COPYRIGHT AND CITATION CONSIDERATIONS FOR THIS THESIS/ DISSERTATION

How to cite this thesis
Causality: Exploratory Data Analysis and Knowledge Discovery.

Pramod Kumar Parida

Department of Electrical & Electronic Engineering Science
University of Johannesburg

This dissertation is submitted for the degree of
Doctor of Philosophy

Faculty of Engineering and the Built Environment
July 2017
This is dedicated to my Bapa, Bau and Bhai . . .
Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in the University of Johannesburg, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

This dissertation contains fewer than 42,500 words including appendices, bibliography, footnotes, tables and equations and has 21 figures.

Pramod Kumar Parida
July 2017
Acknowledgements

I would like to acknowledge my supervisors, Prof Tshilidzi Marwala, for giving me the support and freedom in my studies, and Prof Snehashish Chakraverty, for this opportunity, support, and advice. I would also like to thank all my teachers who are the source of knowledge for the life-long lessons they have given me.

I deeply acknowledge my parents, Bau and Bapa, for making me what I am and for their faith in me in my struggle and success. My brother, Bhai, I thank for his support in every stage of my life. I cannot thank my family enough for their love: they know the depth and extent of my appreciation.

I would like to thank all my friends for their encouragement and support.

It is difficult to address everyone here but, I deeply acknowledge all the people who have helped me, either directly or indirectly.
Abstract

The phenomenon of cause and effect which rules the natural behaviour of the universe is simple in observation but complicated in interdependency. While all action and reaction states observed in time space are easier to work on, still the difficulty lies in the factor relations. Only knowing the facts/features without the time frame as they occurred/observed heightens the complexity of information retrieval. The relation of cause and effect is vital for knowing the past information which constructs the present state, although feature links remain debatable in this case. The study of Causality deals with these exploratory data analysis problems to inform all possible vital facts which can be extracted from the feature sets. Many researchers also consider the Causal Analysis as the golden standard in data mining and analysis.

As is frequently the case, this causal analysis is represented by directed acyclic graphs for simplification of complexity. The directed edges with weighted values inform the flow of information from source/parent to the receiver/child nodes in the graph. The definition of the causal structure for inference analysis is insufficient for many reasons, and the works concluded in this field are inadequate. Most of the techniques proposed provide limited structural analysis, while many others are not able to validate the required criteria for causal analysis. The background study of all the proposed articles with definite contributions towards causality have been studied and are thoroughly analyzed in the literature review.

All the methods proposed yet, use the bivariate model for causal analysis. In this scenario, the model, Linear Non-Gaussian Acyclic Model (LiNGAM) is the first to provide estimation for the most number of features. However, it is not completely effective in analyzing the causal models for datasets of mixed distribution types and also constructing a complete causal model from the estimated results is not possible. While using the fundamental structure of LiNGAM, the estimation process for causal detection is newly introduced by the method Altered-LiNGAM (ALiNGAM) in this work. ALiNGAM uses least square estimation on d-separable sets to find the probable causal directions in the observed feature set. The proposed
method ALiNGAM improves the performance to find causal directions using the LiNGAM system for different types of data which have previously not been supported.

The bivariate models fail to hold on to the primary assumptions on which the causal analysis is established, they are also found to be insufficient for causal analysis and model construction. On the other hand the Multivariate Additive Noise Model (MANM) is an effective system for causal inference, which preserves all the basic assumptions on the system. Furthermore, MANM is capable of handling nonlinear, non-Gaussian mixture models for both correlated and uncorrelated data types, which are not supported by bivariate models. Using this MANM, a new estimation method is proposed for complete causal discovery. The claims about the proposed MANM are verified in experimental and comparison tests with the available methods.

The use of conditional independence in probability analysis for causal studies is insufficient. The causal inference requires qualitative analysis, whereas conditional independence provides quantitative information which is not useful. Therefore, the causal independence is proposed in this work for qualitative analysis on causal models. For estimation of causal directions, the causal influence criteria are newly introduced to measure the effectiveness of causal connections. By knowing the causal connections, the higher valued directions can be found with highest causal effect from parent nodes to child nodes. As a solution to the problem of causal construction, the newly introduced causal levels provide an easier method to arrange the nodes. Using causal influence, causal levels can be detected which show the kinship relation in the feature set. The kinship values can be used to arrange the causal levels, and the complete causal level will be the complete causal model.

This work provides a comparison of previously proposed quantitative analysis with newly developed qualitative analysis using the proposed rules of causal independence, causal influence, and causal levels. The ease and effectiveness for complete causal discovery are analyzed with mathematical and experimental cases.

The work done in this thesis provides new contributions to the causal inference. The defined criteria and rules are added to make it complete and easier to inference and the new methods (ALiNGAM, MANM, Causal Independence, Causal Influence and Causal Levels) are proposed for the same. The valuable facts which worth mentioning are discussed in the conclusion, which also provides suggestions for future work on this topic.
# Table of Contents

**List of Figures**

**List of Tables**

**Nomenclature**

1 Introduction

1.1 An Introduction to Causality

1.1.1 A Graphical Representation of Causal Model

1.2 Problems in Causal Identification

1.3 Motivation and Objectives

1.3.1 Some Unaddressed Issues

1.3.2 Proposed Contributions

1.4 Structure of the Thesis

2 Literature Review

2.1 Introduction

2.2 Understanding Causality: Principles and Problems in Modelling

2.3 Probabilistic Approach of Bayesian Network

2.3.1 MCMC for Bayesian Structure Analysis

2.3.2 Variational Bayesian learning method for DAGs

2.3.3 Causal Inference using Plausible Markov Kernel


<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4</td>
<td>Probabilistic method of Graphical Analysis</td>
<td>19</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Ideal Parents</td>
<td>19</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Intervention Methods on Markov Equivalent Classes</td>
<td>20</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Method for Causal Hierarchy Identification</td>
<td>21</td>
</tr>
<tr>
<td>2.4.4</td>
<td>Markov Blanket for Causal Structure Learning</td>
<td>22</td>
</tr>
<tr>
<td>2.4.5</td>
<td>Information Geometry on Causal Inference</td>
<td>24</td>
</tr>
<tr>
<td>2.4.6</td>
<td>Geometry of faithfulness for Causal Inference</td>
<td>25</td>
</tr>
<tr>
<td>2.5</td>
<td>Statistical Approach for Causal Inference</td>
<td>27</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Kernel Independent Component Analysis</td>
<td>28</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Kernel Mutual Information for Independent Variables</td>
<td>28</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Kernel Constrained Covariance for Dependence Measurement</td>
<td>29</td>
</tr>
<tr>
<td>2.5.4</td>
<td>Measuring Statistical Dependence with Hilbert-Schmidt Norms</td>
<td>29</td>
</tr>
<tr>
<td>2.5.5</td>
<td>Kernel-based Causal Learning Algorithm</td>
<td>29</td>
</tr>
<tr>
<td>2.5.6</td>
<td>Kernel Measures of Independence for non-iid Data</td>
<td>30</td>
</tr>
<tr>
<td>2.5.7</td>
<td>Kernel-based Conditional Independence Test and Application in Causal Discovery</td>
<td>30</td>
</tr>
<tr>
<td>2.5.8</td>
<td>Kernel Independence Test for Random Processes</td>
<td>31</td>
</tr>
<tr>
<td>2.6</td>
<td>Algebraic Approaches to Causal Structure Construction</td>
<td>32</td>
</tr>
<tr>
<td>2.6.1</td>
<td>Linear Non-Gaussian Acyclic Model for Causal Discovery using ICA: ICA-LiNGAM</td>
<td>33</td>
</tr>
<tr>
<td>2.6.2</td>
<td>Distinguishing Causes from Effects using Nonlinear Acyclic Causal Models</td>
<td>34</td>
</tr>
<tr>
<td>2.6.3</td>
<td>Identifiability of the Post-Nonlinear Causal Model</td>
<td>35</td>
</tr>
<tr>
<td>2.6.4</td>
<td>Additive Noise Model for Nonlinear Causal Discovery</td>
<td>35</td>
</tr>
<tr>
<td>2.6.5</td>
<td>Direct Method for Learning a Linear Non-Gaussian Structural Equation Model: DirectLiNGAM</td>
<td>36</td>
</tr>
<tr>
<td>2.6.6</td>
<td>Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models</td>
<td>37</td>
</tr>
</tbody>
</table>
3 ALiNGAM for Nonlinear and Noisy Data 43
3.1 Introduction 43
3.2 Model Selection 45
3.3 Identifiability of the Model 46
3.4 Evaluation over the Model 49
3.5 Experiments, Results and Comparison Test 53
  3.5.1 Results for Synthetic Data 53
    3.5.1.1 Synthetic Dataset-1 53
    3.5.1.2 Synthetic Dataset-2 55
    3.5.1.3 Synthetic Dataset-3 57
    3.5.1.4 Synthetic Dataset-4 57
  3.5.2 Real World Data Test 59
    3.5.2.1 Abalone Data 59
    3.5.2.2 Energy Efficiency Data 60
  3.5.3 Comparison Test 61
3.6 Discussion and Conclusion 64

4 A Multivariate Additive Noise Model for Complete Causal Discovery 66
4.1 Introduction 66
4.2 Multivariate Additive Noise Model: MANM 69
4.3 Causal Identification using MANM 70
4.4 Estimations over MANM 73
4.5 Experiments 75
  4.5.1 Simulated Model Results 75
    4.5.1.1 Dataset 1 (Linear correlated non-Gaussian Model) 76
Table of Contents

4.5.1.2 Dataset 2 (Non-Linear Uncorrelated non-Gaussian Model)  
4.5.1.3 Dataset 3 (Non-Linear Uncorrelated Symmetric Non-Gaussian Mixture Model)  
4.5.1.4 Dataset 4 (Non-Linear Correlated Non-Symmetric Non-Gaussian Mixture Model)  
4.5.1.5 Dataset 5 (Correlated, Uncorrelated mixed datasets of both Gaussian, Non-Gaussian Distributions of Linear and Non-Linear Functions)  

4.5.2 Real World Test Results  
4.5.2.1 Forest Fires Data  
4.5.2.2 Abalone Data  
4.5.2.3 Boston Housing Data  

4.5.3 Comparison Test  
4.5.4 Performance Test  

4.6 Conclusion

5 Inscribing New Rules in Causality: Resolving the Divergence in Theory and Practical Implication  
5.1 Missing Sufficiency for Causal Discovery  
5.2 The Quantitative Analysis of Causality  

5.2.1 Model Selection  
5.2.2 Insufficiency of Bivariate Model  

5.3 Causal Inference: A Qualitative Analysis  

5.3.1 A Multivariate Additive Noise Model  

5.3.1.1 Causal Independence  
5.3.1.2 Causal Influence  
5.3.1.3 Causal Level  

5.4 Conclusion
Table of Contents

6 Conclusion & Future Work 105
   6.1 Conclusion ......................................................... 105
   6.2 Possibilities of Progress .......................................... 106

References 109

Appendix A Publications 114

Index 115
List of Figures

1.1 Multivariate directed acyclic causal model with $n$ number of nodes ........................................... 2

2.1 Causal Models ................................................................................................................................. 14
2.2 The Advancement graph for Articles under different groups where each bar represent an article as represented in Table 2.1 ......................................................... 41

3.1 A four node causal model with three kinship levels ................................................................. 48
3.2 Original and Simulated causal models for Synthetic Dataset-1 .................................................... 55
3.3 Causal models for Original and Simulated DAGs for Dataset-2 .................................................. 56
3.4 Causal models for Original and Simulated DAGs for Dataset-3 .................................................. 58
3.5 Causal models for Original and Simulated DAGs for Dataset-4 .................................................. 59
3.6 Performance of LiNGAM, DLiNGAM and ALiNGAM based on the number of correct directions estimated in simulated datasets ................................................................. 63
3.7 Performance of LiNGAM, DLiNGAM and ALiNGAM based on error rate computed from the number of missing and wrong directions estimated in different datasets ................................................................. 64

4.1 All possible causal relations in $\{X,Y,Z\}$ as shown in (a) and three different Markov equivalent models shown by (b), (c) and (d). ................................................................. 73
4.2 Causal Model for Dataset 1 ................................................................................................................ 77
4.3 Causal Model for Dataset 2 ................................................................................................................ 78
4.4 Causal Model for Dataset 3 ................................................................................................................ 79
4.5 Causal Model for Dataset 4 ................................................................................................................ 80
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>The Original and Simulated Causal Models for Data Set 5.</td>
<td>81</td>
</tr>
<tr>
<td>4.7</td>
<td>Causal relations in natural factors for Forest Fires Data.</td>
<td>82</td>
</tr>
<tr>
<td>4.8</td>
<td>Performance of ICA-LiNGAM, GDS, RESIT and MANM on Simulated Datasets for Model Under Fitting.</td>
<td>89</td>
</tr>
<tr>
<td>4.9</td>
<td>Performance for Model Reverse Fitting for ICA-LiNGAM, GDS, RESIT and MANM over Simulated Datasets.</td>
<td>90</td>
</tr>
<tr>
<td>4.10</td>
<td>Performance for Model Over Fitting on Simulated Datasets for ICA-LiNGAM, GDS, RESIT and MANM.</td>
<td>91</td>
</tr>
<tr>
<td>5.1</td>
<td>A 7 node Directed Acyclic Causal Model with 3 Causal Levels.</td>
<td>103</td>
</tr>
</tbody>
</table>
List of Tables

2.1 The papers reviewed and cited in sections following a year-long increasing development useful for the construction of Directed Acyclic Causal Models 39

2.2 The papers as referred in this Chapter to provide basic ideas on Causality and Bayesian Analysis, for the undirected causal model construction and some others under statistical works for causal inference. 40

