	n	y	.000015	n	y	n	n	y	.0002
y	y	n	n	n	.0000	n	y	n	n	n	.0002
y	n	y	y	y	.00001	n	n	y	y	y	.0001
y	n	y	y	n	.000025	n	n	y	y	n	.0002
y	n	y	n	y	.000025	n	n	y	n	y	.0002
y	n	y	n	n	.0000	n	n	y	n	n	.0001
y	n	n	y	y	.00001	n	n	n	y	y	.0001
y	n	n	y	n	.00001	n	n	n	y	n	.0001
y	n	n	n	y	.00001	n	n	n	n	y	.0001
y	n	n	n	n	.00000	n	n	n	n	n	.996

To store exactly the density function $P(B, E, A, J, M)$ we need a table with $2^5 = 32$ entries.[1]

The density distribution contains information of all aspects of the relationships among the variables. Thus we can compute any probability statement, for example: the probability of burglary given that Mary called, $P(B = y|M = y)$.
To compute the $P(B = y|M = y)$ we need the marginal probability of B,M

$$P(B, M) = \sum_{E,A,J} P(B, E, A, J, M)$$

B	M	Prob
y	y	.000115
y	n	.000075
n	y	.00015
n	n	.99966

$$P(B = y|M = y) = \frac{P(B = y, M = y)}{P(M = y)} = \frac{.000115}{.000115 + .00015} = 0.61$$

As a result, if we want to store the density distribution of n dichotomous variables, say $X_1, X_2, .., X_n$,

[1] This example will be used broadly in this chapter

we need a matrix with 2^n entries. An exponential storage! We can overcome the problem of exponential storage size by exploiting conditional independence. Applying the chain Rule of Probabilities we have :

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i | X_1 = x_1, ...X_{i-1} = x_{i-1}) \tag{2.1}$$

Under the above decomposition schema we don't gain any storage. Since, if we sum the number of parameters are required for each factor of the product we will find just the same number of entries. However, if we suppose that the conditional probability of some variable X_i is not dependent to all predecessors, $X_1, .., X_{i-1}$, but only to a small subset of them, we could achieve a decrease in the size of entries. In other words, for every factor $P(X_i|X_1, ..., X_{i-1})$ we need to find a subset of variables $Pa(X_i) \subseteq \{X_1, ..., X_{i-1}\}$ such that: given of $Pa(X_i)$, X_i becoming independent of all variables in $\{X_1, ..., X_{i-1}\} \setminus Pa(X_i)$ i.e.

$$P(X_i|X_1, ..., X_{i-1}) = P(X_i|Pa(X_i))$$

So now on, instead of specifying the probability of X_i conditional to all possible realizations of its predecessors $X_1, .., X_{i-1}$ we need interest only the possible realizations of the set $PA(X_i)$. The set $PA(X_i)$ is called *Markovian-parents* of X_i.

Definition 0..2. Let $\mathcal{X} = \{X_1, .., X_n\}$ be an ordered set of variables, defining in the D_{X_i} consequentiality, with joint distribution $P^{\mathcal{X}}$ and let $p(X_1, ..., X_n)$ be the joint probability density of these variables. A set of variables $PA(X_i)$ is said to be Markovian-parents of X_i if $PA(X_i)$ is a minimal set of predecessors of X_i that renders X_i independent of all its other predecessors. In other words $PA(X_i) \subseteq \{X_1, ..., X_{i-1}\}$ satisfying

$$p(X_i = x_i | X_1 = x_1, ..., X_{i-1} = x_{i-1}) = p(X_i = x_i | Pa(X_i) = Pa(x_i))$$
$$\forall x_i \in D_{x_i},\ \forall (x_1, .., x_{i-1}) \in D_{x_1} \times .. \times D_{X_{i-1}}$$

and such that no proper subset of $PA(X_i)$ satisfy the above equation.

Then the Joint distribution can factorized as:

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i | Pa(X_i)) \tag{2.2}$$

Under this factorization maybe the number of parameters might have been substantially reduced.We can demonstrate that in example 0..1.

Example 0..3.

$$P(B, E, A, J, M) = P(B)P(E|B)P(A|B, E)P(J|B, E, A)P(M|B, E, A, J)$$

We notice that:

- $P(E|B) = P(E)$
- $P(J|B, E, A) = P(J|A)$
- $P(M|B, E, A, J) = P(M|A)$

So,

$$P(B, E, A, J, M) = P(B)P(E)P(A|B, E)P(J|A)P(M|A)$$

Consequently $Pa(B) = \emptyset,\ Pa(E) = \emptyset,\ Pa(A) = \{B, E\},\ Pa(J) = A$ *and* $Pa(M) = A$
With Conditional Probabilities tables:

B	P(B)
Y	.01
N	.99

E	P(E)
Y	.02
N	.98

M	A	P(M|A)
Y	Y	.9
N	Y	.1
Y	N	.05
N	N	.95

J	A	P(J|A)
Y	Y	.7
N	Y	.3
Y	N	.01
N	N	.99

A	B	E	P(A|B, E)
Y	Y	Y	.95
N	Y	Y	.05
Y	Y	N	.94
N	Y	N	.06
Y	N	Y	.29
N	N	Y	.71
Y	N	N	.001
N	N	N	.999

Under this factorization the model size is reduced .

Without making any reference in Theoretic aspect we will illustrate how possible and simple is to represent probabilistic independencies with Directed Graphs. Maybe one naive idea for the construction of a directed graph from a set of probabilistic independences will be drawing an arc from X_j to X_i iff $X_j \in PA(X_i)$. Under this remark in the example 0..1 we have:

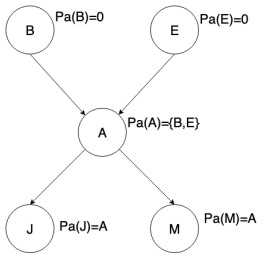


Figure 2.1: DAG of example 0..1

So following this graphical criterion we can construct DAGs. In this particular DAGs each node represent a random variable and arcs represent direct probabilistic dependences. Absence of arc indicates conditional independence. Under the remark $p(X_i|X_1, ..., X_{i-1}) = p(X_i|Pa(X_i))$, which is the basis for the construction of this graphical representations, a variable X_i is conditionally independent of all its non descendants $\{X_1, ..., X_{i-1}\}$ given its parents$Pa(X_i)$.
Thus, the procedure for constructing DAGs under this graphical criterion is the following. First we choose a set of variables that describes the application domain and choose an ordering for the variables. We start with the empty network and add variables to the network one by one according to the ordering. Starting with the pair $\{X_1, X_2\}$ we draw an arrow from X_1 to X_2 if and only if the variables are dependent. Continuing to X_3, we draw no arrow from $\{X_1, X_2\}$ to X_3 in case X_3 is

independent of $\{X_1, X_2\}$; otherwise we examine if X_3 is dependent on X_1 and independent with X_2 or if X_3 is dependent on X_2 and independent with X_1. In the first case we draw an arrow from X_1 to X_3 in the second case from X_2 to X_3, if X_3 depends with X_1 and X_2 then we draw arrows from both $\{X_1, X_2\}$ to X_3. In general in i-th state we find a minimal subset of X_i's predecessor variables, $PA(X_i)$, which makes the X_i independent from all variables already in the network $(X_1, .., X_{i-1})$ such that :

$$p(X_i|X_1, .., X_{i-1}) = p(X_i|Pa(X_i))$$

Then we draw an arc from $Pa(X_i)$ to X_i. The result is a DAG , called Bayesian-Network, in which an arrow from X_i to X_j assigns X_i in the set of Markovian-Parent of X_j. In this procedure we understand the importance of the ordering. For example:

Example 0..4. The Figures of this example is from (Yu, 2008).

In example 0..1 if we choose the ordering B, E, A, J, M we have :
$Pa(B) = \emptyset, Pa(E) = \emptyset, Pa(A) = \{B, E\} Pa(J) = A, Pa(M) = A$
We took the Bayesian Network depicted in Figure 2.2
If we choose the ordering M,J,A,B,E we have :
$Pa(M) = \emptyset, Pa(J) = M, Pa(A) = \{M, J\}, Pa(B) = A, Pa(E) = \{A, B\}$
We took the Bayesian Network depicted Figure 2.3
If we choose the ordering M,J,E,B,A we have :
$Pa(M) = \emptyset, Pa(J) = M, Pa(E) = \{M, J\},$
$Pa(B) = \{M, J, E\}, Pa(E) = \{M, J, E, B\}$
We took the Bayesian Network depicted Figure 2.4

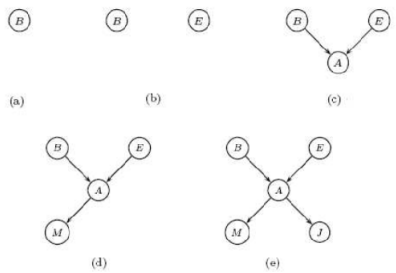


Figure 2.2: Constructed Dag for example 0..1 under the B,E,A,J,M ordering

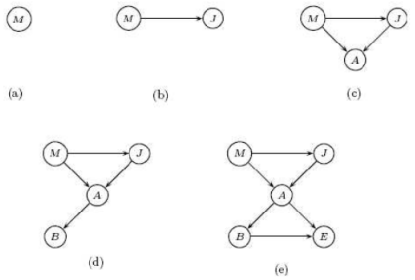


Figure 2.3: Constructed Dag for example 0..1 under the M,J,A,B,E ordering

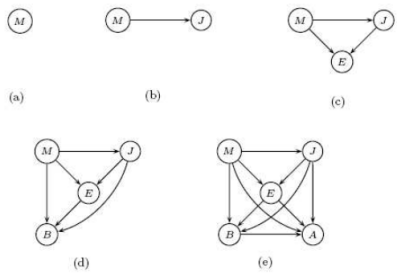


Figure 2.4: Constructed Dag for example 0..1 under the B,E,A,J,M ordering

We therefore conclude that the necessary condition for a DAG to be a Bayesian Network of a probability distribution $P^{\mathcal{X}}$ is p ,density of $P^{\mathcal{X}}$, to admit the product decomposition dictated by $\mathcal{G}$, as given in (2.2).

Definition 0..5. If a probability density p, of a distribution $P^{\mathcal{X}}$, admits the factorization of (2.2) relative to a DAG $\mathcal{G}$, we say that $\mathcal{G}$ represents $P^{\mathcal{X}}$, that $\mathcal{G}$ and $P^{\mathcal{X}}$ are compatible, or that $P^{\mathcal{X}}$ is Markov relative to $\mathcal{G}$.[2]

In example 0..4 we have illustrated the importance of the ordering of the variables. Given different orderings we can create different compatible DAGs with $P^{\mathcal{X}}$,called "Bayesian Networks". The ordering can be used to give us an intuitive interpretation about the meaning of time. The chronological ordering is very important in Causality inference since the causes took place before the effects. Thus given the chronological ordering into the variables we can construct a causal Bayesian network and thus to infer causal relations.

Example 0..6. The density $P(E, B, A, M, J)$ of the example 0..1 admit the factorization 2.2 relative to the below graphs.

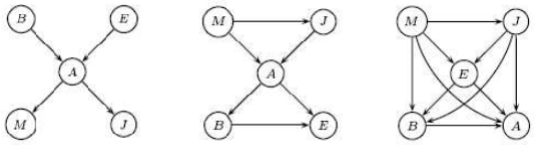


Figure 2.5: Compatible graphs with the distribution $P^{E,B,A,M,J}$

Thus the graphs $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ are compatible with the distribution $P^{E,B,A,M,J}$ of the example 0..1.

Until now we have mentioned a graphical criterion which transform Probabilistic independences into graphs. Under the opposite way of thinking, someone should have asked that given a DAG $\mathcal{G}$ is possible to define a list of (conditional) independences ,"graphical independences". Firstly we need the definition of graphical independences in DAGs.For example in Undirected graphs someone could claim that if in the graph two nodes aren't linked through path this nodes are graphical independent. Secondly we need to examine if this induced graphical independences are the same with the Probabilistic independences of the the compatible distributions $P^{\mathcal{X}}$ with $\mathcal{G}$.

[2]In the appendix of this chapter we use the terminology "G is an I-map of $P^{\mathcal{X}}$."

Example 0..7. In example 0..1 under the representation procedure we take the DAG, depicted in Figure 2.1, which is constructed by the following conditional independence relationships of the distribution $P^{E,B,A,M,J}$:

$$E \perp\!\!\!\perp B, J \perp\!\!\!\perp B|A, J \perp\!\!\!\perp E|A, M \perp\!\!\!\perp B|A, M \perp\!\!\!\perp E|A, M \perp\!\!\!\perp J|A$$

What we need is a procedure to induce independences from the DAG of Figure 2.1 and to examine if the induced graphical independences are satisfied in $P^{E,B,A,M,J}$. Finally we need to examine if there are other distributions $\tilde{P}^{E,B,A,M,J}$ which are compatible with $\mathcal{G}$. Can we model them?

Graphical independences can read from the DAG by using the d-separation criterion.

Interpretation of d-separation

The definition of d-separation can be motivated by regarding DAGs as a representation of causal relationships. Designated a node for every variable and assigning a link between every cause to each of its direct consequences defines a graphical representation of a causal hierarchy, we can think it like a chronological ordering. For example:

Example 0..8. a) To understand the first case of d-separation i.e. what means the nodes A, C in the path: $A \rightarrow B \rightarrow C$ are d-separated by the C, it's wise to see the following example. If "Variable(A)=it's raining", "Variable(B)=The pavement is wet", "Variable(C)=John is slipped on the pavement ". It can be represented by a tree node chain as seen in the Figure 2.6. This means that either rain or wet pavement could cause slipping, however wet pavement is designated as the direct cause; rain could cause someone to slip but not if the pavement is covered. Moreover, knowing the condition of the pavement renders "slipping" and "raining" independent.

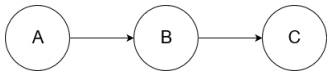


Figure 2.6

b.) We can use the same example to understand the interpretation of the third case of d-separation: What means A and D are not d-separated in the path: $A \rightarrow B \leftarrow D$ by B or any descendant of B. Assume that a "broken pipe" (D) is considered to be another direct cause for wet pavement, as seen in the Figure 2.7. In this case when we know the condition of the pavement i.e. "Wet=Yes" an induced dependency is generated between the two variables that cause the pavement to get wet: "rain" and "broken pipe". Precisely, when the pavement is wet and the "Broken Pipe=Yes" then this information makes the "it's raining" =yes less possible to exist. Although they appear connected in Figure below[3] these propositions are marginally independent and become dependent once we learn that the pavement is wet or that someone broke his leg.

[3] connected means there are a path connecting them.

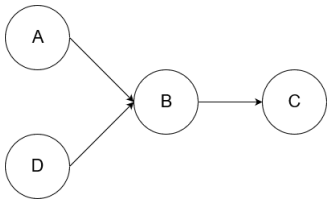


Figure 2.7

c.) Finally for the third case of d-separation e.g. A and D become d-separated in the path $D \leftarrow E \rightarrow A$ by E, is better to examine an other example. If "Variable(E)= Season" "Variable (A)=Crime rate" and "Variable(D)=Ice-cream sales". In Figure 2.8 present a hypothetical example in which the Seasonal variation causally contribute to Ice-cream sales and in Crime-rate. In "season=summer" the Ice-cream sales and the Crime rate are increasing , and the opposite happened when "Season=Winter". So if we haven't any idea about the season variation and we take just data from Ice cream sales and Crime rates per month maybe we find a dependency. However this dependency disappeared when we know the season.

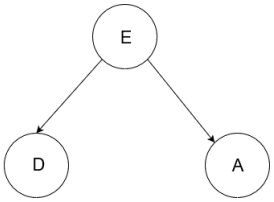
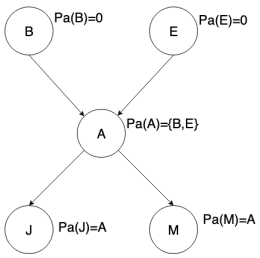


Figure 2.8

From now on we will follow the following notation for the nodes which are d-separated: if A and B are d-separated by C in the graph $\mathcal{G}$ we will denote them as $(A, C, B)_{\mathcal{G}}$ or $A \perp\!\!\!\perp_{\mathcal{G}} B|C$. If two nodes are d-separated by a third we will call them graphical conditional-independent by the third. Also for the Probabilistic conditional independences if X_1 and X_2 are conditional independent given X_3 in the $P^{\mathcal{X}}$ we will denote them as $(X_1, X_3, X_2)_{P^{\mathcal{X}}}$ or $X_1 \perp\!\!\!\perp_{P^{\mathcal{X}}} X_2|X_3$. Examining the graph of the example 0..1:



we take the graphical independences

$$(B, \emptyset, E)_{\mathcal{G}}, (B, A, M)_{\mathcal{G}}, (B, A, J)_{\mathcal{G}}, (B, A, MJ)_{\mathcal{G}}, (E, A, M)_{\mathcal{G}}, (E, A, J)_{\mathcal{G}},$$

$$(E, A, MJ)_{\mathcal{G}}, (M, A, J)_{\mathcal{G}}, (M, AB, J)_{\mathcal{G}}, (M, AE, J)_{\mathcal{G}}, (M, ABE, J)_{\mathcal{G}}$$

This graphical independences we want to include in the compatible distribution $P^{B,E,A,J,M}$ which is used for the construction of the DAG i.e. in our case the distribution of example 0..1.

Theorem 0..9. If sets **X** and **Y** are d-separated by **Z** in a DAG $\mathcal{G}$ then **X** and **Y** are independent by **Z** in every distribution compatible with $\mathcal{G}$. Conversely, if sets **X** and **Y** are not d-separated by **Z** in a DAG $\mathcal{G}$ then there exist at least one distribution compatible with $\mathcal{G}$ which **X** and **Y** are dependent.

Thus, starting with a distribution $P^{\mathcal{X}}$ and construct the compatible DAG $\mathcal{G}$ then every graphical independence in the DAG can be verified by the distribution $P^{\mathcal{X}}$. Moreover the Theorem 0..9 reveals that for a specific DAG $\mathcal{G}$ is possible to find more than one compatible distribution $P^{\mathcal{X}}$ e.g. is possible that we could find more than one different distribution $\tilde{P}^{\mathcal{X}}$ that can construct the $\mathcal{G}$.
A directed consequence of Theorem 0..9 is the following:

Theorem 0..10. For every three disjoint subset of nodes **X,Y,Z** in a DAG and for all probability distribution $P^{\mathcal{X}}$ we have:

1. $(\mathbf{X},\mathbf{Z},\mathbf{Y})_{\mathcal{G}} \Longrightarrow (\mathbf{X},\mathbf{Z},\mathbf{Y})_{P^{\mathcal{X}}}$ whenever the $\mathcal{G}$ and $P^{\mathcal{X}}$ are compatible
2. if $(\mathbf{X},\mathbf{Z},\mathbf{Y})_{P^{\mathcal{X}}}$ holds in all distributions compatible of $\mathcal{G}$, this follows that $(\mathbf{X},\mathbf{Z},\mathbf{Y})_{\mathcal{G}}$.

For example:

Example 0..11. As we see from the graph of the Figure 2.2 is induced the above graphical independences:

$$(B, \emptyset, E)_{\mathcal{G}}, (B, A, M)_{\mathcal{G}}, (B, A, J)_{\mathcal{G}}, (B, A, MJ)_{\mathcal{G}}, (E, A, M)_{\mathcal{G}}, (E, A, J)_{\mathcal{G}},$$

$$(E, A, MJ)_{\mathcal{G}}, (M, A, J)_{\mathcal{G}}, (M, AB, J)_{\mathcal{G}}, (M, AE, J)_{\mathcal{G}}, (M, ABE, J)_{\mathcal{G}}$$

Thus every Distribution $P_{B,E,A,J,M}$(set $\mathcal{X} = \{B, E, A, J, M\}$) is compatible with the graph, Figure 2.2, only if satisfy the above independences i.e. if satisfy:

$$(B, \emptyset, E)_{P^{\mathcal{X}}}, (B, A, M)_{P^{\mathcal{X}}}, (B, A, J)_{P^{\mathcal{X}}}, (B, A, MJ)_{P^{\mathcal{X}}}, (E, A, M)_{P^{\mathcal{X}}}, (E, A, J)_{P^{\mathcal{X}}},$$

$$(E, A, MJ)_{P^{\mathcal{X}}}, (M, A, J)_{P^{\mathcal{X}}}, (M, AB, J)_{P^{\mathcal{X}}}, (M, AE, J)_{P^{\mathcal{X}}}, (M, ABE, J)_{P^{\mathcal{X}}}$$

The connection between d-separation and conditional independence is established through the Theorem 0..9 and become clear the importance of d-separation criterion. A convenient way of characterizing the set of distributions compatible with a DAG $\mathcal{G}$ is to list the set of (conditional) independences that must be satisfied by each compatible distribution.
Now I want to mention two basic points:

- Firstly from the Theorem 0..10 and the example 0..11 each compatible distribution $P^{\mathcal{X}}$ with the graph $\mathcal{G}$ can satisfy more independences than those induced from the graph i.e.

$$\{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{P^{\mathcal{X}}}\} \supseteq \{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{\mathcal{G}}\}$$

- Secondly the ordering which the DAG is constructed does not enter into the definition criterion of the Theorem 0..10; it is only the topology [4] of the resulting graph that determines the set of independences that the probability $P^{\mathcal{X}}$ must satisfy. Under that note the above Theorem following normally.

Theorem 0..12. A necessary and sufficient condition for a probability distribution $P^{\mathcal{X}}$ to be compatible or Markov relative to a DAG $\mathcal{G}$ is that every variable be independent (probabilistic independent in $P^{\mathcal{X}}$) of all its non-descendants in $\mathcal{G}$, conditional on its parents(excluding herself).

As we mention at the beginning of this chapter DAGs facilitate economical representation of joint probability function. This idea is a result of graphical representation of independences. Precisely given a distribution function $P^{\mathcal{X}}$ we generate a DAG $\mathcal{G}$ which is compatible to $P^{\mathcal{X}}$. Using the d separation in this graph we can induce graphical independences which are part of the Probabilistic independences in the compatible $P^{\mathcal{X}}$. i.e.

$$\{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{P^{\mathcal{X}}}\} \supseteq \{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{\mathcal{G}}\}$$

Now the question which is raised normally is that in which cases the opposite way of the above inequality holds?

$$\{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{P^{\mathcal{X}}}\} \subseteq \{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{\mathcal{G}}\}$$

and thus

$$\{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{P^{\mathcal{X}}}\} = \{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{\mathcal{G}}\}$$

In the next chapter we will see that this equation can verified only under two basic assumptions the Markov Property and the Faithfulness.

[4] The shape of the graph i.e. his directed arrows and his skeleton

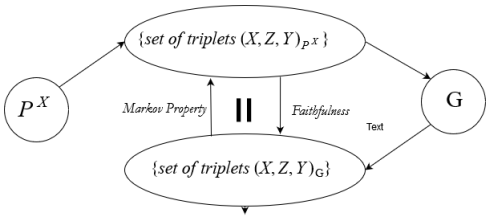
Chapter 3

Markov Property-Faithfulness-Causal Minimality

In the previous chapter we gave a brief introduction about the Bayesian Networks and how these are associated with distribution functions $P^{\mathcal{X}}$. The aim of this chapter is to answer the question when the equation:

$$\{set\ of\ triplets\ (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{P^{\mathcal{X}}}\} = \{set\ of\ triplets\ (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{\mathcal{G}}\} \tag{3.1}$$

holds. For that reason we will illustrate two basic assumption, the Markov property and Faithfulness. If both assumptions hold, then every independence in the Distribution function $P^{\mathcal{X}}$ can be validated from the graph $\mathcal{G}$ and vice verse.



Thus, if the equation (3.1) holds for a $P^{\mathcal{X}}$ and a graph $\mathcal{G}$ then every graphical independence in the graph is also probabilistic independence in $P^{\mathcal{X}}$. This result is very important since we can use the graph $\mathcal{G}$ to extract independences for $P^{\mathcal{X}}$ with out any mathematical calculation.
As a result Pearl used that assumptions to develop his Theory about Causal inference. More precisely, Pearl used Direct graphs under that assumptions to develop transparent and clear justifications for the three basic methods[1] for estimating causal effects that we will present in this dissertation: 1.) conditioning on variables to eliminate non causal associations by blocking all relevant

[1] We will see them in Chapter *Calculating Interventions*.

back-door paths from the causal variable, 2.) conditioning on variables that allow for estimation by a mechanism, 3.) and using an instrumental variable that is an exogenous shock to the causal variable in order to consistently estimate its effect.

1. Markov Property

The Markov property is a commonly used assumption that forms the basis for graphical models.[2] When a distribution is Markovian with respect to a graph, this graph encodes certain independences in the distribution that we can exploit for efficient computation or data storage. The Markov property exists for both directed and undirected graphs and it is well known that these two classes encode different sets of independencies (see chapters Graphical Representation part (I & II)). In causal inference,however, we are mainly interested in directed graphs. While many introductions to causal inference start with the Markov property as the underlying assumption, we will derive it as a property of the graphs.

Definition 1..1. [*Markov property*]
Let $\mathcal{X} = \{X_1, .., X_n\}$ set of variables with a joint distribution function $P^{\mathcal{X}}$ and a DAG $\mathcal{G} = \{\mathcal{X}, E\}$. This distribution is said to satisfy

- the global Markov property with respect to the DAG $\mathcal{G}$ if

$$A, B, \text{ d-separated by } C, A \perp\!\!\!\perp_{\mathcal{G}} B|C \Rightarrow A \perp\!\!\!\perp_{P^{\mathcal{X}}} B|C$$

  for all disjoint sets A , B , C .

- the local Markov property with respect to the DAG $\mathcal{G}$ if each variable is independent of its non-descendants given its parents, and

- the Markov factorization property with respect to the DAG $\mathcal{G}$ if

$$p(\mathbf{x}) = p(x_1, ..., x_n) = \prod_{j=1}^{n} p(x_j | x_{pa_j^{\mathcal{G}}})$$

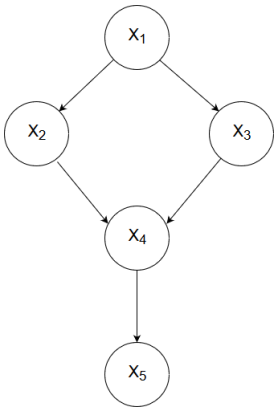
  Let assume that$P^{\mathcal{X}}$ has a density p . *In this Dissertation, we always consider densities with respect to a product measure.*

Theorem 1..2. If $P^{\mathcal{X}}$ has a density p with respect to product measure ,then all Markov properties in Definition 1..1 are equivalent .

***Proof* 3.** The proof can be found in [Lauritzen, 1996] Theorem 3.27 .

Example 1..3. A distribution $P^{\mathcal{X}}$ *with* $\mathcal{X} = \{X_1, .., X_5\}$ is Markov with respect to the graph $\mathcal{G} = (\mathbf{X}, E)$:

[2] Generally we illustrate the Markov Property in the previous example when we examine the compatible distributions in DAGs.



if, according to global and local Markov properties,

$$\begin{cases} X_1 \text{ and } X_4 \text{ d-sep by } \{X_2, X_3\} \Rightarrow X_4 \perp\!\!\!\perp_{P^{\mathcal{X}}} X_1 | X_2, X_3 \\ X_2 \text{ and } X_3 \text{ d-sep by } X_1 \Rightarrow X_2 \perp\!\!\!\perp_{P^{\mathcal{X}}} X_3 | X_1 \\ X_5 \text{ and } X_2, X_3 \text{ d-sep by } X_4 \Rightarrow X_5 \perp\!\!\!\perp_{P^{\mathcal{X}}} X_2, X_3 | X_4 \\ X_5 \text{ and } X_1 \text{ d-sep by } X_4 \Rightarrow X_5 \perp\!\!\!\perp_{P^{\mathcal{X}}} X_1 | X_4 \\ X_5 \text{ and } X_1, X_2, X_3 \text{ d-sep by } X_4 \Rightarrow X_5 \perp\!\!\!\perp_{P^{\mathcal{X}}} X_1, X_2, X_3 | X_4 \end{cases}$$

or, according to Markov factorization property if p is the density of $P^{\mathcal{X}}$

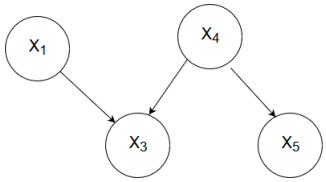
$$p(x_1, ..., x_5) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_4)$$

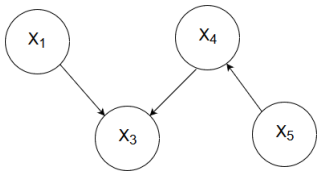
Generally speaking, if a distribution $P^{\mathcal{X}}$ satisfies one of the the above Markov properties with respect to a graph $\mathcal{G}$ then the graph encodes some of the independences of the distribution $P^{\mathcal{X}}$ i.e.

$$\{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{P^{\mathcal{X}}}\} \supseteq \{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{\mathcal{G}}\}$$

The Markov condition relates statements about graph separation to conditional independences. Now it's possible to ask the question: *Can different graphs encode exact the same set of conditional independences?* The answer is given in the examples 1..4, 1..5 . Each of the following two examples contains two graphs. The graphs share exactly the same set of d-separations. Under Markov assumption in these graphs are encoded the same conditional independences :

Example 1..4. .

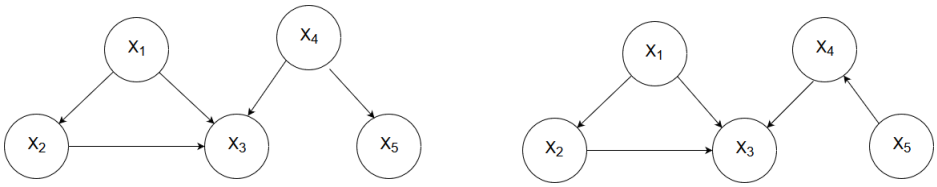




These two graphs induce the same following graphical independences

$$\begin{cases} X_1 \text{ and } X_4 \text{ d-sep by } \varnothing \\ X_1 \text{ and } X_5 \text{ d-sep by } \{\varnothing\}, \{X_4\} \\ X_3 \text{ and } X_5 \text{ d-sep by } X_4 \end{cases}$$

Example 1..5. .



$$\begin{cases} X_1 \text{ and } X_4 \text{ d-sep by } \varnothing, X_2 \\ X_1 \text{ and } X_5 \text{ d-sep by } \varnothing, X_4 \\ X_2 \text{ and } X_4 \text{ d-sep by } \varnothing, X_1 \\ X_2 \text{ and } X_5 \text{ d-sep by } \varnothing, X_4, X_1, \{X_1, X_4\} \\ X_3 \text{ and } X_5 \text{ d-sep by } X_4 \end{cases}$$

These graphs share the same conditional independences however these graphs shares some additional properties. As we have mentioned in the previous chapter given a graph we can find more than one, Markovian, distribution which verify the induced list of d-separations. In the following definition we will see that these graphs share the same Markovian distributions too.

Definition 1..6. [*Markov equivalence of graphs*]
We denote by $\mathcal{M}(\mathcal{G})$the set of distributions $P^{\mathcal{X}}$ that are Markovian with respect to $\mathcal{G}$ i.e.:

$$\mathcal{M}(\mathcal{G}) := \{P^{\mathcal{X}} : P^{\mathcal{X}} \textit{satisfies the global (or local) Markov property with respect to } \mathcal{G}\}$$

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$. This is the case if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy the same set of d-separations, which means the Markov condition entails the same set of (conditional) independence conditions. The set of all DAGs that are Markov equivalent to some DAG is called **Markov equivalence class** of $\mathcal{G}$. It can be represented by a PDAG that is denoted by CPDAG($\mathcal{G}$)=(V,$\mathcal{E}$). It contains the (directed) edge $(i, j) \in \mathcal{E}$ if and only if one member of the Markov equivalence class does.

From the above definition we understand that it is not a trivial procedure to determine when two DAGs are equivalent. Since doing something like that we need to specify precisely all the classes of distributions of the $\mathcal{M}(\mathcal{G})$ which is very difficult. However (Verma & Pearl, 1991) provide a simpler graphical characterization.

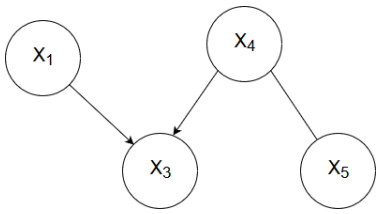
Lemma 1..7. [**Graphical criteria for Markov equivalence**] Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same skeleton and the same immoralities.