3.1 DIRECTION matrix for Dataset-1 54

3.2 DIRECTION matrix for Dataset-2 56

3.3 DIRECTION matrix for Dataset-3 57

3.4 DIRECTION matrix for Dataset-4 58

3.5 DIRECTION matrix for Abalone Data 60

3.6 DIRECTION matrix for Energy Efficient Data 60

3.7 Comparison for causal direction estimations in LiNGAM, DLiNGAM, ALiNGAM 61

3.8 Runtime complexity of LiNGAM, DLiNGAM and ALiNGAM for estimating the causal directions in Simulated Datasets 63

4.1 The $IF$ matrix showing the CIF values for Causal Directions 76

4.2 Maximized CIF values for Dataset 2 78

4.3 Estimated CIF values in Abalone Data 83

4.4 Observed CIF values for causal relations in selected variables of Boston Housing Data 84

4.5 Comparison between ICA-LiNGAM, GDS, RESIT & MANM using Simulated Data Sets 86
4.6 Time complexity for methods ICA-LiNGAM, GDS, RESIT & MANM for solving different node sets and correctly estimated directed edge sets with respect to the assumed edge sets ............................................... 87
# Nomenclature

## Acronyms/Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIS</td>
<td>Annealed Impotence Sampling</td>
</tr>
<tr>
<td>ALiNGAM</td>
<td>Altered-Linear Non-Gaussian Acyclic Model</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>BN</td>
<td>Bayesian Network</td>
</tr>
<tr>
<td>CCA</td>
<td>Canonical Correlation Analysis</td>
</tr>
<tr>
<td>CD</td>
<td>Causal Direction</td>
</tr>
<tr>
<td>CI</td>
<td>Conditional Independence</td>
</tr>
<tr>
<td>CIF</td>
<td>Causal Influence Factor</td>
</tr>
<tr>
<td>CMC</td>
<td>Causal Markov Condition</td>
</tr>
<tr>
<td>COCO</td>
<td>Constrained Covariance</td>
</tr>
<tr>
<td>CPD</td>
<td>Conditional Probability Distribution</td>
</tr>
<tr>
<td>CSC</td>
<td>Cheeseman-Stutz Criterion</td>
</tr>
<tr>
<td>DAG</td>
<td>Directed Acyclic Graph</td>
</tr>
<tr>
<td>DLiNGAM</td>
<td>Direct Linear Non-Gaussian Acyclic Model</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>GDS</td>
<td>Greedy DAG Search</td>
</tr>
<tr>
<td>Nomenclature</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>GES</td>
<td>Greedy Equivalence Search</td>
</tr>
<tr>
<td>GGLAS</td>
<td>Gaussian Graphical Model based on Lasso Regression</td>
</tr>
<tr>
<td>GS</td>
<td>Greedy Search</td>
</tr>
<tr>
<td>HMC</td>
<td>Hybrid Monte Carlo</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HSCI</td>
<td>Hilbert-Schmidt Independence Criterion</td>
</tr>
<tr>
<td>HSN</td>
<td>Hilbert-Schmidt Norm</td>
</tr>
<tr>
<td>IC</td>
<td>Inductive Causation</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
</tr>
<tr>
<td>ICA-LiNGAM</td>
<td>Independent Component Analysis based Linear Non-Gaussian Acyclic Model</td>
</tr>
<tr>
<td>ID</td>
<td>Identification Algorithm</td>
</tr>
<tr>
<td>KC</td>
<td>Kernel Covariance</td>
</tr>
<tr>
<td>KCC</td>
<td>Kernel Constrained Covariance</td>
</tr>
<tr>
<td>KCCA</td>
<td>Kernel Canonical Correlation Analysis</td>
</tr>
<tr>
<td>KCI</td>
<td>Kernel based Conditional Independence</td>
</tr>
<tr>
<td>KERNELICA</td>
<td>Kernel Independent Component Analysis</td>
</tr>
<tr>
<td>KMI</td>
<td>Kernel Mutual Information</td>
</tr>
<tr>
<td>L1MB</td>
<td>$L_1$ regularized Markov Blanket Estimation</td>
</tr>
<tr>
<td>LHS</td>
<td>Left-Hand Side</td>
</tr>
<tr>
<td>LiNGAM</td>
<td>Linear Non-Gaussian Acyclic Model</td>
</tr>
<tr>
<td>MAG</td>
<td>Maximal Ancestral Graph</td>
</tr>
<tr>
<td>MANM</td>
<td>Multivariate Additive Noise Model</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum a Posterior</td>
</tr>
<tr>
<td>Nomenclature</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>MBS</td>
<td>Markov Blanket Search</td>
</tr>
<tr>
<td>MC</td>
<td>Markov Condition</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte Carlo</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimation</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi-Layer Perceptron</td>
</tr>
<tr>
<td>MRF</td>
<td>Markov Random Field</td>
</tr>
<tr>
<td>MWST</td>
<td>Maximum Weight Spanning Tree</td>
</tr>
<tr>
<td>ORDLAS</td>
<td>LARs based Order Search</td>
</tr>
<tr>
<td>PAG</td>
<td>Partial Ancestral Graph</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Component Analysis</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Distribution Function</td>
</tr>
<tr>
<td>PGM</td>
<td>Probabilistic Graphical Model</td>
</tr>
<tr>
<td>PNL</td>
<td>Post Non-Linear</td>
</tr>
<tr>
<td>RESIT</td>
<td>Regression with subsequent Independent Test</td>
</tr>
<tr>
<td>RFE</td>
<td>Recursive Feature Elimination</td>
</tr>
<tr>
<td>RHS</td>
<td>Right-Hand Side</td>
</tr>
<tr>
<td>RKHS</td>
<td>Reproducing Kernel Hilbert Spaces</td>
</tr>
<tr>
<td>SCM</td>
<td>Structural Causal Model</td>
</tr>
<tr>
<td>SEM</td>
<td>Structural Equation Model</td>
</tr>
<tr>
<td>SVR</td>
<td>Support Vector Regression</td>
</tr>
<tr>
<td>TC</td>
<td>Total Conditioning</td>
</tr>
<tr>
<td>TCbw</td>
<td>Total Conditioning Backward Feature Selection</td>
</tr>
<tr>
<td>VB</td>
<td>Variational Bayesian</td>
</tr>
<tr>
<td>VBEM</td>
<td>Variational Bayesian for Expectation Maximization</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 An Introduction to Causality

Any observation in natural phenomena, experiments or the effects in reality is due to a cause. Most of the time though the components or factors are responsible for such causes, the exact relations in them are unknown. There is the possibility that the factors just may not affect each other in a linear process, such as 1 affecting 2, then 2 affecting 3 and then so on. The process may be a more complicated one, where 1 may affect 2, then 3 and 4 in combination with 2 create a new process 5. It should not be forgotten that these processes may extend to a nonlinear form, making it even more complex to trace their relationships. For such processes and the curiosity to understand the cause-effect relations, the term Causality has been coined. In causality, the main purpose relies on finding each relation in factors to explain the process in the best possible way. Not involving in the complexity of representation of multiple features in multiple dimensions, the easiest way of representation is a two-dimensional graphical model with connected and directed edges to provide visualization to such processes. This will be easy to understand that connected edges show whether causality exists in the factors and directions to signify what affects what.

The meaning of causal is very clear in physics by only considering a force and its effect. In other fields, the derivation of variable relations is strictly dependent on characteristics of data and its domains. A more special case to causality, the Granger causality has been widely researched and the time factor relating to it helps to analyze it in simple steps, while the general view of causality is far more complicated and underdeveloped.

Defining a causal model for an irreversible system, it is ... "a directed acyclic graph with nodes representing the variables, connected by weighted directed edges to show the relation
from one to another and weights for the connection strength or the flow of information” (Pearl, 2000). It is not necessary for a causal model to be acyclic, though it can be cyclic for a reversible system. The depth of information is provided in the following sections.

### 1.1.1 A Graphical Representation of Causal Model

In multiple factors, if one affects the other, then there must be something such as energy, property or some influence which is passed from the cause to the effect. A unified name would be the ‘information’, which results in effect when transferred from the cause. Obviously, this information carries some weight, as there are not always the same changes or the same amount of effect created by the same cause-effect relation. So, information varies each time in case of even the same factor relations. This information can be represented in a model by providing the weights for the directed edges, showing the amount of information passed from the source to the receiver. All of these define basic properties of causality and it can be represented in a model as shown in Fig. 1.1.

![Directed Acyclic Causal Model](image)

**Fig. 1.1 Multivariate directed acyclic causal model with n number of nodes**

Fig. 1.1 shows a causal model, where factors are represented by nodes \{N_1, N_2, N_3, ..., N_n\}, an existing causal relation in the nodes are shown by the directed edges and the weights \{W_{12}, W_{13}, W_{14}, ..., W_{(n-1)n}\} on the edges shows the information that is passed on. To make it
1.2 Problems in Causal Identification

easier to understand, the source/cause is called the Parent and the effect/receiver as the Child. So $W_{12}$ shows the amount of information passed from parent $N_1$ to the child node $N_2$ and in a similar way for all the other nodes as shown in the Fig. 1.1. It can be noticed that the causal structure is arranged from the primary cause to the extended effects as shown in Fig. 1.1. This arrangement is termed a causal ordering and represented as $O(.)$. In this case for Fig. 1.1, the orders are $O(N_1) > O(N_2), O(N_3) > O(N_6)$ and similarly for all other nodes as they are represented. In particular, an order is derived from how the factors are related and how they are observed. This will be discussed later in the sections with examples.

Consider a more reasonable question that should arise by now: are the causal models cyclic or acyclic? Most natural processes are cyclic because of the interdependency in all the natural systems. Most of the artificial processes as caused by humans are acyclic. As in case of a causal model which depends on factors, it can be both cyclic and acyclic. Until now all have been busy considering the causal models as acyclic, or more precisely, Directed Acyclic Graphs (DAGs) as represented and solved in most of the papers. Why is this so? Because it is easier to analyze and represent, considering the huge amount of complexity one has to solve in the cyclic processes. In an acyclic model, the feature set complexity is less as information only flows in one direction, whereas in the case of cyclic models the information flow becomes more complicated as the paths cross each other.

This can be explained in this way: in a single-lane road, if all cars are going in the same direction, then there is less chance of an accident. But if in a single lane cars are going in both directions, then there is a high chance of an accident. The accident can be understood as an effect-identification problem in the case of causality. Whenever the relations in factors become cyclic, detecting the direction of information flow and effect identification becomes more complicated.

The above concepts provide the hints of problems which are going to be dealt with while solving causal models. Some major issues that have never been addressed in this regard are discussed in the following section.

## 1.2 Problems in Causal Identification

The following best explains the problems to be faced in contributing towards the development of causal analysis.

The first issue in causal inference which anyone will face is the availability of a good number of works to support the growth and quality in this area. The number of works
1.2 Problems in Causal Identification

done in causality or the number of papers available in causality is very limited. The reason for this is that causality is the general case of the study of causal analysis and inference without the consideration of time. Although the general theory of causality does not explain it explicitly, there are very few methods available which explicitly emphasizes causal inference in observation sets and construction of causal models. Most of the works are published in Artificial Intelligence (AI), and in machine learning-based conferences and journals, of which there are very few and most of them are referred to in this thesis. Most recently (in the past decade) some workshops and open competitions have been held to bring the work into the mainstream science, but there are still not enough.

A most common problem in all proposed methods by Shimizu et al. (2006), Hoyer et al. (2009), Peters et al. (2011), Shimizu et al. (2011), Hyvärinen and Smith (2013), and Peters et al. (2014) is that they use a bivariate model, which is entirely insufficient for causal analysis. This is discussed in detail in Chapter 5 with suitable examples as illustrations. As pointed out the DAG model does provide the ease to solve but not for causal inference, and the methods by which they all try to solve the problem is also another debatable point.

Going back to the properties by which the causal model in Fig. 1.1 has been defined, it is quite clear that no such evidence or significant factor have been provided which exclusively defines causal inference. As the information is defined to provide the quantization for transfer of force produced from the source/parent towards cause/child, the question remains what the significance of this information is if it does not inform us about causality (as it is represented until now). So, the significance of information lies with its quality to identify the causal influence or more specifically, the quantity of information passed from parent to a child does not signify the quality of inference in causality.

**Example:** Consider a case where to address some local issues of a region people have written letters to the state government office and in this process, many individuals have written many letters, but they were unsuccessful in their efforts. When one as a representative of the same group of people writes once or twice, things get fixed quickly. In the first case of individuals, the input information may be huge when grouped together but the influence is very little. In the second case, the community has a large influence, even though the input information is less, or from a smaller source. This is the exact case with causality. The information flowing from parent to child does not exclusively provide information for the goodness of the causal influence as shared in them. Chapters 4 and 5 broadly discuss the properties which define the goodness of causal influence and provides the methods to measure them.
1.3 Motivation and Objectives

The other two problems that arise in the causal analysis are the difference in conditional independence versus causal independence and the construction of the causal model. Whereas the conditional independence is a quantitative property, the causal independence is a qualitative property, and these were never represented or discussed before for causal inference. For causal construction, there is no particular method available to arrange the causal relations to provide a complete causal model. So, researchers have used many conventional methods and combined ideas to overcome this issue of model representation. This issue and solution to it are discussed in subsequent sections.

1.3 Motivation and Objectives

The raised problems and unavailability of papers may be a primary cause for motivation for this work, however the following are stronger reasons which make it a subject of great interest.

1.3.1 Some Unaddressed Issues

The basic propositions on causality and rules for causal structural constructions are explicitly proposed and discussed by Pearl (2000). The theoretical explanation for successful causal discovery primarily uses graph d-separable, Causal Markov Condition, and Markov Equivalent Classes to provide fundamental conditions/criteria for a later estimation process. These primary assumptions provide the reasoning for causal inference and for the development of causal modelling, all of which contribute towards the quantitative analysis of the causal models, which are discussed extensively in Chapter 5.

The models by Shimizu et al. (2006), Zhang and Hyvärinen (2008), Zhang and Hyvärinen (2009), Hoyer et al. (2009), Peters et al. (2011), Shimizu et al. (2011), Hyvärinen and Smith (2013), and Peters et al. (2014) which are proposed to analyze the system for causal inference, emphasize the estimation of path coefficients/connection strength values and external influences, errors, and additive noises. All of these values provide information which is helpful in finding the causal relations in the features, but none of these provide information regarding the magnitude of the influence or the goodness of causal inference.