***Proof* 4.** (Verma & Pearl, 1991) provide a concise characterization, see also (Frydenberg, 1990).

Example 1..4 shows two Markov equivalent graphs .The graphs share the same skeleton and both of them have only one immorality

$$X_1 \rightarrow X_3 \leftarrow X_4$$

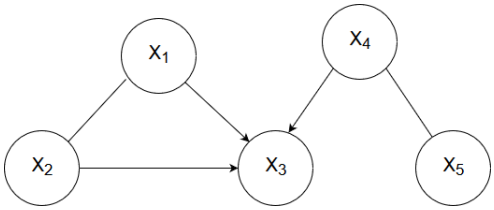
and then the corresponding CPDAG



Example 1..5 shows two Markov equivalent graphs. The graphs share the same skeleton and both of them have two immoralities.

$$X_1 \rightarrow X_3 \leftarrow X_4$$

$$X_2 \rightarrow X_3 \leftarrow X_4$$



But in the next example we must be very careful :

Example 1..8. Two Markov equivalent DAGs (a), (b) in Figure 3.1 are the only two DAGs in the corresponding Markov equivalence class that can be represented by the CPDAG (c) Figure 3.1. The graphs share the same skeleton and both of them have only one immorality:

$$X_1 \rightarrow X_3 \leftarrow X_4$$

But in the CPDAG is required to add some extra arrows for example the arrow $X_2 \leftarrow X_3$ is required to avoid a new v-structure, the $X_2 \rightarrow X_3 \leftarrow X_4$. Furthermore, $X_1 \rightarrow X_2$ prevents the existence of a directed cycle.

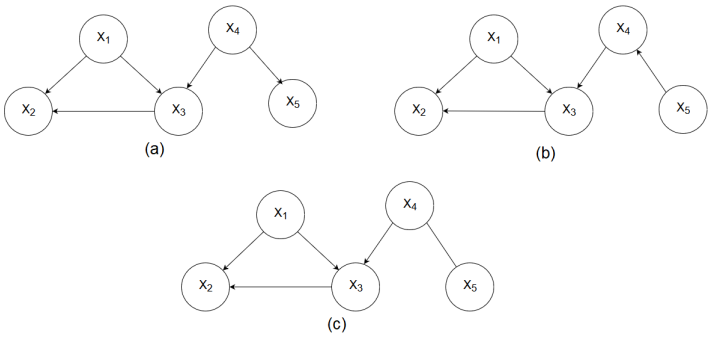


Figure 3.1

Markov property has been proven to be a very useful tool. If we focus in the second Markov property (local), each variable is independent of its non descendants, given its parents. Keeping that in our mind we introduce the graphical concept of a Markov blanket (Pearl, 1988). That becomes relevant when one tries to predict the value of a variable Y from the observed values of all the other variables. One may then wonder what would be the smallest set of variables whose knowledge renders the remaining ones irrelevant for the prediction task.

Definition 1..9. [*Markov blanket*]

Consider a DAG $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ and a node Y. The Markov blanket of Y is the smallest set of variables, $\mathbf{M} \subseteq \mathcal{X}$, such that :

$$Y \; and \; \mathcal{X} \backslash (Y \cup \mathbf{M}) \; d-seterated \; by \; \mathbf{M} \; or \; Y \perp\!\!\!\perp_{\mathcal{G}} \mathcal{X} \backslash (Y \cup \mathbf{M}) | \mathbf{M}$$

If $P^{\mathcal{X}}$ is Markovian with respect to $\mathcal{G}$,then

$$Y \perp\!\!\!\perp_{P^{\mathcal{X}}} \mathcal{X} \backslash (Y \cup \mathbf{M}) | \mathbf{M}$$

Thus, this tool gives us a very powerful technique in regression. In an idealized regression setting we would only need to include the variables in **M** for predicting Y. In other words, if we know the variables in Markov blanket **M**, the other variables do not provide any further information when we predicting the variable Y.

Proposition 1..10. [Markov blanket II]
Consider a DAG $\mathcal{G}$ and a target node Y . Then, the Markov blanket **M** of Y includes its parents, its children, and the parents of its children

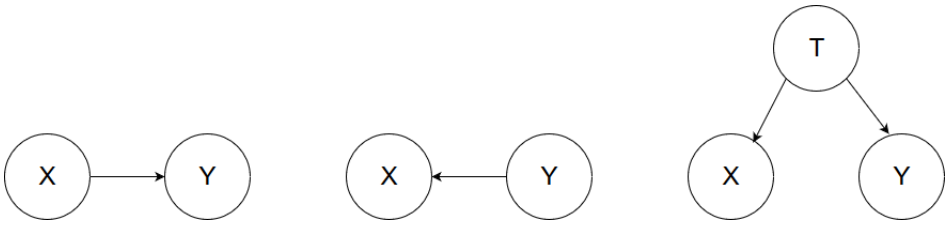
$$\mathbf{M} = PA_Y \cup CH_Y \cup PA_{CH_Y}$$

Until now in the dissertation we have examined and illustrated how to store Probabilistic Knowledge in DAGs i.e. we use graphs as tool for storage Probabilistic independences. Nonetheless, we haven't done any reference to causation. As we will see in the next remark, the graphical repre-

sentation of Probabilistic knowledge is a very useful tool for the illustration of Causality Inference. In 1956 Reichenbach formulated the *Reinchenbach common cause principle*. This principle established a link between statistical properties and causal structures.

Remark 1..11. When the random variables X and Y are dependent, there must be a causal explanation for this dependence:

1. X is (possibly indirectly) causing Y, or
2. Y is (possibly indirectly) causing X, or
3. there is a (possibly unobserved) common cause T that (possibly indirectly) causes both X andY.



Reichenbach common cause principle can be clearly confirmed with the theory of Markov property.

Proposition 1..12. Assume that any pair of variables X and Y can be embedded into a larger system in the following sense: there exists a DAG $\mathcal{G}$ over the collection $\mathcal{X}$ of random variables that contains X and Y. Then the Reichenbach's common cause principle derives from the Markov property: If X and Y are dependent in $P^{\mathcal{X}}$, then there is

- ► either a directed path from X to Y
- ► or from Y to X
- ► or there is a node T with directed path from T to X and from T to Y .

***Proof* 5.** Due to the Markov property, the dependence implies that $\mathcal{G}$ contains a path between X and Y. Otherwise, X,Y would be d-separated by $\emptyset$ and thus X,Y would be independent . Also, the path between X and Y cannot contain a collider, otherwise it would be blocked by the empty set again. The statement can be assumed since any path between X and Y without collider must be of the form

$$X \to ... \to Y$$

$$X \leftarrow ... \leftarrow Y$$

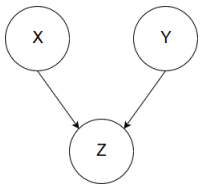
$$X \leftarrow ... \leftarrow T \rightarrow ... \rightarrow Y$$

The dependence between two variables may also arise for a reason different from the ones mentioned in the common cause principle. In real applications, we must be very careful because in Reichenbach's principle, we start with two dependent random variables and obtain a valid statement, but in real statistical experiments, the dependence can appeared in many ways. For example, conditioning in third variables, is something that in literature is called *selection bias*. This can lead to Paradoxes like Berkson's paradox (Berkson, 1947). In this case the dependence between X and Y arise from none of the three conditions of Reichenbach common cause.

Example 1..13. Let X, Y, Z three variables with joint distribution $P^{X,Y,Z}$ and if we assume:

$$\begin{cases} X = N_X \\ Y = N_Y \\ Z = min(X, Y) \end{cases}$$

where $N_X, N_Y \sim Ber(0.4)$. and the graph $\mathcal{G}$



The distribution $P^{X,Y,Z}$ is Markov with respect to graph $\mathcal{G}$ because

$$X \perp\!\!\!\perp_{\mathcal{G}} Y \Rightarrow X \perp\!\!\!\perp_{P^{X,Y,Z}} Y$$

As we can see from the model above, X and Y are independent

$$P(X = 0|Y = 0) = P(X = 0|Y = 1) = 0.6 \text{ and}$$

$$P(X = 1|Y = 0) = P(X = 1|Y = 1) = 0.4$$

However if we condition on Z=1

$$P(X = 1, Y = 1|Z = 1) = P(X = 1, Y = 1|min(X, Y) = 1) =$$

$$P(X = 1, Y = 1|Y = 1, X = 1) = 1 \neq$$

$$\neq P(X = 1, Y = 1) = P(X = 1)P(Y = 1) = 0.16$$

$$X \not\perp\!\!\!\perp Y|Z = 1 \Rightarrow X \not\perp\!\!\!\perp Y|Z$$

As a result X is not independent of Y given Z and non of the three cases of the Reichenbach holds.

2. Faithfulness and causal minimality

In the previous subsection, we discussed the Markov property, which help us when we read off independences from a graph structure. More precisely :

$$\{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{P^{\mathcal{X}}}\} \supseteq \{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{\mathcal{G}}\} \tag{3.2}$$

In this section we will examine in which cases the other direction of the inequality (3.2) holds. This can be achieved if we assume a further assumption the "*Faithfulness*". As a result if we assume the faithfulness and the Markov assumptions we achieve the equality:

$$\{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{P^{\mathcal{X}}}\} = \{\textit{set of triplets } (\mathbf{X}, \mathbf{Z}, \mathbf{Y})_{\mathcal{G}}\}$$

then every **non** graphical independence in the DAG will mean Probabilistic Dependency in $P^{\mathcal{X}}$.Thus with "Faithfulness" assumption we can infer **dependences** from the graph structure.

Definition 2..1. Let $\mathcal{X} = \{X_1, ..., X_n\}$ set of variables with joint distribution P^X and a DAG $\mathcal{G} = (\mathcal{X}, \mathcal{E})$.

- $P^{\mathcal{X}}$is faithful to the DAG $\mathcal{G}$ if

$$A \perp\!\!\!\perp_{P^{\mathcal{X}}} B|C \Rightarrow A \textit{ and } B \textit{ } d-\textit{separated by } C \textit{ or } A \perp\!\!\!\perp_{\mathcal{G}} B$$

  for all disjoint vertex sets A , B , C.

- A distribution $P^{\mathcal{X}}$ satisfies causal minimality with respect to $\mathcal{G}$ if it is Markov with respect to $\mathcal{G}$, but not to any proper subgraph of $\mathcal{G}$.

Faithfulness gives us exactly the opposite side of global Markov property

$$A \textit{ and } B \textit{ d-separated in } \mathcal{G} \textit{ by } C \Rightarrow A \perp\!\!\!\perp B|C \textit{ in } P^{\mathcal{X}}$$

In the following examples, we start with two joint distributions $P^{\mathcal{X}_1}$, $P^{\mathcal{X}_2}$ inducing the same independences. The distributions in both examples satisfy Markov condition and causal minimality with respect to corresponding graph, but not the faithfulness in the first example.

Example 2..2. Consider the joint distribution $P_1^{X,Y,Z}$ which is generated if:

$$\begin{cases} X := N_X \\ Y := aX + N_Y \\ Z := bY + cX + N_Z \end{cases}$$

Where $N_i \sim N(0, \sigma_i), \quad i \in \{X, Y, Z\}$ jointly independent,
and the graph:

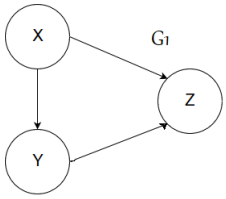


Figure 3.2

For simplicity reasons let $\mathcal{X} := \{X, Y, Z\}$ $\mathcal{G}_1$ is Markov relative to $P_1^{\mathcal{X}}$ since from the graph don't induce any graphical independences. Also the graph satisfies the causal minimality property since if we delete any arrow we lose the Markov property for example if we delete the arrow $X \to Y$ then is generated the $X \perp\!\!\!\perp_{\mathcal{G}} Y$ so it violates the Markov property. Additionally, the Faithfulness is satisfied too.
Now, we tune the parameters of the $P_1^{\mathcal{X}}$ like

$$ab + c = 0$$

With the tuning, we make two paths that cancel each other and create an independence that is not implied by the graph structure. *Spirtes et al. [2000, Theorem 3.2]* .

$$Z = bY + cX + N_z = bY - abX + N_z = b(Y - aX) + N_z = N_Y + N_Z$$

Then we have :

$$\begin{cases} X := N_X \\ Y = aX + N_Y \\ Z = bN_Y + N_Z \end{cases}$$

$$Cov(X, Z) = Cov(N_X, bN_Y + N_Z) = Cov(N_X, bN_Y) + Cov(N_X, N_Z) = 0$$

$$X \perp\!\!\!\perp Z$$

The tuning distribution P_1^X is not faithful with respect to $\mathcal{G}_1$ (figure 3.2)since we obtain X $\perp\!\!\!\perp$ Z, but satisfies the Markov property since from the graph isn't induced any d-separation requirement.

So if we assume faithfulness in $P_1^{\mathcal{X}}$ with respect to graph $\mathcal{G}_1$ then the tuning $ab = c$ is prohibited because it generates a new independence in $P_1^{\mathcal{X}}$. Under this way of thinking ,if we assume faithfulness then it is forbidden to have $a = 0$, $b = 0$ or $c = 0$. As a result the $P_1^{\mathcal{X}}$is faithful with respect to $\mathcal{G}_1$ if and only if $a, b, c, \neq 0, ab \neq -c$.

Example 2..3. Let three variables X,Y,Z with joint distribution $P_2^{X,Y,Z}$ which is determined by :

$$\begin{cases} X := \tilde{N_X} \\ Y = \tilde{a}X + \tilde{b}Z + \tilde{N_Y} \\ Z = \tilde{N_Z} \end{cases}$$

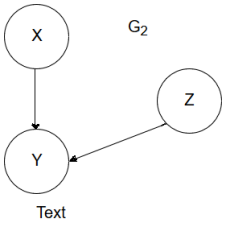
with $\tilde{N}_i \sim N(0, t_i^2)$ jointly independent.
If

$$t_i := \begin{cases} t_X := \sigma_X^2 \\ t_Y = \sigma_Y^2 - \frac{b^2\sigma_Y^4}{b^2\sigma_Y^2+\sigma_Z^2} \\ t_Z = b^2\sigma_Y^2 + \sigma_Z^2 \end{cases}$$

and

$$\begin{cases} \tilde{a} = a \\ \tilde{b} = \frac{b\sigma_Y^2}{b^2\sigma_Y^2+\sigma_Z^2} \end{cases}$$

and let the graph $\mathcal{G}_2$:



The joint distribution is multi-normal $N_k(\mu, \Sigma)$ .To describe a multi-normal distribution we need:

$$Mean = \mu, \textit{Covariance-matrix} = \Sigma, Dimension = k$$

In our case

$$\mu = (0, 0, 0) \text{ and k=3}$$

and

$$\Sigma := \begin{pmatrix} Var(X) & Cov(X,Y) & Cov(X,Z) \\ Cov(Y,X) & Var(Y) & Cov(Y,Z) \\ Cov(Z,X) & Cov(Z,Y) & Var(Z) \end{pmatrix}$$

$$Var(X) = Var(\tilde{N_X}) = \sigma_X^2$$

$$Var(Z) = Var(\tilde{N_Z}) = b^2\sigma_Y^2 + \sigma_Z^2$$

$$Var(Y) = Var(\tilde{a}X + \tilde{b}Z + \tilde{N_Y}) = \tilde{a^2}\sigma_X^2 + \tilde{b^2}Var(Z) + \sigma_Y^2 - \frac{b^2\sigma_Y^4}{b^2\sigma_Y^2 + \sigma_Z^2}$$

$$= \tilde{a^2}\sigma_X^2 + \tilde{b^2}(b^2\sigma_Y^2 + \sigma_Z^2) + \sigma_Y^2 - \frac{b^2\sigma_Y^4}{b^2\sigma_Y^2 + \sigma_Z^2} =$$

$$= a^2\sigma_X^2 + \frac{b^2\sigma_Y^4}{(b^2\sigma_Y^2 + \sigma_Z^2)} + \sigma_Y^2 - \frac{b^2\sigma_Y^4}{b^2\sigma_Y^2 + \sigma_Z^2} = a^2\sigma_X^2 + \sigma_Y^2$$

$$Cov(X,Y) = Cov(\tilde{N_X}, \tilde{a}X + \tilde{b}Z + \tilde{N_Y}) = Cov(\tilde{N_X}, \tilde{a}X + \tilde{b}\tilde{N_Z} + \tilde{N_Y}) = \tilde{a}Var(\tilde{N_X})$$

$$= a\sigma_X^2$$

$$Cov(X,Z) = Cov(\tilde{N_X}, \tilde{N_Z}) = 0$$

$$Cov(Y,Z) = Cov(\tilde{a}X + \tilde{b}Z + \tilde{N_Y}, \tilde{N_Z}) = Cov(\tilde{a}\tilde{N_X} + \tilde{b}\tilde{N_Z} + \tilde{N_Y}, \tilde{N_Z}) =$$

$$= \tilde{b}Var(\tilde{N_Z}) = \frac{b\sigma_Y^2}{b^2\sigma_Y^2 + \sigma_Z^2}(b^2\sigma_Y^2 + \sigma_Z^2) = b\sigma_Y^2$$

$$\Sigma := \begin{pmatrix} \sigma_X^2 & a\sigma_X^2 & 0 \\ a\sigma_X^2 & a^2\sigma_X^2 + \sigma_Y^2 & b\sigma_Y^2 \\ 0 & b\sigma_Y^2 & b^2\sigma_Y^2 + \sigma_Z^2 \end{pmatrix}$$

After all these calculations, we can conclude that the joint distributions after tuning P_1^X example 2..2 and P_2^X example 2..3 are just the same (multi-normal with the same μ and Σ).
It can be checked that the distributions is faithful with respect to $\mathcal{G}_2$ if $\tilde{a}, \tilde{b} \neq 0$ and $t \neq 0$. The joint distribution P_2^X is faithful with respect to $\mathcal{G}_2$ but not with respect to $\mathcal{G}_1$,the graph in example 2..2. Nevertheless, for both models, causal minimality is satistisfied if none of the parameters vanished: the distribution is not Markov to any proper subgraph of $\mathcal{G}_1$ or $\mathcal{G}_2$ since, removing an arrow , would correspond to a new (conditional) independence that does not hold in the distribution.

Proposition 2..4. If P_X is faithful and Markovian with respect to $\mathcal{G}$, then causal minimality is satisfied

***Proof* 6.** If P_X is Markovian with respect to a proper subgraph $\tilde{\mathcal{G}}$ of $\mathcal{G}$, there are two nodes that are directly connected in $\mathcal{G}$ but not in $\tilde{\mathcal{G}}$ of$\mathcal{G}$. Thus, these nodes are d-separeted by $\emptyset$ in new $\tilde{\mathcal{G}}$ and (from Markov condition) imply the corresponding conditional independence statement in P_X , but now P^X cannot be faithful with respect to $\mathcal{G}$.

Proposition 2..5. Consider the random vector $X = (X_1, ..., X_p)$ and assume that the joint distribution P^X has a density with respect to a product measure. Suppose that P^X is Markov with respect to $\mathcal{G}$.Then P^X satisfies causal minimality with respect to $\mathcal{G}$ if and only if $\forall X_j \forall Y \in PA_j^{\mathcal{G}}$ we have that

$$X_j \not\perp\!\!\!\perp Y | PA_j^{\mathcal{G}} / \{Y\}$$

***Proof* 7.** [$\Leftarrow$]If $X_j \not\perp\!\!\!\perp Y | PA_j^{\mathcal{G}}$ .Assume that causal minimality is not satisfied, then we can find a pair of variables $X_j, Y \in PA_j^{\mathcal{G}}$: P_X Markov with respect to proper sub graph $\tilde{\mathcal{G}}$, $\mathcal{G}$ without the link $Y \to X_j$, then X_j, Y d-separated by PA_j in $\tilde{\mathcal{G}} \Rightarrow X_j \perp\!\!\!\perp Y | PA_j^{\mathcal{G}}$
[$\Rightarrow$] Let P^X satisfy causal minimality with respect to $\mathcal{G}$ and assume $X_j \perp\!\!\!\perp Y | PA_j^{\mathcal{G}}$ .Under the assumpion $X_j \perp\!\!\!\perp Y | PA_j^{\mathcal{G}} \Rightarrow p(X_j | PA_j) = p(X_j | PA_j \backslash \{Y\})$.
If P_X has a density, the Markov condition is equivalent to the Markov factorization .

$$p(x) = p(x_j | pa_j \backslash \{y\}) \prod_{i \neq j} p(x_i | pa_i)$$

which implies that P^X is markovian with respect to $\mathcal{G}$, without the link $Y \to X_j$ thus causal minimality is not satisfing.

Part II

Causality Inference with Structural Causal Models

Chapter 4

Structural causal models

In order to deal rigorously with questions of causality, we must have a way to formally set down our assumptions about the causal story behind a data set. To do so ,we introduce Structural causal models (SCMs), also known as (non-parametric) structural equation models (SEMs). In the literature, SEMs have been used for a long time before causality inference in fields like genetics (S. Wright, 1921), econometrics (Haavelmo, 1943), electrical engineering (Mason, 1956); (Mason, 1953) and social sciences (Goldberger & Duncan, 1973); (Duncan, 1975). In this dissertation , we deal with Acyclic SCMs, a special well studied class of SCMs that is closely related to causal Bayesian networks (Pearl, Glymour, & Jewell, 2016). A structural causal model is the attempt to describe the mechanism which nature assigns values to variables of interest.

As we will see in this chapter the purpose of causal inference is to connect graphs that represent causal relations among variables with their joint probability distribution. (S. Wright, 1921) attempt to illustrate this idea using the following example.
He tried to formulate the question how someone can express mathematically the common understanding that symptoms do not cause diseases. Wright for that pusrope used a combination of equations and graphs. For example, if X stands for the desiease and Y stands for a certain symptom of the disease, Wright would write a linear equation

$$y = \beta x + u_Y \tag{4.1}$$

where x stands for the level (or severity) of the disease, y stands for the level (or severity) of the symptom, and u_Y stands for all factors, other than the disease in question, that could possibly affect Y when X is held constant,$X = x$. In interpreting this equation one should think of a physical process whereby Nature examines the values of x and u_Y and, accordingly, assigns variable Y the value $y = \beta x + u_Y$. Similarly, to "explain" the occurrence of disease X, one could write $x = u_X$ where U_X stands for all factors affecting X.
Equation 4.1 still does not properly express the causal relationship implied by this assignment pro-

cess, because algebraic equations are symmetrical objects; if we rewrite 4.1 as

$$x = \frac{y - u_Y}{\beta} \tag{4.2}$$

it might be misinterpreted to mean that the symptom influences the disease. To express the directionality of the underlying process, Wright augmented the equation with a diagram, later called 'path diagram,' in which arrows are drawn from (perceived) causes to their (perceived) effects, and more importantly, the absence of an arrow makes the empirical claim that Nature assigns values to one variable irrespective of another. In Figure 4.1, for example, the absence of arrow from Y to X represents the claim that symptom Y is not among the factors U_X which affect disease X. Thus, in our example, the complete model of a symptom and a disease would be written as in Figure 4.1 The diagram encodes the possible existence of (direct) causal influence of X on Y, and the absence of causal influence of Y on X, while the equations encode the quantitative relationships among the variables involved, to be determined from the data.

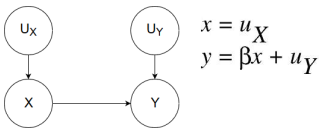


Figure 4.1: Path Diagram for the Wright's example

The parameter β in the equation is called a "path coefficient" and it quantifies the (direct) causal effect of X on Y .
In the previous example we examine the raw idea of Wright in this case where X,Y are linked with linear connection. If we extend this idea we can model this process by writing X as a function (or mechanism) f of Y and some independent noise U_X,$X = f(Y, U_X)$.
Formally, a structural causal model consists of two sets of variables $\mathcal{V}$ (X,Y in Wright's example) and $\mathcal{U}$ (U_X, U_Y in Wright's example) and a set of functions f that assigns each variable in $\mathcal{V}$ a value based on the values of the other variables in the model.

- The variables in $\mathcal{U}$ are called exogenous variables , meaning, roughly, that they are external to the model; we choose, for whatever reason, not to explain how they are caused.
- The variables in $\mathcal{V}$ are endogenous . Every endogenous variable in a model is a descendant of at least one exogenous variable.

Exogenous variables cannot be descendants of any other variables, and in particular, cannot be a descendant of an endogenous variable; they have no ancestors and are represented as root nodes in graphs. If we know the value of every exogenous variable, then using the functions in f , we can determine with perfect certainty the value of every endogenous variable.
In this dissertation we will assume every variable in $\mathcal{U}$ be jointly independent and arbitrary distributed.

$$\forall U_i \in \mathcal{U} \text{ with } U_i \sim P_{U_i}$$

then $P_{U_1,...,U_n}(u_1,...,u_n) = P_{U_1}(u_1)P_{U_2}(u_2)...P_{U_n}(u_n) \forall u_i$

Suppose now that we are interested in studying the causal relationships between a salary X and Education Y. We might assume that Y also depends on, or is caused by, socio-economic levels as captured by a variable Z . In this case, we would refer to X and Y as endogenous and Z as exogenous. This is because we assume that socio-economic levels is an external factor.
From now on we will symbolize a SCM with $\mathcal{S}$. As we see in the Wright's example the SCM was related with the DAG of the Figure 4.1. Thus every SCM is associated with a graphical causal model, referred to informally as a graphical model or simply graph. Graphical models consist of a set of nodes representing the variables in $\mathcal{U}$ and $\mathcal{V}$, and a set of edges between the nodes which are determined by the functions **f**.
Additionally, the graphical model $\mathcal{G}$ of an SCM $\mathcal{S}$ contains one node for each variable in $\mathcal{S}$. If, in $\mathcal{S}$, the function f_X for a variable X contains the variable Y so that $X = f_X(Y, ..)$, then, there will be a directed edge from Y to X in $\mathcal{G}$. We will deal primarily with SCMs for which the graphical models are DAGs. Due to the relationship between SCMs and graphical models , we can give a graphical definition of causation, which states: "If, in a graphical model, a variable X is the child of another variable Y , then Y is a direct cause of X ; if X is a descendant of Y , then Y is a potential cause of X."
Let's illustrate the mathematical definition of SCM :

Definition 0..1. A **Structure Causal Model (SCM)** is defined as a tuple $\mathcal{S} := (S, P_U)$,where $S := (S_1, ..., S_p)$ is a collection of p equations

$$S_j : X_j = f_j(PA_j, U_j), j = 1, ..., p \tag{4.3}$$

$PA_j \subseteq (X_1, ..., X_p) \backslash \{X_j\}$ are called parents of X_j and $P_U = P_{U_1,...,U_p}$ is the joint distribution of the noise variables, which we require to be jointly independent. P_U is a product distribution.
The graph $\mathcal{G}$ of a structural causal model is obtained simply by drawing direct edges from each parent to its direct effects, i.e., from each variable $X_k \in PA_j$ occurring on the right-hand side of equation (4.3) to X_j.We henceforth assume this graph to be acyclic. We sometimes call the elements of PA_j not only parents but also direct causes of X_j and we call X_j a direct effect of each of its direct causes.
An augmented graph $\mathcal{G}^a$ of a structural causal model can be obtained simply by drawing direct edges from each augmented parent(noise variables U_j and PA_j) to its direct effects X_j.[1]

Example 0..1. $Let \mathcal{S}_1 := (S_1, P_1^U)$ where U_i jointly independent .

$$\mathcal{S}_1 = \begin{cases} X_1 = f_1(X_3, U_1) \\ X_2 = f_2(X_1, U_2) \\ X_3 = f_3(U_3) \\ X_4 = f_4(X_2, X_3, U_4) \end{cases}$$

[1] Generally, in acyclic SCMs in which the noise variables are independent, the augmented graph and graph do not differ when it comes to the information they provide. For this reason, we will use whichever suits as best.

represented in graph $\mathcal{G}_1$ Figure 4.2.
We can assign exact functions in $\mathcal{S}_1 := (S_1, P_1^U)$ and distributions in noise variables, i.e.

$$\mathcal{S}_1 = \begin{cases} X_1 = 2X_3 - 0.5U_1 \\ X_2 = (0.5X_1)^2 + U_2 \\ X_3 = U_3 \\ X_4 = X_2 + 2sin(X_3 + U_4) \end{cases}$$

where $U_i \overset{P}{\sim} N(0,1)$

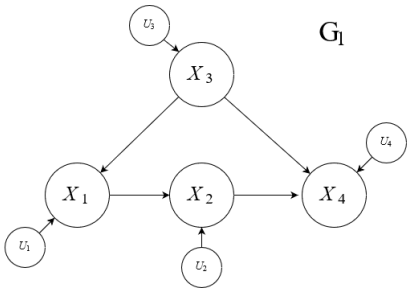


Figure 4.2

SCMs are the key to formalize causal reasoning. With the next Proposition we will show that an SCM is capable to entail observational distributions. However as we say in the previous chapter a SCM unlike usual probabilistic models, they can additionally entail intervention distributions and counterfactuals.

Proposition 0..2 (Entailed distributions). A SCM $\mathcal{S}$ defines a unique distribution over the variables $(X_1, ..., X_p)$ such that $X_j = f_j(PA_j, U_j)$ in distribution, for j=1,...,d. We refer to it as the entailed distribution $P_X^{\mathcal{S}}$.

***Proof* 8.** The proof is in (Bongers, Peters, Schölkopf, & Mooij, 2016).
A plan of the proof, avoiding mathematical details, formalizes the procedure of sampling *n* data points from a joint distribution.
i.e. Firstly, we generate $U_1, ..., U_n \sim P_U$ and then subsequently use the structural assignments **f** to generate i.i.d data points $X_1, ...X_n \sim P_X$.
We recall that our definition of an SCM includes the requirement that the underlying graph is acyclic. Thus, the graph has at least one topological ordering. Using one of the topological ordering π , we can write each node j as a function of the noise terms U_k with $\pi(k) < \pi(j)$. We can substitute the structural assignments recursively into each other and can therefore write each node X_j as a unique function of all noise terms $(U_k)_{k \in AN_j}$ that belong to the ancestors of X_j.That is ,

$$X_j := g_j((U_k)_{k \in AN_j}) = g_j((U_k)_{k:\pi(k)<\pi(j)})$$

U_k is a random variable and g_j a measurable function, so X_j is a random variable .

Generally speaking, the existence of entailed distribution is a piece of a general property of SCMs, called the solvability ,and it is not refered only to acyclic SCM , (Bongers et al., 2016).
For better comprehension, we can see example 0..1 and how this distribution generating process works in R .

```
set.seed(1)
%Firstly generate from X3,\pi(1)=3 i.e. 3 is source node
X3 <- runif(100)-0.5
%continuing with X1 \pi(2)=1
X1 <- 2*X3+rnorm(100)
X2 <- (0.5*X1)^2+rnorm(100)^2
X4 <- X2+2*sin(X3+rnorm(100))
```

1. Interventions

1..1 Introduction to Interventions logic

A commonsensical idea about causation is that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control. Roughly speaking, if C is genuinely a cause of E, then if we can manipulate C in the right way , this should be a way to manipulate or change E. For this reason it is important to find way of representing manipulations in our causal theory .We define this way as interventions. When we intervene to fix the value of a variable, we curtail the natural tendency of that variable to vary in response to other variables in nature. Let's see some manipulation examples in practice .

1. Let us examine the correlation between ice cream sales and crime rate. One possible intervention in this system is to set ice cream sales low. One way to do so is e.g. say, by shutting down all ice cream shops. When we examine correlations in this new manipulated system (under intervention) we find that crime rates are totally independent (i.e., uncorrelated) to ice cream sales.

2. Assume that we have a sequence of upright Domino blocks. A possible system intervention is to keep e.g. the second block fixed in an upright position (e.g. say, by glueing it to the floor). Then we can conclude that each block only directly causes the next neighboring stone to topple.

Consequentially, the ultimate aim of many statistical studies is to predict the effects of interventions.

- When we collect data on factors that are associated with air pollution in Athens, we are actually searching for something we can intervene upon in order to decrease air pollution.
- When we perform a study on a new drug , we are trying to identify how a patient's illness responds when we intervene upon it by medicating the patient.
- When we research how aggressive acts by youngsters is affected by violent PC-games, we are trying to determine whether reducing children access to violent games will result in aggression reduction.

In statistics we use the randomize trials if we want to illustrate interventions .

Randomize Trials

In a properly randomized controlled experiment, all factors that influence the outcome variable are either static, or vary at random, except for one. Any change in the outcome variable must be due to that one input variable. For this reason, the randomized controlled experiment is considered "*the golden standard of statistics*".
(Peters, Janzing, & Schölkopf, 2017), an early example of a randomized trial was performed by James Lind. During the eighteenth century, Great Britain lost more soldiers from scurvy than from enemy action; vitamin C and its relation to scurvy was still unknown. The Scottish physician James Lind (1716-1794) worked as a surgeon on a ship and reports the trial as follows [cited after Bhatt, 2010]:

On the 20th of May 1747, I selected twelve patients in the scurvy, on board the Salisbury at sea. Their cases were as similar as I couldhave them. They all in general had putrid gums, the spots and lassi- tude, with weakness of the knees.... Two were ordered each a quart of cyder a day. Two others took twenty-five drops of elixir vitriol three times a day.... Two others took two spoonfuls of vinegar three times a day.... Two of the worst patients were put on a course of sea-water.... Two others had each two oranges and one lemon given them every day.... The two remaining patients, took ... an electary recommended by a hospital surgeon.... The consequence was, that the most sudden and visible good effects were perceived from the use of oranges and lemons; one of those who had taken them, being at the end of six days fit for duty.