While arranging the variables in the causal model, methods by Eliden et al. (2007), He and Geng (2008), Pellet and Elisseeff (2008), Shpitser and Pearl (2008), Zhang (2008), Janzing et al. (2012), and Uhler et al. (2013) and use different graphical construction techniques
to represent it. Question arises that when the complete causal relations are available for the causal structure construction, how it can be used for causal modelling? Primarily, the kinship relations show the change of causal influences in different relations. However, none of the available papers provide information on the arrangement of kinship relations and the successive causal influences in different stages while representing a causal model. All these questions make it difficult to estimate complete causal model from observational sets, and this is what makes it more interesting to work on.

1.3.2 Proposed Contributions

The method multivariate additive noise model can easily analyze multiple variables in less time and with more efficiency. Such a model is crucial for system analysis and provides an edge over general bivariate causal models for relation analysis. The detection of independent confounders in the variables will be easy in this case with the presence of exogenous and independent noises. This type of system can be handled by using matrix manipulation, multivariate regression analysis, and maximum likelihood estimation techniques. This also includes a measurement system to track the impact of information flowing in each node. This can be estimated by the Causal Influence values which will be the measures of effectiveness of a causal relation. Causal influences provide the critical information on how strongly the parent node affects the child nodes in causal models.

The next improvement will be to draw the Causal Levels while the kinship relation is drawn. Though the effectiveness of the kinship relation in the variable is untraceable in real world applications, the significant improvement in outcome is comparable with existing methods. This will be a breakthrough in this field if this can be achieved. Most of the proposed methods produce significantly low results while constructing the causal structure. This is because the requirement for the construction of a perfect causal model cannot be done using any of the original methods proposed until now. One solution for this is the independent sub-structural analysis and cumulative construction of the full causal model. Many of the prior methods have failed in this regard, and the model itself is not fully detectable in any of the cases. The causal level is an effective solution to this problem of causal structural construction. Causal levels can be detected using the causal influence values and by arranging the causal model using causal levels, the kinship relations in the features can be represented.
1.4 Structure of the Thesis

This work provides an in-depth view of the basics/fundamentals of causality with a steady growth towards the advancements, contribution, problem identification and new criteria for complete causal analysis in any system. The thesis is modelled to provide a gradual improvement of understanding on the topic i.e. causality, leading to a stage where one can easily access the necessity of the newly amended rules for better inference.

Starting from the fundamentals of causality, Chapter 1 provides a concise introduction to the subject of discussion. Suitable examples and model representation make it easier to understand the topic without any background knowledge. Also, it discusses the issues in the causal analysis which have not been addressed before and provides the solutions for those. It summarizes the rationale leading to the choice of the subject and the contributions it makes towards the subject.

In Chapter 2, the progress in the causal analysis is discussed by providing a complete and systematic analysis of selected research from the past few decades, having powerful methods to infer the area of study. In this case, all selected works are categorized into three groups, based on techniques used to construct the causal model. To provide a full comparative study under categories of probabilistic, statistical and algebraic approaches, under-lying difficulties, limitations, merits and disadvantages in applying these techniques are discussed. It will be possible to choose and use the appropriate method for a better implication.

The Linear non-Gaussian acyclic model (LiNGAM) (Shimizu et al., 2006) is efficient in finding the linear causality in non-Gaussian datasets. But real world data possesses nonlinear causality which is not scoped by most of the models. In Chapter 3, the ALiNGAM method is proposed to find the causal orderings while the data is nonlinear and noisy. It show that the basic model of LiNGAM can be used to analyze nonlinear data for effective error minimization and to reduce the noises in the model. While using the primary structure of the original LiNGAM, it introduce a new estimation process to provide the Altered-LiNGAM method to find nonlinear causality. The proposed Altered-LiNGAM or ALiNGAM imposes specific conditions on the system to maximize the probable causal directions and provides complete causal inference on observed datasets.

Explaining causal reasoning in the form of directed acyclic graphs (DAGs) yield nodal structures with multivariate relationships. In real world phenomena, these effects can be seen as a multiple feature dependency with unmeasured external influences or noises. The bivariate models for causal discovery fail to find the multiple feature dependency criteria in the causal models. In Chapter 4, the Multivariate Additive Noise Model (MANM) is proposed
to solve these issues while analyzing and presenting a multi-nodal causal structure. The new criteria of causal independence for qualitative analysis of causal model and Causal Influence Factor (CIF) for the successful discovery of causal directions in the multivariate system are introduced. The score of CIF provides the information for the goodness of causal inference. The identifiability of the model for linear, nonlinear causal relations is verified in simulated, real world datasets as well as the constructibility for complete causal model construction. In the comparison test, MANM is compared with Independent Component Analysis based Linear Non-Gaussian Acyclic Model (ICA-LiNGAM) (Shimizu et al., 2006), Greedy DAG Search (GDS) (Chickering, 2002), and Regression with Subsequent Independent Test (RESIT) (Peters et al., 2014).

The general propositions and fundamental assumptions given for causal inference to validate causality are somehow incomplete in providing all the necessary information for structural construction. Some major issues for qualitative analysis on causal structures are: how the flow of information i.e. connection strengths/path coefficients vary from the parent node to the descendant nodes; how to measure the magnitude of causal influence in a model; and the issue of the arrangement of nodes for known/estimated kinship relations. None of the questions raised above have been addressed yet, and all of them are very significant for successful evaluation of causal models. These unsolved factors can boost the implication process for causal inference, which is a gap in theoretical representation and practical evaluation. Chapter 5 introduces new rules and values into causality to resolve the flaws in fundamental theory as explained above. It introduces the qualitative analysis on causality using causal independence, causal influence, and causal levels, which significantly boost the analysis values and constructibility on the same.

Chapter 6 concludes this study and provides concise but important findings and contributions towards causal studies. Chapter 6 also discusses the possibilities of further improvement in this subject which will extend the utilization of causality into many different domains of analysis and research.

Appendix 1 contains the list of publications and communicated papers. From this research work Four papers have been produced from which two are already published and other two are under consideration.

The following Papers which are published:


The following Papers are submitted in journals and under review:


Chapter 2

An Overview of Recent Advancements in Causal Studies

2.1 Introduction

The recent approach in observational data inference is to study the flow of information in the variables, based on their occurrence in the time space. This phenomenon is classified under Causal Studies. Many of the methods have been proposed to describe causal inference by explaining the flow of information from one variable to another and the path of flow by directions. Considering the time series into count, this is an easier task to model a causal structure, as there is a record of the occurrence of features in time and the only task is to show the relation in features as in a DAG. To remove this record over time and then to know the kinship relations in the factors by arranging them in a DAG is a tough task.

The most popular analysis to time series data is Granger Causality Granger (1969), which has been widely used by economists. While discussing a more general view of this system of study, the basic structure of causality appears not well exposed and even underdeveloped. The researchers are trying to provide a good outlay of this topic. Starting with the most commented and debated view provided by Judea Pearl (Pearl, 2000) is "causal and counterfactuals". An observation regarding the view given by Pearl is that "...the direct effect of observation is causal, while uncertainties in possible outcomes of present causal relations lead to counterfactuals". The causal model given by Pearl (2000) is the Structural Causal Model (SCM), which is all about a probabilistic approach to causal analysis, although Pearl (2009) referred to it as the statistical approach to causal inference. In this review, an
A simple but groundbreaking method of Structural Equation Model (SEM) was proposed by Bollen (1989). The linear structure of SEM is an effective model which provides wide usability, but the method sticks to its fundamental properties of path finding. The methods of Bayesian Network (BN) given in Pearl (2000) and Spirtes et al. (1993) are widely used, developed and implemented in many fields of studies. Probability analysis is one key feature of BN. Considering the relation of conditional dependence given by Bayes law, BN works well to optimize randomly proposed distribution to the original distribution using Markov Chain Monte Carlo (MCMC) with the random samplers such as Metropolis, Metropolis-Hastings, and Greedy Hill Climbing. The book by Gilks et al. (1996) on MCMC with different samplers is an introductory book about how to implement the ideas in the real world with practical examples. In addition, it comes with working algorithms which makes it easier for beginners. Some recent algorithms which show variation to MCMC are the Hidden Markov Model (HMM) (Ghahramani, 2001) as well as Hybrid Monte Carlo (HMC) methods (Duane et al., 1987). A practical approach on causal, correlation and artificial intelligence is discussed by Marwala (2015).

Quite different in implementation from probabilistic methods, statistical methods rely on the independence of observed features by computing relations in terms of covariates and cross covariates. As most statistical relations use correlations in variables to check positive or negative effects of variation on one another, this system of analysis cannot be used to validate causal relations. The empirical estimations using kernel structure provide more accurate estimates to independent feature selection. The papers on this use are: canonical methods by Bach and Jordan (2002) as well as Gretton et al. (2005c), kernel methods by Gretton et al. (2003); Sun et al. (2007) and Zhang et al. (2012) and Hilbert-Schmidt independence criterion based methods by Gretton et al. (2005a); Zhang et al. (2009) and Chwialkowski and Gretton (2014), which scopes for the detection of independent features required for causal modelling.

There are quite a few methods which have been developed more recently which are different from the probabilistic method of BN. Many of these methods have a central model which uses Independent Component Analysis (ICA) to analyze causal models. The most recognizable models are the additive noise model by Hoyer et al. (2009) as well as Mooij et al. (2011) and non-Gaussian models by Shimizu et al. (2006), Shimizu et al. (2011) as well as Hyvärinen and Smith (2013), featuring methods for machine implementations but lacking in constructing the full causal model.
This Chapter is modelled based on the best available and implementable methods for the purpose of causal inference. In this, papers are selectively categorized into three major groups of Probabilistic, Statistical and Algebraic approaches. Under these three groups, the conclusions were drawn about methods for their limitations, as well as their advantages and drawbacks. The following sections, provide a detailed discussion of these groups and the conclusion. First, the discussion begins with some basic ideas to understand causality.

2.2 Understanding Causality: Principles and Problems in Modelling

The known view of causality is provided by Pearl (2000) in his book describing causality with counterfactuals. This is debatable among statisticians and economists, however philosophy and psychology have been utilizing it. However the present focus is not on counterfactuals. In physics, the meaning of causal is very clear when one considers the matter of force and its effect. The mainstream of science had never shown any interest in causality and its usefulness until Granger Causality made its way into economics. Granger Causality is mostly used for analyzing time series data and future predictions. If one needs a more generalized version of Granger Causality then consider Causality.

Consider a set of \( n \) variables \( \{ V_1^m, V_2^m, \ldots, V_n^m \} \), where each variable occurs \( m \) times. Trying to find relations in these variables using operations like regression analysis, finding the correlation or performing a cluster analysis can be easily done by any statistical approach. But some major problems remain unsolved as in the case of the following:

- How are all these variables related to each other, if these need to be shown in a directed acyclic graph representing the causal relation?

- Without having any prior information about what these variables represent for or from where the dataset is drawn, how correctly can these be arranged in a DAG?

- Can the kinship relations be shown if there is any in the variables, when the time factor is unobserved?

- If the occurrence of variables is measured in time, then what is the probability of the earlier affecting later?

- What is the level of accuracy for possible graphical results using real world data, knowing that there is no way to verify the accuracy of the outcomes?
A possible explanation to all of the above questions is contained in a causal model. So, what exactly is a causal model? A causal model for an irreversible system is “… a directed acyclic graph with nodes representing the variables, connected by weighted directed edges to show the relation from one to another and weights for the connection strength or the flow of information” (Pearl, 2000). It is not necessary for a causal model to be acyclic, though it can be cyclic for a reversible system. If it is clear what exactly causality means, what information is required to construct it? Take a look at the following points:

- A multivariate dataset without any prior information for the worst-case scenarios.
- A dataset of categorical type, but not a classification one.
- Having enough data points/instances for each variable in comparison to the number of variables, to perform any kind of machine-implementable analysis.

A question that arises is why such a particular dataset is chosen. By choosing a multivariate set, it can be made sure that all the distributions differ from one another. In a classification dataset, only classifications can be performed, but in a categorical set there are many options. A large dataset in the sense of more data points than the number of variables can increase the reserve criteria for training, testing and validation sets. One may consider the case of a neural network for a separate number of sets, as it is helpful to have different sets to test and validate in any experiment. Even considering all these assumptions, it does not significantly contribute to the possibility of getting a good approximation for causal models. If a model works for synthetic data, then the possibility of getting the same kind of accuracy on real world data is less because there are always unobserved variables, the incompleteness of data or other unknown effects left unmeasured, which affect the evaluation of causal model. So, in some way the addition of the errors into the system contribute to those unmeasured effects.

Next what kind of system assumptions are needed? A system can be either noiseless or with an additive noise model. Selecting any one of the two options does not help much. They both have merits and demerits. Selecting an isolated error system can make the system easier to handle, and only need the estimations for path coefficients. And even it makes more sense, the errors do not help to construct the causal model or to find the path coefficients. In the case of an additive noise model, one or more variables can be considered as independent noises if these can be found out. Knowing where from and what exactly the data is, newly added errors in the model remain unsolved even if those have effects on the model. To make it more confusing there are observed and unobserved confounders which can be added into the model. Figures 2.1a show all the assumptions and system types required to construct the causal models.
2.2 Understanding Causality: Principles and Problems in Modelling

(a) A noiseless causal model

(b) An additive noise causal model

Fig. 2.1 Causal Models

Fig. 2.1a shows a causal model without any external influences and Fig. 2.1b shows a model with independent noises. The nodes represented by variable set \([v_1, v_2, v_3, v_4, v_5]\), path coefficients/connection strengths are shown by \([w_{32}, w_{34}, w_{21}, w_{25}, w_{45}]\) for model 2.1a and \([w_{25}, w_{24}, w_{53}, w_{51}, w_{41}, w_{13}]\) for model 2.1b with errors \([e_1, e_2, e_3]\).

To obtain such causal models, some good approximations and assumptions for the primary system are needed, which are ultimately constructed to an expected causal model. In this thesis, different methods published over the past few decades, trying to achieve the desired causal structure were studied. In this Chapter, the methods available to date are considered for this selective study and discussed. In subsequent sections, all the methods are categorized into three different groups for the choice of methods/techniques these have used to solve the causal problem.

In this research, it has been found that researchers perform only three kinds of operations with some variations to the original in each of the cases to determine the required causal structure. Three of the original ideas contain a widely-used version of probability analysis through BN, a more recently used but old technique of covariance structure and the more recently introduced methods using Independent Component Analysis (ICA). These papers are grouped as the Probabilistic approach for probability uses, the Statistical approach for covariance structure analysis and Algebraic approach for using matrix analysis in the form of ICA. In the following sections, a clear and concise analysis of the working principles, advantages, limitations, and demerits for the above-mentioned groups are provided.
2.3 Probabilistic Approach of Bayesian Network

The fact about methods under this group is that they have been used for a long time for the evaluation of parameters and path analysis while estimating DAGs. The Bayesian method has undergone a significant number of modifications and developments over time since it was proposed by Pearl (1998). There are many papers available related to BN and its implementations. This will not be discussed in this study. The discussion below is only relevant to causal inference based on BN methods.

As BN is classified under probabilistic methods, the basic working principle is probability estimation. Given a prior distribution for the dataset of interest and a current state of observation, BN estimates the likelihood or the posterior conditioning on the prior probabilities. This is done by the Bayes law of conditional distribution theorem, the first and principle theorem for all BN methods.

**Theorem 1 Bayes Theorem:** Let X be the set of data of our assumption with a prior probability distribution $P(X)$ and observation for data set $O$ can be expressed through the conditional probability or the likelihood of $P(O|X)$. Now the effect of observation on prior is a posterior distribution $P(X|O)$ can be given by Bayes theorem as

$$P(X|O) = \frac{P(O|X)P(X)}{P(O)}.$$

The BN works on Bayes theorem by using Markov Model, which set rules for BN.

**Markov Model:** A directed graph without having any bi-directed edges is a Markov model/network and such models having the same comparable structures called Markov equivalent classes. So, what exactly is required for Markov model and how does it work? Most of the time the conditional dependence or independence (in general always conditional independence) are needed to be estimated for successive observations and the fellow descendants, while provided with a prior probability. This needs the Markov model to be conditional independence with the variables, where it may be observational or a prior. Consider the case of $P(X)$, which is a prior distribution and consider the set of successive observations $O_1, O_2, ..., O_n$. The task is to find $P(O_i|X)$ for $i = 1, 2, ..., n$. The joint distribution can be decomposed as conditional distribution, and this decomposition of conditional distribution for causal inference or graphical inference is studied under graph d-separation for Markov equivalent classes.
**D-separation:** The original definition for d-separable graphs is given in the work of Pearl (1998). What is the use of d-separation? Construct two sets so that one contains all the paths/edges, and other one has all the nodes. Now by conditioning on a node which is the link between two other nodes, the path can be separated from any three sets of connected nodes with two paths. The possible combinations of three nodes with two edges which form the V-structures, can be considered as these three cases of (1) \( a \rightarrow c \rightarrow b \), (2) \( a \leftarrow c \rightarrow b \), and (3) \( a \rightarrow c \leftarrow b \), where \((a, b, c)\) is a set of variables such that only node \( c \) can make the graph d-separable. In cases (1) and (2), conditioning on node \( c \) can make the graph d-separable, but in case of (3) it is also needed considering all the predecessors of \( c \), so that the variables can be marginally independent to each other. The V-structure of type (3) is called a collider.

After all the above ideas let us move towards the method which can work on these. The next section describes the Markov Chain Monte Carlo (MCMC) technique.

### 2.3.1 MCMC for Bayesian Structure Analysis

MCMC is one of the most used BN simulation methods than some more recently introduced versions of HMM and HMC. In causal inference MCMC has a greater impact than any other methods, considering that the others are relatively new. In simple words, MCMC is an optimization technique which works with some other sampling methods such as Greedy Hill climbing, Metropolis-Hasting algorithm to optimize the conditional probabilities and to achieve the target distribution by forming successive chains. There are many papers available on MCMC and on its different uses. This discussion is more about a summary of all these implementations. A more recent paper on this topic is by Friedman and Koller (2003), which discusses use of MCMC with other optimizers and effective maximization of samplers to get the targeted distribution quickly. Below follows a general discussion on the limitations, merits, and demerits of MCMC.

**Advantages of using BN on MCMC:**

- It only requires probability estimations, but function estimation is a much harder task, which is not necessary.

- It is simple and easy to use for large-scale data analysis.

**Disadvantages:**
• Needs good proposal distribution and parameter regularization for better approximation to reduce error and faster simulations to minimize time complexity, or go for a trial and error process.

• Other methods of Maximum Likelihood Estimation (MLE) and/or Expectation Maximization (EM) need to be considered to estimate the parameters of the model.

• End results do not provide a directed acyclic graph for causal modelling.

Facts mentioned above are helpful for using the MCMC technique with BN for causal analysis only. The general framework of MCMC for the purpose of causal inference is not discussed here.

### 2.3.2 Variational Bayesian learning method for DAGs

In BN analysis, a significant amount of effort lies on the computation of marginal likelihoods to estimate the posterior distributions for DAGs. The methods like Expectation Maximization (EM) with Maximum Likelihood (ML), Maximum a Posterior (MAP), Bayesian Information Criterion (BIC), Cheeseman-Stutz (CS) criterion and Annealed Impotence Sampling (AIS) are used to estimate parameters of likelihood estimation function. The paper by Beal and Ghahramani (2004) presented the method called Variational Bayesian (VB) to estimate the marginal likelihoods by optimizing the lower bounds of approximation. A VB for expectation maximization is called the VBEM algorithm, which is applied on conjugate-exponential families of distributions over the hidden variables to tighten the lower bounds of likelihood approximation. It provides a good comparison over EM, BIC, CS, and AIS that of VB. The paper provides all the advantages and disadvantages of VB algorithm, however, an use of this method for causal inference is provided in the below.

**Usefulness for causal studies:**

• The method estimates the posterior distribution by maximizing the marginal likelihoods similar to the EM algorithm. The approximation to posterior distribution is comparable to MCMC.

• In fact, MCMC does not do the parameter estimation, whereas the VB can perform parameter estimations for the model.
2.3 Probabilistic Approach of Bayesian Network

- Estimation to a hidden variable is a useful feature and using this the errors in the model can be found. But factorizing the hidden variables and the parameters of the model is a complicated process where the posterior distribution is not simple.

- The distributions are restrictive for the selection of conjugate-exponential families.

- A complete causal structure is unpredictable by selecting this method.

2.3.3 Causal Inference using Plausible Markov Kernel

In the Markov model, the features responsible for the direct cause of some action/effect is called a Markov kernel. As in the case of the Markov blanket, the parents are the Markov kernels. If the Markov kernels are known for the variables in the model, by comparing their conditional distributions the causal relations can be found. Based on this idea Sun et al. (2006) proposed the method of Plausible Markov kernel for causal inference. A threshold is always required to compare these conditional probability distributions. So, the choice of threshold must be to maximize the conditional levels to get the best Markov kernels. The idea is to maximize the entropy to find the plausible Markov kernels. The entropies of choice are moments such as expectations, variance and covariance. The plausible Markov kernels are traced as follows: given the joint distribution for child and parents, the most plausible Markov kernel of a child is the conditional probability that maximizes the entropy of the child using the constraints on the expectation and variance of the child and the cross-covariance in the child and all its parents. The problem in estimation of probability functions need to be optimized and can be solved easily. How this can be used to find causal reasoning?

Usability of plausible Markov kernel:

- Low in complexity and easy to use.

- Need computation only for joint and conditional probability distributions.

- Requires prior knowledge of the parent structure of variables to get the plausible Markov kernels.

- A complete causal model with path coefficients remains untraceable.
2.4 Probabilistic method of Graphical Analysis

Graphical analysis is another way to infer Causality. The main idea behind all graphical models for the causal study is to get DAGs by using BN theories or probability theory. However, in addition to this some other methods are also used to optimize the likelihood function, or for a good approximation to proposal distribution using nonlinear Gaussian functions; or only by maximizing the conditional distribution functions on edge formation for the causal structure. But the only common factor in all these methods is the use of probability analysis, a general frame for evaluating edges or directions. For this purpose, this section introduces a number of major papers for causal inference based on graphical analysis, which have shown some significant improvements.

2.4.1 Ideal Parents

A paper by Eliden et al. (2007) proposed a method to find the ideal parent structures to form DAGs. It provides a BN-based analysis utilizing conditional distribution on Gaussian data to produce the DAG and EM algorithm to estimate the parameters of the model. The BN optimization for conditional distribution uses the general method of Greedy Hill climbing, and $C_2$ similarity scaling function handles the pruning of edges. They even used sigmoid functions for non-linearity and log-likelihood distributions to evaluate Conditional Probability Distribution (CPD) functions. The CPDs are useful to get the ideal parents and to introduce new variables through the ideal parent’s method. The technique for finding ideal parent also utilizes some other methods and long graphical evaluation processes. But they manage to reduce the overall complexity of the algorithm by giving the regularization criteria and acceptance conditions. One will find it very interesting to check how the lower complexity of the algorithm can be achieved while utilizing multiple methods in combination to produce the DAG with an unknown level of accuracy. The question follows whether it is really helpful for causal analysis or only to form the DAGs.

Advantages:

- Works for linear, nonlinear and noisy data types, even can estimate missing values in the dataset. So, dealing with data types is not a problem using this method.
- Can identify parent structure by adding hidden variables and ideal parent search algorithm.
2.4 Probabilistic method of Graphical Analysis

**Limitation:** The method is suitable for a single layer structural causal study only by creating an ideal parent and child structure.

**Disadvantages:**

- Complex and long processes in comparison to other methods.
- Requires prior input of a graphical model for estimation and convergence of the model.
- Unable to find the path coefficients which can be used as a criteria to check the casual construction.
- Needs regularization and acceptance features like candidate parent selection.
- Unable to construct the multilayer structure for causal inference.

2.4.2 Intervention Methods on Markov Equivalent Classes

The paper of active learning for causal networks by He and Geng (2008), which uses intervention technique on Markov equivalence classes to produce subgraphs with directed edges. The idea is to generate v-structures of subgraphs using Markov property, which translates to find directions in the model using conditional probabilities. The question is how helpful it is. Analysis of a larger set is a difficult and complex process. If v-structures can be generated with Markov equivalence property, then by combining small structure a full causal model can be formed. Using randomized experiments for those classes which are independent of parents and quasi-experiments for those dependent on parents, they constructed structures by calculating the pre- and post-intervention distributions. The model is optimized using batch and sequential interventions to create chains of components containing sub-structures/sub-classes of Markov equivalent classes. The method sounds good for theories but is it helpful.

**Advantages:**

- The graphical implement is simple and easy as it only requires estimation of subclasses of v-structures at a time.
- There is no need for function and parameter estimations.

**Limitation:** Only works with Markov equivalent classes.
Disadvantages:

- Complexity increases factorial (!) times with the increase in the number of nodes in the observational set, resulting in very high complexity.
- Can only form chain structures with optimization of sub-classes for chain components.
- The model assumes existence of no latent variables.
- Multi-layered causal model is not achievable.

2.4.3 Method for Causal Hierarchy Identification

A method for complete identification of causal hierarchy was proposed by Shpitser and Pearl (2008) using probability for graphical interventions. The defined causal hierarchy contains an associative relation, causal-effect relation and the counterfactuals. They derived these observations by producing joint distributions for associative relation, external interventions using functional distributions for causal-effect relations, and experimental results by performing a possible span on distributions derived from observations and interventions for counterfactuals. First, the associative relation only depends on joint distributions of variables, follows a direct probability calculation. The second part of causal-effect relation is derived by checking the front-door criterion and back-door criterion for nodes containing middle nodes or descendants respectively in the graphs. They provided an identification algorithm called ID to find a complete causal-effect relation in the graph. ID algorithm checks the front-door and back-door criterion for identification of c-components, and then it does the next step to calculate the distributions over the sub-graphs to find the complete causal model. IDC is the next to ID, which checks for the conditional effects in the model. They claimed that ID is complete. The last part of the causal hierarchy is the counterfactuals which are identified by the proposed make-cg algorithm. The make-cg algorithm works by measuring the effects of observation and interventions on produced causal-effect model. It computes over all possible distributions for different effects of both observation and intervention and other given criteria. The confounders play their role for probability calculations on causal-effect relation and possible selection of counterfactuals. All the explanations for why to or not to use this method is discussed below,

Advantages:
2.4 Probabilistic method of Graphical Analysis

• This method contains both the features of observational analysis and interventional analysis over the data to derive the causal diagram.

• It can compute the counterfactuals for experimental conclusions if interested in parallel effects or uncertainties of possibilities.

• The algorithms are based on probability calculations like joint distributions, conditional distributions, and conditional marginalization. Therefore, computations are easy to perform.

Disadvantages:

• This method needs regularization and conditionalization for harsh-less handling for the algorithm.

• The complexity of the algorithm is unknown and is higher than most efficient methods.

• Probability is not enough to conclude over the model that the graph does not show the path coefficients for connections.

• It requires an initial input of graph for the later formation of structures.

2.4.4 Markov Blanket for Causal Structure Learning

A Markov blanket is a causal graph where the strongly relevant connections inside that blanket can provide the information for the relevant connection, and no other variable can provide any of the required information outside that blanket. So, in a connected graph, the strong relevant variables for the child node are the parents and the child’s of the child. The strongly relevant feature only depends on the conditional features of relevant features only. So, in a Markov blanket, all the connected feature are strongly relevant to each other. Based on this fundamental property of strong relevance of features, Pellet and Elisseeff (2008) proposed the method of causal structural learning from Markov blankets. This approach deals with all the Markov properties of a causal graph-like conditional independence, faithfulness, relevance, d-separable, v-structure and Markov blanket. They combined all these to perform a grow-shrink algorithm to search the moral graphs from the Markov blanket. The moral graph is an undirected causal graph where only the relevant nodes are connected by the edges. By performing a feature selection, the moral graphs can be found from the Markov blanket.
On the moral graph, the d-separable feature can be used to remove triangles and to find the collider sets.