Unfortunately,many questions do not lend themselves to randomized controlled experiments. We cannot control the weather, so we cannot randomize the variables that affect wildfires. Even randomized drug trials can run into problems when participants drop out, fail to take their medication, or misreport their usage. Precisely (Rubin, 1974) in the preface of his paper mentions :
"*(a) the cost of performing the equivalent randomized experiment to test all treatments is prohibitive*

(e.g., 100 reading programs under study); (b) there are ethical reasons why the treatments cannot be randomly assigned (e.g., estimating the effects of heroin addiction on intellectual functioning); or (c) estimates based on results of experiments would be delayed many years (e.g., effect of childhood intake of cholesterol on longevity)."

In cases where randomized controlled experiments are not practical, researchers instead perform observational studies, in which they merely record data, rather than controlling it. The problem of such studies is that it is difficult to distinguish the genuine association "causal" from false associations "not causal".

1..2 Interventions with SCM

SCM give us the opportunity to model interventions in the system even if those interventions are very difficult or impossible to happen when we record the data . Hence, we construct intervention distributions from an SCM $\mathcal{S}$. They are obtained by making modifications to $\mathcal{S}$ and considering the new entailed distribution $\tilde{P}_X$.

As we saw in the previous section every endogenous variable X_j connects with parent variables PA_j under functional relationship $X_j = f_j(PA_j, U_j)$.Moreover, the functional characterization provides a convenient language for specifying how the resulting distribution would change in response to external interventions. This is accomplished by encoding each intervention as an alteration on a selected subset of functions e.g. $f_j \to \tilde{f}_j$ while keeping the other functions intact. For example : Let

$$\mathcal{S} = \begin{cases} X_1 = f_1(U_1) \\ X_2 = f_2(X_1, U_2) \\ X_3 = f_3(U_3) \\ X_4 = f_4(X_2, X_3, U_4) \end{cases}$$

If we want to intervene in X_2 and assign new mechanism $\tilde{f}_2$ and keep all the other f_i mechanisms intact.

$$\tilde{\mathcal{S}} = \begin{cases} X = f_1(U_1) \\ X_2 = \tilde{f}_2(\tilde{PA}_2) \\ X_3 = f_3(U_3) \\ X_4 = f_4(X_2, X_3, U_4) \end{cases}$$

We can consider this procedure as the Randomized trial where all the factors follow the original mechanisms , except for one. The central idea is that when we intervene to one mechanism e.g. $f_k \to \tilde{f}_k$ the others are not affected by this change e.g. this stay intact. [2]

Once we know the identity of the mechanisms altered by the intervention and the nature of the alteration, the overall effect of the intervention can be predicted by modifying the corresponding

[2] all this is results of a general principle of (physically) independent mechanisms. (Peters et al., 2017) p.19

Principle 1..1. The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.

equations in the model and using the modified model to compute a new entailed probability function just in the same way as before we entail the observational distribution.

Definition 1..1 (Intervention Distribution). Consider a distribution P_X that has been generated from an SCM $\mathcal{S} := (S, P_U)$.We can then replace one (or more) structural equation **without generating cycles in the graph** and obtain a new SCM $\tilde{\mathcal{S}}$. We call the distribution in the new SCM **intervention distribution** and we say that the variables whose structural equation have replaced have been **intervened on**. We denote the new distribution by

$$P_X^{\tilde{\mathcal{S}}} = P_X^{\mathcal{S};do(X_j=\tilde{f}(\tilde{PA}_j,\tilde{U}_j))}$$

The set of noise variables in $\tilde{\mathcal{S}}$ now contains both some "new" ˜U's and some "old" U's, all of which are required to be jointly independent. Some types of interventions are:

- When $\tilde{f}(\tilde{PA}_j, \tilde{U}_j) = c$, namely $P_X^{\mathcal{S};do(X_j=c)}$, we call it **atomic intervention** .
- When the marginal distribution of the intervened variable has positive variance is called **stochastic intervention**.

Example 1..2. Let SCM:

$$S = \begin{cases} X = 2U_X \\ Y = 3X + U_Y \end{cases}$$

with $U_X, U_Y \sim \mathcal{N}(0, 1)$ jointly independent

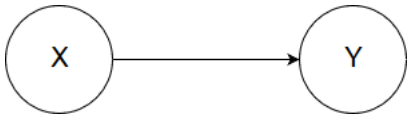


Figure 4.3

Suppose that $P_{X,Y}^{\mathcal{S}}$ is the induced distribution of SCM $\mathcal{S}$. Then the marginal distribution of Y:

$$P_Y^{\mathcal{S}} = \mathcal{N}(0, 37)$$

Now if we want to intervene on X and precisely on $do(X = 2)$, the SCM is modified to :

$$\tilde{\mathcal{S}}_1 = \begin{cases} X = 2 \\ Y = 3X + N_Y \end{cases}$$

and now the new marginal distribution of Y :

$$P_Y^{\tilde{\mathcal{S}}_1} = P_Y^{\mathcal{S};do(X=2)} = \mathcal{N}(6, 1)$$

Following the same concept, we have :

$$P_Y^{\mathcal{S};do(X=3)} = \mathcal{N}(9, 1)$$

In this example if we are interested in the conditional distribution of Y given X=x and the comparison with intervention marginal distribution,do(X=x):

$$P_Y^{\mathcal{S}|X=2} = \mathcal{N}(6,1) = P_Y^{\mathcal{S};do(X=2)} \neq$$

$$P_Y^{\mathcal{S}|X=3} = \mathcal{N}(9,1) = P_Y^{\mathcal{S};do(X=3)}$$

Intervening on X changes the distribution of Y.
But on the other hand, if we intervene on X we have:

$$P_X^{\mathcal{S}} = \mathcal{N}(0,4) = P_X^{\mathcal{S};do(Y=1)} = P_X^{\mathcal{S};do(Y=7)} = P_X^{\mathcal{S};do(Y=123456789)}$$

On the other hand:

$$\mathcal{N}(0,4) \neq P_X^{\mathcal{S}|Y=1} \neq P_X^{\mathcal{S}|do(Y=1)}$$

because

$$P_{X,Y}^{\mathcal{S}} = \mathcal{N}\left[\begin{pmatrix}E[X]\\E[Y]\end{pmatrix}, \quad \begin{pmatrix}Var(X) & Cov(X,Y)\\Cov(X,Y) & Var(Y)\end{pmatrix}\right] =$$

$$= \mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \quad \begin{pmatrix}4 & 12\\12 & 37\end{pmatrix}\right]$$

so $[\mu, \Sigma] = \left[\begin{pmatrix}0\\0\end{pmatrix}, \quad \begin{pmatrix}4 & 12\\12 & 37\end{pmatrix}\right]$ and Cor= $\frac{12}{2\sqrt{37}}$. Also, we can prove that:

$$P_X^{\mathcal{S}|Y=y} = \mathcal{N}(E[X] + \frac{Cov(X,Y)}{Var(Y)}(y - E[Y]), (1 - Cor(X,Y)^2)Var[X]) =$$

Hence,

$$P_X^{\mathcal{S}|Y=1} = \mathcal{N}(\frac{12}{37}, (1 - \frac{12^2}{148})4)$$

As a comparison, no matter how strongly we intervene on Y, the distribution of X remains invariant.

This example motivates two basic ideas.

- Firstly, it confirms our intuition about relation cause-effect . No matter how strongly we intervene on the effect, the distribution of cause remains invariant e.g: “The more someone whitens its teeth will not have any effect on his smoking habits”
- and secondly, it gives us an idea about the difference of *intervene on* and *condition on* variables. Precisely the deference Seeing versus Doing

Let’s start with the difference between *intervene on* and *condition on*.

Generally,

- $P(Y = y|X = x)$) is the probability that Y = y conditional on finding X = x, and reflects the population distribution of Y among individuals whose X value is x.
- $P(Y = y|do(X = x))$ is the probability that Y = y when we intervene to make X = x represents the population distribution of Y if everyone in the population had their X value fixed at x.

Example 1..3. Let's go back to the swimsuit sales - violent crimes rates example. Let X:=swimsuit sales, Y:=temperature, Z:=violent crimes. We have mentioned that the swimming sales is associated with the violent crime rates($X \not\perp\!\!\!\perp Z$). However, if we were to intervene to make swimsuit sales low (say, by shutting down all summer clothing shops), this intervention corresponds to do($X = x$), we would have the new SCM with graph shown in Figure 4.4 (b). When we examine correlations in this new SCM, we find that crime rates are, of course, totally independent of swimsuit sales ($X \perp\!\!\!\perp Z$)since the latter is no longer associated with temperature. If we consider the corresponding SCM without intervention on swimming suites:

$$S := \begin{cases} X = f_1(Y, N_1) \\ Y = f_2(N_2) \\ Z = f_3(Y, N_3) \end{cases}$$

with graph Figure 4.4(a). On the other hand the SCM with intervention on swing suites is:

$$\tilde{\mathcal{S}} := \begin{cases} X = x \\ Y = f_2(N_2) \\ Z = f_3(Y, N_3) \end{cases}$$

with graph Figure 4.4(b)

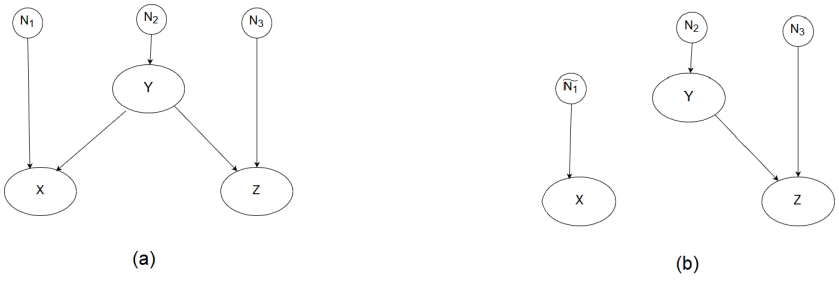


Figure 4.4: Graph of SCM $\mathcal{S} := (S, P^N)$ (a) and $\tilde{\mathcal{S}} := (\tilde{S}, P^N)$ (b)

In the corresponding intervention distribution

$$P_{X,Y,Z}^{\mathcal{S};do(X=x)}$$

we would find X $\perp\!\!\!\perp$ Z since $Z = g(N_3, N_2)$.
This holds independently of the distribution of $\tilde{N}_X$. In other words, even if we vary the level at which we hold X constant, that variation will not be transmitted to variable crime rates.*We say there is no causal effect from X to Z.*

Thus, motivated the last statement of the previous example we define the existence of a total causal effect [(Pearl, 2009), total causal effect].

Definition 1..2. [**Total causal effect**] Given a SCM $\mathcal{S}$, there is a total causal effect from X to Y if and only if :
$X \not\perp\!\!\!\perp Y$ in $P_{\mathbf{X}}^{\mathcal{S};do(X:=\tilde{U}_X)}$ for some random variable $\tilde{U}_X$.

Equivalent statements for total causal effect are:

Proposition 1..4. [Total causal effects] Given a SCM $\mathcal{S}$, the following statements are equivalent:

1. There is a causal effect from X to Y .
2. There are x^a and x^b, such that $P_Y^{\mathcal{S};do(X:=x_a)} \neq P_Y^{\mathcal{S};do(X:=x_b)}$
3. There is x^a, such that $P_Y^{\mathcal{S};do(X:=x_a)} \neq P_Y^{\mathcal{S}}$
4. X$\not\perp\!\!\!\perp$ Y in $P_{X,Y}^{\mathcal{S};do(X:=\tilde{U}_X)}$ for any $\tilde{U}_X$ whose distribution has full support.

In general, we have that

Proposition 1..5. (Graphical criteria for total causal effects)

- If there is no directed path from X to Y , then there is no total causal effect.
- Sometimes there is a directed path but no total causal effect.

Remark 1..6. [Correct SCM] Formally, we say that a SCM $\mathcal{S}$ over $X = (X_1, ..., X_p)$ is a correct model (the correct SCM) for the underlying data generating process if the observational distribution is correct and all interventional distributions $P_{\mathbf{X}}^{\mathcal{S};do(\mathbf{X}:=\tilde{U}_X)}$ correspond to distributions that we obtain from randomized experiments . Importantly, a SCM is therefore falsifiable (if we can do the randomized experiments).

.

2. Counterfactuals

Before giving the definition of counterfactuals, let‘s see some examples of counterafactual statements :

- While driving home last night, I came to a fork in the road, where I had to make a choice: to take the freeway (X = 1) or go on a surface street (X = 0) . I took surface street , only to find out that the traffic was touch and go. As I arrived home, an hour later, *I said to myself , I should have taken the freeway*. i.e *If I had taken the freeway, I would have gotten home earlier*.

- We assume playing football-bets and there is the football match AEK-OFI. We bet the exact score 0-1. 4 minutes before the finish of the match, AEK scored to go up 1-0 . We make the hypothesis : *If I had reversed the betting score in betting slip, my chances would have been good to win.*

- Finally , *If he had eaten more at breakfast, he would not have been hungry at 11 am*.

These reactions are typical counterfactual statements. The above statements called Counterfactuals because follow the same logic[3]. This statement incorporates the observed data (e.g. my bet was 0-1 and the score 4 minutes before the finish was 1-0)into the model , and then analyses an intervention distribution (e.g. I had reversed the bet), in which the rest of the environment remains unchanged (score 4 minutes before finish the match 0-1).
We use counterfactuals to emphasize our wish to compare two outcomes (e.g., driving times,betting outcomes) under the exact same conditions, differing only in one aspect.
We saw in the previous subsection how structural causal models can be used to induce distributions e.g. observational or interventional which correspond to the new observed distribution if we manipulate the system. In this section, we show that, by using the same operation in a slightly different context, we can use SCMs to define what counterfactuals stand for and how to read counterfactuals from a given model.

Definition 2..1. [**Counterfactuals**] Consider a SCM $\mathcal{S} = (S, P_U)$ over nodes **X**. We define a counterfactual SCM by replacing the distribution of noise variables:

$$\mathcal{S}_{X=x} = (S, P_U^{\mathcal{S}|X=x})$$

where $P_U^{\mathcal{S},X=x} := P_{U|X=x}$[4] The new set of noise variables does not need to be mutually independent anymore. Counterfactual statements can now be seen as do-statements in the new counterfactual SCM.

We demonstrate this definition on a simple causal model:

Example 2..1.

$$\mathcal{S} := \begin{cases} X = 7U_X \\ Y = 9X + U_Y \end{cases}$$

with graph:

[3] To illustrate the logic behind the counterfactual statement I use the 2nd counterfactual statement

[4] (Peters et al., 2017) In the continuous case, this definition comes with measure theoretic problems since usually the conditional distribution is only defined up to nullsets. To make our life easier, we restrict counterfactuals to the discrete case, that is, when the noise distribution has a probability mass function. In the case of continuous variables with density, we condition not on X=x but on $X \in A$ with $P(X \in A) > 0$ instead.

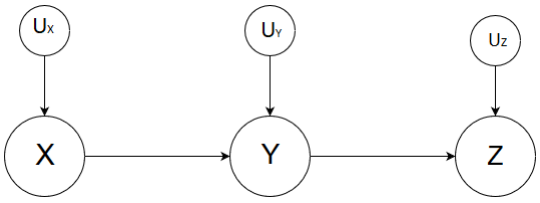
with $U_X, U_Y \sim$ Discrete-uniform in (-10,10).Now we assume observing $(X, Y) = (14, 122)$. Then $P_U^{\mathcal{S}|X=14,Y=122} = P_{U|X=14,Y=122}$puts a point of mass on $(U_X, U_Y) = (2, -4)$. We therefore, have the *counterfactual statement*(in the context of $(X, Y) = (14, 122)$:

- "Y would have been 3 had X been 2. This sentence is interpreted as :Y would have been 3 had X been set to 2,$do(X = 2)$. Or:
  $P_Y^{\mathcal{S}|\mathbf{X}=(14,122);do(X=2)}$ has a point mass on 3 $\Leftrightarrow P^{\mathcal{S}|\mathbf{X}=(14,122);do(X=2)}(Y = 3) > 0$ "

Example 2..2.

$$\mathcal{S} := \begin{cases} X = U_X \\ Y = X^2 + U_Y \\ Z = 2Y + U_Z \end{cases}$$

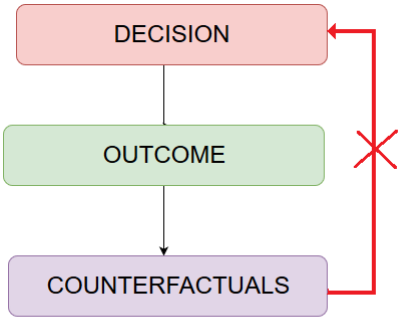
with graph



with $U_X, U_Y, U_Z \sim \mathcal{U}(-2, 2)$.Now we assume observing $(X, Y, Z) = (1, 2, 4)$ and then $(U_X, U_Y, U_Z) = (1, 1, -1)$ we therefore have the counterfactual statement:

- Y would have been 4, had X been 4 ,namely $P_{Y=4}^{\mathcal{S}|(X,Y,Z)=(1,2,4);do(X=4)} > 0$
- Z would have been 2, had X been 10 namely $P_{Z=2}^{\mathcal{S}|(X,Y,Z)=(1,2,4);do(X=10)} > 0$

Pearl [2009](Pearl, 2009) uses the somewhat simpler notation $Z_x(\mathbf{u})$, where the subscript x denotes the intervention do (X : = x) and u represents the additional information about the error terms, which he calls **u** , under $\mathbf{X} = \mathbf{x}$.
The following example illustrates the following idea : counterfactuals cannot provide any information about our decisions before the outcome becomes true.



Example 2..3. A patient with blocked heart arteries goes to the hospital and dies(B=1) after the doctor suggests the treatment (T=1). Let's assume that the correct SCM has the form

$$\mathcal{S} := \begin{cases} T = U_T \\ B = TU_B + (1-T)(1-U_B) \end{cases}$$

with $U_B = Ber(0.03)$ .The question: What did have happened if the doctor had decided to give the treatment $T = 0$? Doctor act optimally?
We notice that if (B,T)=(1,1) then $(U_B, U_T) = (1, 1)$.We ,therefore, have the counterfactual statement $T = 0$ then :

$$P_{B=1}^{\mathcal{S}|(B,T)=(1,1);do(T=0)} = P^{\mathcal{S}|(B,T)=(1,1)}(B = 1|do(T = 0)) = 0$$

and

$$P_{B=0}^{\mathcal{S}|(B,T)=(1,1);do(T=0)} = P^{\mathcal{S}|(B,T)=(1,1)}(B = 0|do(T = 0)) = 1$$

$$P_{B}^{\mathcal{S}|(B,T)=(1,1);do(T=0)} = Ber(0)$$

As a result if the doctor had change the treatment the patient will live. However he act optimally ? We can answer this question if we compare the two above intervention probabilities:

$$P(B = 0|do(T = 0)) = 0.03$$

$$P(B = 0|do(T = 1)) = 0.97$$

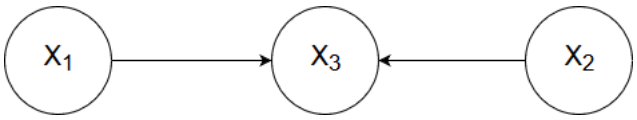
As a result, the doctor acted optimally with the decision $T = 1$ since $P(B = 0|do(T = 1)) > P(B = 0|do(T = 0))$. However, if he knew the outcome of his decision, $B = 1$, then he would change his opinion to $T = 0$ because $P^{\mathcal{S}|(B,T)=(1,1)}(B = 0|do(T = 0)) = 1$. For that reason, we say that we cannot provide details about the role of counterfactuals in our law system. The counterfactual statement requires knowledge which we do not have. For counterfactual statements, there is no apparent correspondence in the real world. But if there is none, these statements may be considered as being not falsifiable and therefore as non-scientific according to Popper [e.g. Popper, 2002].

Can we have two SCMs with the same observational distribution, same interventional distributions , same graphs but different counterfactual statemens?
The answer is yes :

Example 2..4. We define two different SCMs.

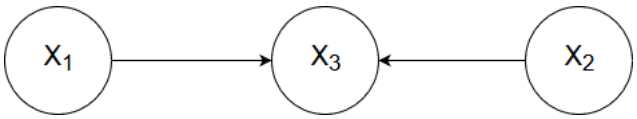
$$\mathcal{S}_1 := \begin{cases} X_1 := U_1 \\ X_2 := U_2 \\ X_3 := (\mathcal{I}_{U_3>0}X_1 + \mathcal{I}_{U_3=0}X_2)\mathcal{I}_{X_1 \neq X_2} + U_3\mathcal{I}_{X_1=X_2} \end{cases}$$

Where $\mathcal{I}_{U_3>0} = 1$ if $U_3 > 0$ and $U_1, U_2 \sim Ber(0.5), U_3 \sim U(\{0,1,2\})$ and graph:

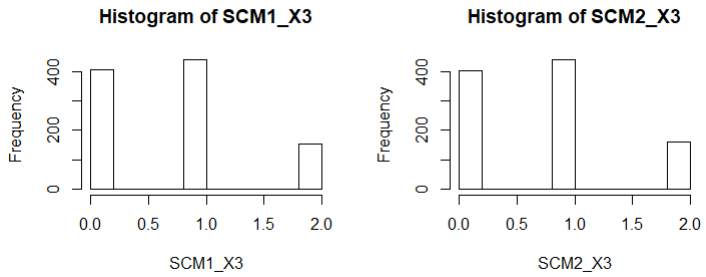


$$\mathcal{S}_2 := \begin{cases} X_1 := U_1 \\ X_2 := U_2 \\ X_3 := (\mathcal{I}_{U_3>0}X_1 + \mathcal{I}_{U_3=0}X_2)\mathcal{I}_{X_1 \neq X_2} + (2 - U_3)\mathcal{I}_{X_1=X_2} \end{cases}$$

With graph:



Both of two SCMs generate the same observational distribution $X_1, X_2 \sim Ber(0.5)$ and in X_3 irrespective the $(\mathcal{I}_{U_3>0}X_1 + \mathcal{I}_{U_3=0}X_2)\mathcal{I}_{X_1 \neq X_2}$ we have $U_3 \sim (2 - U_3) \sim$ U({ 0,1,2 }). The last statement ,that X_3 follows the same marginal distribution in $\mathcal{S}_1$ and $\mathcal{S}_2$, can be validated also if we simulate same data from that marginal distribution in every SCM:



Also its easy to validate that for any possible intervention they entail the same intervention distributions. But the two models differ in a counterfactual statement. We suppose having observed $(X_1, X_2, X_3) = (1, 0, 0)$ and we are interested in the counterfactual question: what would X_3 have been if X_1 had been 0? Under $(X_1, X_2, X_3) = (1, 0, 0)$ we have $(U_1, U_2, U_3) = (1, 0, 0)$,but we

can prove different "predict" values for X_3 under a counterfactual change of X_1 in the $\mathcal{S}_1$ we have $X_3 = 0$ and in the $\mathcal{S}_2$ $X_3 = 2$.

Counterfactual statements depend strongly on the structure of the SCM.Both SCMs correspond to the same causal graphical model , and in this sense, causal graphical models are not rich enough to predict counterfactuals. This the resson we use SCM and not just graphs is causal research .Also this example shows that we require assumptions that let us distinguish between SCM with this problems.

3. SCM and Causal Assumptions

In chapter 3 we described the Markov Property, Causal minimality and faithfulness assumption . In this section we will examine how the induced graph and observed distribution from a SCM is related with some of the above assumptions.

Proposition 3..1. Assume that $P_X$is generated by an acyclic SCM with graph $\mathcal{G}$. Then, P_X is Markov with respect to $\mathcal{G}$.

Proof. The proof is in (Pearl, 2009). □

As we mention in the Chapter **??** given a joint distribution function P_X we can generate several SCM. How rich is the class of that SCM? The answer is given in the above Proposition.

Proposition 3..2. Consider $X_1, .., X_p$ and let P_X have a strictly positive density with respect to Lebesgue measure and assume it is Markov with respect to $\mathcal{G}$. Then there exists a SCM(S,P^N) with graph $\mathcal{G}$ that generates the distribution P^X.

So given a distribution P_X over random variables $X = X_1, ..., X_p$. We can find more than one graphs which are markov equivalent w.r.t a distribution and from proposition 3..2 this leads to more than one SCM which induce the distribution and the graphs respectively. This can answer the question *how many different SCMs entail a distribution.*

Chapter 5

Calculating Intervention Distributions

In this chapter we will assume that we know the true causal structure. Hence using the underlying causal structure we will examine methods calculating Intervention distributions. Thus an other possible title for this chapter could have been: "*Using the known underlying causal structure calculating Intervention Distribution*".

1. Calculating intervention distribution from SCM

In chapter 1..2 we saw how to generate,"induce", an intervention distribution from a SCM. Given a SCM $\mathcal{S}$ if we are interested in intervening on X_k we transform the structure equation S_k of X_k i.e. :

$$X_k := f(PA_k, U_k) \rightarrow X_k := \tilde{f}(\tilde{PA_k}, \tilde{U_k})$$

and the other structure equations remain invariant. Thus we obtain to a new SCM $\tilde{\mathcal{S}}$. In the new SCM the other equations remain invariant thus the parents for every variable $X_j \neq X_k$ are the same i.e. $\forall j \neq k \; PA_j^{\tilde{\mathcal{S}}} = PA_j^{\mathcal{S}}$ [1]. Now if we are interested about the density of $X_j \neq X_k$ condition on $PA_j^{\mathcal{G}}$ in the new intervention SCM we have :

$$p^{\tilde{\mathcal{S}}}(x_j | pa_j^{\mathcal{G}}) = p^{\mathcal{S}}(x_j | pa_j^{\mathcal{G}}) \tag{5.1}$$

The above result is one of the most useful and appealing results of the SCM. Specifically if we intervene on a variable, then the other mechanisms[2] remain invariant. Precisely consider the following

[1] If $\mathcal{G}$ is the induced graph from the $\mathcal{S}$ then $PA_j^{\mathcal{S}} = PA_j^{\mathcal{G}}$. Thus in our case we have $PA_j^{\tilde{\mathcal{G}}} = PA_j^{\mathcal{G}}$

[2] the other functions f_k e.g. the other structural equations S_k

SCM

$$\begin{cases} X := f_1(U_X) \\ Y := f_2(X, U_Y) \\ Z := f_3(Y, X, U_Z) \end{cases}$$

Now if we intervene on Y i.e. do(Y=y) the resulting SCM will be:

$$\begin{cases} X := f_1(U_X) \\ Y := y \\ Z := f_3(Y = y, X, U_Z) \end{cases}$$

As we see the intervention changes only the equation assignment S_Y (more precisely the mechanism which generates the variable Y) but this change will not affect the other causal mechanisms S_X, S_Z because the functional forms f_1, f_3 remain invariant. Only the inputs of the structure assignments of Z will change. Thus the condition density of X and Z in the new SCM will be just the same i.e.

$$p^{\tilde{\mathcal{S}}}(x|pa_X^{\mathcal{G}}) = p^{\mathcal{S}}(x|pa_X^{\mathcal{G}}), p^{\tilde{\mathcal{S}}}(z|pa_Z^{\mathcal{G}}) = p^{\mathcal{S}}(z|pa_Z^{\mathcal{G}}) \tag{5.2}$$

The equation 5.1 is very important for the computation of intervention distributions even though we have never seen data from them.
Consider a SCM $\mathcal{S}$ with structural assignments S_j

$$X_j := f_j(X_{PA_j}, U_j) \quad j = 1, .., p$$

and density $p^{\mathcal{S}}$. Because of the Markov property ,we have :

$$p^{\mathcal{S}}(x_1, ..., x_p) = \prod_{j=1}^{d} p^{\mathcal{S}}(x_j|x_{pa_j})$$

Now consider the SCM $\tilde{S}$ is evolved from S after $do(X_k := \tilde{N_k})$, where $\tilde{N_k}$ allows for the density $\tilde{p}$. Again, from the Markov assumption we conclude that :

$$p^{\mathcal{S}:do(X_k=\tilde{N_k})}(x_1, ..., x_p) = \prod_{j\neq k} p^{\mathcal{S}:do(X_k:=\tilde{N_k})}(x_j|x_{pa_j}) p^{\mathcal{S}:do(X_k:=\tilde{N_k})}(x_k) =$$

$$= \prod_{j\neq k} p^{\mathcal{S}:do(X_k:=\tilde{N_k})}(x_j|x_{pa_j}) \tilde{p}(x_k) =$$

$$= \prod_{j\neq k} p^{\mathcal{S}}(x_j|x_{pa_j}) \tilde{p}(x_k)$$

In the last step, we use the equation 5.1. Hence:

$$p^{\mathcal{S}:do(X_k=\tilde{N_k})}(x_1, ..., x_p) = \prod_{j\neq k} p^{\mathcal{S}}(x_j|x_{pa_j}) \tilde{p}(x_k) \tag{5.3}$$

The equation 5.3 allows us to compute an interventional statement (left-hand side) from observational quantities (right-hand side). As a special case if we define an deterministic function for the $\tilde{N}_k$ i.e $\tilde{N}_k = a$ with probability 1, we obtain:

$$p^{\mathcal{S}:do(X_k=a)}(x_1,...,x_p) := \begin{cases} \prod_{j\neq k} p^{\mathcal{S}}(x_j|x_{pa_j}) & \text{if } x_k = a \\ 0 & \text{otherwise} \end{cases}$$

In practice the equation 5.3 are widely applicable but sometimes a bit cumbersome to use. Hence we will learn about some practical alternatives. We motivate our perspective with the following example.

Example 1..1 (Kidney stone)**.** We record the recovery rates of 700 patients. The first half was treated with open surgery (treatment T = a , 78% recovery rate) and the other half with percutaneous nephrolithotomy (T = b , 83 % recovery rate), a surgical procedure to remove kidney stones by a small puncture wound. Also observing the data in more detail, we can categorize kidney stones of the patients into small and large stones.
We don't know the correct SCM but let's assume to be known with graph is depicted in the Figure 5.1:

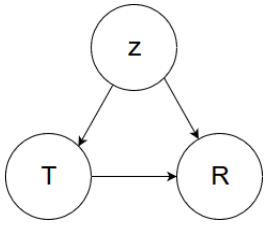


Figure 5.1: Kidney stone Graph

where

$$\cdot \begin{cases} T\text{=treatment} \\ Z\text{=size of the stone} \\ R\text{=recovery} \end{cases}$$

Note that, all the variables are binary.
Generally if we want to describe this clinical trial under a structural-causal perspective we can do it, using the following the sense. Firstly let assume that the observed data come from a SCM $\mathcal{S}$[3]. We consider now two other SCMs $\mathcal{S}_A$ which depicts the correct SCM after intervention on T=A, do(T=A), and $\mathcal{S}_B$ which depicts the system after intervention on T=B, do(T=B). In practice, the intervention SCMs $\mathcal{S}_A$ is obtained if all the patients take the treatment A and SCMs $\mathcal{S}_B$ is obtained if all the patients take the treatment B. In real life having these experiments the same time is obvious impossible. However the SCM solve that issue using the equation 5.3 and the observed data. Let us call the corresponding resulting probability distributions P^{S_A} and P^{S_B} from the SCM $\mathcal{S}_A, \mathcal{S}_B$

[3]we we give the treatment A in the half-population and treatment B in the other half. As a result the process can be described as if we took data from from a SCM $\mathcal{S}$ being produced if we intervene into the starting correct SCM on the Treatment under a randomize way.

respectively and P^S the probability distributions from the observed SCM $\mathcal{S}$. If a patient is diagnosed with Kidney Stone without any reference about their size, then the doctor should base his choice about the treatment on the comparison between $E^{S_A}[R] and E^{S_B}[R]$:

$$E^{S_A}[R] = 1P^{S_A}(R=1) + 0P^{S_A}(R=0) = P^{S_A}(R=1) = P^{S:do(T=A)}(R=1)$$

$$E^{S_B}[R] = 1P^{S_B}(R=1) + 0P^{S_B}(R=0) = P^{S_B}(R=1) = P^{S:do(T=B)}(R=1)$$

But how we can estimate this quantities only with observational data from $\mathcal{S}$.

$$P^{S_A}(R=1) = P^{S_A}(R=1, T=A) = \sum_{z=0}^{1} P^{S_A}(R=1, T=A, Z=z)$$

The first equation is applied because $P^{S_A}(T=A) = P^S(T=A|do(T:=A)) = 1$

$$= \sum_{z=0}^{1} P^{S_A}(R=1|T=A, Z=z) P^{S_A}(T=A, Z=z) \tag{5.4}$$

Since$P^{S_A}(T=A, Z=z) = P^{S_A}(Z=z|T=A)P^{S_A}(T=A) = P^{S_A}(Z=z|T=A) = P^{S_A}(Z=z)$ Hence the 5.4 equals:

$$= \sum_{z=0}^{1} P^S(R=1|T=A, Z=z) P^S(Z=z)$$

$$= \sum_{z=0}^{1} \underbrace{P^{S_A}(R=1|T=A, Z=z)}_{p^{S_A}(r|pa_r)=p^S(r|pa_r)} \underbrace{P^{S_A}(Z=z)}_{p^{S_A}(z|pa_z)=p^S(z|pa_z)}$$

So,

$$P^{S_A}(R=1) = \sum_{z=0}^{1} P^S(R=1|T=A, Z=z) P^S(Z=z)$$

analogously

$$P^{S_B}(R=1) = \sum_{z=0}^{1} P^S(R=1|T=B, Z=z) P^S(Z=z)$$

Now we can estimate $P^{S_A}(R=1)$ and $P^{S_B}(R=1)$ from the empirical data.

	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

$$P^{S_A}(R=1) = P_S(R=1|T=A,Z=0)P_S(Z=0) + P_S(R=1|T=A,Z=1)P_S(Z=1)$$

$$= 0,93\frac{(81+234)}{700} + 0,73\frac{343}{700} = 0,832$$

Respectively

$$P^{S_B}(R=1) = 0.87\frac{(81+234)}{700} + 0.69\frac{343}{700} = 0,782$$

The difference :

$$P^{S_A}(R=1) - P^{S_B}(R=1) = 0.83 - 0.78$$

Is called the **average causal effect (ACE)** for binary treatments.

In this example, the difference between intervene on and condition on gets clear:

$$P^{S}(R=1|S=A) - P^{S}(R=1|S=B) = 0.78 - 0.83$$

In the previous example, when we compute the $P^{S_A}(R=1)$ we choose the variable Z to condition on. We will see later that this procedure is not a random choice because it lead us to the truncated factorization :

$$p^{\mathcal{S}:do(X_k=a)}(x_1,...,x_p) := \begin{cases} \prod_{j\neq k} p^{\mathcal{S}}(x_j|x_{pa_j}) & if \;\; x_k = a \\ 0 & \text{otherwise} \end{cases}$$

This procedure is called "adjusting" on Z. But in the previous example, this choice was not a very complicated thing ,we have only three variables, in comparison with other complicated problems such as those with one hundred variables .

Definition 1..1 (Valid adjustment set). Consider a SCM $\mathcal{S}$ over a set of Variables **X** and let $Y \notin PA_X$. We call a set $Z \subseteq \mathbf{X}\backslash\{X,Y\}$ a valid adjustment set for the ordered pair (X,Y) if

$$p^{\mathcal{S};do(X=x)}(y) = \sum_z p^{\mathcal{S}}(y|x,z)p^{\mathcal{S}}(z) \tag{5.5}$$

Here, the sum (could also be an integral) is over the range of Z, that is, over all values z that Z can take.
In the case $Y \in PA_X$ $p^{\mathcal{S};do(X=x)}(y) = p^{\mathcal{S}}(y)$ [4]

In the Example 1..1 as it mentioned we use the variable Z="Size of the stone" as a valid adjust-

[4]To understand the restriction $Y \notin PA_X$ we can assume that we have only two Variables the X and Y and $Y \in PA_X$,$Y \to X$, Y is the cause and X is the effect. As we mention before if we intervene on the effect the cause are not affected. Using this perspective we can generalize it into models with more variables.

ment set for the (T,R). Since with this set we transform unknown intervention densities distributions like, $p^{\mathcal{S}_A}$ and $p^{\mathcal{S}_B}$, into observed densities distributions $p^{\mathcal{S}}$. Also this valid adjustment set was helped us to compute the Average Causal Effect. An appealing result of the Definition 5.5 is that, if the valid adjustment set is the empty set, so $Z = \emptyset$, then the Equation 5.5 transforms into:

$$p^{\mathcal{S};do(X=x)}(y) = p^{\mathcal{S}}(y|x) \tag{5.6}$$

As a result conditioning on is the same procedure as adjusting with empty set. But we have seen that simple conditioning led to false conclusions like the Example 1..1. The reason why we have that is because the empty set isn't a valid adjustment set. In such a case, we say that the causal effect are in confounded .

Definition 1..2 (confounding). Consider an SCM $\mathcal{S}$ over Variables **X** where in the induced graph there is a directed path from X to Y ,X,Y $\in$ **X**. The causal effect from X to Y is called confounded if

$$p^{S;do(X=x)}(y) \neq p^{S}(y)$$

Otherwise, the causal effect is called "unconfounded."

The previous discussion shows that not all sets are valid adjustment sets. But how can we find this valid adjustment sets? We can answer if we understand the properties which we require to have this set. i.e.

$$p^{\mathcal{S};do(X=x)}(y) = \sum_z p^{\mathcal{S};do(X=x)}(y|x,z)p^{\mathcal{S};do(X=x)}(z) =$$

$$= \sum_z p^{\mathcal{S}}(y|x,z)p^{\mathcal{S}}(z)$$

So we require

$$p^{\mathcal{S},do(X=x)}(y|x,z) = p^{\mathcal{S}}(y|x,z) \;\; and \;\; p^{\mathcal{S},do(X=x)}(z) = p^{\mathcal{S}}(z) \tag{5.7}$$

Thus we need to address the question which conditionals distributions remain invariant under the intervention do(X := x). The answer is given in the following proposition

Proposition 1..2 (Valid adjustment sets). Consider an SCM over variables **X** with X,Y $\in$ **X** and Y $\notin PA_X$. Then, the following three statements are true.

- "**parent adjustment**"

$$Z := PA_X$$

is a valid adjustment set for(X,Y).

- "**backdoor criterion**": Any $Z \subseteq X \backslash \{X, Y\}$ with
  - Z contains no descendant of X and
  - Z blocks all paths from X to Y entering X through the back-door (as a result all the paths from X to Y that has an direct arrow into X i.e $X \leftarrow ...$)

  is a valid adjustment set for (X,Y).

- "**toward necessity** ": Any $Z \subseteq X \backslash \{X, Y\}$ with
  - Z contains no descendant of any node on a directed path from X to Y (except for descendants of X that are not on a directed path from X to Y)
  - Z blocks all non-directed paths from X to Y

  is a valid adjustment set for(X,Y).

The proof of this proposition are in the appendix 8.D.
In the following example we are searching for valid adjustment sets using the theory of Proposition 1..2.

Example 1..3. Let the graph of the figure 5.2.

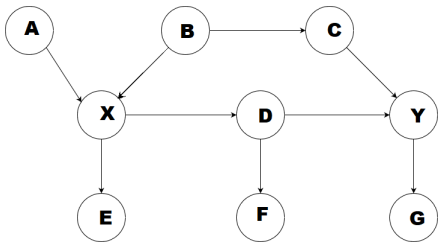


Figure 5.2: Example in valid adjustment sets

Back-door-criterion: Examining the adjustment sets for (X, Y) using the back door criterion. Assume that **Z** is the valid adjustment set then **Z** can't contain the descendants of X so $\{E, D, F, G, Y\} \notin$ **Z**. Additionally the **Z** must blocks all the paths from X to Y entering X. We have only one back door path the $X \leftarrow B \rightarrow C \rightarrow Y$. Using the d-separation theory the sets $\{B\}, \{C\}, \{B, C\}$ blocks the back door path. Also if we add the A in any of these sets nothing change since the two hypothesis of the back-door criterion is satisfied. Hence the sets $\{B\}, \{C\}, \{B, C\}, \{A, B\}, \{A, C\}, \{A, B, C\}$ are valid adjustment sets under the back door criterion.
Toward-necessity: Examining the valid adjustment set using the toward necessity criterion. Again **Z** can't contain any descendant of X except the descendants of X that are not on the the directed paths from X to Y, in our example the node E. Thus $\{D, F, G, Y\} \notin$ **Z**. Additionally the **Z** must blocks

all the non-directed paths. The sets $\{B\}, \{C\}, \{B, C\}$ has this property. Again if we add any of the nodes E or A in these sets the two hypothesis of the toward necessity is not violated. As a result the valid adjustment set under the towards necessity is: $\{B\}, \{C\}, \{B, C\}, \{A, B\}, \{A, C\}, \{A, B, C\}$, $\{E, B\}, \{E, C\}, \{E, B, C\}, \{A, E, B\}, \{A, E, C\}, \{A, E, B, C\}$

1..1 Do-calculus

This subsection generally is very technical and also is out of the general framework of this dissertation. But it is mentioned only for intuition reasons. As a result I will focus on the basic theory on headlines and I will try to avoid proofs and examples and technical details.
With the Proposition 1..2 we saw how intervention distribution $p^{S,do(X:=x)}$ is computed from the observational distribution and the graph structure with the valid adjustment formula but we can compute intervention distributions $p^{S,do(X:=x)}$ in other ways than the adjustment formula. Let us therefore call an intervention distribution $p^{S,do(X:=x)}$ identifiable if it can be computed from the observational distribution and the graph structure. For example if there is a valid adjustment set for (X,Y) then the $p^{S,do(X:=x)}$ is certainly identifiable. Consider a SCM over variables **X**. (Pearl, 2009) in [Theorem 3.4.1] has developed the so-called do-calculus that consists of three rules.
Given a graph $\mathcal{G}$ and disjoint subsets **X,Y,Z,** and **W**, we have :

1. "Insertion/deletion of observations":

$$p^{S,do(\mathbf{X}=\mathbf{x})}(\mathbf{y}|\mathbf{z},\mathbf{w}) = p^{S,do(\mathbf{X}=\mathbf{x})}(\mathbf{y}|\mathbf{w})$$

   if **Y** d-separates **Z** given **X,W** in a graph where incoming edges in **X** have been removed.

2. "Action/observation exchange":

$$p^{S,do(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})}(\mathbf{y}|\mathbf{w}) = p^{S,do(\mathbf{X}=\mathbf{x})}(\mathbf{y}|\mathbf{z},\mathbf{w})$$

   if **Y** d-separates **Z** given **X,W** in a graph where incoming edges in **X** and outgoing edges from **Z** have been removed.

3. "Insertion/deletion of actions":

$$p^{S,do(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})}(\mathbf{y}|\mathbf{w}) = p^{S,do(\mathbf{X}=\mathbf{x})}(\mathbf{y}|\mathbf{w})$$

   if **Y** d-separates **Z** given **X,W** in a graph where incoming edges in **X** and **Z(W)** have been removed. Here, **Z(W)** is the subset of nodes in **Z** that are not ancestors of any node in **W** in a graph that is obtained from G after removing all edges into **X**.

Theorem 1..4. The following statements can be proved

1. The rules are complete (Shpitser & Pearl, 2006), that is all identifiable intervention distributions can be computed by an iterative application of these three rules.

2. In fact, there is an algorithm, proposed by(Tian & Pearl, 2002) that is guaranteed (Huang & Valtorta, 2012), (Shpitser & Pearl, 2006) to find all identifiable intervention distributions. distributions.

1..2 Adjusting in linear Gaussian System

In the example 1..1 we saw how to summarize the causal effect from variable X to Y into a single real number called average causal effect (ACE). But in that example the data was binary, what can be done in the case of continuous random variables? As a first approximation we may look at the expectation of this distribution and then take the derivative with respect to x:

$$\frac{\partial E^{S;do(X=x)}[Y]}{\partial x} \tag{5.8}$$

In this section we will define and deal with Linear-Gaussian SCM. Since using the approximation 5.8 it can be proved that this models have some appealing properties. Consider an SCM $\mathcal{S}$ over variables **X** with Gaussian distributed error terms U_i and linear functions f_i this SCM called linear Gaussian systems. For example:

Example 1..5. let the following linear Gaussian SCM :

$$\mathcal{S} := \begin{cases} X = aZ + U_x \\ Z = U_Z \\ Y = bX + eK + U_Y \\ K = cX + dZ + U_K \end{cases}$$

where $U_i \sim N(0,1)$.

Hence let a linear Gaussian SCM $\mathcal{S}$ over **X** Variables. Now for $\mathbf{Z}, X, Y \subseteq \mathbf{X}$ assuming that **Z** is a valid adjustment set for (X,Y). As we mention U_i is Gaussian distributed thus **X** follows a Gaussian distribution then the conditional Y|X=x,**Z**=**z** follows a Gaussian distribution too; its means is

$$E^{\mathcal{S}}[Y|X = x, \mathbf{Z} = \mathbf{z}] = ax + b^t\mathbf{z}$$

for some a and **b**.Proof in the appendix 8.E. For example:

Example 1..6. in the linear Gaussian SCM of 1..5 example.

$$E^{\mathcal{S}}[K|Y = y, Z = z] = E^{\mathcal{S}}[cX + dZ + U_K|Y = y, Z = z] =$$

$$= E^{\mathcal{S}}[c(aY + U_x) + dZ + U_K | Y = y, Z = z]$$

$$= E^{\mathcal{S}}[caY + dZ + cU_x + U_k | Y = y, Z = z] =$$

$$= cay + dz$$

Generally the linear Gaussian SCM have some appealing properties if we use the approximation $\frac{\partial E^{S;do(X=x)}[Y]}{\partial x}$ as a measure of the causal effect between two variables X and Y. We will prove that the value of $\frac{\partial E^{S;do(X=x)}[Y]}{\partial x}$ is constant. Precisely:
If $\mathcal{S}$ linear-Gaussian SCM, with zero mean for every error term, over the variables **X** and X,Y,**Z**$\subset$ **X**. Assume that **Z** is a valid adjustment set for the causal effect from X to Y Then if X|Y=y,**Z**=**z** follows Gaussian distribution with mean:

$$E[Y|X = x, \mathbf{Z} = \mathbf{z}] = ax + \mathbf{b}^t\mathbf{Z}$$

Then:

$$\frac{\partial E^{S;do(X=x)}[Y]}{\partial x} = \alpha \tag{5.9}$$

***Proof* 9.** From the adjustment formula

$$p^{S,do(X=x)}(y) = \int_Z p^S(y|x,z)p_S(z)dz$$

$$\int_Y y p^{S,do(X=x)}(y)dy = \int_Y y \int_Z p^S(y|x,z)p^S(z)dzdy$$

$$E^{S;do(X=x)}[Y] = \int_Z \int_Y y p^S(y|x,z)p^S(z)dydz \tag{5.10}$$

$$E^{S;do(X=x)}[Y] = \int_Z p^S(z) \int_Y y p^S(y|x,z)p^S(z)dydz$$

$$E^{S;do(X=x)}[Y] = \int_Z p^S(z) E^S[Y|X=x, Z=z]dz$$

$$E^{S;do(X=x)}[Y] = \int_Z p^S(z)(ax+bz)dz$$

$$E^{S;do(X=x)}[Y] = ax\int_Z p^S(z)dz + b\int_Z zp^S(z)dz$$

$$E^{S;do(X=x)}[Y] = ax + bE^S[Z]$$

The equation 5.10 holds because:

$$\int_Z \int_Y |yp^S(y|x,z)p^S(z)|dydz < \infty$$

and

$$\frac{\partial E^{S,do(X=x)}[Y]}{\partial x} = \frac{\partial(ax+bE^S[z])}{\partial x} = a$$

So using the formula

$$E[Y|X=x, \mathbf{Z}=\mathbf{z}] = ax + b^t\mathbf{z}$$

we can estimate the value of causal effect just regressing Y on X and Z and then reading off the regression coefficient for X. For example:

Example 1..7. Let the following Linear-Gaussian SCM:

$$\mathcal{S} := \begin{cases} A = 0.1U_A \\ B = 0.3U_B \\ X = 2A + 3B + 0.4U_X \\ E = 10X + 0.2U_E \\ D = 4X + 0.5U_D \\ F = 5D + 0.6U_F \\ C = 6B + 0.7U_C \\ Y = 7D + 8C + 0.8U_Y \\ G = 9Y + 0.9U_G \end{cases}$$

with graph depicted in Figure 5.2. In this example we will simulate an i.i.d. sample of size n=1000 in R. Then we will estimate the value of equation 5.9 by regressing Y on X and an adjustment set

Z. If Z is a valid adjustment set, we obtain an unbiased estimator. In the example 5.2 we calculate all the valid adjustment sets we choose randomly some of them them. Hence in the following code, the adjustment set $Z = \emptyset$ leads to a biased estimator ; however the all the correct adjustment give us the same, approximately, estimator, the regression coefficient of X.

```
> set.seed(1); n <- 1000
> A <- 0.1*rnorm(n)
> B <- 0.3*rnorm(n)
> X <- 2*A + 3*B + 0.4*rnorm(n)
> E <- 10*X+0.2*rnorm(n)
> D <- 4*X + 0.5*rnorm(n)
> F <- 5*D + 0.6*rnorm(n)
> C <- 6*B + 0.7*rnorm(n)
> Y <- 7*D + 8*C + 0.8*rnorm(n)
> G <- 9*Y + 0.9*rnorm(n)
> #Biased estimator
> lm(Y~X)$coefficients
(Intercept)           X
 -0.2866196 40.6366330
> #Unbiased estimator
> lm(Y~X+B)$coefficients
(Intercept)           X           B
 -0.182269  26.933440  51.842398
> lm(Y~X+E+C)$coefficients
(Intercept)           X           E           C
-0.0706424 25.7778289   0.1796941   8.1865134
> lm(Y~X+A+B)$coefficients
(Intercept)           X           A           B
-0.1812445 26.8643103   0.6503057  52.0507571
> lm(Y~X+E+C+B)$coefficients
(Intercept)           X           E           C           B
-0.07159035 25.39788082 0.19634634  8.01064772  1.86867165
> lm(Y~X+E+C+B+A)$coefficients
(Intercept)           X           E           C           B
-0.07128965 25.39088911 0.19503153  8.01048766  1.93036023
          A
 0.18948257
```

Someone maybe notice that the unbiased estimator of example 1..7 took values between 28-29 and also notice that if we multiply the coefficients of the direct path between X to Y has the same value. As a result the coefficients of the path $X \to D \to Y$, 4*7=28. This is not a random result, the approximation 5.9 has an very interesting second physical interpretation. Specifically if there is exactly one directed path from X to Y, then the approximation 5.9 equals with the product of the

path coefficients. If there is no directed path, then a = 0 and if there are different direct paths, if n:=size of the different direct paths, and a_i is the product of the path coefficients in the i path then equals with $\sum_{i=1}^{n} a_i$, . This is a result of the Wright's formula (S. Wright, 1934). For example:

Example 1..8. In this example we illustrate different cases of the above result in different Linear-Gaussian SCM and estimate the unbiased estimators for the value 5.9 using R. In the first two cases the coefficient of the direct path between X to Y is equal with 2.

$$\mathcal{S}_1 := \begin{cases} X = 0.1U_X \\ Y = 2X + 0.3U_Y \end{cases} \qquad \mathcal{S}_2 := \begin{cases} A = 0.2U_A \\ X = 5A + 0.1U_X \\ Y = 2 * X + 3Z0.3U_Y \end{cases}$$

with $U_i \sim \mathcal{N}(0, 1)$ and graphs depicted in the Figure 5.3 (a) for $\mathcal{S}_1$ and (b) for $\mathcal{S}_2$.

Figure 5.3

We calculate the approximation 5.9 using the appropriate valid adjustment sets for the (X,Y). As a result following the Proposition 1..2 in the first case the valid adjustment set Z is the empty set and in the second $Z = \{A\}$. In the next code we observe that the regression coefficient of X is very close to 2, hence is equals with the path coefficient.

```
> # physical interpretation of Causal−Effect coefficient in linear Gaussian SCM
> X <− 0.1∗rnorm(n)
> Y <−2∗X + 0.3∗rnorm(n)
> lm(Y~X)$coefficients
 ( Intercept )              X
1.301087e−05 1.918806e+00
> # add an node but the coefficient of the direct path remains the same
> A<− 0.2∗rnorm(n)
> X <− 5∗Y+0.1∗rnorm(n)
> Y <−2∗X + 3∗A + 0.3∗rnorm(n)
> lm(Y~X+Z)$coefficients
( Intercept )              X              A
0.001704358 1.998656072 3.022121130
```

The next SCM has a little more complex structure. However we can calculate again the coefficient

of the equation 5.9.

$$\mathcal{S}_3 := \begin{cases} X = 0.1U_X \\ D = 6X + 0.2U_D \\ B = 2X + 0.3U_B \\ Y = 7D + 3B + 0.4U_Y \\ C = 4X + 5Y + 0.5U_C \end{cases}$$

with $U_i \sim \mathcal{N}(0, 1)$ and graphs depicted in the Figure 5.4.

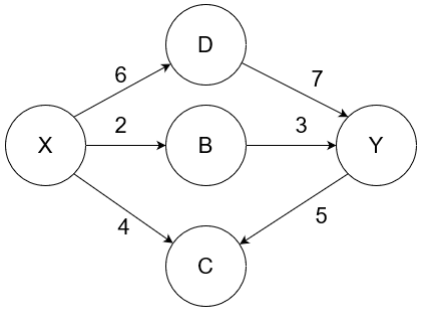


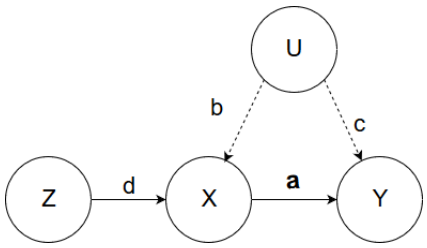
Figure 5.4

The coefficient of the equation 5.9 must be equal with $2 * 3 + 6 * 7 = 48$. We can validate this result simulates some data from these structure and compute the coefficient of X in the regression of Y on Z and X, where Z is the valid adjustment set. In this case the valid adjustment set is the empty set. Since the nodes X, B, D, Y is in nodes of the direct paths of X to Y and C is descendant of Y as a result the valid adjustment set Z=∅.

```
> X<− 0.1∗rnorm(n)
> D<−6∗X +0.2∗rnorm(n)
> B<−2∗X +0.3∗rnorm(n)
> Y<−7∗D + 3∗B +0.4∗rnorm(n)
> C<−4∗X +5∗Y + 0.5∗rnorm(n)
> #unbiased  estimator
> lm(Y~X)$coefficients
( Intercept )              X
 0.06191674 47.57132224
> lm(Y~X+B)$coefficients
( Intercept )              X              B
 0.01219116 42.15274757  2.96985795
```

1..3 Instrumental Variables

An other very useful application in the Linear Gaussian SCM is the theory of Instrumental variables (P. Wright, 1928) are widely used in practice (Imbens & Angrist, 1994), (Bowden & Turkington, 1990).
Consider a linear-Gaussian SCM with the following corresponding graph :



Assuming that we want to estimate the causal effect of X on Y. Here, the coefficient 'a' is the quantity of interest (because the SCM is linear-Gaussian and as we discuss in section 1..2 this coefficient equals with the causal effect of X on Y). However we will assume also that U is unobserved variable, called hidden common cause, hence the coefficient is not directly accessible. However (U,U_X) is independent of Z[5], so we can regard the value $bU + U_X$ in equation 5.11 as a noise term.

$$X = dZ + bU + U_X \tag{5.11}$$

is the same with:

$$X = dZ + \tilde{U_X} \tag{5.12}$$

So, therefore we can consistently estimate the coefficient d, also we have access to $\tilde{Z} = dZ$. Thus:

$$Y = aX + cU + U_Y = adZ + (ab + c)U + N_Y = a\tilde{Z} + \tilde{U_Y}$$

it is clear that we can then consistently estimate a. Thus, we first regress X on Z and then regress Y on the predicted values of X (predicted from the first regression). This method is commonly referred to as "two-stage-least-squares". It makes heavy use of the following assumptions

- linear SCMs
- non zero d
- the independence between U and Z.
- the absence of a direct influence from Z to Y.

[5] (U,U_X) is d-separated from Z given $\emptyset$ using the Markov property (U,U_X) is independent of Z

Chapter 6

Identifiability

1. Introduction In Structure Identification

The analysis which took place in the previous chapters is based on a known "correct" SCM. In this chapter we will focus on methods and algorithms which estimate the correct SCM. Precisely, this estimation process is called ("*structure identification*"). However, this identification procedure is impossible to work without some basic assumptions.

The process of "*structure identification*" is an extremely difficult problem, and there is not direct answer to that if we don't assume some basic assumptions. Until now, we have seen some of these assumptions like the Markov Property and the Faithfulness. As a result, the following question might arise. If these two assumptions are sufficient to the structure identification, or alternatively, given a distribution P_X over random variables $X = \{X_1, ..., X_p\}$, *how many SCMs can be entailed from this distribution if we assume faithfulness and Markov property.* If the answer is "one", then our task is to find this one SCM. But if the answer is more than one, the problem becomes harder, since we need to answer one additional question: how we can specify the true SCM from the class of equivalent structures. Unlike, the answer is more than one. In Proposition 3..2 we saw that if we find a DAG $\mathcal{G}$ Markov with the distribution P_X, then we can find a corresponding SCM that entails the distribution P_X and the graph $\mathcal{G}$ at hand. Also, in example 2..2-2..3 it is mentioned that we can find more than one graph Markov with respect to the distribution P_X. As a result, a distribution could have been generated from many SCMs with different graphs.

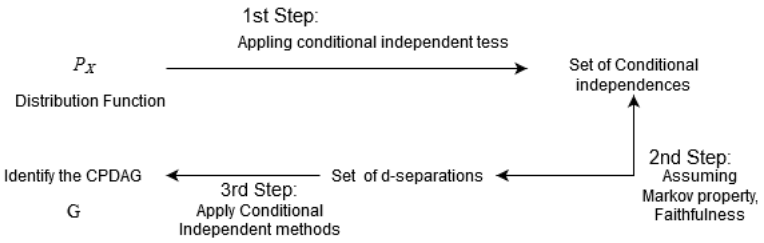
In the beginning of this chapter, we will describe how big the class of equivalent structures is, if we have assumed the Markov property and the Faithfulness. Then, we will describe techniques to identify this class of equivalent structures. Finally, we will add some further assumptions about the functional and the distributional structure of the model, for example the linearity of the functions assignments or Gaussian error terms, to obtain a precise identifiable result.

2. Structure Identification using Faithfulness

If the distribution P_X is Markovian and faithful with respect to the underlying DAG $\mathcal{G}$, we have an one-to-one correspondence between d-separation statements in the graph and the corresponding conditional independence statements in the distribution. Thus, we could stay focus on the class of graphs which induces a set of d-separations that is equal to the set of conditional independence of P_X. As it has been mentioned that this class of graphs is the Markov equivalent class of $\mathcal{G}$. Thus, this gives the answer to the previous question of how big the class which we will try to identify is. However, the following problem is generated: are we able to distinguish any two different graphs from this "Markov equivalent class" only using a data set? We are not, without further assumptions about the structural and the distributional form of the models. However, there are a lot of algorithms which identify the Markov-equivalent class CPDAG($\mathcal{G}$). This class of algorithms called Independence-based methods or Constraint-based methods.

Lemma 2..1. Assume that P_X is Markov and faithful with respect to $\mathcal{G}$. Then, for each graph $\mathcal{G}_1 \in CPDAG(\mathcal{G})$, we find an SCM that generates the distribution P_X . Furthermore, the distribution P_X is not Markov and faithful to any graph $\tilde{\mathcal{G}} \notin CPDAG(\mathcal{G})$.

The Lemma 2..1 is the key of Constraint -based-methods. Precisely, in this algorithmic category we assume faithfulness and try to estimate the correct Markov equivalence class of graphs with respect to P_X.



2..1 Constraint-based methods

These methods, algorithms, will return an equivalent class of graphs, usually in the form of a CPDAG. The searching process for the equivalent class is called *learning process*. In this category of the algorithms, the learning process has two phases. In the first phase, the algorithm looks for conditional-independences using independence-tests and it outputs the skeleton of the Markov-equivalent class[1]. In the second phase, it tries to orient as many edges as many possible by following a set of rules. The most common algorithms are:

- IC (inductive causality)
- SGS(Spirtes, Glymour and Scheines),
- PC(Peter and Clark)

[1] All the members of the class have the same skeleton

- IA or IAMB,TPDA,RAI.

In this dissertation, we will focus on the first three algorithms. Before we analyse the algorithms, we should mention some important lemmas and statements for the algorithmic processes. In the first phase the algorithm searches for the skeleton. The following lemma plays a significant role in this procedure:

Lemma 2..2. The following two statements hold.

- Two nodes X,Y in a DAG $\mathcal{G} = (\mathbf{X}, \mathcal{E})$ are adjacent if and only if they cannot be d-separated by any subset $S \subseteq \mathbf{X} \setminus \{X, Y\}$.
- If two nodes X,Y in a DAG $\mathcal{G} = (\mathbf{X}, \mathcal{E})$ are not adjacent, then they are d-separated by either PA_X or PA_Y.

***Proof* 10.** (Pearl & Verma, 1991)

In the second phase, the algorithm continues with the "orientation". As we have mentioned in Lemma 1..7, the graphs of the Markov Equivalent class have the same skeleton and the same v-structures. This statement has great importance in that part. Specifically:
Orientation of Edges: As it is mentioned, the first phase of the algorithm outputs a graph-skeleton. For every two nodes which are not directly connected in the obtained skeleton, we can find a set that d-separates these nodes. Let **S** be the set that d-separates X and Y. We further suppose that the skeleton contains the structure $X - Z - Y$ with no direct edge between X and Y. The structure $X - Z - Y$ is a possible v-structure and can therefore be oriented as $X \rightarrow Z \leftarrow Y$ if and only if $Z \notin \mathbf{S}$. After the orientation of v-structures, we may be able to orient some further edges in order to avoid cycles, for example. There is a set of such orientation rules that has been shown to be complete and is known as Meek's orientation rules (Meek, 2013).

2..2 SGS-Algorithm

SGS algorithm (Spirtes, Glymour, & Scheines, 1993):

1. Form the complete undirected graph $\mathcal{G}_0$ on the vertex set **X**.
2. For each pair of vertices X and Y, if there exist a subset $\mathbf{S} \subseteq \mathbf{X} \setminus \{X, Y\}$ such that X and Y are d-separated given **S**, remove the edge between X and Y from $\mathcal{G}_0$.
3. Let $\mathcal{G}_0^{(1)}$ be the undirected graph resulting from step 2. For each triple of vertices X, Y, Z such that the pair X and Y and the pair Y and Z are each adjacent in $\mathcal{G}_0^{(1)}$ but the pair X and Y are not adjacent in $\mathcal{G}_0^{(1)}$ (i.e. X-Y-Z), orient
   X - Y - Z as $X \rightarrow Y \leftarrow Z$ if and only if there is no subset S of $\{Y\} \cup \mathbf{X} \setminus \{X, Z\}$ that d-separates X and Z.
4. repeat

- If $X \to Y$, Y and Z are adjacent, X and Z are not adjacent, and there is no arrowhead at Y, then orient Y -Z as $Y \to Z$.
- If there is a directed path from X to Y, and an edge between X and Y, then orient X - Y as $X \to Y$.

until no more edges can be oriented.

Verma and Pearl (1990) have suggested a variation of the SGS algorithm the IC-algorithm.

2..3 IC-Algorithm

1. We starts with the empty graph $\mathcal{G}_\emptyset$.
2. For each pair of variables X,Y searches for a subset of nodes $S_{X,Y}$ such that they are conditionally independent given on $S_{X,Y}$. If no such subset exists it adds an undirected edge between X,Y.
3. Continuing with the 3rd step of SGS algorithm.
   Let $\mathcal{G}_0^{(1)}$ be the undirected graph resulting from step 2 .For each triple of vertices X,Y,Z such that the pair X and Y and the pair Y and Z are each adjacent in $\mathcal{G}_0^{(1)}$ (written as X-Y-Z) but the pair X and Y are not adjacent in $\mathcal{G}_0^{(1)}$, orient
   X - Y - Z as $X \to Y \leftarrow Z$ if and only if there is no subset S of $\{Y\} \cup \mathbf{X} \setminus \{X, Z\}$ that d-separates X and Z.
4. repeat
   - If $X \to Y$, Y and Z are adjacent, X and Z are not adjacent, and there is no arrowhead at Y, then orient Y -Z as $Y \to Z$.
   - If there is a directed path from X to Y, and an edge between X and Y, then orient X - Y as $X \to Y$.

   until no more edges can be oriented.

Complexity-Problem: In the Step 2 of the SGS algorithm for each pair of adjacent variables we have to check all the possible subsets of the remaining variables. However, that has exponential growth with respect to the number of variables. Also, we can't decrease the size of this search because we will lose the effectiveness of the algorithm. Two variables X, Y can be conditional dependent on a set S but independent on a superset or subset of S.

2..4 PC-Algorithm

Algorithm

1. Form the complete undirected graph $\mathcal{G}^{(n)}$ on the vertex set **X**.

2. set n=0
   repeat

      repeat

         Select an ordered pair of variables X and Y that are adjacent in $\mathcal{G}^{(n)}$ such that $|\mathbf{Adj}_{\mathcal{G}^{(n)}}(X)\backslash\{Y\}| \geq$n [2], i.e. this set has cardinality greater than or equal to n , and a subset **S** of $\mathbf{Adj}_{\mathcal{G}^{(n)}}(X)\backslash\{Y\}$ of cardinality n,$|\mathbf{S}| = n$, and if X and Y are d-separated given **S** delete edge X-Y from $\mathcal{G}^{(n)}$ and record **S** in**Sepset**(X,Y) and **Sepset**(Y,X);
         We continue until all ordered pairs of adjacent variables X and Y such that $\mathbf{Adj}_{\mathcal{G}^{(n)}}(X)\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets **S** of $\mathbf{Adj}_{\mathcal{G}^{(n)}}(X)\backslash\{Y\}$ of cardinality n have been tested from d-separations.;
         n=n+1;

   We finish when each ordered pair of adjacent vertices X, Y the $\mathbf{Adj}_{\hat{\mathcal{G}}}(X)\backslash\{Y\}$ is of cardinality less than n i.e. $|\mathbf{Adj}_{\hat{\mathcal{G}}}(X)\backslash\{Y\}| \leq n$.