They used feature selection criteria which select only the features having strong relevance among themselves. For feature selection, they used the Recursive Feature Elimination (RFE) method in their algorithm over Support Vector Regression (SVR). By using Total Conditioning (TC) criteria they are able to condition features having partial correlations and improve the significant level of the algorithmic outcomes. The significance level of TC algorithm was tested using student t-test. In cases where there are fewer of samples, the TC unexpectedly fails. So, they proposed an alternative algorithm called Total Conditioning Backward Feature-Selection (TCbw) to overcome the missing link estimation problems for fewer sample size. Both models are tested for different real-world datasets and against the PC algorithm. The performance of the models is shown for the Alarm network, Insurance, Hailfinder, Carpo and a subset of Diabetes datasets and the comparison with methods such as PC, GS and Bach-Jordan. The estimation results for causal structure formation are compared with PDAG and CPDAG methods. The usefulness of the proposed technique will be examined next.

**Advantages:**

- A possible estimation of the causal structure with all possible directed edges.
- Works for all sorts of dataset.
- No requirement for function estimations in the model.

**Disadvantages:**

- Gaussianity of the data is assumed for the process and fails for inconsistent and nonlinear distributions of the data.
- Causal structure estimation is fairly complex as compared to others.
- Cannot be used to find the path coefficients in the structure.
- Too many conditionalizations required in order to perform the operations.
- Require input of target variables for feature selection process.
- Needs larger datasets to produce better results.
2.4 Probabilistic method of Graphical Analysis

2.4.5 Information Geometry on Causal Inference

If all the possible information can be extracted from data, then that will help to produce a causal model by using the geometry of alignment. Janzing et al. (2012) provided a theoretical modelling for causal inference by using information geometry. Information geometry is the combination of different techniques which check the relation in variables by measuring the differences in their individual probability and conditional probability. An interval containing the individual probability and conditional probability which provide different information on the relation in the two candidate variables can be used to find the alignment in them. The geometric alignment provides the distribution of these variables. So, using a definite function for the distributions, the covariance structure of these causal effect pairs can be found which reflect the relationship.

Advantages:

- The method uses uniform or Gaussian or any other distribution to see the geometry of alignment that is the orthogonality of data when spanned over the probability regime.
- It is easy enough to find the covariance relation in the paired set for the choice of distribution function over the data.
- It can be used for both noisy and deterministic (noiseless or isolated) systems.
- It works faster than most of the compared methods.

Limitation: Only two variable sets at a time can be processed as the method works in pair set for inference analysis.

Disadvantages:

- It cannot be used for finding causal relations in a multivariate dataset. A full causal structure will not be obtained.
- The flow of information through the directed path is untraceable.
- A set of Gaussian or uniform distribution type is not always possible for working on.
- It can cause a wrong direction prediction depending on the distribution.
2.4.6 Geometry of faithfulness for Causal Inference

A probability distribution which imposed Conditional Independence (CI) only for the Markovian equivalent classes i.e. classes which hold Markov properties is called faithful. That implies all the conditional independence directions generated are only formed under Markovian classes and have Markov properties. Based on the faithfulness of a distribution, Uhler et al. (2013) proposed the methods and conditions to infer causal structure. It provides inference based on weak, strong and $\lambda$-strong faithfulness conditions for CI. These faithfulness conditions can be derived using correlation and conditional correlations in the interested features for Markovian classes. For a Gaussian distribution and assumptions of uniform or random distribution for path coefficients and errors in intervals [-1,1] and [0,1] respectively, the covariance and conditional covariance structures are beneficial for finding the $\lambda$-strong and restricted-strong faithfulness conditions. This conditional faithfulness can be used to extract the directed paths for the causal model in the form of DAGs. The paper focused on the unfaithfulness of distributions under different assumptions and combinatorial polynomials. The results are shown for the use of PC-algorithm under faithfulness assumptions to derive the causal structures. The highlights of the studies and its usefulness for the causal inference are discussed below:

Advantages:

- It is a very distinct and different approach to causal inference although the core of the working principle is the same as BN, but the faithfulness criteria on the interval are interesting.

- One can model for analysis of different causal structures based on the faithfulness criteria which are easy to use as shown for the Gaussian distribution cases, although the paper does not provide an algorithm to work on.

- It works well to extract structures like trees, cycles, and bipartite graphs.

Disadvantages:

- The faithfulness feature is very restrictive to the small and large size of nodal graphs. This means the unfaithfulness of the structure increases as the size of the nodes varies from medium to small or medium to large.
• The results can be strongly unfaithful for complex structures with more edges and nodes.

• PC-algorithm is not enough to analyze larger Markovian classes of DAGs.

• The system is extremely complex.

• The assumption of Gaussian type distribution does not always hold.

• Interval assumptions for path coefficient and errors are very restrictive.

• Parameter estimation for this type of consideration is very inconsistent.

All the above-discussed papers are studied on a systematic annual review for the sole purpose of methods based on causal inference. As all these papers are classified under probabilistic methods, a detailed clarification have been provided for the causal study. Some of the other papers which are studied based on unidirectional structure formation are as follows:

• "Bayesian methods for learning graphical models with incompletely categorical data" by Geng et al. (2003). It is constructed to estimate the posterior means of the incomplete categorical data for graphical models. It uses EM algorithm for posterior mean estimation for graphs with hyper Markov properties.

• "Causal reasoning with ancestral graphs" by Zhang (2008). This paper provides all the explanations require representing the causal structure as an ancestral graph. An ancestral graph is arranged in order from Grandparent, parent, child to grandchild. Zhang provided the ancestral structures like Maximal Ancestral Graph (MAG) and Partial Ancestral Graph (PAG). This paper provides a theoretical approach to DAG, MAG and PAG.

• "Learning structurally consistent undirected Probabilistic graphical models" by Roy et al. (2009). It provides the method called the Markov Blanket Search (MBS) for Probabilistic Graphical Models (PGMs), based on the Markov Random Field (MRF) structure learning feature. The MRF is a more refined version of the Markov Blanket Canonical Parametrization. The performance of the MBS algorithm was compared to other undirected graph formation methods such as the Graphical Gaussian Model based on lasso regression (GGLAS), Full DAG search, LARs Based order search (ORDLAS), DAG search using Sparse Candidate for pruning (SPCAND) and L1 regularized Markov Blanket Estimation (L1MB).
2.5 Statistical Approach for Causal Inference

- "A conditional independence algorithm for learning undirected graphical models" by Borgelt (2010). It introduces a new and easier method to find the undirected graphical model under the probabilistic graphical model. The method is an alteration to Cheng-Bell-Liu's algorithm. It provides new techniques for the processes of learning undirected graphs using a conditional independence algorithm by performing operations such as drafting, thinning, moralizing and additional thinning processes.

The following section is designed to provide a clear view of statistical approaches on causal inference. In this section, all the covariance structures and Kernel base covariance structure methods necessary are included to provide a clear explanation.

2.5 Statistical Approach for Causal Inference

The causal inference defined using statistical analysis is quite debatable when compared to probabilistic approaches. Many argue there are no such reasons for distinguishing between these two models. However, the main focus of the probabilistic approach requires the estimation of probability distributions such as joint, marginal, posterior and conditional probabilities to show the dependence or independence of features in a dataset of interest. On the other hand, the statistical approaches depend on correlation and covariance structures of the data to draw the empirical conclusions and to show the independent relations. The most debatable factor in the statistical approach is the use of correlation structure i.e. "causal relation may have a correlation, but correlation does not imply causal relation". So, depending on correlation the right causal conclusions cannot be drawn. The covariance itself is not a very strong criterion to provide inference on cause-effect relations. The newly developed methods of kernel structures and norm spaces on covariates provide the boost to mutual information and independence criteria. The sole criteria of these techniques are to provide the independence structures in the variables and using these dependence measures the causal structures can be modelled.

A very sustainable progress has been shown in signal processing for separating features from mixture models and to trace original sources. This phenomenon produced some of the well-defined and wide implacable methods for the separation of independent features in the observed sets. The following papers are discussed to provide advancements on statistical approaches to find independence in causal models. The papers discussed in this section provide the ideas for generating the set of independent features for the choice of data types and the successive refinement in their development over time.
2.5 Statistical Approach for Causal Inference

2.5.1 Kernel Independent Component Analysis

For the non-Gaussian data structure, the Independent Component Analysis (ICA) provides an excellent tool for parameter estimations of the function under consideration. The Reproducing Kernel Hilbert Space (RKHS) can be used to map the bounded linear functional in the Hilbert space domain. Based on RKHS, the minimization of contrast function is achieved using ICA to derive the canonical correlations. The paper by Bach and Jordan (2002) provides a method called Kernel Independent Component Analysis (KERNELICA), which optimizes the canonical correlation using ICA in RKHS. The basic approach of canonical correlation analysis (CCA) is very similar to the Principal Component Analysis (PCA). Using a linear transformation, the PCA is converted to a vector of uncorrelated components, while CCA tries to convert the vector with correlated components. They defined contrast functions for ICA using a feature mapping technique called $\mathcal{F}$-correlation obtained from computing canonical correlation in RKHS. Using CCA in feature space, the vector transformation is reduced to eigenvalue problem. Solution for the eigenvalue problem is obtained using Cholesky decomposition. The KERNELICA, which is the proposed method and is named for using kernel techniques which has two parts. The first one which uses the kernel CCA is KERNELICA-KCCA and a generalized version of the first one which generalizes the kernel variance called KERNELICA-KGV. The methods are fully tested and compared with existing methods. This provides a detailed kernelized criterion based on ICA and CCA.

2.5.2 Kernel Mutual Information for Independent Variables

Kernel mutual information (KMI) is a method to find independent features in the model, does the same with little variation from the basic proposal of KERNELICA. KMI is also a contrast function for ICA which provides stronger upper bounds for mutual information on kernel density estimations. Gretton et al. (2003) introduced two quantities for the identification of independent variables on KMI: first, the kernel covariance (KC) and Kernel mutual information (KMI), where both the values become zero if and only if the variables are independent. Where KC can only be tested for the independence of random variables, the KMI also provides the upper bound for mutual information on parzen window estimation. The method scopes for empirical estimations on independent feature extraction for both KC and KMI. For experimental use, it is implacable through Amari divergence (Cichocki and Amari, 2002). The results of KERNELICA and KMI are very comparative depending on the domain’s nature.
2.5.3 Kernel Constrained Covariance for Dependence Measurement

Both the methods discussed above draw empirical conclusions on the independence of variables on RKHS and upper bounds for mutual information. If this idea can be extended to a universal RKHS for the selection of independent features, then both population and empirical conclusion can be easily made. Based on this idea, Gretton et al. (2005c) proposed the method of constrained covariance (COCO), which being zero only for the independence of universal kernels. So, the universal RKHS must be dense in continuous function space, which is satisfactory for Gaussian and Laplace functions. The basic definition of COCO is derived as the supremum over the covariance function of selected features. These assumptions lead to a positive correlation of population COCO with empirical COCO. The results satisfactorily conclude that COCO better maximizes kernel-based criteria of dependence for variable selection than KGV. The proposal is tested on data of FMRI for humans, after and before the breathing results.

2.5.4 Measuring Statistical Dependence with Hilbert-Schmidt Norms

All the above methods such as KCC, KGV, KMI and COCO, try to find the independence in variables, have regularization criteria over RKHS and provide slower learning rates. In the paper by Gretton et al. (2005a) for measuring statistical dependence between features using Hilbert-Schmidt norm (HSN) provides a very comparative result with non-regularization and faster learning rates. Hilbert-Schmidt norm is defined as the trace of the product of Gram matrix, which is also the method for empirical estimation. The rate of convergence of empirical estimation towards population estimation is $\frac{1}{\sqrt{n}}$, where $n$ is the number of variables in the set. Criteria for independence is given by Hilbert-Schmidt Independence Criterion (HSIC) as the squared Hilbert-Schmidt norm over the associated cross-covariance operator, where the cross-covariance operator is the inner product of the separable Hilbert-Schmidt norm. The performance is measured by comparative analysis of different methods such as FastICA, Infomax, RADICAL, CFICA, COCO, KGV, KCCA, KMI and HSIC using the Amari divergence method.

2.5.5 Kernel-based Causal Learning Algorithm

Using a certain variation of the original HSIC and HSN criteria, Sun et al. (2007) proposed a method of causal learning through kernel-based HSN and empirical estimations over population. The causal learning can be achieved through three steps: (1) selecting the
connections using HSN for which the connected variables are conditionally independent of every neighbourhood containing the two variables; (2) any set of three variables is directed on the basis of votes which count the possible number of directions found in them; and (3) employing the Inductive Causation (IC) to construct causal structures only for those edges which have the highest statistical dependence to form the directions. The method provides features for the detection of causal structure based on statistical dependence in the features. It is compared with other similar methods such as constraint-based PC algorithm, BN-PC, exhaustive search (ES), Greedy Search/Hill-climbing (GS), Markov Chain Monte Carlo (MCMC), maximum weight spanning tree algorithm (MWST) and mixture models like MWST+K2 and MWST+GS for benchmark analysis.

2.5.6 Kernel Measures of Independence for non-i.i.d Data

All the above methods are formalized for independent and identically distributed (i.i.d) samples to find the independent features, none of which deals with the non-i.i.d data types. Zhang et al. (2009) proposed the method for causal structure formation using the undirected graphs for non-i.i.d datasets. The purpose of this criterion is to form structures using HSIC on RKHS and arrange the independent features in sequential clusters. Structures are formed using conditional independence for maximal cliques observed from the sets of random variables and then by performing an empirical estimation for the choice of joint kernels on the maximal cliques. The structured HSIC method provides scope for 2D and 3D image formation using the clusters of dependent variables in undirected graphical structures. Most importantly, structured HSIC is applied to non-i.i.d data types and that gives access to ICA analysis, as structured HSIC can be used as a contrast function for ICA. An extension to this can be the formulation of PCA using structured HSIC to analyze clustered time series data. Structured HSIC is tested with spectral clustering and HMM for exponential and Gaussian RBF datasets.

2.5.7 Kernel-based Conditional Independence Test and Application in Causal Discovery

To obtain causal information using Conditional Independence (CI) criteria is a very difficult process for larger datasets. But this can be solved by using kernel feature for conditional independence testing. The paper by Zhang et al. (2012) proposed the kernel-based conditional independence test (KCI) to provide a faster access to CI for larger datasets and smaller sample
sizes. KCI is computed over CI variables to estimate a kernel matrix through eigenvalue
decomposition and associated asymptotic distributions. The characterization of CI is done
using HSIC on RKHS, where the characterization of CI is given as the cross product over
cross-covariance operators defined in RKHS. It also provides estimation towards conditional
independence and unconditional independence tests for KCI. The null hypothesis for asymp-
totic distribution is approximated using two parameter Gamma distribution. Another problem
solved using KCI is the dimension of dependent variable, by observing type I and type II
errors for asymptotic distributions of the type i.i.d Gaussian, and two parameter Gamma
distributions. The novelty of the proposal is tested for both synthetic and real world datasets
and estimated causal models are provided in the paper.

2.5.8 Kernel Independence Test for Random Processes

A major group of papers proposed on statistical dependence using HSIC while worked on i.i.d
sets fail for randomly drawn datasets from random processes. This paper by Chwialkowski
and Gretton (2014) proposed statistical test statistics for HSIC between any two random
variables. Test statistics are the null hypothesis where \( H_0 \) is the independence of two random
variables, tested using auxiliary kernel function and characterized by normalized \( V \)-statistic
distribution. The null distribution is provided using \( \gamma \) estimation having wrong p-value and it
is removed by shifting \( H_0 \) under the empirical distribution estimators called 'Shift HSIC'. The
method also provides a better estimation for type I error, while methods proposed previously
have large type I errors in the distribution estimations. It also provides a better compatibility
for noisy time series datasets with low granulations. The results produced by Shift HSIC
provide undirected causal models that are causal cliques. The proposal is tested for artificial
and real world time series datasets and has a high significance level when compared to that
of other bootstrap methods of the same class.

The papers studied in this section provide the kernel-based methods for the selection of
independent variables from the sample set of i.i.d., non-i.i.d and time series data analysis.
Some papers relating to above methods for a better understanding to some of the concepts
discussed are given in the following:

- Gretton et al. (2005b) broadly discussed and proposed two functional constrained
covariance (COCO) and kernel mutual information (KMI), to measure independence
random variables. This work is an extension of works discussed in 2.5.2 and 2.5.3. It
provides a more in-depth analysis and comparison to other preexisting methods such
as KCC, KGV, ICA and RKHS.
• The convergence of finite samples associated with the kernel function of canonical correlation analysis (CCA) is established towards their population, but it is not provided for infinite sample space. Fukumizu et al. (2007) provided the proof for infinite sample convergence of the kernel CCA. They also have shown the convergence of regularization coefficient involved in kernel CCA.

• The kernel methods for machine learning provides great tools for statistical learning and a wide range of usability. Hofmann et al. (2008) provides a wide range of machine-implementable methods based on kernel approaches for the statistical process.

• Petrović and Dimitrijević (2011) discusses the invariance of causal relations between flow of information depending on types of convergence. They considered the statistical concept of causality as defined by Granger: while Granger’s causality based on time series, this paper focuses on continuous time processes.

• Song et al. (2012) introduced a new method for feature selection based on dependence maximization between the selected features and labels of an estimation problem using HSIC. The method is tested for artificial and real world datasets.

In the following section, the papers related to the grouping of Algebraic approach to causal inference are discussed:

2.6 Algebraic Approaches to Causal Structure Construction

The last category of papers under three groups, this grouping under Algebraic approaches for causal inference contains all the methods proposed other than probability and statistical methods. This group of papers present methods of algebraic operations performed on matrices, which are basically methods related to matrix operations, which are required to be solved for random datasets for the sole purpose of causal structure construction. Like Kernel methods, the algebraic approaches are formed to target the related features of the variable sets converted into matrices and using different matrix operations on these matrices the required triangular or independent equations can be obtained. Use of different matrix operations like ICA and PCA in formatting the independent equations can help to resolve the dependent structures. Most of the recent developments include a wide use of ICA with non-Gaussian nature of the data to construct the causal structures. Non-Gaussian data assumption features the most
useful properties to extract information. Although methods like ICA are classified under statistical techniques due to the use of matrix operations, we classified it under the group of algebraic tools.

All the following papers studied show the development of algebraic methods in the field of causal inference. A very steady progress can be seen through a year-long study of the papers provided in this section.

2.6.1 Linear Non-Gaussian Acyclic Model for Causal Discovery using ICA: ICA-LiNGAM

The development of extraction of information is done through the imposition of the different conditions on generating processes or observing the behaviour of the data. And Shimizu et al. (2006) imposed conditions on data to observe causal features. The proposed method called the Linear Non-Gaussian Acyclic Model, is known as LiNGAM. The model is based on three fundamental assumptions given as (1) data is generated through a linear process, (2) there exist no unobserved confounders, and (3) external influences are generated from non-Gaussian distributions and have non-zero variances. The causal model is identified using the method of Independent Component Analysis (ICA). Using ICA, they estimated the lower triangular matrices to show the causal related features and noise in the assumed model. They also performed pruning on edges to provide the best possible causal directions in the variables. The permutation over edges provides more accurate ordering of nodes in the causal model. The final results produced by the method contain causal ordering, errors in the system and coefficient matrix. This leads to the first and most effective modelling for causal learning. As this provides all the necessary outcomes to construct the causal structure, the following points provide the advantages and disadvantages of the modelling:

**Advantages:**

- Use of linear model for structural learning provides easier implication and non-Gaussian property provides flexibility for choice of data for external influences.
- Provides estimations for coefficients and possible causal ordering for causal construction.

**Disadvantages:**
2.6 Algebraic Approaches to Causal Structure Construction

- Assumption of having all the confounders observed may not support the discovery of all available confounders.
- Linearity of the model does not hold for all types of real-world datasets.

### 2.6.2 Distinguishing Causes from Effects using Nonlinear Acyclic Causal Models

In the above, the linearity of the model is constructed over the non-Gaussian disturbance and computed using ICA. Zhang and Hyvärinen (2008) proposed a method to distinguish cause from effects using a two-step nonlinear acyclic causal model. In the first stage of construction, each observed variable is generated by a nonlinear function of its parents with additive noise, followed by a nonlinear distortion. And in the second stage, nonlinearities of sensor distortions are applied on the parents formed in the first stage. For a two-variable case, they defined a Post-Non-Linear (PNL) Independent Component Analysis (ICA) to get causal-effect pairs, whenever non-linearity involved are invertible for the assumed model. For the discovery of directions in the causal model, they proposed a two-step method in which the first step is to use nonlinear ICA to find a PNL mixing model and then in the second step they verified the produced results using a kernel-based statistical test proposed by Gretton et al. (2007). The generated nonlinearities of independent parameters are produced through the multi-layer perceptrons (MLPs). The proposed nonlinear causal discovery method is applied to the "CauseEffectPairs" task dataset of Pot-luck challenge. Some of the important aspects in using this method are discussed below:

#### Advantages:

- The consideration of distortions and inner disturbance as functions is very useful to analyze causal system in a more detailed manner.
- Implication of the two-step method for causal structure and direction identification provides better results.

#### Disadvantages:

- The process of non-linearity in assumption is not very flexible and is quite a long process, not for computationally faster implementation.
- Does not provide sufficient outcome to construct a complete causal DAG.
2.6.3 Identifiability of the Post-Nonlinear Causal Model

The previous paper discussed provides a touch of post-non-linear effects of inner noise for inference on causal structures. But this phenomenon is not fully discussed and observed to its full efficiency level by any previous papers. Zhang and Hyvärinen (2009) discussed all the post-nonlinear (PNL) effects and how to measure them by different methods in this work. It provides a deeper analysis for the identifiability of PNL in two variable cases which is one of the major problems in PNL. They imposed that the identification of causal structure is fairly assured when PNL is used for the two-variable causal model. By identifying the equivalent classes in the causal structure for the Markov model, full detection of two variable models is possible. Considering the same model as discussed in Section 2.6.2, they formulated the PNL model for two variable causal structures. For models having more than two variables, they used a two-step PNL causal model verification method. They first used a conditional independence-based method to get d-separables for Markov equivalent classes. Then they used a PNL causal model to identify the directions in the causal structure by estimating disturbances for plausible causal structures. The proposal is tested for simulated datasets and a theorem is provided for all such conditions and cases where the PNL is unable to detect causal models.

Advantages:

• The complete causal structure is identifiable in case of two variable model, with very effective and easily detectable features.

• Easy to avoid an exhaustive search for directions over all possible causal structures and do not require any extra high-dimensional statistical tests to find the mutually independent variables in the causal model.

Disadvantages:

• Method is not successful for mixture models, meaning variables from different distributions cannot be solved using PNL.

2.6.4 Additive Noise Model for Nonlinear Causal Discovery

The relationship in between two variables is always understood through simple linear function, while in reality many of them follow a nonlinear relation. Often nonlinearity of the causal
relation is hard to derive and furthermore it is hard to work on. Hoyer et al. (2009) proposed the idea of handling and deriving nonlinear causal relations. They considered the linear relation in any two variables as the function of a nonlinear quantity with an error. They observed the consistency of direction in two variables by comparing them each as a nonlinear functional represented by separate two equations. By using the Hilbert-Schmidt independence criteria, they derived the independent components in the set and by performing a hypothesis test, the accuracy of the method is verified. The results for synthetic data are good whereas the real world data is very comparable with other methods of the same class. Some of the observations are discussed below:

Advantages:

- The proposed method is very simple to perform and quite accurate.
- Very good for smaller causal models.

Disadvantages:

- The model has a very low feasibility as it only holds for smaller networks, mostly less than 7 nodes.
- Comparing two variables at a time using this model is quite a slow process.
- Using other methods in combination such as HSIC and hypothesis testing are more likely to increase the step size and complexity of the method.

2.6.5 Direct Method for Learning a Linear Non-Gaussian Structural Equation Model: DirectLiNGAM

After the introduction of the Linear non-Gaussian acyclic model (LiNGAM) with ICA solvable method by Shimizu et al. (2006), the intervention method showed the usefulness of LiNGAM in causal discovery without any of prior knowledge. This method is known as ICA-LiNGAM, which leads for the proposition of some other method in the first stage of evolution. Next, Shimizu et al. (2011) introduced a direct evaluation method for causal models in a finite step computation preserving all the criteria of ICA-LiNGAM. As the intervention method of ICA-LiNGAM does not assure the correct evaluation of causal ordering, they introduced a new method to find the causal order with minimum error. They used regression-based
residual calculation for finding the independent features for comparative analysis. A finite step method for causal analysis is provided in the paper which is also compatible to perform on availability of prior information, producing better results than non-prior information models. Performance of the proposed model is tested for simulated data types of symmetric and asymmetric distributions of same kinds. The real-world data test results are good for a finite step model. The following points are discussed for comparable utilization:

**Advantages:**

- The method uses LiNGAM for causal estimations, which guarantees the right convergence in a fixed number of steps when the method is best followed.
- It produces results for causal ordering along with disturbance, directions and latent variables.

**Disadvantages:**

- Mixed model estimations are not provided or discussed in the paper and even incompatible for result analysis.
- Produces better results with prior information, which is not always possible for real world analysis.

### 2.6.6 Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models

Based on the likelihood estimations of Linear acyclic non-Gaussian model for causal directions, Hyvärinen and Smith (2013) presented the method of pairwise likelihood estimation on LiNGAM. They proposed a cumulant-based measure to analyze the likelihood ratios using a first order approximation. They show implications for more than two variable cases and cyclic nonlinear models. The likelihood ratio model uses the log-pdf to approximate the regressional residuals for the estimation of directions in between variables. The positive value reflects that the directions in any two variables may show the reverse direction for the negative value, for the same two variable case. Estimation for first order models is considered for sparse, skewed and noisy disturbance effects on the model. Variables having nonlinear relations are even thoroughly tested and observed for different cases like the first order. The
simulation results are tested for LiNGAM and DirectLiNGAM methods. For a different case scenario, they used the different methods to compare their results. Real word data test is carried out using FMRI dataset, and the efficiency of their proposed model for estimation of entropy is approximately 69% which is an observable difference as compared to other existing methods, which are approximately 61% (maximum). Some of the major aspects are discussed below:

**Advantages:**

- Effectively uses properties to estimate directions in any two variables.
- Can handle models of nonlinear and cyclic types with distortions like sparse, skewed and noisy.

**Disadvantages:**

- Multivariate analysis is not possible and very difficult for larger datasets.
- Accuracy level of the model is unknown as its results are different from methods LiNGAM and DirectLiNGAM.
- Causal structure estimation is not possible using this method though it can produce all the possible directions available in the variables.

### 2.7 A Technical Analysis of the Work

In above sections, all the selected papers have been discussed under three major groups to provide the methods these proposed for the analysis of causal DAGs. As this Chapter serves as a review of many different but most effective methods proposed for causal analysis over the past decades, the significance lies in tracking the effectiveness of their influence on later developments. In this scenario, this section is constructed to provide the technical arrangement of sections and sub-sections, and most of all the papers that have been used to provide the three categories for their use of different methods. This study strictly follows a year-long development and citation of works under each of the groups.