3. For each triple of vertices X,Y,Z such that the pair X and Y and the pair Y and Z are each adjacent in $\hat{\mathcal{G}}$ but the pair X and Z are not adjacent in $\hat{\mathcal{G}}$, orient
   X - Y - Z as $X \rightarrow Y \leftarrow Z$ if and only if Y is in **Sepset**(X,Y).

4. repeat

      If $X \rightarrow Y$, Y and Z are adjacent, X and Z are not adjacent, and there is no arrowhead at Y, then orient Y -Z as $Y \rightarrow Z$.

      If there is a directed path from X to Y, and an edge between X and Y, then orient X - Y as $X \rightarrow Y$.

   until no more edges can be oriented.

The second part of the PC algorithm is much more complicated that the IC, SGS. Hence, we illustrate an example:

Example 2..3. In the Figure 6.1 the true graph $\mathcal{G}$ is depicted with (a). The PC algorithm starts with an complete undirected graph $\mathcal{G}_0$, (b)-graph in the Figure 6.1.

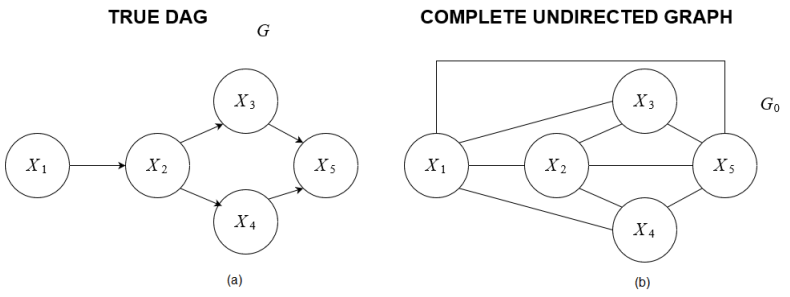


Figure 6.1

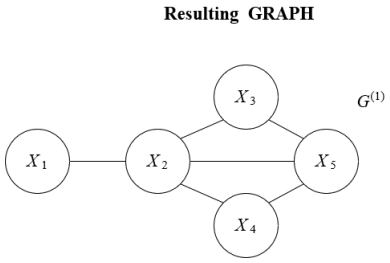
[2]The set $\mathbf{Adj}_{\mathcal{G}^{(n)}}(X)$ includes all the adjacent vertices with X in the graph $\mathcal{G}^{(n)}$.

For n=0, all the pairs in the complete graph are adjacent. Hence, we select any possible random pair of variables, i.e. X_i, X_j, with $|\mathbf{Adj}_{\mathcal{G}^{(0)}}(X_i) \setminus \{X_j\}| \geq 0$. We search for an appropriate subset $\mathbf{S} \subseteq \mathbf{Adj}_{\mathcal{G}^{(0)}}(X_i) \setminus \{X_j\}$ with cardinality $|S| = 0$. However, we can't find such S with the ability to d-separate any X_i, X_j. Thus, we continue with n=1.

Set n=1. As previously, all the pairs of variables in the complete graph are adjacent. Hence we select a pair at random, i.e. X_i, X_j, with $|\mathbf{Adj}_{\mathcal{G}^{(0)}}(X_i) \setminus \{X_j\}| \geq 1$. However, when we search for **S** with $|\mathbf{S}| = 1$ we take:

1. $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_3 | X_2$
2. $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_4 | X_2$
3. $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_5 | X_2$

As a result, we take the $\mathcal{G}^{(1)}$:



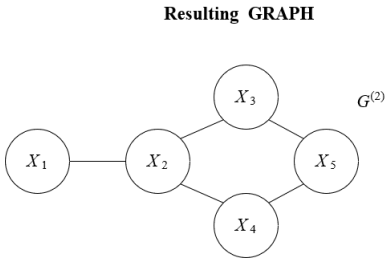
We have tested each pair X_i, X_j and **S** for d-separations. Hence we continue with n=2.

Set n=2

In this case we have only

1. $X_2 \perp\!\!\!\perp_{\mathcal{G}} X_5 | \{X_4, X_3\}$

As a result we take the graph $\mathcal{G}^{(2)}$



For n=3. We can't find in the $\mathcal{G}^{(2)}$ any pair X_i, X_j with $|\mathbf{Adj}_{\mathcal{G}^{(0)}}(X_i) \setminus \{X_j\}| \geq 3$,the max is 2. As a result this part of the algorithm get finished.

Complexity: PC starts with a fully connected undirected graph and step-by-step increases the size of the conditioning set S, starting with $\#S = 0$[3]. At iteration k, it considers sets S of size #S =

[3]#S:=Size of the set S

k, using the following neat trick: If someone had to test whether X and Y is d-separated by S, then he has only to check sets S that are subsets either of the neighbours of X or of the neighbours of Y; this idea is based on Lemma 2..2$-(ii)$ and clearly improves the computation time, especially for sparse graphs.
More precisely, the complexity of the algorithm in the worst case i.e the number of conditional independence tests required by the algorithm is:

$$2\binom{n}{2}\sum_{i=0}^{k}\binom{n-1}{k} \leq \frac{n^2(n-1)^{k-1}}{(k-1)!}$$

which k=$max(|\mathbf{Adj}_{\mathcal{G}}(X_i)|)\forall i$ so k is the maximal number of adjacent vertexes in the true DAG $\mathcal{G}$ and n the number of all variables. Then the bound is given in the (Spirtes et al., 1993).

3. Score-Based Methods

Until now, we have described how to use independent statements with the aim of constructing an appropriate class of graphs. In this section, we will illustrate algorithms under a different perspective. In this methods instead of trying to construct a graph from independent-statements, we will test how different graphs fit the data. The rationale behind this algorithms is that graph-structures encoding the wrong conditional independences will yield bad model fits. The roots of Score based methods in causality start with (Geiger & Heckerman, 1994), (Heckerman, Meek, & Cooper, 1999), (Chickering, 2002). Also there are algorithms which combine Score based Methods and Independence-based methods called Max-Min Hill-Climbing algorithm (Tsamardinos, Brown, & Aliferis, 2006).
Best Scoring Graph: Give data $D = (\mathbf{X}^1, ..., \mathbf{X}^n)$ from a vector $\mathbf{X}$ of variables, that is, a sample containing n i.i.d. observations, the idea is to assign a score $S(D, \mathcal{G})$ to each graph $\mathcal{G}$ and search over the space of DAGs to find the graph with the highest score:

$$\tilde{\mathcal{G}} := \underset{\mathcal{G}DAG\ over\ X}{argmax}\ S(D, \mathcal{G}) \tag{6.1}$$

There are several possibilities to define such a scoring function S. Often a parametric model is assumed (e.g., linear Gaussian equations or multinomial distributions), which introduces a set of parameters $\theta \in \Theta$. Now we present some of them: **(Penalized) likelihood**: For each graph we may consider the maximum likelihood estimator $\tilde{\theta}$ for θ. We may then define a score function by the Bayesian Information Criterion (BIC)

$$S(D, \mathcal{G}) = \log p(D|\tilde{\theta}, \mathcal{G}) - \frac{|parameters|}{2} log(n) \tag{6.2}$$

where n is the sample size and the $\log p(D|\tilde{\theta}, \mathcal{G})$ is the log-likelihood.

We can understand how difficult it is to search the space of all DAGs from the Figure 6.2 since

the size of different graphs is growing super-exponentially with respect of the number of variables :

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	9942119532215951589522891459235452451655502687858830501478 3
19	3327719012271075917361775733112611258835830762584219025835467735 05
20	2344880451051088988152559855229099188899081192234291298795803236068491263

The number of DAGs depending on the number p of nodes, taken from `http://oeis.org/A003024` (Feb 2015).

Figure 6.2

For that reason Greedy-search algorithms can be applied with in order to solve the equation 6.1. At each step there is a candidate graph and a set of neighbouring graphs. For all these neighbour graphs, we compute their score and consider the best-scoring graph as the new candidate. If none of the neighbours obtain better score, the search procedure get terminated. But how this neighbour graphs are produced ? Starting from a graph $\mathcal{G}$, we define all the neighbour graphs as those which can be obtained by removing, adding or reversing one edge. So the new graphs defer just to one edge with the starting graph $\mathcal{G}$. The problem in this methods is that we don't know whether we have captured a local maximum or the total maximum.

4. Additive Noise Models with Continues-Variables

Until now we saw how we can identify the unknown Markov equivalent graph just having assume Markov property and the faithfulness. But as it's mentioned if we don't assume further assumptions we can't identify the correct, unknown, SCM. Generally, we didn't impose any restriction in the functions assignments f_j appearing in the SCM. In this section we will be restricted in a special SCM classes the Additive Noise Models Definition 4..1. Hence, we will examine if we can take identifiability results in this cases. Let's start with the first category of this restricted models with the ANMs SCM.

Definition 4..1 (ANMs). We call a SCM $\mathcal{S}$ ANM if the structural assignments are of the form :

$$X_j = f_j(PA_j) + U_j \quad for \quad j = 1, ..., p$$

that is, if the noise is additive. For simplicity, we further assume that the functions f_j are differentiable and the noise variables U_j have a strictly positive density. In this dissertation we will focus, also, on continuous densities

Although, if we are restricted in the additive noise models then, what can we say about identifiability? Again the identifiability is not resulted in all ANMs. As we can see in (Peters et al., 2017) the linear Gaussian SCMs, for example, is not identifiable. But fortunately this is just an exception. For almost all other combinations of functions and distributions, we obtain identifiability. All the non identifiable cases have been characterized in(Zhang & Hyvarinen, 2012) and (Peters, 2015).

4..1 Linear Additive Noise Models

Let's start from the identifiability in linear Gaussian models with two variables (cause-effect). As we mention before in this special cases of ANMs we can not achieve identifiability. For example:

Example 4..1. Consider two SCMs

$$\mathcal{S}_1 \begin{cases} X := U_X \\ Y := 0.8X + U_Y \end{cases}$$

where U_X, U_Y are independent with $var(U_X) = 1, Var(U_Y) = 0.6^2$ so that $Var(X) = 1, Var(Y) = 1$ and zero means with graph is depicted in Figure 6.3-(a). And the 2nd model:

$$\mathcal{S}_2 \begin{cases} X := 0.8X + U_X \\ Y := U_Y \end{cases}$$

where U_X, U_Y are independent with $var(U_Y) = 1, Var(U_X) = 0.36$ so that $Var(X) = 1, Var(Y) = 1$ and zero means and graph is depicted in Figure 6.3-(b).

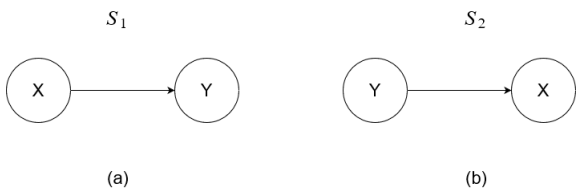


Figure 6.3

If U_i are further-more assumed to be Gaussian, the two models provide the same joint-Gaussian distribution density of the observed variables X and Y with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and covariance matrix $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$. In Figure 6.4 we can verify that since it's depicted the join-density-distributions $p(X, Y)$,by scatterplots. As a result when we observe data from that distribution like the Figure 6.4 it is impossible to identify the SCM which generate that data.

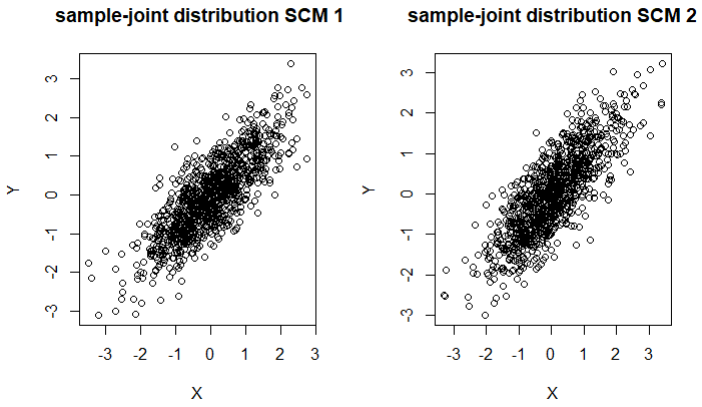


Figure 6.4

However if we assume the same SCM with the only difference the error terms, U_i, to be non Gaussian distributed i.e Uniform or the exponential distribution. Then can we identify the correct SCM from the joint density distribution p(X,Y)? Figures depicts the scatter plots like before give an intuition about the answer. Precisely in Figure 6.5 we assumed uniform distributed error terms. As we can see the sampled-p(X,Y) is different in these two cases. So we have more chances to identify the SCM.

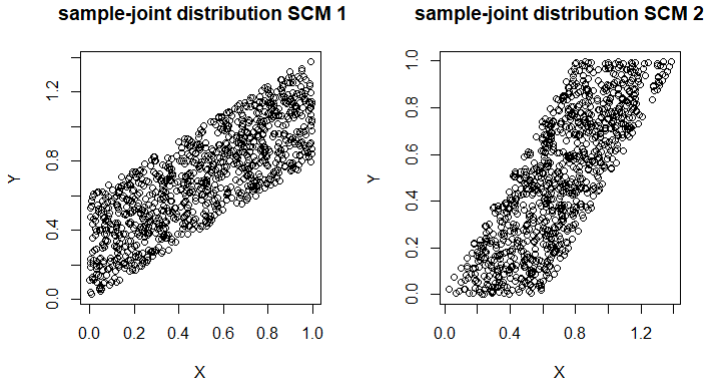


Figure 6.5

Also assuming exponentially distributed error terms as in the Figure 6.6 again the sampled-p(X,Y) have differences.

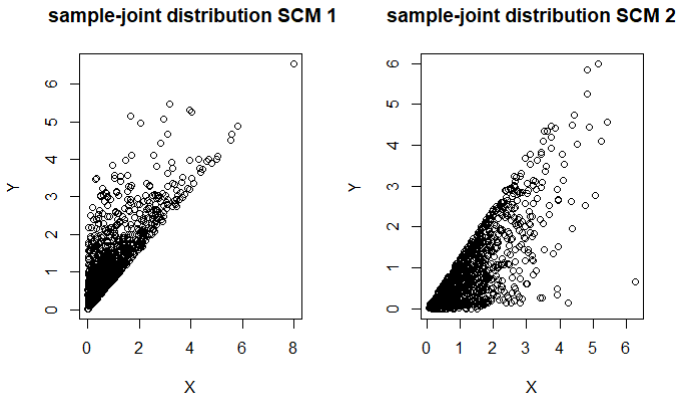


Figure 6.6

But why we have so different results? The answer is simple If we have a SCM over the variables X,Y with functional assignments $X = U_X$ and $Y = aX + U_Y$ with $U_Y \perp\!\!\!\perp X$ then if we can find a_1 such as $X = a_1 Y + \tilde{U}_X$, $Y = \tilde{U}_Y$ and $\tilde{U}_X \perp\!\!\!\perp Y$ then is impossible to identify the true SCM. Generally the example 4..2, from (Peters, 2008), can confirm this claim for the linear Gaussian bivariate models:

Example 4..2. Let,

$$Y = \phi X + U,\ U \perp\!\!\!\perp X$$

where X and U are normally distributed with mean zero and σ^2 be the variance of noise variable. (Peters, 2008) proves that:

$$X = \tilde{\phi} Y + \tilde{U}, \tilde{U} \perp\!\!\!\perp Y$$

with $\tilde{\phi} = \frac{\phi Var[X]}{\phi^2 Var[X] + \sigma^2} \neq \frac{1}{\phi}$ and $\tilde{U} = X - \tilde{\phi} Y$

Hence from this example becomes clear that, is impossible to distinguish data from this two models. Thus the linear Gaussian SCMs is not identifiable.
However if we consider non-Gaussian noise the structural equation model becomes identifiable.

Theorem 4..3. Identifiability of linear non-Gaussian models Assume that $P_{X,Y}$ admits the linear model

$$Y = aX + N_Y \quad U_Y \perp\!\!\!\perp X$$

with continuous random variables X, N_Y, and Y . Then there exist b$\in \mathbb{R}$ and a random variable N_X such that

$$X = bY + N_X \quad U_X \perp\!\!\!\perp Y$$

if and only if U_Y and X are Gaussian distributed.

The proof of the previous Theorem is based on on a characterization of the Gaussian distribution that was proved independently by (SKITOV, 1962), (Darmois, 1953).

Theorem 4..4 (Darmois-Skitovic). Let $X_1, ..., X_p$ be independent, non-degenerate random variables. If there are non-vanishing coefficients $a_1, .., a_p$ and $b_1, ..., b_p$ $(a_i, b_i \neq 0)$such that the two linear combinations

$$l_1 = a_1 X_1, ..., a_p X_p$$

$$l_2 = b_1 X_1, ..., b_p X_p$$

are independent, each X_i is normally distributed.

This result holds in the multivariate case, too. (Shimizu, Hoyer, Hyvärinen, & Kerminen, 2006) prove it using the theorem 4..4.

Theorem 4..5. [(Shimizu et al., 2006)] Assume an SCM with graph $\mathcal{G}_0$.

$$X_j = \sum_{k \in PA_j^{\mathcal{G}_0}} b_{j,k} X_k + U_j, \quad j = 1, ..., p$$

where all U_j are jointly independent and non-Gaussian distributed with strictly positive density. Additionally, for each j $\in \{1...p\}$ we require $b_{jk} \neq 0$ for all k$\in PA_j^{\mathcal{G}}$. Then, the graph $\mathcal{G}$ is identifiable from the joint distribution.

(Shimizu et al., 2006) called this special class of SCM a linear non-Gaussian acyclic model (LiNGAM)and provide a practical method based on Independent Component Analysis, ICA, that can be applied to a finite amount of data.
We saw in the theorem 4..3 that, is impossible to distinguish two linear Gaussian SCM. However in (Peters, Bühlmann, & Meinshausen, 2016) show that restricting the noise variables to have the same variance is sufficient to recover the graph structure.

Proposition 4..6. Consider an SCM with graph $\mathcal{G}$ and assignments

$$X_j := \sum_{k \in PA_j^{\mathcal{G}}} b_{j,k} X_k + U_j, \quad j = 1, ..., d$$

where all U_j are i.i.d. and follow a Gaussian distribution. In particular, the noise variance σ^2does not depend on j. Additionally, for each j $\in \{1...p\}$ we require $b_{j,k} \neq 0$ for all k$\in PA_j^{\mathcal{G}}$. Then, the graph $\mathcal{G}$ is identifiable from the joint distribution.

4..2 Non-linear Additive Noise Models

Until now we have considered only the linear Gaussian and non-Gaussian additive noise SCM. We now describe non-linear Gaussian additive noise models (ANMs). As we can see we have similar

results like before. However since it is out of the main framework of this dissertation I will illustrate only the Basic Theorems and the References for the readers who want to take deeper knowledge in that domain.

Theorem 4..7 (Identifiability of nonlinear Gaussian ANMs). 1. Let $P_X = P_{X_1,...,X_d}$ be induced by an SCM with

$$X_j := f(PA_j) + N_j$$

with normally distributed noise variables $N_j \sim N(0, \sigma_j^2)$ and three times differentiable functions f_j that are not linear in any component in the following sense. Denote the parents PA_j of X_j by $X_{k_1}, .., X_{k_l}$, then the function $f_j(x_{k_1}, .., x_{a-1}, \cdot, x_{a+1}, .., x_{k_l})$ is assumed to be non-linear for all a and some $x_{k_1}, .., x_{a-1}, x_{a+1}, .., x_{k_l} \in \mathbb{R}^{l-1}$

2. As a special case, let $P_X = P_{X1,...Xd}$ be induced by an SCM with

$$X_j := \sum_{k \in PA_j} f_{j,k}(X_k) + N_j$$

with normally distributed noise variables $N_j \sim N(0, \sigma_j^2)$ and three times differentiable, nonlinear functions $f_{j,k}$. This model is known as a causal additive model (CAM).

In both cases (1) and (2), we can identify the corresponding graph $\mathcal{G}$ from the distribution P_X. The statements remain true if the noise distributions for source nodes, that is, nodes without parents, are allowed to have a non-Gaussian density with full support on the real line $\mathbb{R}$.

The proof can be found in [(Peters, Mooij, Janzing, & Schölkopf, 2014), Corollary 31].

4..3 Additive Noise Models-Methods

We will start with the bivariate case of this type of algorithms. This method tests the independence of residuals and is a special case of the regression with subsequent independence test (RESIT) algorithm.

1. Regress Y on X; that is, use some regression technique to write Y as a function $\tilde{f}_Y$ of X plus some noise.
2. Test if $Y - \tilde{f}_Y(X)$ is independent of X.
3. Repeat the procedure with exchanging the roles of X and Y.
4. If the independence is accepted for one direction and rejected for the other, infer the former one as the causal direction.

This procedure has two major problems: First we need regression methods for general cases linear, non-linear models. Second independence techniques for general cases non just Gaussian terms. For

the first problem if the function is linear we can use regression techniques which minimize the least square error. For the Second problem (Mooij, Janzing, Peters, & Schölkopf, 2009) use the Hilbert-Schmidt Independence Criterion (HSIC) ,which proved by (Gretton et al., 2008) as an method for searching independences. We can apply this algorithm in the following example:

Example 4..8. Consider a joint density-distribution $p_{X,Y}$ induced by the following SCM:

$$\mathcal{S} \begin{cases} X = U_X \\ Y = 2X + U_y \end{cases}$$

$$\textit{with } X \sim \mathcal{U}(-1,1),\ U_Y \sim \mathcal{U}(-0.4,0.4)$$

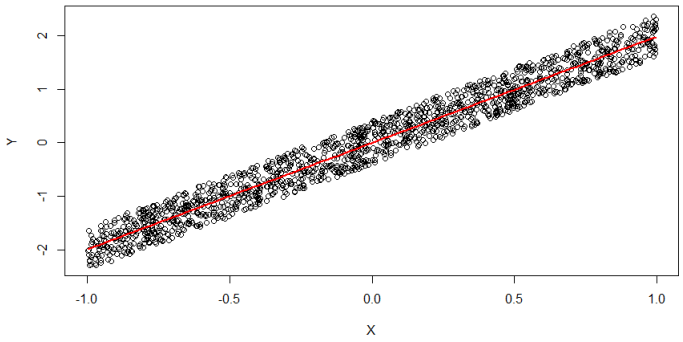
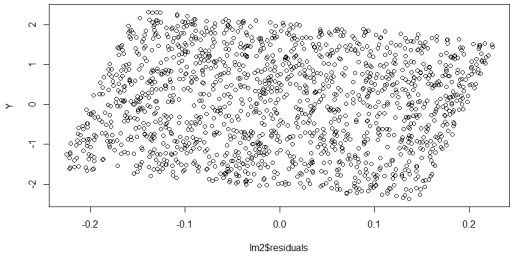
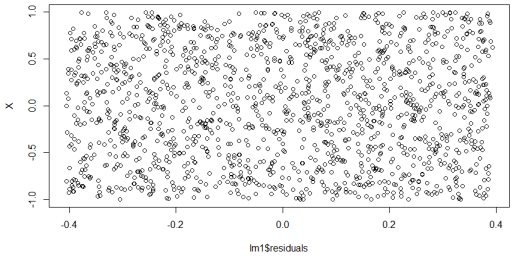


Figure 6.7

If the SCM is unknown then our task is to identify the SCM from the $p_{X,Y}$. In Figure 6.7 is depicted a sample of this $p_{X,Y}$ so given this sample we must identify the SCM. Theorem 4..3 states that we can identify the SCM since the distribution of (U_X, U_y) is non-Gaussian. We can confirm that with the following procedure: Firstly we apply regression of Y on X using the least square estimator and take $Y = \hat{\beta}_1 X + \hat{\gamma}_1 + \hat{U}_Y$. Secondly we apply the regression of X on Y using the least square estimator again and take $X = \hat{\beta}_2 Y + \hat{\gamma}_2 + \hat{U}_X$. However in the first case the resulting noise $\hat{U}_Y$ would not be independent of Xand in the second cases the resulting noise $\hat{U}_X$ would not be independent of Y. In Figure 6.8 this result is verified. In the (a) plot of the Figure 6.8 is depicted in x-axis the Residuals of estimation of X on Y and in the y axis the Y and in (b) plot depicted in x-axis the Residuals of the estimation of Y on X and in the y axis the X.Clearly as we can see in the second case the X is independent with U_Y , $X \perp\!\!\!\perp U_Y$, but in the first the Y is dependent with U_X $Y \not\perp\!\!\!\perp U_X$:



(a)



(b)

Figure 6.8

Until now we saw the bivariate case but what happens in the multivariate case. In the multivariate case the ANMs can be learned with score-based methods that are combined with a greedy search technique. This methods first proposed for searching Linear Gaussian models with same Variances for the error terms or Non linear Gaussian models. The procedure is same with the bivariate case. Given data $\mathcal{D} = \{X^1, ..., X^n\}$ from a vector $\mathbf{X} = \{X_1, ..., X_d\}$ of variables, that is, a sample containing n i.i.d. observations. For a given graph structure $\mathcal{G}$, we regress each variable on its parents and obtain the score

$$log(p(\mathcal{D}|\mathcal{G})) = \sum_{j=1}^{d} -log(\tilde{var}[R_j])$$

where $\tilde{var}[R_j]$ is the empirical Variance of the residual R_j obtained from the regression of variable X_j on its parents.

Chapter 7

Simulation Study

In this section we will compare the algorithms of the previous chapter in Linear Gaussian models , is called LIGAM, with same variance in the Error terms. We will concentrate only in this type of SCM for two basic reasons. First is the regression techniques. The LIGAM is linear model with Gaussian distributed error terms we can apply least square regression techniques. Second is the independence tests since we didn't illustrate non linear independence test in this dissertation,like Hilbert-Schmidt Independence criterion (HSIC), we need the Gaussian distributed error terms to apply $\chi^2 - test$.

1. Simulate LIGAMs with same Variance for the error terms

In this section we will apply the algorithms of the previous chapter in simulated data. Hence we will estimate DAGs from the data. The simulated data from a linear Gaussian models is based on the approach of (Colombo & Maathuis, 2014). Firstly we will generate a random weighted DAG with a given number of vertices p and an expected neighborhood size $E[N]$. For that purpose we first construct an adjacent matrix A as follows:

- Fix the ordering of variables
- Fill the adjacent matrix A with zeros .
- Replace every matrix entry below the diagonal by independent realizations of Bernoulli random variables with success probability s where $0 < s < 1$. We will call s the sparseness of model.
- Replace each entry with 1 in the adjacency matrix by independent realizations of a *Uniform*$[(0.1, 1)]$random variable.

Finally all the entries of matrix A are zero or in the range [0.1,1]. The corresponding DAG draws a directed edge from node i to node j if i < j and $A_{ji} \neq 0$. DAGs that are created in this way have the following property: $E[N_i] = s(p-1)$where N_i is the number of neighbors of a node i.

The matrix A will be used to generate the data as follows. The value of the random variable X_1 , corresponding to the first node, is given by

$$X_1 = U_1 \, , \, \epsilon_1 \sim N(0,1)$$

and the values of the next random variables (corresponding to the next nodes) can be computed recursively as

$$X_i = \sum_{k=1}^{i-1} A_{ik} X_k + U_i, \;\; for \; i = 2, .., p \; U_i \sim N(0,1)$$

where all U_i are mutually independent random variables.

1..1 PC-Algorythm

An other very important thing in PC algorithm is the choice of the significant level α. In (Kalisch & Bühlmann, 2007) provide a significant level for Consistency in High-Dimensional Data. Unfortunately , this value is not constructive, since it depends on the unknown lower bound of partial correlations between the variables.For that reason we fitted a wide range of parameter settings and compared the quality of fit for different significance levels.
To test how well the algorithms works in each significant level we follow an approach suggested by (Tsamardinos et al., 2006) the Structural Hamming Distance (SHD). Roughly speaking, this counts the number of edge insertions, deletions and flips in order to transfer the estimated CPDAG into the correct CPDAG. Thus, a large SHD indicates a poor fit, while a small SHD indicates a good fit.

We fitted 80 replicates to all combinations of

- [1] $\alpha \in \{0,00005, 0,0001, 0,0005, 0,001, 0,005, 0,01, 0,05, 0,1\}$
- $p \in \{7, 15, 40, 70, 100\}$
- $E[N] \in \{2, 5\}$

[1] α:= significant level

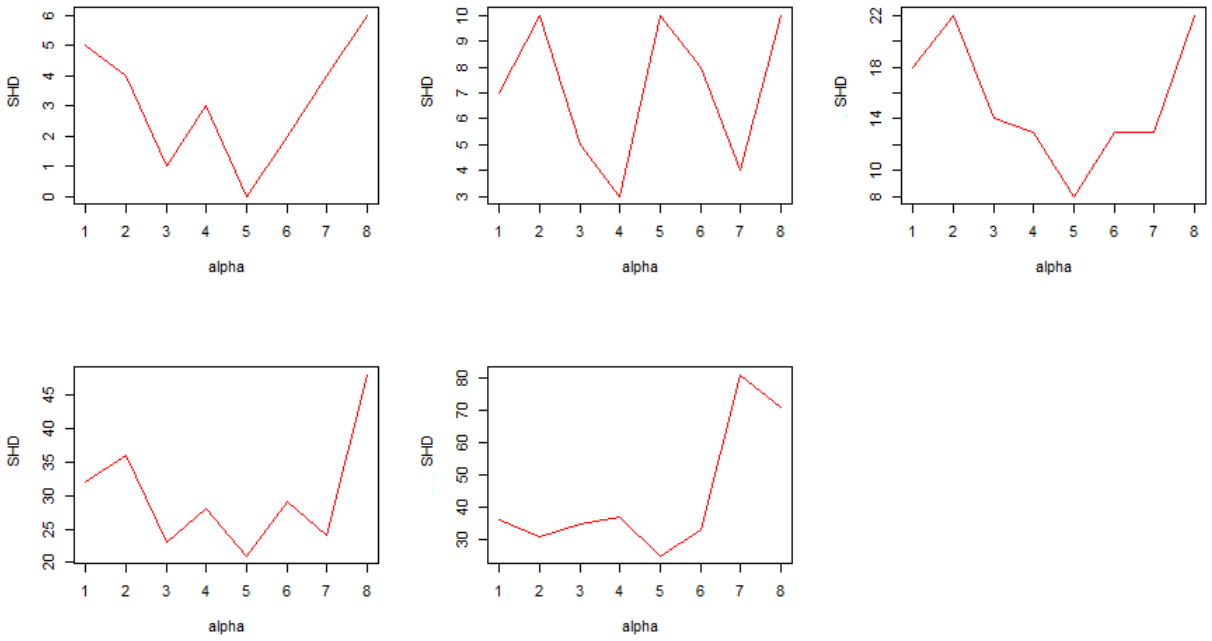


Figure 7.1: SHD for p $\in \{7, 15, 40, 70, 100\}$respectively for $E[N] = 2$

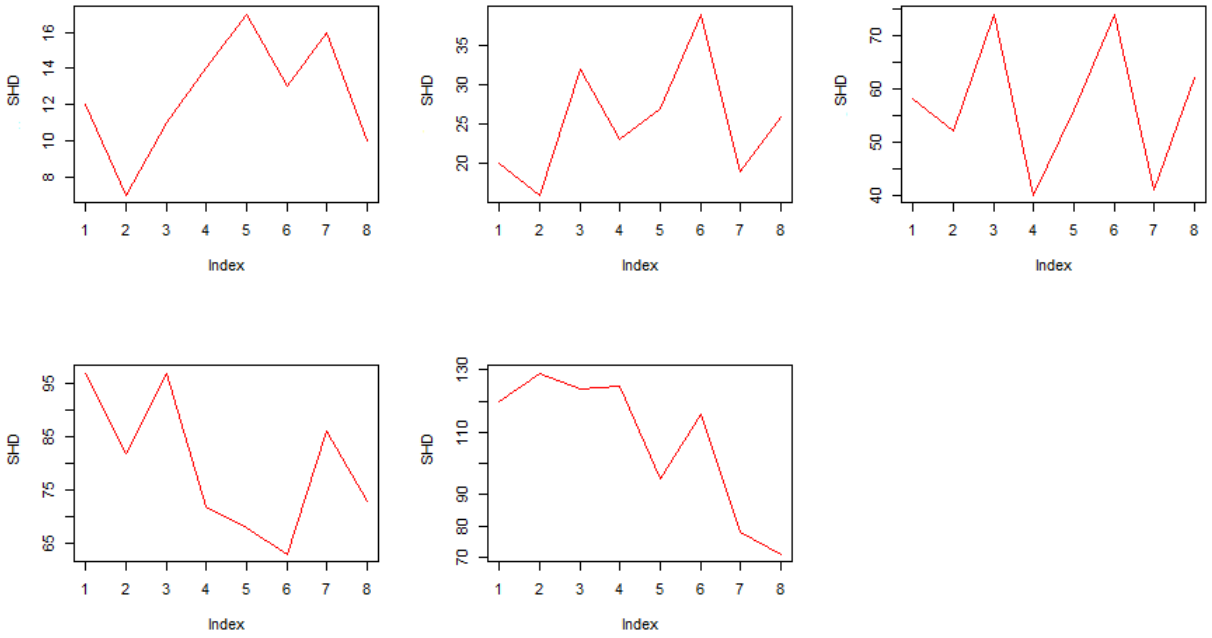
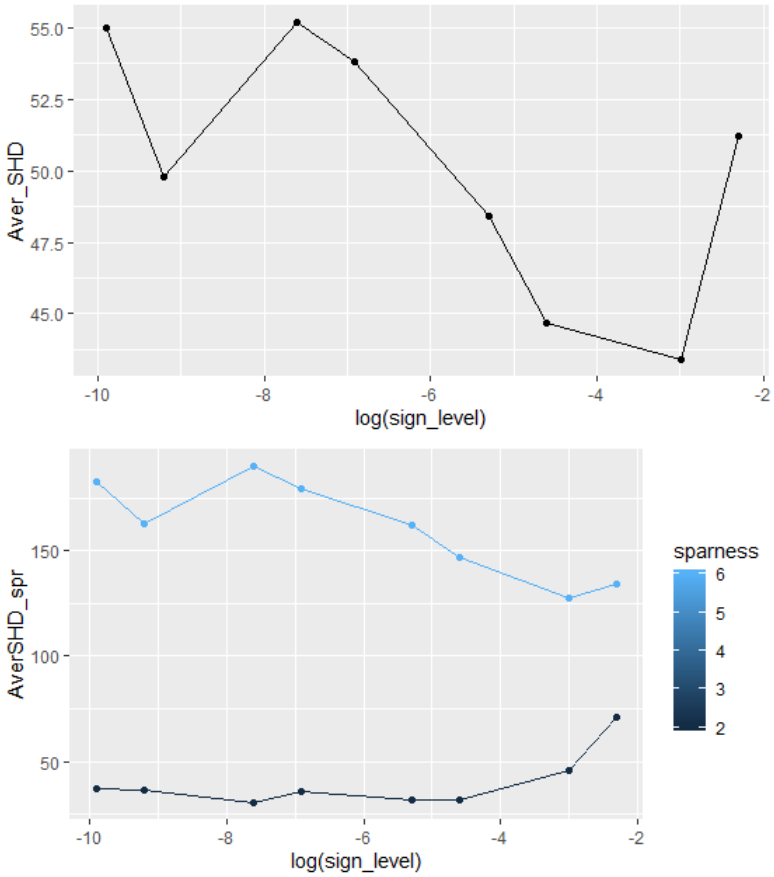


Figure 7.2: SHD for p $\in \{7, 15, 40, 70, 100\}$respectively for $E[N] = 5$

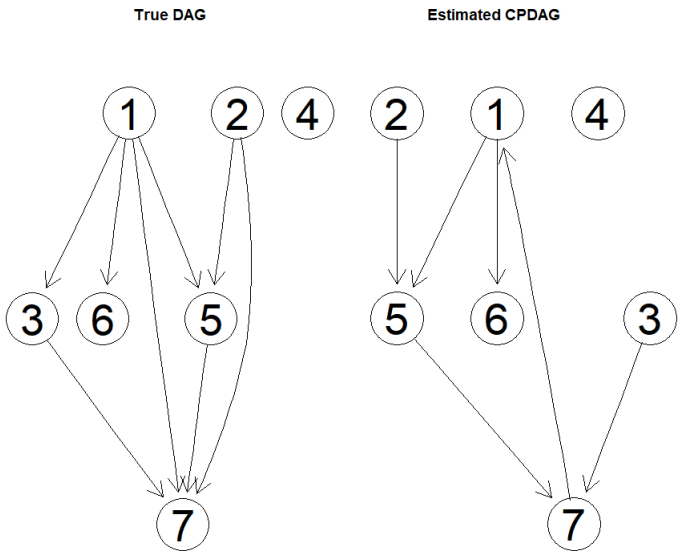


We can see that the average SHD achieves a minimum in the region around $\alpha = 0,01$ and $\alpha = 0,05$. For higher or lower significance levels, the average SHD increases. Hence we will be concentrated in cases where the significant level took minimum score in SHD. Lets see same examples of PC-algorithm for p=7 p=15,p=40 ,p=100 in case where $E[N] = 2$ and $E[N] = 5$ respectively.

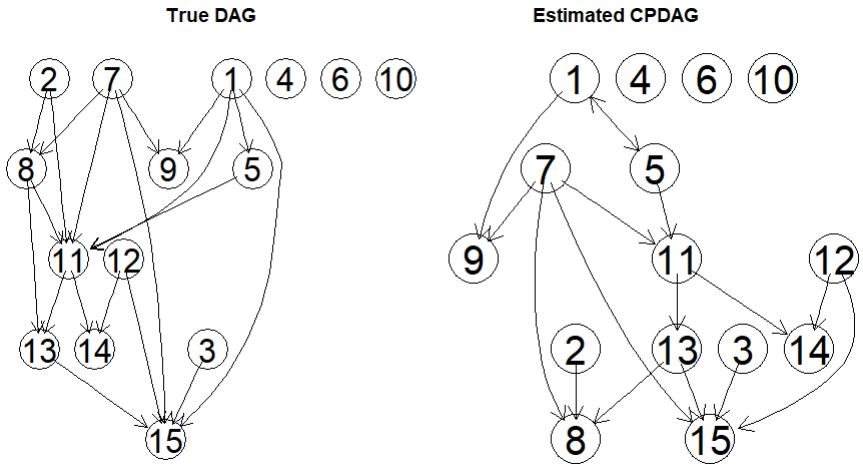
Lets start with :

$$p = 7, E[N] = 2,$$

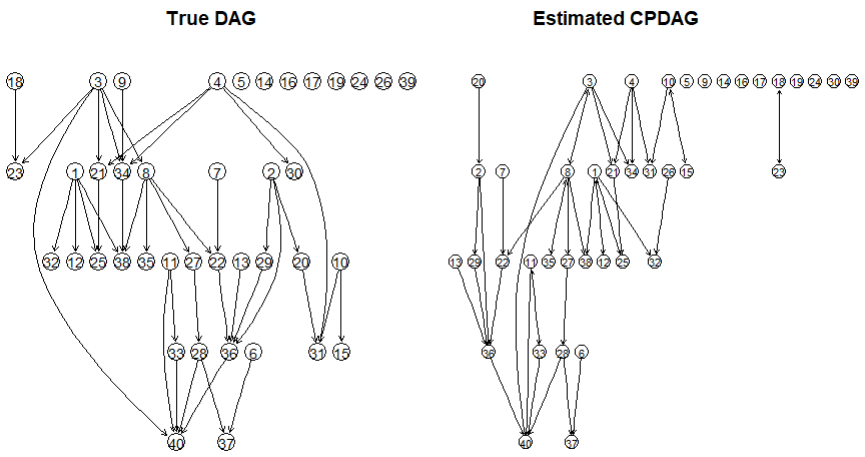
In this case we can see in figure 7.1 the minimum are catches in $a = 0.05$.



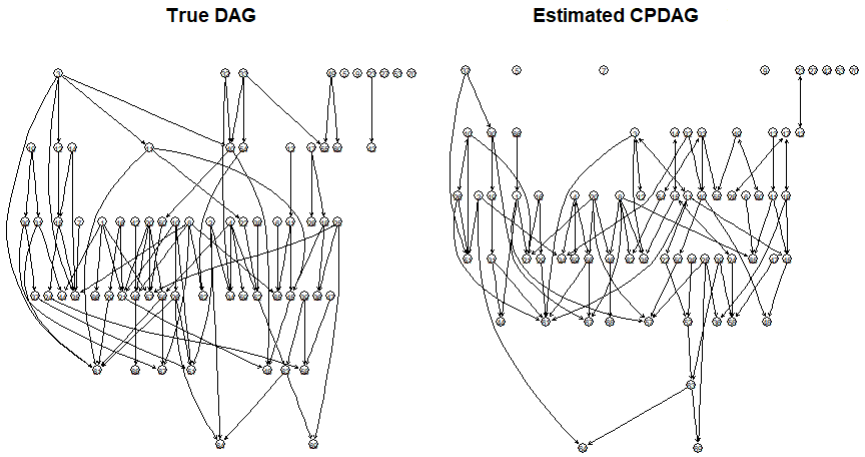
$$p = 15, E[N] = 2, a = 0.05$$



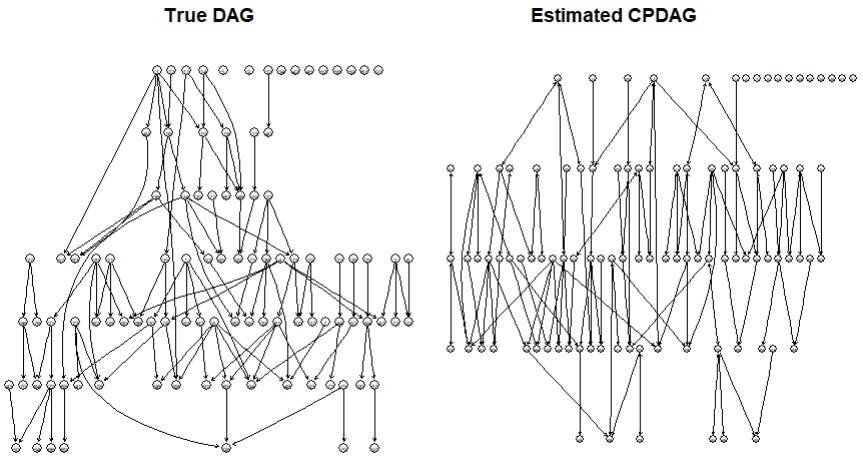
$$p = 40, E[N] = 2, a = 0.05$$



$$p = 70, E[N] = 2, a = 0.05$$

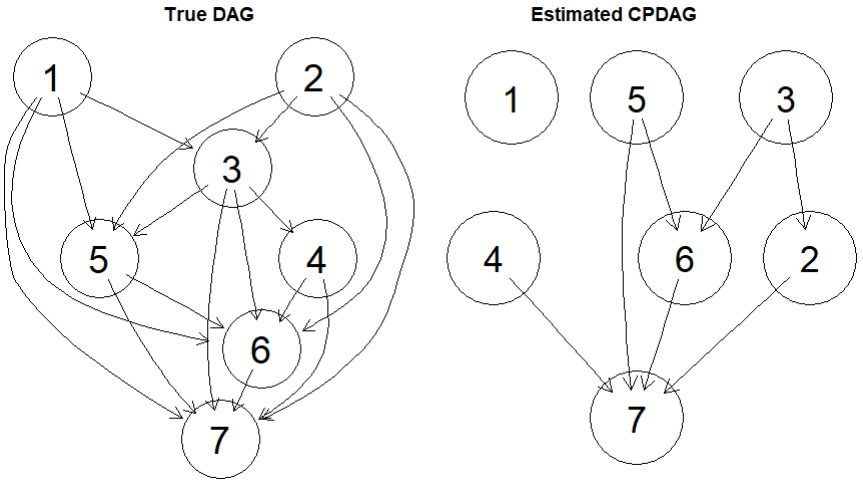


$$p = 100, E[N] = 2, a = 0.05$$

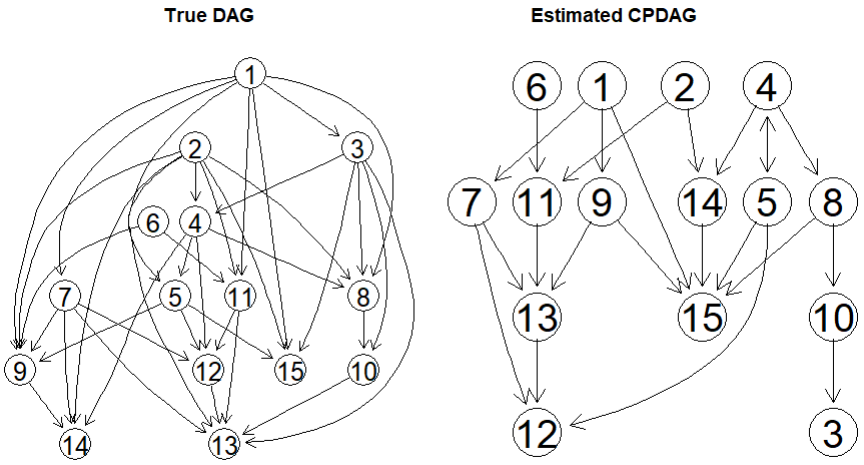


For simplicity reasons because the DAG be very complicated in $E[N] = 5$ we mention only the cases $p \in \{7, 15, 40\}$

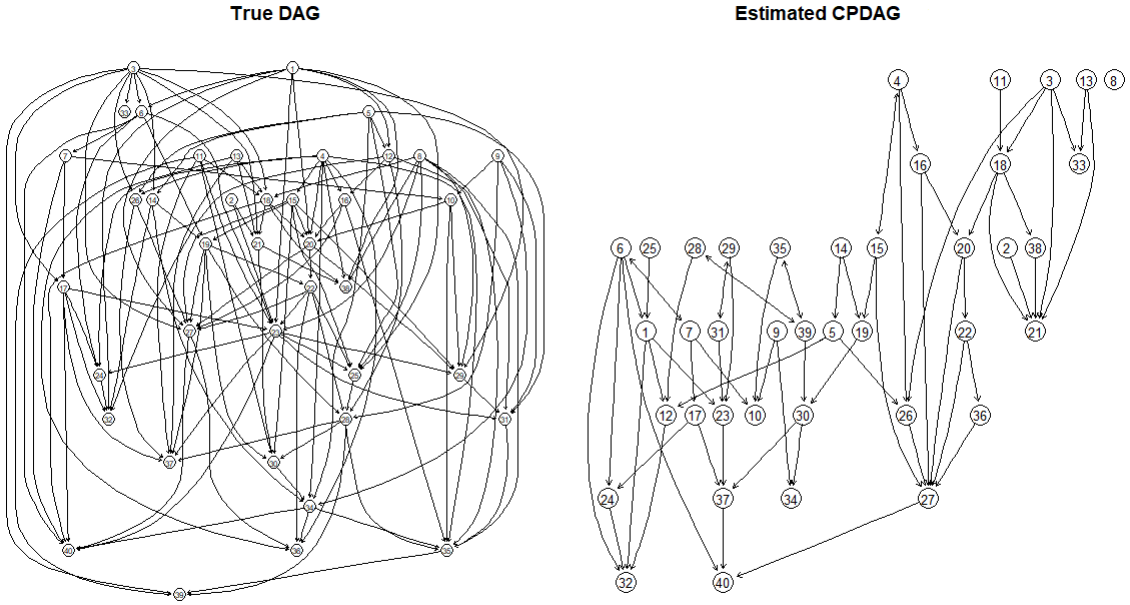
$$p = 7, E[N] = 5, a = 0.05$$



$$p = 15, E[N] = 5, a = 0.05$$



$$p = 40, E[N] = 2, a = 0.05$$



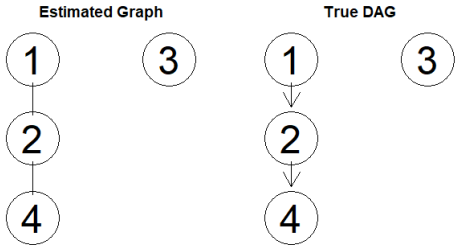
2. Estimating the Size of the Causal Effect When The Causal Structure Is Known

Until now we have assumed the following, the observational data are multivariate Gaussian also are faithful to the true (but unknown) underlying causal DAG and finally we know all the variables

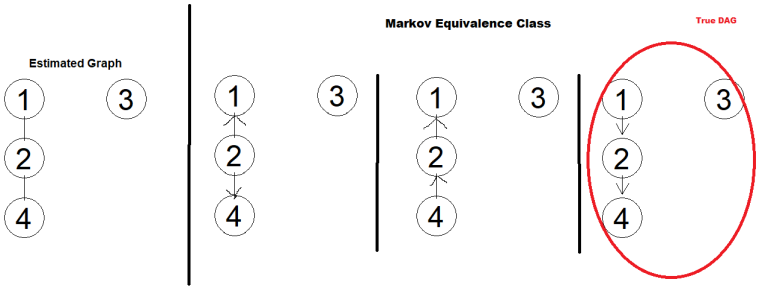
in the model, since there aren't hidden variables. Under these assumptions we will use algorithms to estimate the total causal effect of a variable x on y, total causal effect is defined via Pearl's do-calculus as $\frac{\partial E[Y|do(X=x)]}{\partial x}$ (this value does not depend on x, since under Gaussianity assumption the conditional expectations $E[Y|do(X=x)]$ is linear).
The algorithm IDA(pcalg-package in R) perform this procedure.

IDA

We have data from four Gaussian variables X_1, X_2, X_3, X_4 [2]. Assuming that the causal structure is unknown and we want to infer the causal effect of node 1 *to* 2. First, we estimate the equivalence class of DAGs ,or CPDAG, using pc-algorithm.
For example:



Comparing the true DAG with the CPDAG in figure above. The CPDAG and the true DAG have the same skeleton. Two edges in the CPDAG are bi-directed as a result we can generate several graphs Markov equivalent with the CPDAG. Consequently, since we have 2 bi-directed edges we can generate $2^2 = 4$ possible DAGs but we search the markov equivalent DAGs so it's three since the directionality $1 \rightarrow 2 \leftarrow$ is rejected because is generated a new d-separation.
For each DAG $\mathcal{G}$ in the equivalence class, we apply the theory of do-calculus to estimate the total causal effect of X on Y. This can be done via a simple linear regression: if Y is not a parent of X , we take the regression coefficient of X in the regression $lm(y \sim x + pa(x))$, where pa(x) denotes the parents of x in the DAG G; if Y is a parent of X in G, we set the estimated causal effect to zero.



If the PC-algo estimates correct the CPDAG, as in this case, one of the DAGs in the Markov

[2] In the graph denote $X_1 = 1, X_2 = 2, X_3 = 3, X_4 = 4$

Equivalent class is the true. Since we do not know which is, we can't estimate the true total causal effect of X on Y. However, we can return the entire multiset of k estimated effects, k=the size of different DAGs in Markov equivalent class [3] ,i.e. in the above example we have only 3 Markov equivalent DAGs the DAG with $1 \rightarrow 2 \leftarrow 4$ is rejected since is generated a new v-structure. Therefore, the function ida (with option method = "global") will produce 3 possible values of causal effects (one for each DAG). As we can see in the figure above in two cases the Variable X_2 is in the Parent set of X_1 hence the causal effect is zero. i.e.

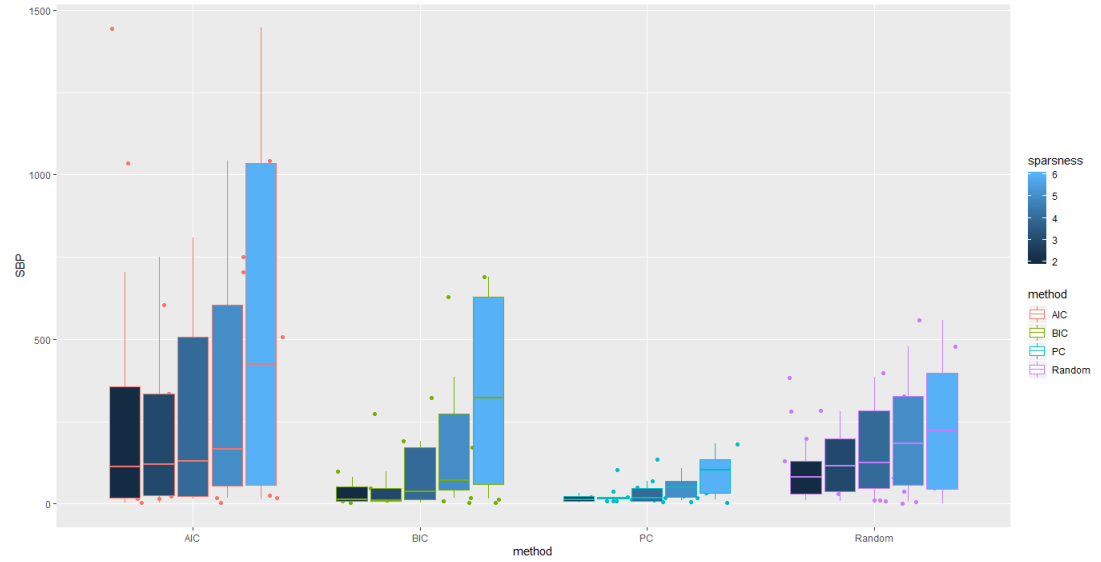
```
> ida (1,2, cov(d), pcfitIDA@graph , method = "global", verbose = FALSE)
[1] 0.0000000 0.0000000 0.3633399
```

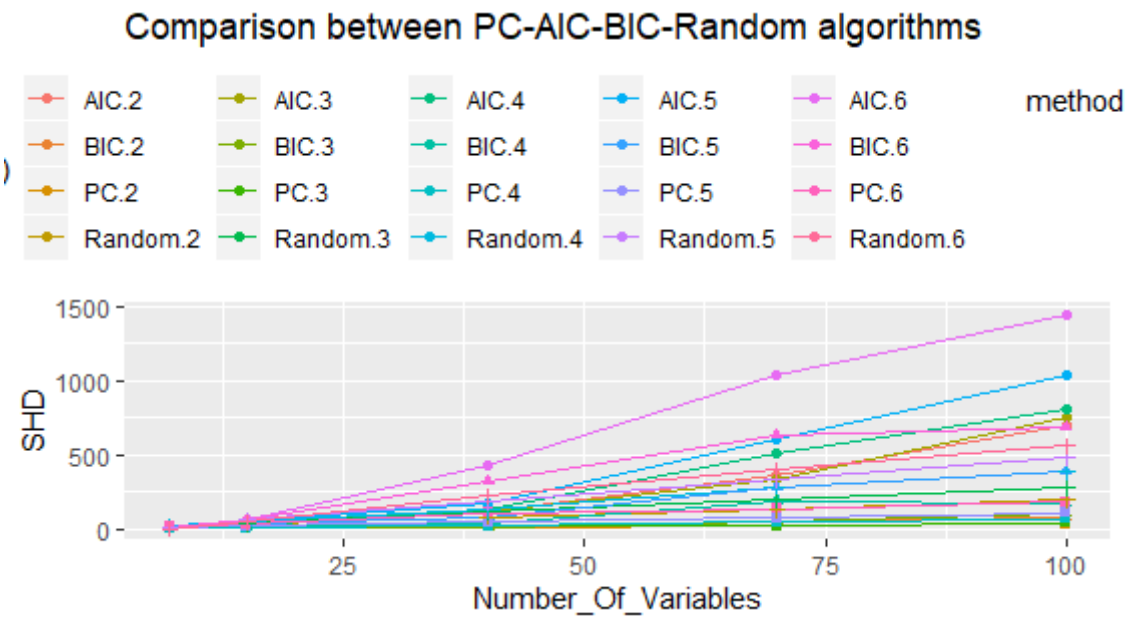
We can verify using linear regression of X_1 and X_2 in the estimated data from the true DAG.

```
Call:
lm(formula = data$X2 ~ data$X1)

Residuals:
      Min        1Q    Median        3Q       Max
-2.67604 -0.69420  0.03133  0.73969  3.06729

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06930    0.03194    2.17   0.0303 *
data$X1      0.36334    0.03042   11.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.009 on 998 degrees of freedom
Multiple R-squared:  0.125,    Adjusted R-squared:  0.1242
F-statistic: 142.6 on 1 and 998 DF,  p-value: < 2.2e-16
```

[3] Markov equivalent graph have the same skeleton and the same v-structures hence we be careful

3. Benchmarks





4. Code

Evaluating the PC algorithm using different significant levels in different levels of changing

```
install .packages("BiocManager")
BiocManager:: install ("Rgraphviz")
BiocManager:: install ("RBGL")
```

```
library (pcalg)
library (Rgraphviz)
library (RBGL)

library (dHSIC)
library (mgcv)


set .seed(42)
# sparsness
s1 =c(2,6)
# number of random variables
p1 =c(7, 15 , 40 , 70 , 100)
# sparsness of the graph
a1 =c(0.00005,0.0001 ,0.0005,0.001,0.005,0.01,0.05,0.1)

shd.PC_sign<-numeric(24)

#sumpe size
n <- 1000
set .seed(42)

i=1
for(s2 in s1)
  {
  for (p2 in p1)
    {
  for (a2 in a1)
     {
    s3 = s2/(p2-1)
    # # # # # # # # # # # # # # #
    # generate the random graph #
    # # # # # # # # # # # # # # #

    g <- randomDAG(p2,s3)
    # generate random samples
    d <- rmvDAG(n,g)

    # # # # # # # # # # # # # # # # # # # # # # #
    # estimate of the CPDAG with PC-algorythm #
    # # # # # # # # # # # # # # # # # # # # # # #
     suffStat <- list (C = cor(d), n = nrow(d))
```

```
      pcfit  <- pc( suffStat ,  indepTest=gaussCItest ,  p=p2,  alpha  =a2)

      # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
      #          Calculate  the  Hamming distance for  each  method           #
      # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

      shd.PC_sign[i]  <- shd(g, pcfit )

      i=i+1

      }
    }
}

 library ( ggplot2 )

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# Calculate  the  average  Hamming Distance for all  values  of  sparsness
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

AvSHD=numeric(length(a1))
for(k  in  1:length(a1)){
for(j  in  1:length(PC_sign$number_of_variables)){
  if(PC_sign$sign_level[j]==a1[k]){AvSHD[k]=AvSHD[k]+PC_sign$PC[j]}
}
}
AvSHD=AvSHD/(length(s1)*length(p1))

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# Calculate  the  average  Hamming Distance for the  different  values  of  sparsness
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

AvSHDspr=matrix(0,2,length(a1))
for(t  in  1:2){
  for(  l  in  s1){
    for(k  in  1:length(a1)){
      for(j  in  1:length(PC_sign$number_of_variables)){
    if(PC_sign$sparness[j]==s1[t]&PC_sign$sign_level[j]==a1[k])
    {AvSHDspr[t,k]=AvSHDspr[t,k]+PC_sign$PC[j]}
      }
    }
  }
}
```

```
AvSHDspr=AvSHDspr/length(p1)

PC_sign <- data.frame(PC=shd.PC_sign,
                      number_of_variables=rep(p1,each=length(a1)),
                      sign_level=rep(a1),
                      sparness=rep(s1,each=length(a1)*length(p1)),
                      Aver_SHD=rep(AvSHD),
                      AverSHD_spr=c(rep(AvSHDspr[1,],length(p1)),rep(AvSHDspr[2,],length(p1))))
PC_sign

#use log( significant levels ) for better visual representation
  plot_logsig_PC <- ggplot(PC_sign,aes(x=log(sign_level), y=AverSHD_spr,
      group=sparness) ) +
  geom_line(aes( color=sparness))+
  geom_point(aes( color=sparness))
  plot_logsig_PC
```

IDA

```
set.seed(42)
#size of the sample
n=1000
# fix the sparsness and the number of variables at 2 and at 4.
s=2
s3 = 2/(4-1)
# # # # # # # # # # # # # # # # # # # # # # # # # # # #
# generate the random graph and simulate the data #
# # # # # # # # # # # # # # # # # # # # # # # # # # # #

g <- randomDAG(4,s3)
# generate random sample
d <- rmvDAG(n,g)

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
#  estimate of the CPDAG using PC-algorythm with significant level a=0.05 #
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
 suffStat <- list(C = cor(d), n = nrow(d))
pcfitIDA <- pc( suffStat , indepTest=gaussCItest , p=4, alpha =0.05)
par(mfrow=c(1,2))
plot(pcfitIDA,main = "Estimated Graph")
plot(g,main = "True DAG")
```

```
# # # # # # # # # # # # # # # # # # # # # # # # # # # #
# Estimate the causal effect of random X2 in X1 #
# # # # # # # # # # # # # # # # # # # # # # # # # # # #
ida (1,2, cov(d), pcfitIDA@graph , method = "global", verbose = FALSE)

data<- data.frame(d)
regression <- lm(data$X2~data$X1)
summary(regression)
```

Benchmarks

```
library (pcalg)
library (Rgraphviz)
library (RBGL)

library (dHSIC)
library (mgcv)


set .seed(42)
# sample size
n <- 1000
# sparsness of the graph
s2=c (2,3,4,5,6,7)
# number ofRandom Variables
p1 =c(7, 15 , 40 , 70 , 100)
# significant levels
a1 =c(0.00005,0.0001 ,0.0005,0.001,0.005,0.01,0.05,0.1)

#Evaluation Vectors with the Hamming Distance
shd. val_SB_BIC<-numeric(30)
shd. val_SB_AIC<-numeric(30)
shd. val_PC<-numeric(30)
shd. val_RND<-numeric(30)

set .seed(42)

i=1
# We chech
```

```
for (s2 in s2)
{
for (p2 in p1)
{
  s3 = s2/(p2-1)
  ##################
  # generate the random graph #
  ##################

  g <- randomDAG(p2,s3)
  # generate random samples
  d <- rmvDAG(n,g)

  ##################################################
  # estimate of the CPDAG with Score based methods based on BIC #
  ##################################################
  score <- new("GaussL0penObsScore", data = d)
  ges. fit .BIC <- ges(score)
  ges. fit .BIC.graph<-as(ges. fit .BIC$essgraph,"graphNEL")

  ##################################################
  # estimate of the CPDAG with Score based methods based on AIC #
  ##################################################
  score <- new("GaussL0penObsScore", data = d, lambda =1)
  ges. fit .AIC <- ges(score)
  ges. fit .AIC.graph<-as(ges. fit .AIC$essgraph,"graphNEL")

  ##########################
  # estimate of the CPDAG with PC-algorythm #
  ##########################
  suffStat <- list (C = cor(d), n = nrow(d))
  pcfit <- pc( suffStat , indepTest=gaussCItest , p=p2, alpha =0.005)

  ##########################
  # Random Choice For appropriate DAG #
  ##########################
  RDAG <- randomDAG(p2,s3)

  ##################################################
  #          Calculate the Hamming distance for each method          #
  ##################################################
  ges. fit .BIC$essgraph<-as(ges. fit .BIC$essgraph,"graphNEL")
  shd. val_RND[i]<-shd(g,RDAG )
  shd. val_PC[i] <- shd(g, pcfit )
```

```
  shd.val_SB_BIC[i] <- shd(g,ges.fit.BIC.graph)
  shd.val_SB_AIC[i] <- shd(g,ges.fit.AIC.graph)
  i=i+1
}
}

shd.val_SB_AIC=shd.val_SB_AIC[-c(26:30)]
shd.val_SB_BIC=shd.val_SB_BIC[-c(26:30)]
shd.val_PC=shd.val_PC[-c(26:30)]
shd.val_RND=shd.val_RND[-c(26:30)]

df2 <-
    data.frame(SHD=c(c(shd.val_SB_AIC),c(shd.val_SB_BIC),c(shd.val_PC),c(shd.val_RND)),
                    method=rep(c("AIC", "BIC","PC","Random"),
                        each=length(shd.val_SB_BIC)),
                    Number_Of_Variables=p1,
                    sparsness=c(rep(c (2,3,4,5,6) ,each=length(p1))))
df2


library (ggplot2)

graph_benchmarks1 <- ggplot(df2,
    aes(x=Number_Of_Variables,y=SHD,group=interaction(method,sparsness)))+
geom_line(aes( color= interaction (method,sparsness)))+
  geom_point(aes(color= interaction (method,sparsness),shape=method))+
  theme(legend.position="top")+
   ggtitle ("Comparison of PC-AIC-BIC-Random algorithms")
graph_benchmarks1

box_plot_benchmarks <- ggplot(df2, aes(x=method, y=SHD, color=method,fill=sparsness,
    group= interaction (method,sparsness)))+
  geom_jitter ()+
  geom_boxplot()
box_plot_benchmarks
```

Appendix

Chapter 8

Appendix

8.A Conditional-Independences

If X, Y, Z are random variables with a joint distribution P,we say that X is *conditional independent* of Y under P, and write $(X \perp\!\!\!\perp Y|Z)_P$, if, for any measurable set A in the sample space of X, there exists a version of conditional probability $P(A|Y, Z)$ which is a function of Z alone. Usually P will fixed and omitted from the notion.
Formally:

Definition 8.A.1 (Conditional Independence.)**.** Let X, Y, Z be random variables on $(\Omega, \mathcal{A}, P)$. We say that X is (Conditional) independent of Y given Z ,$(X \perp\!\!\!\perp Y|Z)_P$ if for all $A_X \in \sigma(X)$ and all $A_Y \in \sigma(Y)$,

$$\mathbb{E}(\mathbb{1}_{A_X \cap A_Y}|Z) = \mathbb{E}(\mathbb{1}_{A_X}|Z)\mathbb{E}(\mathbb{1}_{A_Y}|Z) \ \ a.s.$$

Using standard tools from measure theory, we can deduce equivalent forms for the abovedefinition.

Proposition 8.A.2. Let X, Y, Z be random variables on $(\Omega, \mathcal{A}, P)$. Then the following are equivalent.