Table 2.1 shows all the papers those are referred in this Chapter and as they have been cited in different groups on a year-wise basis.
Table 2.1 The papers reviewed and cited in sections following a year-long increasing development useful for the construction of Directed Acyclic Causal Models

<table>
<thead>
<tr>
<th>Groups</th>
<th>Section</th>
<th>Citation</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic</td>
<td>2.3.1</td>
<td>Freiedman and Koller</td>
<td>2003</td>
</tr>
<tr>
<td></td>
<td>2.3.2</td>
<td>Beal and Ghahramani</td>
<td>2004</td>
</tr>
<tr>
<td></td>
<td>2.3.3</td>
<td>Sun et al.</td>
<td>2006</td>
</tr>
<tr>
<td></td>
<td>2.4.1</td>
<td>Eliden et al.</td>
<td>2007</td>
</tr>
<tr>
<td></td>
<td>2.4.2</td>
<td>He and Geng</td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td>2.4.3</td>
<td>Shpitser and Pearl</td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td>2.4.4</td>
<td>Pellet and Elisseeff</td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td>2.4.5</td>
<td>Janzing et al.</td>
<td>2012</td>
</tr>
<tr>
<td></td>
<td>2.4.6</td>
<td>Uhler et al.</td>
<td>2013</td>
</tr>
<tr>
<td>Statistical</td>
<td>2.5.1</td>
<td>Bach and Jordan</td>
<td>2002</td>
</tr>
<tr>
<td></td>
<td>2.5.2</td>
<td>Gretton et al.</td>
<td>2003</td>
</tr>
<tr>
<td></td>
<td>2.5.3</td>
<td>Gretton et al.</td>
<td>2005</td>
</tr>
<tr>
<td></td>
<td>2.5.4</td>
<td>Gretton et al.</td>
<td>2005</td>
</tr>
<tr>
<td></td>
<td>2.5.5</td>
<td>Sun et al.</td>
<td>2007</td>
</tr>
<tr>
<td></td>
<td>2.5.6</td>
<td>Zhang et al.</td>
<td>2009</td>
</tr>
<tr>
<td></td>
<td>2.5.7</td>
<td>Zhang et al.</td>
<td>2012</td>
</tr>
<tr>
<td></td>
<td>2.5.8</td>
<td>Chwialkowski and Gretton</td>
<td>2014</td>
</tr>
<tr>
<td>Algebraic</td>
<td>2.6.1</td>
<td>Shimizu et al.</td>
<td>2006</td>
</tr>
<tr>
<td></td>
<td>2.6.2</td>
<td>Zhang and Hyvärinen</td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td>2.6.3</td>
<td>Zhang and Hyvärinen</td>
<td>2009</td>
</tr>
<tr>
<td></td>
<td>2.6.4</td>
<td>Hoyer et al.</td>
<td>2009</td>
</tr>
<tr>
<td></td>
<td>2.6.5</td>
<td>Shimizu et al.</td>
<td>2011</td>
</tr>
<tr>
<td></td>
<td>2.6.6</td>
<td>Hyvärinen and Smith</td>
<td>2013</td>
</tr>
</tbody>
</table>

This Chapter also provides some basics regarding Causality and Bayesian Network use, to infer causality in Sections 2.1 and 2.2 respectively. In Sections 2.4 and 2.5, some papers have been referred which proposed the methods to construct undirected causal models under probabilistic approach and some papers containing techniques for deeper analysis of statistical methods respectively. Table 2.2 shows the papers those have been referred either as basic ideas, Undirected Causal Model construction or as related works under statistical approaches on causal analysis.

To provide a steady development in the field over the past decades, many papers are selected and categorized on their merit to infer causality. All the referred papers are ranked using a scoring system for (1) the technique they use, (2) handling, (3) complexity and (4) constructibility towards the causal model formation. These four parameters decide what score
Table 2.2 The papers as referred in this Chapter to provide basic ideas on Causality and Bayesian Analysis, for the undirected causal model construction and some others under statistical works for causal inference.

<table>
<thead>
<tr>
<th>Type of Work</th>
<th>Category</th>
<th>Citation</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Ideas</td>
<td>SEM</td>
<td>Bollen</td>
<td>1989</td>
</tr>
<tr>
<td></td>
<td>Causality</td>
<td>Pearl</td>
<td>1998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pearl</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pearl</td>
<td>2009</td>
</tr>
<tr>
<td></td>
<td>Bayesian Analysis</td>
<td>Spirtes et al.</td>
<td>1993</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gilks et al.</td>
<td>1996</td>
</tr>
<tr>
<td>Undirected</td>
<td>Probabilistic Methods</td>
<td>Geng et al.</td>
<td>2003</td>
</tr>
<tr>
<td>Causal Models</td>
<td></td>
<td>Zhang</td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Roy et al.</td>
<td>2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Borgelt</td>
<td>2010</td>
</tr>
<tr>
<td>Relative Works</td>
<td>Statistical Approaches</td>
<td>Gretton et al.</td>
<td>2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fukumizu et al.</td>
<td>2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hofmann et al.</td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Petrović and Dimitrijević</td>
<td>2011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Song et al.</td>
<td>2012</td>
</tr>
</tbody>
</table>

An paper gets as each of the criteria carries 10 points. An paper can achieve a score from 1 to 10 for each criterion, depending on the methods they use, how easy or difficult they are, the level of complexity in implication and ability to provide all relative information to construct a complete causal structure. Based on their scores the advancement graph is plotted to show the relative progress over the years in the selective groups. Fig. 2.2 shows the progress on causal analysis over the years in three groups.

The papers under their respective groups are arranged as they cited in the sections, shown in Fig. 2.2. Each Bar represents an paper and the scores they achieved for each of the criteria as considered for scoring. It is clear that algebraic methods score the highest of all, as they are technically sound, easier to handle, low in complexity and provide much of the required information to construct a causal model.

2.8 Conclusion

The recent advancements in the field of causal inference are based on methods (papers) with high impact on the structural analysis of causal orderings. The main goal of this is to
Fig. 2.2 The Advancement graph for Articles under different groups where each bar represent an article as represented in Table 2.1

provide a systematic review which is achieved in the discussed sections. The papers included are based on their ability to infer causality and make a significant contribution towards further developments in relative categories. Papers are discussed under selective, categorized groups for their central architecture for using particular methods to solve causal structure construction problems.

One can follow a year-wise development in any method studied in this Chapter, each providing all the information to implement them in different fields of study. Since it is required to suit a wide range of the concept, this study aims to provide pre-implementation aspects for better handling and fitting of the selected model. Under technical analysis, all the papers which are referred in this study process under different categories vitally contributes for their knowledge base towards the construction of the causal models. The tabulations are useful for easier access to the referred papers under three major groups. The advancement
graph shows our selection of papers as they were cited in continuous sections following a year-long development. Although none of the methods satisfies all the imposed criteria for a complete causal model, the steady development is a crucial factor shown in this work. The paper provides all the aspects necessary for the primary intervention to the subject of causality. This Chapter covers a wide range of papers in this study process, from the basics of the materials to the advanced methods providing background knowledge for related topics which is useful for an in-depth analysis of the subject.
Chapter 3

Altered-LiNGAM (ALiNGAM) for Solving Nonlinear Causal Models when Data is Nonlinear and Noisy

3.1 Introduction

The real word phenomena of dependability can be described in terms of causality, but to date, the available models are insufficient to perform on a wide range of observational data types. The causal models represented by acyclic graphs with directed edges and connection strengths are general models for structural explanations. The structural equation model (Bollen, 1989), serves as a path-finding tool for its simple linear structure. The probabilistic model given by Pearl (2000); Spirtes et al. (1993) for Bayesian network works well to describe a fair number of features of the causal structure, using the Markov network and its properties. The method provides estimation over causal models without having any prior information on the structures.

The paper by Shimizu et al. (2006) provides a better tool to describe linear causality in non-Gaussian and continuous datasets using a Linear Non-Gaussian Acyclic Model (LiNGAM). But LiNGAM is designed to work with independent component analysis (ICA), which makes it more complicated and inefficient for a wide range of data types. The gradient-based method of ICA fails to converge to global minimum and has a lower accuracy rate for the small number of nodal structures. The model even fails for mixture models of non-Gaussian data itself, and multivariate nonlinear sets are beyond its scope. A successful upgrade to (Shimizu et al., 2006), the method of Direct-LiNGAM (Shimizu et al., 2011) estimates the causal
model within a fixed step size and can find the global minimum. The method also provides the tool for the use of any prior available knowledge regarding the structure. Still, the method fails for non-Gaussian mixture models and produces better results when prior information is known. A very recent approach of likelihood estimation over LiNGAM by Hyvärinen and Smith (2013) can produce results for model estimations when the system is nonlinear and cyclic. But the accuracy level is very low at 69% and even if it produces all the causal directions then complete causal structure construction as well is not possible.

The method by Hoyer et al. (2009) was proposed to analyze nonlinear causality using an additive noise model, which is estimated over functions of probability densities of nodes. The method is tested using Gaussian distributions and independent non-Gaussian noises. The dependency in the nodes is verified using Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005b) for Gaussian kernel sets approximated by gamma distributions. So, the true outcomes of the model for a nonlinear data are unknown and the exact probability densities for observed variables are also unknown, as they are approximated through gamma distributions. The real-world test results are performed on bivariate sets, and there is no preferred full causal ordering. The method of HSIC (Gretton et al., 2005a) provides the criteria to check the independent variables present in the system, by taking the sum of the squared singular values of cross-covariance operators. Indirectly HSIC depends on variable correlation and covariance factors to derive the independent features. The method is found to be insufficient for a higher number of nodes as it only can estimate when the number of nodes is less than seven.

In this Chapter, a method is proposed to find nonlinear causality when data is nonlinear or mixed non-Gaussian and noisy. The structure of LiNGAM is used to find causal ordering in the mentioned data types, resulting in an easier method for error minimization. By reducing the noise factors affecting the model, the causal directions and causal orders are successfully estimated for any given sets of data. The Probable values for causal directions provide accurate causal directions and help to discover the causal ordering with their kinship relations. The model ALiNGAM uses the least square estimation and has a much lower computational complexity than other methods. A different estimation technique is proposed from LiNGAM and also the method ALiNGAM can analyze acyclic models which provide better inference for causal analysis. It is also discussed why there is no possibility for the formation of cycles in causal models estimated by ALiNGAM and how it can estimate over every nodal direction resulting into acyclic causal models. The performance of ALiNGAM is compared with LiNGAM and DLiNGAM to provide the usefulness and novelty of the proposed one. In the comparison test, a table is provided to show the estimated results (number of directions) using simulated causal models and figures illustrate the accuracy and effectiveness in causal
direction estimation. The complexity of the proposed algorithm is discussed with comparable methods of the same class.

This thesis introduces a new concept of causal levels, to arrange and represent the estimated causal directions in a directed acyclic causal models. Although researchers have used the sub-structural arrangement methods to construct the causal model, there is no effective tool available to do the same. The causal levels are provided which make it easier to arrange the features as nodes using their inter-dependency or kinship relation. By arranging them in subsequent levels of similar groups with similar kinship relations, the causal directions can be represented into causal orderings and then into a directed acyclic causal model.

In the following sections, the discussion is started by providing the basic model for LiNGAM with selective assumptions in Section 3.2. Section 3.3, best explains the identifiability of the complete causal model using the proposed method and provides the tool for effective error minimization using the probable causal direction method. The methods for successful evaluation of causal directions and algorithms are discussed in Section 3.4. The experimental set-ups, results for synthetic, real world datasets and comparisons with selected methods are shown in Section 3.5.

### 3.2 Model Selection

The method LiNGAM (Shimizu et al., 2005) was proposed to analyze causal inference when the nodes of the graph are generated from the Gaussian process with non-Gaussian independent noises. The general structure of LiNGAM (Shimizu et al., 2006) is used with some specific system assumptions to make it efficient and easier for nonlinear causal analysis. The system assumptions are as follows: consider the directed acyclic graph (DAG) $G$, where each node $i$ is observed as a variable $x_i$. This DAG can be represented by an adjacency matrix of $(m \times m)$ for $C = \{c_{ij}\}$, where each $c_{ij}$ is the connection strength from a variable $x_i$ to $x_j$. By defining the order $O'$ of nodes such that none of the later nodes have direction towards any of the earlier nodes and then $O(i) > O(j)$ denotes a causal ordering, where variable $x_i$ is observed prior to variable $x_j$. A directed path from $x_i$ to $x_j$ is a sequence of directed edges where $x_j$ is reachable from $x_i$ using these paths. Since in the real world, there are always some unmeasured or unknown effects of environmental factors affecting the system, consider the following model added with external noise as:

$$ x_j = \sum_{o(j) < o(i)} c_{ij} x_i + e_j \quad \text{for} \quad (j \neq i) $$

(3.1)
where $c_{ij}(j \neq i)$ is the connection strength whenever $x_i$ has a directed edge towards $x_j$ and $e_j$ is an external influence or error/noise. The error $e_j$ is a continuous random variable with non-Gaussian distribution. The assumption is that, there is no latent confounding variables leading $e_j$ to be independent of each other. Assume that $x_j$ is equal to $e_j$ if no other observed variable $x_i(j \neq i)$ has a directed edge to $x_j$ in the model, so all $c_{ij}(j \neq i)$ are zeros. In these cases, the noise $e_j$ is detected as $x_j$, where either $x_j$ is an exogenous variable or $x_j$ is a noise.

Furthermore, all the variables are generated through a nonlinear process, and they have nonlinear causality. The assumption of graph $G$ is DAG and all the variables represented as nodes follow a particular causal ordering in the graph. This leads to the fact that there is only one way to represent a particular node as a combination of any other nodes and external noises. As shown in eq. 3.1, the node $x_j$ can only be represented by $x_i$ not in a reverse way from $x_j$ to $x_i$ and this is true for all the nodes.

Assuming all the above assumptions are true and by considering two opposite models for analyzing the causal directions in any two particular set of nodes, the actual causal order can be found in the system. The minimization of the error, probable causal directions and identifiability are discussed in the next section.

### 3.3 Identifiability of the Model

The model shown in eq. 3.1, represents the existence of causal direction $\{x_i \rightarrow x_j\}$, which can be explained using Causal Markov Condition (CMC). For a graph $\mathcal{G}$ with node set $\mathcal{N}$, the node $x_j \in \mathcal{N} \setminus \{x_j\} \subseteq \mathcal{N}$ then CMC provides that $x_j$ is independent of all the variables in $\mathcal{N}$ which are not direct causes of $x_j$. In the case of eq. 3.1, $\{x_i \rightarrow x_j\} \Rightarrow Pr(x_j|\text{Parent}(x_j))$. But this process needs computation for probability density functions of parent and child nodes, which is difficult to approximate in case of nonlinear noisy models. In the following discussions, a better criteria is provided for the identification of causal directions in bivariate models.

The Left-Hand Side (LHS) of eq. 3.1 is the child node $x_j$ which is approximated by the combination of parent node $x_i$ and $e_j$ on the Right-Hand Side (RHS) of eq. 3.1. So, the division of these sample values to their respective population values will provide the Probable Causal Direction (PCD). The probable values of causal directions (CDs) for eq. 3.1 can be
3.3 Identifiability of the Model

given as:

\[
\begin{align*}
\text{LHS} & \quad \text{RHS} \\
\frac{x_j}{x_j} & = \frac{|\sum c_{ij}x_i|}{|\sum c_{ij}x_i + e_j|} + \frac{|e_j|}{|\sum c_{ij}x_i + e_j|}, & (3.2) \\
1 & = CD_{\text{Parent}=x_i} + CD_{\text{Noise}=e_j}; \\
\text{Prob}(\text{Child}=x_j) & = \text{Prob}(\text{Parent}=x_i) + \text{Prob}(\text{Noise}=e_j),
\end{align*}
\]

where Prob represents the probable values for causal direction (PCD) in the observed node set and CD represents the causal direction.