1. $(X \perp\!\!\!\perp Y|Z)_P$
2. For all $A_X \in \sigma(X)$,$\mathbb{E}(\mathbb{1}_{A_X}|Y, Z) = \mathbb{E}(\mathbb{1}_{A_X}|Z)$
3. For all real, bounded and measurarble functions $f(X)$,
   $\mathbb{E}(f(X)|Y, Z) = \mathbb{E}(f(X)|Z)$ a.s..
4. For all real, bounded and measurable functions $f(X), g(Y)$,
   $\mathbb{E}(f(X)g(Y)|Z) = \mathbb{E}(g(X)|Z)$ a.s..

If Z is trivial we say that X is independent of Y , and write $X \perp\!\!\!\perp Y$.The notation is due to (Dawid, 1979) who studied the notion of conditinal independence in a systematic fashion. (Dawid, 1980) gives a formally treatment.

When X, Y and Z are discrete random variables the condition for the $X \perp\!\!\!\perp Y|Z$ simplifies as

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$$

where the equation holds for z with $P(Z = z) > 0$.
When the three variables admit a joint density, p, with respect to product measure, we have

$$X \perp\!\!\!\perp Y|Z \iff p_{XY|Z}(x, y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z) \tag{8.1}$$

where the equation is to hold almost surely with respect to P. If all densities are continuous ,the equality in 8.1 must hold for all z with $p_Z(z) > 0$. Here it is understood that all functions on a discrete space are considered continuous functions. The condition 8.1can be rewritten as :

$$X \perp\!\!\!\perp Y|Z \iff p_{XYZ}(x, y, z) = p_{XZ}(x, z)p_{YZ}(y, z) \tag{8.2}$$

and this equality must hold for all values of z when the densities are continuous.
The relation $X \perp\!\!\!\perp Y|Z$ has the following properties, where h denotes an arbitrary measurable function on the sample space of X:

1. if $X \perp\!\!\!\perp Y|Z$ then $Y \perp\!\!\!\perp X|Z$
2. if $X \perp\!\!\!\perp Y|Z$ and $U = h(X)$ then $U \perp\!\!\!\perp Y|Z$
3. if $X \perp\!\!\!\perp Y|Z$ and $U = h(X)$, then $X \perp\!\!\!\perp Y|(Z, U)$
4. if $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp WY|(Y, Z)$,then $X \perp\!\!\!\perp (W, Y)|Z$

If we use p as a general symbol for the probability density of the random variables corresponding to its arguments, the following statements are true:

$$X \perp\!\!\!\perp Y|Z \iff p(x, y, z) = \frac{p(x, z)p(y, z)}{p(z)} \tag{8.3}$$

$$X \perp\!\!\!\perp Y|Z \iff p(x|y, z) = p(x|z) \tag{8.4}$$

$$X \perp\!\!\!\perp Y|Z \iff p(x,z,y) = p(x|z)p(z|y) \tag{8.5}$$

$$X \perp\!\!\!\perp Y|Z \iff p(x,z,y) = h(x,z)k(y,z) \text{ for some h,k} \tag{8.6}$$

$$X \perp\!\!\!\perp Y|Z \iff p(x,z,y) = p(x|z)p(y|z) \tag{8.7}$$

The equation above hold from a set of triplets (x, y, z) with probability zero. If the densities are continuous functions (in particular if the state spaces are discrete), the equations hold whenever the quantities involved are well defined, i.e. when the densities of all conditioning variables are positive.

Another property of the conditional independence relation is often used:

5) if $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ then $X \perp\!\!\!\perp (Y, Z)$

However (5) does not hold universally but only under additional conditions - essentially that there be no non-trivial logical relationship between Y and Z. A trivial counterxample appears when $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = \frac{1}{2}$.We have however

Proposition 8.A.3. If the joint density of all variables with respect to a product measure is positive and continuous, then the statement (5) will hold true

8.B Graphical Representation of Dependency Knowledge

If we have acquired a body of knowledge Z and we wish to assess the truth of a proposition X, it is important to know whether it would be worthwhile to consult another proposition Y, which is not in Z. In other words, before we consult Y it is important to know if its true value can potentially generate new information relative to X, information not available from Z.

What logic would facilitate this type of reasoning?

A powerful formalism for relevant information is provided by probability theory, where the notion of relevance is identified with dependence or, more specifically, conditional independence.
An other representation, more abstract, are graphs. In general, graphs offer a useful representation for a variety of phenomena. For example: family relations, electric circles ,communication networks. But, what is the feature that makes this phenomenon graphical related? When we deal with a phenomenon where the notion of neighbourhood or connectedness is explicit, as the above, we

have no problem configuring a graph which represents the main feature of the phenomenon. However, when modelling relevancies it is hard to distinguish the direct neighbours from the undirected. For example:

Example 8.B.1. Firstly it's easy to understand the influence of the seasonal variation and the wetness on the slipperiness of the pavement. Namely, the knowledge of season (i.e. Season=winter) would be relevant to the slipperiness of the pavement. In addition, if we know the pavement is wet,Wet=yes, this information is relevant too. But if we want to infer a graph which depicts this flow of relevance information, then we may conclude that: the seasonal variation and wetness are directed neighbours of the slipperiness of the pavement. Is this appropriate?

The next chapters examine the feasibility of devising graphical representation for the relational structures. Precisely the next chapter examines the *undirected graph* as a graphical representation. In this representations the notion of neighbourhood is not specified in advance. Rather, what is given explicitly is the relation of " betweenness" .In other words, we are given the theoretic background to test whether any given subset S of elements intervenes in a relation between element X and Y.
As we discuss above given a probability Distribution P we can easily test when a Variables S intervenes in the relation of Variables $\{X, Y\}$. Namely, given a distribution P and and any three variables X, Y, Z it easy to verify whether X is independent of Y given Z. However, P does not dictate which variables should be regarded directed neighbours. But the above remark maybe can help us so as to formulate the notion of betweenness.

8.B.1 An Axiomatic Basis For Conditional Independence

In this chapter we will illustrate the notion of conditional independence in discrete variables. All the theoretic aspects is from (Pearl, 1988).

We will consider a finite set $\mathcal{U}$ of discrete random variables, where each variable $X \in \mathcal{U}$ may take on values from a finite domain D_X. We will use capital letters for variable names, e.g. (X, Y, Z), and lowercase letter for their specific values, e.g. (x, y, z). The sets of variables will be denoted by boldface (**X,Y,Z**) and their specific values, assignments, will be denoted by boldfaced lowercase letters (**x,y,z**).

Example 8.B.2. If **Z** stands for the set of variables $\{X, Y\}$, then z repressents the configuration $\{x, y\} : x \in D_X, y \in D_y$.

Greek letters will be used to represent individual variables. We shall repeatedly use the short notation $P(x)$ for the probabilities

$$P(X = x), \quad x \in D_X \,,$$

and will write $P(\mathbf{z})$ for the probabilities

$$P(\mathbf{Z}{=}\mathbf{z}) = P(X = x, Y = y), \;\; x \in \; D_X, y \in D_Y.$$

Definition 8.B.3. Let $\mathcal{U} = \{\alpha, \beta, ...\}$ be a finite set of variables with discrete values. Let $P(.)$be a joint probability function over the variables in $\mathcal{U}$, and let **X,Y** and **Z** stands for any three subsets of variables in $\mathcal{U}$. **X** and **Y** are said to be *conditionally independent* given **Z** if

$$P(\mathbf{x}|\mathbf{y},\mathbf{z}) = P(\mathbf{x}|\mathbf{z}) \quad whenever \quad P(\mathbf{y},\mathbf{z}) > 0$$

We will use the notion $I(X, Z, Y)_P$ to denote the independence of **X** and **Y** given **Z**, thus

$$I(\mathbf{X},\mathbf{Z},\mathbf{Y})_P \quad \text{iff} \quad P(\mathbf{x}|\mathbf{y},\mathbf{z}) = P(\mathbf{x}|\mathbf{z}) \tag{8.8}$$

for all values **x,y,z** such that $P(\mathbf{y},\mathbf{z}) > 0$. Unconditional independence (also called marginal independence) will be denoted by $I(\mathbf{X}, \emptyset, \mathbf{Y})$ i.e.

$$I(\mathbf{X}, \emptyset, \mathbf{Y})_P \text{ iff } \; P(\mathbf{x}|\mathbf{y}) = P(\mathbf{x}) \;\; \text{whenever } P(\mathbf{y}) > 0 \tag{8.9}$$

Note that $I(\mathbf{X},\mathbf{Y},\mathbf{Z})_P$ implies the conditional independence of all pairs of variables $\alpha \in \mathbf{X}$ and $\beta \in \mathbf{Y}$ but the converse is not necessarily true.

$$I(\mathbf{X},\mathbf{Y},\mathbf{Z})_P \Rightarrow I(\alpha, \mathbf{Y},\beta)_P \forall \alpha \in \mathbf{X}, \beta \in \mathbf{Y}$$

$$I(\mathbf{X},\mathbf{Y},\mathbf{Z})_P \not\Leftarrow I(\alpha, \mathbf{Y},\beta)_P \forall \alpha \in \mathbf{X}, \beta \in \mathbf{Y}$$

This rules out some logical and functional relationships:
The conditional independence $I(X, Z, Y)_P$ satisfies the following functional properties ,as we can see in (Lauritzen, 1996).

1. $I(\mathbf{X},\mathbf{Y},\mathbf{Z})_P \Longleftrightarrow P(x, y|z) = P(x|y)P(y|z)$
2. $I(\mathbf{X},\mathbf{Y},\mathbf{Z})_P \Longleftrightarrow \exists f, g : P(x, y, z) = f(x, z)g(y, z)$
3. $I(\mathbf{X},\mathbf{Y},\mathbf{Z})_P \Longleftrightarrow P(x, y, z) = P(x|y)P(y, z)$

Generally these proofs can be derived from the numeric representation of P. We now ask that, what logical conditions, being avoided of any reference to numerical forms, should constrain the relationship $I(\mathbf{X},\mathbf{Z},\mathbf{Y})_P$ if in some probability model P it stands for the statement "**X** is independent of **Y**,given that we know **Z**"

Theorem 8.B.4. Let **X,Y,Z** be three disjoint subsets of variables from $\mathcal{U}$. Suppose we learn about a conditional independence $I(\mathbf{X},\mathbf{Y},\mathbf{Z})_P$. Can we conclude other independence properties that must hold in the distribution?

- Symmetry $I(\mathbf{X,Z,Y})_P \iff I(\mathbf{Y,Z,X})_P$
- Decomposition $I(\mathbf{X,Z,Y} \cup \mathbf{W})_P \implies I(\mathbf{X,Z,Y})_P \& I(\mathbf{X,Z,W})_P$
- Weak Union $I(\mathbf{X,Z,Y} \cup \mathbf{W})_P \implies I(\mathbf{X,Z} \cup \mathbf{W,Y})_P$
- Contraction $I(\mathbf{X,Z,Y})_P \& I(\mathbf{X,Z} \cup \mathbf{Y,W})_P \implies I(\mathbf{X,Z,Y} \cup \mathbf{W})_P$

If P is strictly positive,then a fifth condition holds:

- Intersection $I(\mathbf{X,Z} \cup \mathbf{W,Y}) \& I(\mathbf{X,Z} \cup \mathbf{Y, W}) \implies I(\mathbf{X,Z,Y} \cup \mathbf{W})$

As we discuss in the introduction of this chapter Conditional independences is a very useful formal tool to represent relevant information. This is the best way "the key" to understand the intuition behind the above axioms.
If we give the meaning of $I(\mathbf{X,Z,Y})_P$:= as the information of **Y** is irrelevant about **X** given the knowledge base of **Z**. More details about the interpretation can be found in (Pearl, 1988).

- *Symmetry* $I(\mathbf{X,Z,Y})_P \iff I(\mathbf{Y,Z,X})_P$
  The symmetry axiom states that, in any state of knowledge Z, if Y tells us nothing new about X, then X tells us nothing new about Y.
- *Decomposition* $I(\mathbf{X,Z,Y} \cup \mathbf{W})_P \implies I(\mathbf{X,Z,Y}) \& I(\mathbf{X,Z,W})_P$
  The decomposition axiom asserts that if two combined items of information are judged irrelevant to X, then each separate item is irrelevant as well.
- *Weak Union* $I(\mathbf{X,Z,Y} \cup \mathbf{W})_P \implies I(\mathbf{X,Z} \cup \mathbf{W,Y})_P$
  The weak union axiom states that learning irrelevant information W cannot help the irrelevant information Y become relevant to X.
- *Contraction* $I(\mathbf{X,Z,Y})_P \& I(\mathbf{X,Z} \cup \mathbf{Y,W})_P \implies I(\mathbf{X,Z,Y} \cup \mathbf{X})_P$
  The contraction axiom states that if we judge W irrelevant to X after learning some irrelevant information Y, then W must have been irrelevant before we learned Y.
- *Intersection* $I(\mathbf{X,Z} \cup \mathbf{W,Y})_P \& I(\mathbf{X,Z} \cup \mathbf{Y, W})_P \implies I(\mathbf{X,Z,Y} \cup \mathbf{W})_P$
  unless Y affects X when we know W or if W is irrelevant to X when we know Y, then neither W nor Y (nor their combination) is relevant to X.

Together, the weak union and contraction properties mean that irrelevant information should not alter the relevance status of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant.

8.B.2 On the logic of representing Dependencies by Undirected Graphs

In this section we are restricted in the undirected graphs however in the next sections we will expand our examination into the directed too. As we mention before, maybe an alternative representation

of relevance information is graphs. If that is true we need a graphical procedure with the ability to depict the relevance information in graphs, working with the same logic as conditional independences in Probabilities. Also the results about the relevance information taken from graphs must be the same as the results taken from the Probability Independences. For that reason, we need a mechanism which can match the conditional independences and the graphical conditional independences. For that purpose we use the theory of Dependency models.

Definition 8.B.5. A *dependency model* $\mathcal{M}$ over a set of object $U = \{\alpha, \beta, ...\}$ is a subset of triplets **(X,Y,Z)** where the X, Y, and Z are three disjoint subset of elements of U. The triplets in $\mathcal{M}$ represent independences, that is, if **(X,Y,Z)** $\in \mathcal{M}$ asserts that X and Y interact only via Z or "X is independent of Y given Z". This statement is also written via $I(X,Z,Y)_{\mathcal{M}}$ with the subscript to be optional clarifying only the type of dependency.

Any Probabilistic model is a Dependency Model because every triplet **(X,Y,Z)** can test the validity of $I(\mathbf{X,Y,Z})_P$ using eq 8.1. Our task now is to characterize the dependency models captured by Undirected graphs. So, firstly we need to define the undirected graphs. Undirected Graph $\mathcal{G} = (V, E)$ is characterized by the set of nodes V and the set of edges E that connect certain pair of nodes in V. An example of undirected graph is depicted in the Figure 8.B.1.

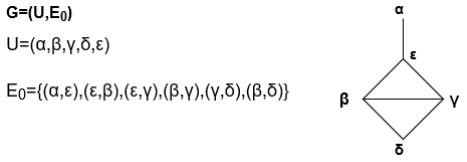


Figure 8.B.1: Example of Undirected Graph.

As it is mentioned we need to examine if the Dependency Models can be described in terms of undirected graphs. By a graphical representation of a Dependency model $\mathcal{M}$ we mean a direct correspondence between the elements in U, of $\mathcal{M}$, and the nodes in V, of $\mathcal{G}$, such that the topology[1] of $\mathcal{G}$ reflects some properties of $\mathcal{M}$. Ideally if a subset of nodes, Z, intercepts all paths between nodes of X and nodes of Y,write $(\mathbf{X,Z,Y})_{\mathcal{G}}$, then this interception should correspond to a conditional independence in $\mathcal{M}$ between **X,Y** given **Z** .Namely

$$(\mathbf{X,Z,Y})_{\mathcal{G}} \Longleftrightarrow (\mathbf{X,Z,Y})_{M} \tag{8.10}$$

Example 8.B.6. In this example we illustrate the meaning of the phrase "A subset set of nodes ,Z, intercepts all paths between nodes of X and nodes of Y,writes $(\mathbf{X,Z,Y})_{\mathcal{G}}$ ". In the graph of figure 8.B.1 if X stands for vertex α and **Y** stands for $\{\delta, \gamma\}$ then $Z_1 = \{\epsilon\}$ and $\mathbf{Z}_2 = \{\beta, \epsilon\}$ intercepts all paths between nodes of X and nodes of Y , namely the the paths $\alpha-\epsilon-\gamma, \alpha-\epsilon-\beta-\gamma, \alpha-\epsilon-\beta-\delta$

This correspondence would provide a clear graphical representation for the notion X does not effect on Y directly because the variables in Z mediated them. Unlike the requirement 8.10 is too strong. There often is no way of using vertex separation in graphs to display all the independence and dependencies embodied in a Dependency Model $\mathcal{M}$.

[1] the way which construct the graph

Definition 8.B.7. A graph $\mathcal{G}$ is a dependency map (D-Map)of a dependency model $\mathcal{M}$ if there is a one to one correspondence between the elements of U and the nodes V of $\mathcal{G}$, such that for every three disjoint subset **X,Y,Z** of elements we have

$$(\mathbf{X,Z,Y})_{\mathcal{G}} \Longleftarrow (\mathbf{X,Z,Y})_{M} \tag{8.11}$$

Similarly a graph $\mathcal{G}$ is an independency Map , I-Map, of $\mathcal{M}$ if

$$(\mathbf{X,Z,Y})_{\mathcal{G}} \Longrightarrow (\mathbf{X,Z,Y})_{M} \tag{8.12}$$

$\mathcal{G}$ is said to be perfect Map of $\mathcal{M}$ if it is both I-Map and D-Map.

An D-map guarantees that vertices found connected are indeed dependent in $\mathcal{M}$; it may ,however,display a pair of dependent variables in $\mathcal{M}$ as a pair separated vertices. An I-Map guarantees that vertices found to be separated correspond to independent variables but does not guarantee that all those shown to be connected are in fact dependent.

Example 8.B.8. Given a graph where every node is connected with all the other nodes "complete" is an I-Map. $The\ graph\ \mathcal{G}$ haven't any requirement about any correspondence in $\mathcal{M}$. For this every $\mathcal{M}$ is appropriate .
Just with the same logic empty graph is trivial D-Map.

8.B.3 Markov Networks

So in the previous section we define the mechanism which examines the validity of irrelevance information in graphs " called Vertex separation". If we assume as a dependency model $\mathcal{M}$ any arbitrary Probability distribution P we would like to learn if there are exist a perfect Map between undirected graph $\mathcal{G}$ and the P. Unfortunately in (Pearl & Paz, 1986) is proven that there isn't.

Lemma 8.B.9. There are probability distribution P for which no graph can be both D-map and I-map.

Proof. Graph separation always satisfies $(X, S_1, Y)_{\mathcal{G}} \Longrightarrow (X, S_1 \cup S_2, Y)_{\mathcal{G}}$ for any two subset S_1, S_2 of vertices. However this is not happened in all distributions, P. Some P's may induce the $(X, S_1, Y)_P$ and not $(X, S_1 \cup S_2, Y)_P$. Let P which induce the $(X, S_1, Y)_P$ and not $(X, S_1 \cup S_2, Y)_P$ the D-mapness forces G to display S_1 as a cut-set separating X and Y while I-mapness prevent $S_1 \cup S_2$ from separation of X and Y. As a result the aren't any graph which can satisfy these two requirements simultaneously. □

This can be easily explained by the flowing example:

Example 8.B.10. Consider the following experiment: We toss two Coins and a bell rings whenever the outcomes of the coins are the same. If we haven't any information about when the bell rings, the outcomes of the coins are mutually independent. So if there is appropriate graph, perfect-Map, then in the graph there aren't any link between this two variables. However, if we know when the bell

rings ,then the outcome of the one coin should change our opinion about the other coin, so become dependent. As a result this dependency must be depicted in the graph with a link, this is unable.

So, being unable to provide an graphical representation for all independences. But we can make the following compromise. Given any arbitrary distribution P, we can construct an I-map $\mathcal{G}$ of P that has the minimum number of edges. But what we mean with the state minimum number of edges?

Definition 8.B.11. A graph $\mathcal{G}$ is a minimal I-Map of P if no one edge of $\mathcal{G}$ can be deleted without destroying its I-Mapness . We call such Graph a *Markov Network*.

Theorem 8.B.12. Every P has a unique minimal I-map $\mathcal{G}_0$ producing by connecting only pairs of (α, β) for which :

$$(\alpha, U - \alpha - \beta, \beta)_P \text{ is False}$$

Proof. Follows directly from the proof of theorem 8.16 □

The above definition tell us how to construct an edge minimal Graph with the following properties :

- I-Mapness
  such that each time we observe a vertex x separated from y by a subset **S** of vertices,we can be guaranteed that variables x and y are independent in P given the values of variables in **S**.
- Minimality
  such that the set of neighbors assigned by the minimal $\mathcal{G}_0$ to each x coincides exactly with the smallest set of variables needed to shield x from the influence of all the other variables in the system. This Set of variables called Markov Boundary.

Definition 8.B.13. A Markov Boundary $B_P(a)$ of a variable α is a minimal subset S that renders α independent of all other variables i.e.

$$(\alpha, S, U - S - \alpha)_P, \alpha \notin S \tag{8.13}$$

and simmultaneously, no propes subset S' of S satisfies $(\alpha, S', U - S' - \alpha)_P$. If no S satisfies the 8.13, define $B_P(\alpha) = U - \alpha$

Theorem 8.B.14. Each variable α has a unique Markov boundary $B_P(\alpha)$ that coincides with the set of vertices $B_{\mathcal{G}_0}(a)$ adjacent to α in the Markov Net $\mathcal{G}_0$.

Proof. The proof of theorem 8.B.14 follows immediately from the proof of Theorem 8.B.20. □

8.B.4 Axiomatic Characterization of Graph Isomorph Dependencies

In the previous section we briefly illustrate the representation of Probability independences by Graphs. We conclude that: It is impossible given any distribution P to find graph which depict

all the probabilistic independences. As we see a probability Distributions is a specific type of Dependency model. The following question can be raised can we describe the family of Dependency Models which can be represented by graphs? In this section we will examine other types of Dependency models with special properties. Specifically we will try to formulate the class of Dependency-Models that can be described by the graph representation. So in the following section we introduce and establish an axiomatic characterization for the family of relations that are isomorphic with the vertex separation.

Definition 8.B.15. A dependency model $\mathcal{M}$ is said to be *graph isomorfic* if there exists an undirected graph $\mathcal{G} = (U, E)$ that is perfect Map of $\mathcal{M}$ i.e. for every disjoint subsets**X,Y,Z** of U we have:

$$(\mathbf{X,Z,Y})_{\mathcal{G}} \Longleftrightarrow (\mathbf{X,Z,Y})_{M}$$

Theorem 8.B.16. A necessary and sufficient condition for a dependency model $\mathcal{M}$ to be graph isomorph is that $I(X, Z, Y)_{\mathcal{M}}$ satisfies the following five independent axiom

1. Symmetry $I(\mathbf{X,Z,Y})_{\mathcal{M}} \Longleftrightarrow I(\mathbf{Y,Z,X})_{\mathcal{M}}$
2. Decomposition $I(\mathbf{X,Z,Y} \cup \mathbf{W})_{\mathcal{M}} \Longrightarrow I(\mathbf{X,Z,Y})_{\mathcal{M}} \& I(\mathbf{X,Z,W})_{\mathcal{M}}$
3. Intersection $I(\mathbf{X,Z} \cup \mathbf{W,Y})_{\mathcal{M}} \& I(\mathbf{X,Z} \cup \mathbf{Y,}\ \mathbf{W})_{\mathcal{M}} \Longrightarrow I(\mathbf{X,Z,Y} \cup \mathbf{W})_{\mathcal{M}}$
4. Strong Union $I(\mathbf{X,Z,Y})_{\mathcal{M}} \Longrightarrow I(\mathbf{X,Z} \cup \mathbf{W,Y})_{\mathcal{M}}$
5. Transitivity : $I(\mathbf{X,Z,Y})_{\mathcal{M}} \Longrightarrow I(\mathbf{X,Z,}\gamma)_{\mathcal{M}}$ or $I(\mathbf{X,Z,}\gamma)_{\mathcal{M}}\ \forall\gamma \notin \mathbf{X} \cup \mathbf{Z} \cup \mathbf{Y}$

Proof. [⇐]The axioms are clearly satisfied for vertex separation in graphs. So we prove the " necessary" part of the theorem. The logical independence of the five axioms can be demonstrated by letting U contain for elements and showing that is always possible to contrive a subset I of triplets that violets one axiom and satisfies the other four.

[⇒] the equation of Intersection and Strong Union imply the converse of Decomposition. Meaning I is completly defined by th set of triplets $I(\alpha, \mathbf{Z}, \beta)$ in which α and β are individual elements of U :

$$I(X, Z, Y) \Longleftrightarrow I(\alpha, \mathbf{Z}, \beta) \forall \alpha \in \mathbf{X}\ and\ \beta \in \mathbf{Y} \tag{8.14}$$

[→] If I(**X,Z,Y**)where$\mathbf{Y} = \{y_1, ..., y_n\}$ and $\mathbf{X} = \{x_1, ..., x_n\} \Longrightarrow [Decomposition]$
$I(\mathbf{X}, \mathbf{Z,Y} - \{y_1\})\ \&\ I(\mathbf{X,Z}, y_1) \Longrightarrow [Decomposition]$
$I(\mathbf{X,Z}, y_i\})\forall y_i \in \mathbf{Y} \Longrightarrow [Decomposition, Symetry]$
$I(x_i, \mathbf{Z}, y_i)\forall x_i \in \mathbf{X} \& y_i \in \mathbf{Y}$
[←] If $I(x_i$**,Z,**$y_i)\forall x_i \in \mathbf{X} \& y_i \in \mathbf{Y}$.
Then $I(x_1$**,Z,**$y_1)\ \&\ I(x_1$**,Z,**$y_2) \Longrightarrow [StrongUnion]$

$I(x_1\mathbf{,Z} \cup \{y_2\}, y_1)$ & $I(x_1\mathbf{,Z} \cup \{y_1\}, y_2) \Longrightarrow [Intersection]$
$I(x_1\mathbf{,Z}, \{y_1, y_2\}) \Longrightarrow [Repeat\ with\ the\ other\ y_i]$
$I(x_1\mathbf{,Z}, \mathbf{Y}) \Longrightarrow [Symetry]$
$I(\mathbf{Y}, \mathbf{Z}, x_1) \Longrightarrow [follow\ the\ same\ prosedure\ with\ x_i]$
I(**X,Z,Y**)

We must show for any set I of triplets (**X,Y,Z**) satisfing the above axioms, there exist a Graph such that (**X,Z,Y**) is in I iff **Z** is an cutset in $\mathcal{G}$ that separates **X** from **Y**.We show that $\mathcal{G}_0 = (U, E_0)$ is an such graph, where $(\alpha, \beta) \notin E$ iff $I(\alpha, Z, \beta)$. Under the Remark above it is sufficientto show only:

$$I(\alpha, S, \beta) \Longleftrightarrow ((\alpha, S, \beta))_{\mathcal{G}} \text{ where a,b, S } \subseteq U \tag{8.15}$$

The converse [$\Leftarrow$]follows automatically from the construction of $\mathcal{G}_0$. The other direction [$\Rightarrow$] is proved by descending induction. For $|S| = n - 2$ the theorem holds automatically, because of the contraction of $\mathcal{G}_0$. Assume the theorem holds for all S size $|S| = k \leq n - 2$. Let S' be any set of size $|S'| = k - 1$. For $k \leq n - 2$, there exist an element $\gamma \notin S' \cup \alpha \cup \beta$ and using the Strong union , we have $I(\alpha, S', \beta) \Rightarrow I(\alpha, S' \cup \gamma, \beta)$. By the transitivity axiom we have $I(\alpha, S', \beta) \Rightarrow I(\alpha, S', \gamma)\ or\ \ I(\gamma, S', \beta)$. Applying Strong union again we have , $I(\alpha, S', \gamma) \Rightarrow I(\alpha, S' \cup \beta, \gamma)$ The middle argument $S' \cup \beta$ and $S' \cup \gamma$ are both of size k, so by the induction hypothesis we have $(\alpha, S' \cup \beta, \gamma)_{\mathcal{G}_0}$ and $(\alpha, S' \cup \gamma, \beta)_{\mathcal{G}_0}$. By the intersection property of vertex separation in graphs we have $(\alpha, S', \beta \cup \gamma)_{\mathcal{G}_0} \Rightarrow [Decomposition](\alpha, S', \beta)_{\mathcal{G}_0}\&(\alpha, S', \gamma)_{\mathcal{G}_0}$

□

Having a complete characterization for vertex separation in graphs allows us to test whether given a dependency-Model lend itself to graphical representation.In fact, it is easy to show that probabilistic models may violate both of last two axioms. In the example 8.B.10 we can observe this failure. Strong Union is violated ,namely $(X, \emptyset, Y)_P \not\Rightarrow (X, Z, Y)_P$. Also Transitive is violated too, namely $(X, \emptyset, Y)_P \not\Rightarrow (X, \emptyset, Z)_P$ or $(Z, \emptyset, Y)_P$.

8.B.5 Graphoids and Semi-Graphoids

As we see, we have fail to provide isomorphic graphical representation for every Probabilistic model , Lemma 8.B.9 and example 8.B.10. For this reason we settle the following compromise: instead of complete graph isomorphism, we will consider only minimal I- Maps. In this chapter we will extend this idea, not only in the Probabilistic models P but in special category of Dependency models $\mathcal{M}$,the *Graphoids*.

Definition 8.B.17. A graph $\mathcal{G}$ is a minimal I-Map of a dependency model $\mathcal{M}$ if deleting any edge of $\mathcal{G}$ would make $\mathcal{G}$ cease to be an I-map. We call such a graph a Markov Network of $\mathcal{M}$.

Theorem 8.B.18. Every Dependency Model $\mathcal{M}$ satisfying symmetry, decomposition, and intersection has a unique minimal I-map $\mathcal{G}_0 = (U, E_0)$ produced by connecting only those pairs (α, β) for

which $I(\alpha, U - \{\alpha, \beta\}, \beta)_{\mathcal{M}}$ is false i.e.

$$(\alpha, \beta) \notin E_0 \Longleftrightarrow I(\alpha, U - \{\alpha, \beta\}, \beta)_{\mathcal{M}} \tag{8.16}$$

Proof. This proof is in (Pearl & Paz, 1986).
First we will prove that: $\mathcal{G}_0$ is an I-Map using descending induction.
Let $n = |U|$. For $|S| = n - 2$ the I-mapness of $\mathcal{G}$, is guaranteed by its method of contraction 8.16.
Assume the theorem hold for every $S' = k \leq n - 2$ and let S be any set with size $S = k - 1$ and $(x, S, y)_{\mathcal{G}}$. We distinguish two sub cases: $x \cup S \cup y = U$ and $x \cup S \cup y \neq U$.

If $x \cup S \cup y = U$ then either $|x| \geq 2$ or $|x| \geq 2$. Assume without loss of generality, that $|y| \geq 2$,$y = \gamma \cup y'$.
From $(x, S, y' \cup \gamma)_{\mathcal{G}} \Rightarrow (x, S, \gamma)_{\mathcal{G}} \& (x, S, y')_{\mathcal{G}}$.
Then $(x, S, \gamma)_{\mathcal{G}} \Rightarrow (x, S \cup y', \gamma)_{\mathcal{G}}$ & $(x, S, y')_{\mathcal{G}} \Rightarrow (x, S \cup \gamma, y')_{\mathcal{G}}$
We can observe that the two separating sets above, $S \cup \gamma$ and $S \cup y'$, are in size k . Therefore by the induction hypothesis we have
$(x, S \cup \gamma, y')_{\mathcal{G}} \Rightarrow I(x, S \cup \gamma, y')$ and $(x, S \cup y', \gamma)_{\mathcal{G}} \Rightarrow I(x, S \cup y', \gamma)$

If $x \cup S \cup y \neq U$ then there are exist at least one element δ which is not in $x \cup S \cup y$. For any such δ if we use Strong Union of Vertex separation and we have $(x, S, y)_{\mathcal{G}} \Rightarrow (x, S \cup \delta, y)_{\mathcal{G}}$ and if we use the transitive of vertex separation
$(x, S, y)_{\mathcal{G}} \Rightarrow (x, S, \delta)_{\mathcal{G}}$ or $(\delta, S, y)_{\mathcal{G}} \Rightarrow$
$(x, S, \delta)_{\mathcal{G}} \Rightarrow [StrongUnion](x, S \cup y, \delta)_{\mathcal{G}}$
or
$(\delta, S, y)_{\mathcal{G}} \Rightarrow [StrongUnion](\delta, S \cup x, y)_{\mathcal{G}}$
The separation sets above are at least $|S| + 1 = k$ in size. Therefore by the induction hypothesis :

$$(x, S \cup \delta, y)_{\mathcal{G}} \& (x, S \cup y, \delta)_{\mathcal{G}} \Rightarrow I(x, S \cup \delta, y) \& I(x, S \cup y, \delta) \tag{8.17}$$

OR

$$(x, S \cup \delta, y)_{\mathcal{G}} \& (\delta, S \cup x, y)_{\mathcal{G}} \Rightarrow I(x, S \cup \delta, y) \& I(\delta, S \cup x, y) \tag{8.18}$$

Applying the intersection property to eather 8.17 , 8.18 yields $I(x, S, y)$, which establish the I-Mapness of $\mathcal{G}_0$.
.
Next we show that $\mathcal{G}_0$ is an edge minimal and unique. Indeed if deleting an edge $(\alpha, \beta) \notin E_0$ leaves α separating from β by the complementary set $U - \alpha - \beta$, if the resulting graph is still an I-map, we can conclude $I(\alpha, U - \alpha - \beta, \beta)$. However from the method of constructing $\mathcal{G}_0$ and from $(\alpha, \beta) \in E_0$ we know that $(\alpha, U - \alpha - \beta, \beta)$ is not in I. Thus $(\alpha, U - \alpha - \beta, \beta)_{\mathcal{G}_0} \not\Rightarrow I(\alpha, U - \alpha - \beta, \beta)$.Thus no edge can deleted from $\mathcal{G}_0$ and its minimality and uniqueness are established. □

Definition 8.B.19. A Markov Blanket $BL_I(\alpha)$ of an element $\alpha \in U$ is any subset **S** of elements

for which

$$I(\alpha, S, U - \mathbf{S} - \alpha) \text{ and } \alpha \notin \mathbf{S} \tag{8.19}$$

A set called a Markov Boundary of α, denoted $B_I(\alpha)$, if it is a minimal Markov blanket of αi.e.not of the proper subsets satisfy 8.19.

The Boundary $B_I(\alpha)$ is be interpreted as the smallest set of elements that shield α from the influence of all other elements.

Theorem 8.B.20. Every element $\alpha \in U$in a dependency model $\mathcal{M}$ satisfying symmetry, decomposition, intersection and weak union,called Semi-Graphoid, has a unique Markov boundary $B_I(\alpha)$. Moreover, $B_I(\alpha)$ coincides with the set of vertices $B_{\mathcal{G}_0}(\alpha)$ adjacent to α in the minimal I-Map $\mathcal{G}_0$.

Proof. [$\Rightarrow$] Let $\mathbf{BL}^*(\alpha)$ stands for the set of all Markov blankets satisfing 8.19. $B_I(\alpha)$ is unique because every element of $\mathbf{BL}^*(\alpha)$ is in the form $B_I(\alpha) \cap S$ for some set S.
Proof
If $\emptyset \in \mathbf{BL}^*(\alpha)$ then $B_I(\alpha) = \emptyset$.
If $\emptyset \notin \mathbf{BL}^*(\alpha)$and $S_1, S_2 \in \mathbf{BL}^*(\alpha)$ and $S_1 \cap S_2 = \emptyset \Rightarrow$
$I(\alpha, S_1, U - S_1 - \alpha), I(\alpha, S_2, U - S_2 - \alpha)$
$I(\alpha, S_2, U - S_2 - \alpha) \Rightarrow [S_1 \cap S_2 = \emptyset]$
$I(\alpha, S_2, S_1 \cup (U - S_2 - S_1 - \alpha)) \Rightarrow [WeakUnion]$
$I(\alpha, S_2 \cup (U - S_2 - S_1 - \alpha), S_1) \Rightarrow I(\alpha, U - S_1 - \alpha, S_1)$
$FromI(\alpha, S_1, U - S_1 - \alpha)\&I(\alpha, U - S_1 - \alpha, S_1) \Rightarrow [Intersection]$
$I(\alpha, \emptyset, U - \alpha) \Rightarrow \emptyset \in \mathbf{BL}^*(\alpha)$ Contradiction.
Moreover $B_I(\alpha)$ equals to the intersection of all members of $BL_I(\alpha)$.
[$\Leftarrow$] Conversely every Markov blanket $BL \in BL_I^*(\alpha)$ remains in $BL_I^*(\alpha)$ after we add to it an arbitrary set of elements S' ,not cantaining α,follows from the weak union property .

$$BL \in BL_I^*(\alpha) \Rightarrow I(\alpha, BL, U - BL - \alpha) \Rightarrow I(\alpha, BL, S' \cup (U - S' - BL - \alpha)) \Rightarrow$$

$$\Rightarrow [WeakUnion] I(\alpha, BL \cup S', U - BL - \alpha) \Rightarrow$$

$$BL \cup S' \in BL_I^*(\alpha)$$

In particular if there is an element β outside $B_I(\alpha) \cup \alpha$ then $U - \alpha - \beta$ is in $BL^*(\alpha)$. From this we conclude that for every element β outside $B_I(\alpha)$we have $I(\alpha, U - \alpha - \beta, \beta)$, meaning β cannot be connected to a is $\mathcal{G}_0$. Thus

$$B_{\mathcal{G}_0}(\alpha) \subseteq B_I(\alpha)$$

To prove the other direction $\supseteq$ it is sufficient to show that $B_{\mathcal{G}_0}$ is in $BL^*(\alpha)$, but this follows from the fact that $\mathcal{G}_0$, as an I-Map. □

8.C Graphical Representation of Dependency Knowledge on DAGs Part II

8.C.1 Dependence semantics for Bayesian-Networks

In this chapter we will examine Bayesian-Networks. Generally speaking Bayesian Networks are DAGs in which their nodes represents variables, their arcs signify the existence of directed causal influence between the linked variables, and the strengths of these influences are expressed by conditional probabilities. The semantics of Bayesian-Networks demands a clear correspondence between the topology of the DAG and the dependence relationship portrayed by it. In the previous section we see that this correspondence in Markov-Networks, Undirected graphs, was based on a simple separation criterion, vertex separation. More precise, if the removal of a subset Z of nodes from the network render nodes X and Y disconnected were proclaimed to be independent given Z. However in DAGs we use a slightly more complex separability criterion, called d-separation which takes into consideration the directionality of the arrows in the Graph.

Definition 8.C.1. If **X,Y** and **Z** are three disjoint subsets of nodes in a DAG $\mathcal{D}$, then **Z** is said to *d-separate* **X** from **Y**,denoted $(\mathbf{X},\mathbf{Z},\mathbf{Y})_{\mathcal{D}}$ if there is no path between a node **X** and a node in **Y** along which the following two conditions hold:

1. every node with converging arrows is in **Z** or has an descendent in **Z**
2. every other node is outside **Z**

Precisely, as Judea Pearl mentions in the book (Pearl et al., 2016) p.51.
"The rule of d-separation for determining conditional independence by graphs (Definition 8.C.1) was introduced in (Pearl, 1986) and formally proved in (Verma & Pearl, 1990b) using the theory of graphoids (Pearl & Paz, 1986)."

Definition 8.C.2. A DAG $\mathcal{D}$ is said a I-map of a dependency model $\mathcal{M}$ if every d-separation condition displayed by $\mathcal{D}$ corresponds to a valid conditional independence relationships in $\mathcal{M}$if for every three disjoint sets of vertices **X,Y,Z** we have

$$(\mathbf{X},\mathbf{Z},\mathbf{Y})_{\mathcal{D}} \Longrightarrow I(\mathbf{X},\mathbf{Z},\mathbf{Y})_{M} \tag{8.20}$$

A DAG is a minimal I-map of $\mathcal{M}$ if none of its arrows can be deleted without destroying its I-mapness.

Now let's define *Bayesian-Networks*:

Definition 8.C.3. Given a probability distribution P on a set of variables U, a DAG $\mathcal{D} = (U, \mathcal{E})$ is called a Bayesian-Network of P iff $\mathcal{D}$ is a minimal I-Map of P.

As mentioned in (Verma & Pearl, 1990b) there is a procedure which produces a minimal I-map of any dependency model which is a semi-graphoid. Before we illustrate this algorithm is important

to define the notion of stratified protocol and the tail boundary. A tail boundary of a variable X is any set of lesser variables (with respect to the ordering) rendering X independent of all other lesser variables. A stratified protocol of a dependency model contains two things: an ordering of the variables, and a function that assigns a tail boundary to each variable X.

Definition 8.C.4. Let $\mathcal{M}$ be a dependency model defined on a set $\mathbf{U} = \{X_1, .., X_n\}$ of elements, and let d an ordering $(X_1, ..., X_i, ..)$ of the elements of $\mathbf{U}$. The *boundary strata* of $\mathcal{M}$ relative to d is an ordered set of subset of $\mathbf{U}$, $\{B_1, .., B_i, ...\}$ such that each B_i is a Markov Boundary of X_i w.r.t. the set of $U_{(i)} = \{X_1, .., X_{i-1}\}$ i.e. B_i is the minimal set satisfying $B_i \subseteq U_{(i)}$ and $I(X_i, B_i, U_{(i)} - B_i)$.
The DAG created by designated each B_i as a parents of vertex X_i is called boundary DAG of $\mathcal{M}$ relative to d.

However we have a problem in the complexity of the algorithm that examines every possible boundary-strata so as to find the Markov Boundary in every possible ordering. If we define a particular dependency model over n variables there are n! different possible orderings and for each ordering can be up to $\prod_{i=1}^{n} 2^{i-1}$ different sets of tail boundaries. Because under some ordering d the i-variable have $\sum_{k=0}^{i-1} \binom{i-1}{k} = 2^{i-1}$ different possible Boundaries strata and then multiply to find every possible Boundary for every variable i : $\prod_{i=1}^{n} 2^{i-1} = 2^{\sum_{i=1}^{n}(i-1)} = 2^{n\frac{n-1}{2}}$. So, $n!2^{n\frac{n-1}{2}}$ different Boundary strata.

Theorem 8.C.5. Let $\mathcal{M}$ be any semi-graphoid (i.e. the dependency model satisfies the axioms Symmetry, Decomposition, weak union, Contraction). If $\mathcal{D}$ is boundary DAG of $\mathcal{M}$ relative to any ordering d, then $\mathcal{D}$ is a minimal I-Map of $\mathcal{M}$.

Proof. (Verma & Pearl, 1990a) □

As we see from the previous chapter every probability distribution P is a semi-graphoid. Eventually:

Corollary 8.C.6. Given a Probability Distribution $P(X_1, .., X_n)$ and any order d of the variables, the DAG created be designating as parents of X_i any minimal set $Pa(X_i)$ of predecessors satisfying :

$$P(X_i|Pa(X_i)) = P(X_i|X_1, .., X_{i-1}), Pa(X_i) \subseteq \{X_1, .., X_{i-1}\}$$

is a Bayesian network, minimal I-map, of P. If P is strictly positive, then all the parents sets are unique and the Bayesian-Network is unique (given d).

Although the structure of Bayesian Network depends strongly on the node ordering d used constructing it, each network nevertheless is an I-Map of the underling distribution P. This means all conditional independences is valid in P and hence are independent of the constructing ordering.

Corollary 8.C.7. Given a DAG $\mathcal{D}$ and a probability distribution P, a necessary and sufficient condition for $\mathcal{D}$ to be a Bayesian network of P is that each variable X_i be conditionally independent

of all its non descendants, given it's parents $Pa(X_i)$, and that no proper subset of $Pa(X_i)$ satisfy the condition

Corollary 8.C.8. In a Bayesian-Network, the union of the following three types of neighbours is sufficient for forming the Markov Blanket of node X: the direct parents of X, the direct successors of X and all the direct parents of X's children (direct successors).

Under the above theorems we solve the problem of finding minimals I-maps from a distribution P. The algorithm consists of the following steps: first assign a total ordering d to the variables of P. For each variable i of P ,identify a minimal set of predecessors S_i; that renders i independent of all its other predecessors (in the ordering of the first step). Assign a direct link from every variable in S_i; to i. The resulting DAG is an I-map of P, and is minimal in the sense that no edge can be deleted without destroying its I-mapness. The input list L for this construction (we called it Markov Boundary) consists of n conditional independence statements, one for each variable, all of the form $I(i, S_i, U_{(i)} - S_i)$.

Definition 8.C.9. A conditional independence statement σ logically follows from a set Σ of such statements if σ holds for every distribution that obeys Σ. In such case we also say that σ is a valid consequence of Σ.

Hence we can understand how difficult is someone to conclude if some independence statement σ is a valid consequence of a set of independences L. However d-separation is offers us an easiest and much delicate solution. If we prove that DAG graphically verifies every conditional independence statement that logically follows from input list L(Markov Boundary) then we can conclude if a statement σ is a valid consequence of an input list L and then σ holds in every distribution obeys L. Equivalently, every graphically-unverified statement in DAG is not a valid consequence of L. We will call this completeness of d-separation criterion.

Clearly, the constructed DAG represents more independences than those listed in the input list, namely, all those that are graphically verified by the d-separation criterion. The above analysis guarantees that all graphically-verified statements are indeed valid in P i.e., the DAG is an I-map of P. However, we will shows in the next section that the constructed DAG has an additional property; it graphically-verifies every conditional independence statement that logically follows from L (i.e. holds in every distribution that obeys L). Hence, we cannot hope to improve the d-separation criterion to display more independences, because all valid consequences of L (which defines $\mathcal{D}$) are already captured by d -separation.

8.C.2 Completeness of d-Separation

A Bayesian network can be viewed as an inference instrument for deducing new independence relationships from those used in constructing network. The topology of network is assembled from the list of independence statements that comprise the boundary strata. This input list implies a host of additional statements, many of which can be deduced from the network by graphical criteria such as d-separation For example the network in Figure8.C.1 :

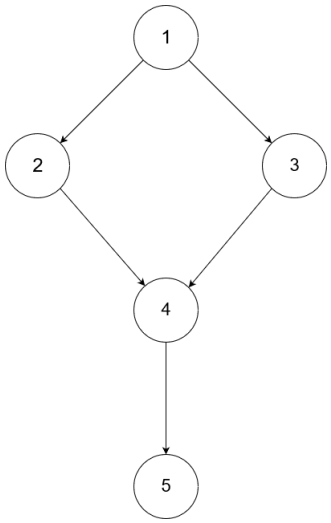


Figure 8.C.1

Is contracted from the Boundary-Strata

$$B_2 = \{1\} \ , \ B_3 = \{1\} \ , \ B_4 = \{2, 3\} \ , \ B_5 = \{4\}$$

Represented the input list

$$L = \{I(2, 1, \emptyset), I(3, 1, \emptyset), I(4, 23, 1), I(5, 4, 123)\}$$

New independence relationships all of them valid consequence of L, can be deduced from the network e.g. $I(5, 23, 1)$ or $I(3, 124, 5)$. This raises the following questions :

1. Can be d-separation be improved ? Can a more sophisticated criterion reveal additional independences of the input information ?
2. Are there valid consequences that escape graphical representation algorithm ?

The answer of two question is no. every valid consequence of the input information L must shown up as a d-separation condition in the DAG build from L. This follows from the next theorem

Theorem 8.C.10. For any DAG $\mathcal{D}$ there exists a probability distribution P such that $\mathcal{D}$ is a perfect map of P relative to d-separation i.e. P embodies all the independences portrayed in $\mathcal{D}$,and no other.

Proof. (Geiger & Pearl, 1990) □

Corollary 8.C.11. Given a list L of independence relationships in the form of boundary strata, a Bayesian Network combined with d-separation criterion constitutes a polynomially sound and complete inference mechanism relative to the closure of L i.e. it identifies in polynomial time every conditional independence relationship that follows logically from those in L.

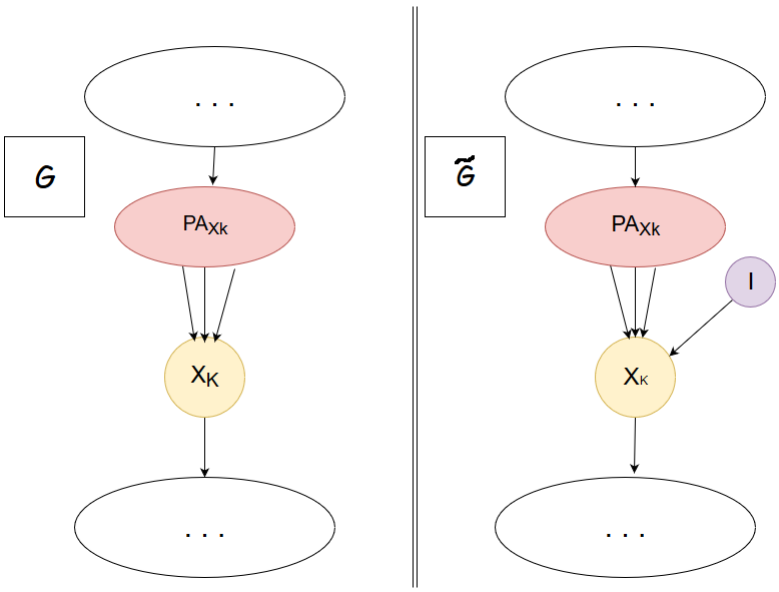
From this corollary is makes us clear that its impossible for some valid consequence σ of input list L to escape detection by d-separation.

8.D Valid Adjustment Set

Our goal in this chapter is to prove the proposition 1..2. To do so we must see another way to do interventions in the system.
Concider a SCM S with structural assignments

$$X_j = f_j(PA_j, N_j)$$

and the intervention $do(X_k = x_k)$. Now let's consider an alternative (but sometimes more appealing) account of intervention which treats the intervention as a variable within the system (Pearl, 2009). We can construct a new SCM $\tilde{S}$ that equals S with the only difference $\tilde{S}$ has one more variable I, which represents whether we have intervention in X_k .Also I is in the set of parents of X_k and does not have any other neighbors.The new structural assignments



$$I := N_I$$

$$X_k := \begin{cases} f_k(PA_k, N_k) & if \quad I = 0 \\ x_k & if \quad I = 1 \end{cases}$$

$$X_j = f_j(PA_j, N_j) \quad j \neq k$$

where $N_I \sim Ber(0.5)$, for example.

So, if I=0

$$\underbrace{p^{\tilde{S}}(x_1,...,x_p|I=0)}_{I \text{ has not parents ``intervntion=condition''}} = p^{\tilde{S};do(I=0)}(x_1,...,x_p) =$$

$$= p^{S}(x_1,...,x_p)$$

which corresponds in observation settings .
If I=1

$$\underbrace{p^{\tilde{S}}(x_1,...,x_p|I=1)}_{I \text{ has not parents ``intervntion=condition''}} = p^{\tilde{S},do(I=1)}(x_1,...,x_p) =$$

$$= p^{S,do(X_k=x_k)}(x_1,...,x_p)$$

Now we can use the Markov Condition to solve the problem of the searching of valid adjustment set .We can observe that :
if the set A and I is d-separated by a set of variables **B** then:

$$A \perp\!\!\!\perp_{\tilde{\mathcal{G}}} I|\mathbf{B} \Rightarrow p^{\tilde{\mathcal{S}}}(a|\mathbf{b}, I=0) = p^{\tilde{\mathcal{S}}}(a|\mathbf{b}, I=1)$$

$$\Rightarrow p^{\mathcal{S}}(a|b) = p^{\mathcal{S};do(X_k=x_k)}(a|b)$$

and after the discussion above

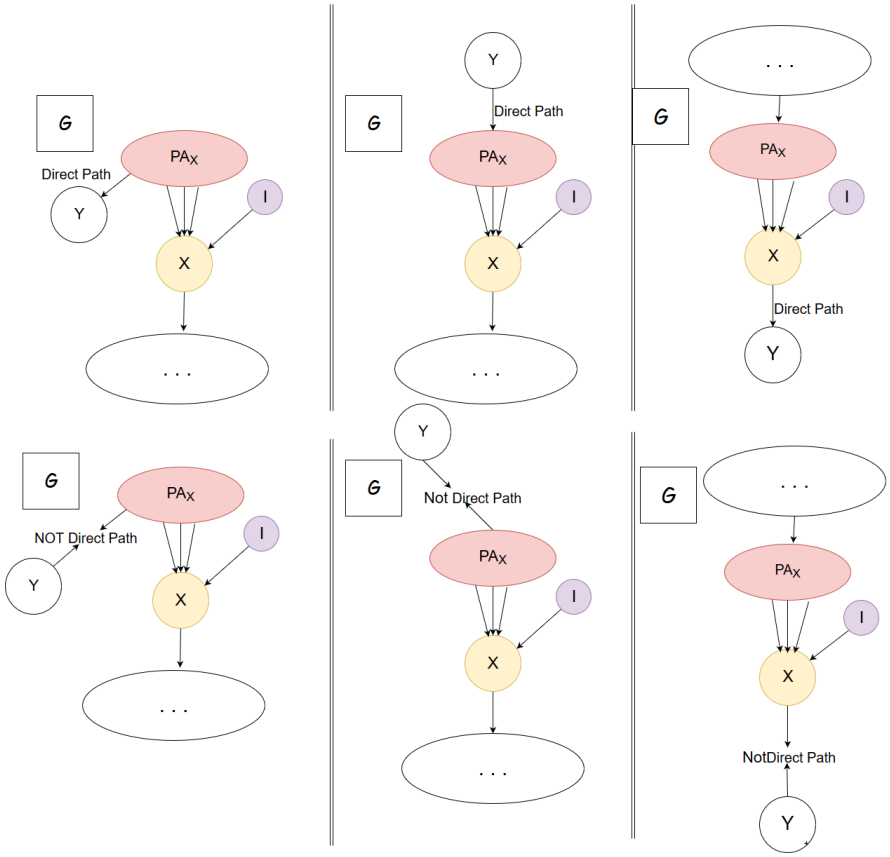
$$p^{S}(a|b) = p^{S,do(X_k=x_k)}(a|b)$$

So from equations 5.7 we searching sets **Z** witch satsfies :

$$Y \; d-sep_{\tilde{G}} \; I \; given\{\mathbf{Z}, X\} \;\; and \;\; Z \; d-sep_{\tilde{G}} \; I \tag{8.21}$$

8.D.1 Parents Adjustment

Lets see the first case :
If $Z = PA_X$ then we can see in the figure bellow all the possible graphs:



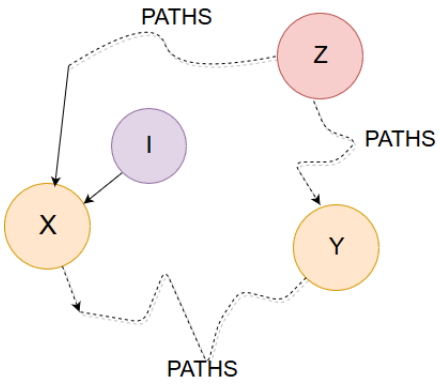
we can see in all possible cases .

$$X \; d-sep_{\tilde{G}} \; I \mid PA_X, Y \quad and \quad PA_X \; d-sep_{\tilde{G}} \; I \tag{8.22}$$

As a result once we identify the parents of X, the rest of the graph can be discarded, and the causal effect can be evaluated from the adjustment formula. However in most practical cases, the set of Xs parents will contain unobserved variables that would prevent us from calculating the conditional probabilities in the adjustment formula. For that reason we have the others criteria like

8.D.2 Back Door Criterion

According to the back door condition (i) of the definition of the back-door criterion, no node in Z is a descendant of X . So, all paths between I and each node $z \in Z$ have (at least) a collider ,the variable X(see the following figure).



Using the Markov condition we have: *Zd-separate* $I \Rightarrow Z \perp\!\!\!\perp I$.
Thus

$$p^{\tilde{S},do(I=1)}(z) = p^{\tilde{S}}(z|I=1) = p^{\tilde{S}}(z|I=0) = p^{S}(z)$$

Assuming the case which all back-door paths from X to Y blocked, then all paths from I to Y must go through the children of X , as a result this paths would be blocked if we condition on X. Hence using the Markov property $Y \perp\!\!\!\perp_{\tilde{G}} I|X \Rightarrow Y \perp\!\!\!\perp I|X$.
In the general case all back-door paths from X to Y aren't blocked but according to the back-door condition (ii), Z blocks every path between X and Y that contains an arrow into X. Thus the previous thinking $Y \perp\!\!\!\perp_{\tilde{G}} I|X \Rightarrow Y \perp\!\!\!\perp I|X$ holds only if we condition and on Z ,because Z blocks every back door path. Hence using the Markov condition: $Y \perp\!\!\!\perp_{\tilde{G}} I|(X,Z) \Rightarrow Y \perp\!\!\!\perp I|(X,Z)$

$$p^{\tilde{S},do(I=1)}(y|x,z) = p^{\tilde{S}}(z|I=1) = p^{\tilde{S}}(y|x,zI=0) = p^{S}(y|x,z)$$

8.D.3 Toward Necessity

The third statement "**toward necessity** " is very important because all valid adjustment sets can be characterized. The proof in (Shpitser, VanderWeele, & Robins, 2012).

8.E Linear-Gaussian Systems

A random vector $\mathbf{X} = \{X_1, .., X_n\}$ is multivariate normal if any linear combination of that random variables $X_1, X_2, .., X_n$ is normal distributed. In other words:

$$a_1X_1 + a_2X_2 + .. + a_nX_n$$

has a normal distribution for any constants $a_1, .., a_n$.
Let Linear-Gaussian SCM $\mathcal{S}$ over variables $\mathbf{X}$ with normally distributed error terms U_i. Without loss of generality we can assume zero mean in every error term . Hence the random vector consisting with

all the error terms, **U**, follows multi-normal distribution with mean the zero-vector, $0_n := (0, .., 0)$ where n:=size of different variables in **X**, and variance-covariance matrix the diagonal [2] Σ.

$$\Sigma = \begin{pmatrix} Var[U_1] & 0 & .. & 0 \\ 0 & Var[U_2] & .. & 0 \\ 0 & 0 & .. & Var[U_n] \end{pmatrix}$$

Every variable $X_i \in \mathbf{X}$ can be written as linear combination of the error terms, since the SCM is acyclic, so the random vector **X** follows Gaussian distribution, also every subset of **X** follows the Gaussian distribution, with mean again the zero-vector, $0_n := (0, .., 0)$, and covariance matrix $\tilde{\Sigma}$. Let $K \subset \mathbf{X}$ since the random vector $K \subset \mathbf{X}$ we have that $K \sim \mathcal{N}(0_{k_1+k_2}, \Sigma_1)$. Now consider partitioning of K into two components X_1 and X_2 of dimension k_1 and k_2 respectively, that is,

$$K = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Let we want to deduce the conditional distribution of X_1 given that $X_2 = x_2$. First write

$$\Sigma_1 = \begin{pmatrix} \Sigma_{X_1X_1} & \Sigma_{X_1X_2} \\ \Sigma_{X_2X_1} & \Sigma_{X_2X_2} \end{pmatrix} \tag{8.23}$$

where $\Sigma_{X_1X_1}$ is $k_1 \times k_1$, $\Sigma_{X_2X_2}$ is $k_2 \times k_2$ and $\Sigma_{X_1X_2}$=$\Sigma^t_{X_2X_1}$ and

$$\Sigma_1^{-1} = V_1 = \begin{pmatrix} V_{X_1X_1} & V_{X_1X_2} \\ V_{X_2X_1} & V_{X_2X_2} \end{pmatrix}$$

so that $\Sigma_{X_1X_1}V_1 = I_{k_1+k_2}$ ($I_{k_1+k_2}$ is the $k_1 + k_2 \times k_1 + k_2$ identical matrix)gives:

$$\begin{pmatrix} \Sigma_{X_1X_1} & \Sigma_{X_1X_2} \\ \Sigma_{X_2X_1} & \Sigma_{X_2X_2} \end{pmatrix} \begin{pmatrix} V_{X_1X_1} & V_{X_1X_2} \\ V_{X_2X_1} & V_{X_2X_2} \end{pmatrix} = \begin{pmatrix} I_{k_1} & 0_{k_1} \\ 0_{k_2} & I_{k_2} \end{pmatrix}$$

giving that, and after some calculations, we can prove that

$$X_2|X_1 = x_1 \sim \mathcal{N}(-V^{-1}_{X_2X_2}V_{X_2X_1}x_1, V_{X_2X_2})$$

or using the terms of Σ_1 matrix we can prove that:

$$X_2|X_1 = x_1 \sim \mathcal{N}(\Sigma_{X_2X_1}\Sigma^{-1}_{X_1X_1}x_1, \Sigma_{X_2X_2} - \Sigma_{X_2X_1}\Sigma^{-1}_{X_1X_1}\Sigma_{X_1X_2})$$

so the expectation mean is $E[X_2|X_1 = x_1] = \Sigma_{X_2X_1}\Sigma^{-1}_{X_1X_1}x_1$
if set $A = \Sigma_{X_2X_1}\Sigma^{-1}_{X_1X_1}$ then $E[X_2|X_1 = x_1] = Ax_1$

[2] diagonal since we have assumed uncorrelated error terms

Bibliography

Berkson, J. (1947). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*.

Bongers, S., Peters, J., Schölkopf, B., & Mooij, J. M. (2016). Theoretical aspects of cyclic structural causal models. *arXiv preprint arXiv:1611.06221*.

Bowden, R. J., & Turkington, D. A. (1990). *Instrumental variables*. Cambridge University Press.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, *3*(Nov), 507–554.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, *15*, 3921-3962. Retrieved from http://jmlr.org/papers/v15/colombo14a.html

Darmois, G. (1953). Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, 2–8.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(1), 1–31. Retrieved from http://www.jstor.org/stable/2984718

Dawid, A. P. (1980, 05). Conditional independence for statistical operations. *Ann. Statist.*, *8*(3), 598–617. Retrieved from https://doi.org/10.1214/aos/1176345011 doi: 10.1214/aos/1176345011

Duncan, O. D. (1975). *Introduction to structural equation models*. Academic Press, New York.

Frydenberg, M. (1990). The chain graph markov property. *Scandinavian Journal of Statistics*.

Geiger, D., & Heckerman, D. (1994). Learning gaussian networks. In *Uncertainty proceedings 1994* (pp. 235–243). Elsevier.

Geiger, D., & Pearl, J. (1990). On the logic of causal models. In *Proceedings of the fourth annual conference on uncertainty in artificial intelligence* (pp. 3–14). Amsterdam, The Netherlands, The Netherlands: North-Holland Publishing Co. Retrieved from http://dl.acm.org/citation.cfm?id=647231.719429

Goldberger, A., & Duncan, O. (1973). *Structural equation mod-els in the social sciences*. Seminar Press, New York.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2008). A kernel statistical test of independence. In *Advances in neural information processing systems* (pp. 585–592).

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, *11*(1), 1–12. Retrieved from http://www.jstor.org/stable/1905714

Heckerman, D., Meek, C., & Cooper, G. (1999). A bayesian approach to causal discovery. *Computation, causation, and discovery*, *19*, 141–166.

Huang, Y., & Valtorta, M. (2012). Pearl's calculus of intervention is complete. *arXiv preprint arXiv:1206.6831*.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475. Retrieved from http://www.jstor.org/stable/2951620

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, *8*(Mar), 613–636.

Lauritzen, S. L. (1996). Graphical models ,oxford: Clarendon press..

Mason, S. J. (1953). Feedback theory-some properties of signal flow graphs. *Proceedings of the IRE*, *41*(9), 1144–1156.

Mason, S. J. (1956). Feedback theory: Further properties of signal flow graphs.

Meek, C. (2013). Causal inference and causal explanation with background knowledge. *CoRR*, *abs/1302.4972*. Retrieved from http://arxiv.org/abs/1302.4972

Mooij, J., Janzing, D., Peters, J., & Schölkopf, B. (2009). Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 745–752).

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, *29*(3), 241–288.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). New York, NY, USA: Cambridge University Press.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J., & Paz, A. (1986). Graphoids: Graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z? In *Ecai*.

Pearl, J., & Verma, T. (1991). A theory of inferred causation. , 441–452.

Peters, J. (2008). *Asymmetries of time series under inverting their direction* (Unpublished doctoral dissertation). Universität Heidelberg Heidelberg, Germany.

Peters, J. (2015). On the intersection property of conditional independence and its application to causal discovery. *Journal of Causal Inference*, *3*(1), 97–108.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(5), 947–1012.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

Peters, J., Mooij, J. M., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, *15*(1), 2009–2053.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. , *66*(5), 688–701.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*(Oct), 2003–2030.

Shpitser, I., & Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the national conference on artificial intelligence* (Vol. 21, p. 1219).

Shpitser, I., VanderWeele, T., & Robins, J. M. (2012). On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*.

SKITOV, V. (1962). Linear combinations of independent random variables and the normal distribution law. *Selected Translations n Mathematical Statistics and Probability*, *2*, 211–28.

Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation, prediction, and search.

Tian, J., & Pearl, J. (2002). *Studies in causal reasoning and learning* (Unpublished doctoral dissertation). (AAI3070088)

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, *65*(1), 31–78.

Verma, T., & Pearl, J. (1990a). Causal networks: Semantics and expressiveness*. *Machine Intelligence and Pattern Recognition*, 69–76. Retrieved from http://dx.doi.org/10.1016/B978-0-444-88650-7.50011-1 doi: 10.1016/b978-0-444-88650-7.50011-1

Verma, T., & Pearl, J. (1990b). Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition* (Vol. 9, pp. 69–76). Elsevier.

Verma, T., & Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the sixth annual conference on uncertainty in artificial intelligence*. Elsevier Science Inc.

Wright, P. (1928). *The tariff on animal and vegetable oils*. Macmillan. Retrieved from https://books.google.gr/books?id=zJBBAAAAIAAJ

Wright, S. (1921). Correlation and causation.

Wright, S. (1934). The method of path coefficients. *The annals of mathematical statistics*, *5*(3), 161–215.

Yu, H. (2008). Introduction to bayesian networks.

Zhang, K., & Hyvarinen, A. (2012). On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*.