In eq. 3.2, it can be seen that the cause i.e. LHS is an effect of parent node and additive noises on RHS and the probability of causal directions are decided from their values as discussed in following points.

1. If Prob\((x_i) = CD_{x_i} > \text{Prob}(e_j) = CD_{e_j}\) then the direction \(\{x_i \rightarrow x_j\}\) will be found.

2. If Prob\((e_j) = CD_{e_j} > \text{Prob}(x_i) = CD_{x_i}\) then the node \(x_j\) is observed as an external noise.

For causal discovery and direction detection, only step 1 is considered, which will maximize the probable causal direction criteria.

Consider the graph \(G\) where the node \(i\) is a parent, \((j, k)\) are descendants of \(i\) and \(l\) is the descendant of \((j, k)\), are generated through nonlinear processes. The observed datasets are given as \(x_i, x_j, x_k\) and \(x_l\) with the causal ordering as \(O_i, O_j, O_k\) and \(O_l\) respectively. The causal orderings are as follows, \(\{O_j < O_l > O_k\}\) and \(\{O_j > O_k < O_l\}\) and the directions in \(G\) is as following: \(x_j \leftarrow x_l \rightarrow x_k\) and \(x_j \rightarrow x_l \leftarrow x_k\). A graphical representation of \(G\) is given in Fig. 3.1.

In Fig. 3.1, the causal model shows a four-nodal structure with three kinship relations. The causal model preserves all the criteria for graph \(G\), while the causal ordering in \(G\) is transformed into causal kinship relations with their respective levels and names as marked in Fig. 3.1.
3.3 Identifiability of the Model

Using the model defined in eq. 3.1, the graph $G$ can be represented by its causal directions and causal orderings as in the following equations:

\[
\begin{align*}
x_j &= \sum_{o(j)<o(i)} c_{ij}x_i + e_j \quad \text{for } (j \neq i), \\
x_k &= \sum_{o(k)<o(i)} c_{ik}x_i + e_k \quad \text{for } (k \neq i), \\
x_l &= \sum_{o(l)<o(j)} c_{lj}x_j + e_l \quad \text{for } (l \neq j), \\
x_l &= \sum_{o(l)<o(k)} c_{lk}x_k + e_l \quad \text{for } (l \neq k).
\end{align*}
\]

(3.3)

For estimation of probable causal directions in eq. 3.3, this can be represented using eq. 3.2 as below,

\[
\begin{align*}
Prob(x_j) &= Prob(x_i) + Prob(e_j) \\
Prob(x_k) &= Prob(x_i) + Prob(e_k) \\
Prob(x_l) &= Prob(x_j) + Prob(e_l) \\
Prob(x_l) &= Prob(x_k) + Prob(e_l)
\end{align*}
\]

(3.4)

Considering cases in eq. 3.4, it is clear that the total probable value on left-hand side of equation is 1, which is approximated over the combined probable values of parent nodes and errors of the system. Furthermore, assuming that the system is DAG, there is no bi-directional in any cases. The two-step process discussed for directions $x_i \rightarrow x_j$ are true for all the nodes in the graph, which leads to the successful evaluation of causal directions in the model only by checking the probable values of errors in the considered set of nodes. Notice that,
3.4 Evaluation over the Model

the minimization of the error in any case, maximizes the probable causal directions, which ensures the existence of causal directions.

For causal orderings, all the evaluated causal directions must result in an acyclic model. For this checking the direction in \( x_j \) and \( x_k \), the probable values in these two variables can be written as,

\[
Prob(x_j) = Prob(x_k) + Prob(e_j), (x_k \rightarrow x_j)
\]

\[
Prob(x_k) = Prob(x_j) + Prob(e_k), (x_j \rightarrow x_k).
\]

In either of the cases of two of the above, the resulting graph is acyclic and may be noted that the descendant nodes generated from same parent may produce causal direction in them. The best-case results may show that the parent node \( x_i \) is entirely independent of all its descendant nodes, while \( x_l \) is dependent on every parent node. But in all cases the acyclic nature of the graph remains unchanged. The evaluation of the model is discussed in the next section.

3.4 Evaluation over the Model

The conventional or pre-existing methods for causal directions require computation over the probability density functions to estimate the non-linearity of the model. While using the probability for evaluation, the problem starts with the proposal density functions. These methods try to approximate the nonlinear densities using the proposal density, which is not possible by any means while using a linear non-Gaussian model as a proposal for approximation. If the probability feature is not to be used then the matrix method can be used to solve this issue with the help of independent component analysis method (ICA) (Shimizu et al., 2006, 2005). But ICA depends on covariance structure to find independent features.

In this proposal to find causal directions in graph \( G \), the model given by eq. 3.3 is to be evaluated, which ultimately decomposes into eq. 3.4. For the successful evaluation of causal directions in any two sets of nodes, the direction is to be estimated in both the ways (\( \leftrightarrow \)). Consider the case for \( x_j \) and \( x_l \) and the causal model for both possible causal directions in them as below:

\[
x_l = c_{jl} x_j + e_l \quad (x_j \rightarrow x_l),
\]

\[
x_j = c_{lj} x_l + e_j \quad (x_l \rightarrow x_j).
\]
3.4 Evaluation over the Model

Here requirement is to find the direction from \( x_j \) to \( x_l \) and the reverse direction from \( x_l \) to \( x_j \). This process is valid for all the nodes of graph \( G \). The probable values for causal directions in \( x_j \) and \( x_l \) can be given by

\[
\text{Prob}(x_l) = \text{Prob}(x_j) + \text{Prob}(e_l) \quad \text{and} \quad \text{Prob}(x_j) = \text{Prob}(x_l) + \text{Prob}(e_j).
\]  

(3.6)

For minimization of error, the minimum of \( \text{Prob}(e_l) \) and \( \text{Prob}(e_j) \) are to be estimated. The minimum of error probable values can be estimated by solving the eq. 3.5. A Least Square Estimation Method is used to solve the equations by taking the partial derivatives over unknowns. The generalized matrices solvable for any two arbitrary set of nodes \( x \) and \( y \) using eq. 3.5 are given as follows:

\[
\begin{bmatrix}
\sum_{m}^x \sum_{m}^x \\
\sum_{m}^x - 1
\end{bmatrix}
\begin{bmatrix}
e_{x,y} \\
e_{y}
\end{bmatrix}
= \begin{bmatrix}
\sum_{m}^x \sum_{m}^y \\
\sum_{m}^y - 1
\end{bmatrix}
\]

for \( (x \rightarrow y) \)

\[
\begin{bmatrix}
\sum_{m}^y \sum_{m}^y \\
\sum_{m}^y - 1
\end{bmatrix}
\begin{bmatrix}
e_{y,x} \\
e_{x}
\end{bmatrix}
= \begin{bmatrix}
\sum_{m}^y \sum_{m}^x \\
\sum_{m}^x - 1
\end{bmatrix}
\]

for \( (y \rightarrow x) \)

(3.7)

Equation 3.7 can be used to solve the unknowns in the model for evaluation processes. After this, eq. 3.2 can be used to find the probable values of errors for both cases and the minimum will provide the most probable value for causal direction.

Non Existence of Cycles in Estimated Models: Consider a DAG \( \mathcal{G} \) with \( n \) number of nodes. All the \( n \) nodes are observed as variables \( X_i \), \((i = 1, 2, \ldots, n)\) in a data set. \( X \) is the matrix of variables of dimension \((m \times n)\), where \( m \) is the number of observations or instances and \( n \) is the number of variables. The directions found from evaluation process are stored in a matrix \( \text{Direction} \), the dimension of which is \((n \times n)\). The algorithm is designed to store binary numbers 1 if a causal direction exists and 0 if there are no directions in the observed variables. In the case of conventional methods, the produced direction or path coefficient matrices are either upper or lower triangular. The proposed model estimates the same \( nC_2 \) number of causal direction as other methods, but instead of producing a triangular matrix, the proposed model finds every possible direction at random in each of the observed nodes. Because of this estimation process the question arises whether there is any possibility of ALiNGAM estimating cycles in the model. The answer is 'No', as in causal analysis the parent is independent of all child nodes, and our assumption of no later node has directed
edge towards prior ones leading to the conclusion that cycles are not possible to estimate. This process can be better understood in the following example:

**Example:** Consider a small causal model consisting of three nodes \( \{A, B, C\} \) and let the model be of type \( \{A \rightarrow B, B \rightarrow C\} \). When AliNGAM estimates causal directions, assuming the estimation is started from node \( A \). Then AliNGAM is going to estimate the causal directions in the following order \( \{A \rightarrow B, C\}, B \rightarrow C\} \). It can be observed that none of the directions are estimated twice, because while estimating directions from \( \{A \rightarrow B, C\} \) the method also estimated the directions \( \{B \rightarrow A, C \rightarrow A\} \). As node \( A \) is generated using independent random distribution, there is only one way to generate and represent it. Being a parent node which generates \( B \) and then from \( B \) then node \( C \) is generated. There is every possibility that any method will estimate a direction \( \{A \rightarrow C\} \). Any method which estimates a direction \( \{C \rightarrow A\} \), which is not considered in the original model and is a reverse direction from what is assumed, is a wrong direction. These methods are found to be insufficient for estimation of causal directions. As in every direction estimation, the probable causal direction values are checked; there is no possibility of estimation of the direction \( \{C \rightarrow A\} \) over the direction \( \{A \rightarrow C\} \). The choice of node \( A \) as a parent can influence all the child nodes and notice that these influences may be minor for subsequent child nodes. This means the probable causal direction values for \( \{A \rightarrow C\} \) may not be as good when compared to \( \{A \rightarrow B\} \) or \( \{B \rightarrow C\} \), but there is no chance that the probable values of \( \{C \rightarrow A\} \) will be higher than the probable values of \( \{A \rightarrow C\} \). The direction \( \{A \rightarrow C\} \) is considered as an extra direction, which does not affect the acyclic nature of the graph. So, the model ALiNGAM never estimates cycles in causal models, and never produces wrong or reverse directions.

The above explanation for the example set of triplets can extend to a multivariate case, but the results remain the same. The method ALiNGAM does not support for cyclic direction estimation in causal models, which preserve the assumption of a causal model being a DAG.

The estimated directions can be arranged from the most independent parent node to the least independent descendant nodes. In this process, the directions which are produced due to a common parent structure can be removed and they can be arranged in causal levels as shown in Figure 3.1. The algorithm for the proposed model ALiNGAM is given below.

The complexity of the proposed bivariate model is discussed below:
Algorithm 1 Altered-LiNGAM for Non-Linear Causal Analysis

 Require: A matrix $X$ of size $(m \times n)$.
1: for $i = 1$ to $(n - 1)$ do
2:  for $j = i + 1$ to $n$ do
3:    Solve for Equation 3.7, where $x = X[i]$ and $y = X[j]$.
4:  end for
5: end for
6: Find $Min = \min(CD_{ij}, CD_{ji})$, where $CD_{ij} = \text{Prob}(e_{X[j]} | x) \& CD_{ji} = \text{Prob}(e_{X[i]} | y)$.
7: if $(Min = CD_{ij})$ then
8:  Update $DIRECTION[i, j]$
9: else if $(Min = CD_{ji})$ then
10:    Update $DIRECTION[j, i]$
11: end if
12: return $(DIRECTION)$

Computational Complexity

Most of the algorithms have higher complexity and have limitations when evaluating the causal models. The estimated maximum computational complexity of $O(n^4)$ is found for methods Hoyer et al. (2009); Hyvärinen and Smith (2013); Shimizu et al. (2006, 2011), which is significantly high considering the number of nodes these can process. Some of the methods are more suitable for a lesser number of nodes ($n < 5$), while others show improvement in the direction of estimation for the higher number of nodes ($n > 7$). As this proposed method computes directions for two nodes at a time by checking opposite directions in them, a total of $\binom{n}{2}$ combinations of nodes can be found for a $n$ variable dataset. This is twice the number of equations than nodes i.e. $2\binom{n}{2}$ are to be solved to find all the probable causal directions. When a pair is only visited once through the whole process, the complexity of the model is less than $O(n^2)$. The change in the number of nodes does not affect the performance or evaluation process of the model.

In the next section, the experimental set-ups, data generating process, real world data test and comparison test results are provided to validate the proposed model.
3.5 Experiments, Results and Comparison Test

To be able to find true causal directions when all the assumptions hold true for the model, the proposed method is tested on various simulated and real world datasets. Considering the fact that data are correlated, for most of the cases, the system becomes noisy. In this scenario, the system holds all its properties of DAG, and there should be no bi-directed edges. Next, the processes to generate synthetic nonlinear and non-Gaussian mixture model datasets and the results using the proposed method are provided.

3.5.1 Results for Synthetic Data

To test and validate the model the simulations are performed using the proposed method on different types of datasets with a varying number of features. The results for synthetic nonlinear and noisy datasets with 5 and 7 nodes, and non-Gaussian symmetric and non-symmetric mixture models with 6 and 8 nodes are estimated. For test purposes, the produced synthetic datasets are drawn very carefully such that none of the descendant variables will be able to imitate their parent variables. So, the functional relation in the variables never changes in their data states. First, let us start with synthetic datasets with simple nonlinear non-Gaussian noisy relations.

3.5.1.1 Synthetic Dataset-1

A nonlinear causal structure of 5 nodes with 4 kinship levels are generated. All the five nodes are nonlinear and acyclic. Consider the node set \{A, B, C, D, E\} with causal directions \{B ← A → C\}, \{B → D ← C\} and \{D → E\}. The nodes are generated as follows, \( A = \text{Exp}(r = 1.5) \), \( B = 1/\exp(A) + e_B \) with \( e_B = \text{Unif}(-1,2) \), \( C = 2 \cdot \tan(A) + e_C \) with \( e_C = T(df = 2) \), \( D = \cos(B/2) + 2(1/\sin(C)) + e_D \) with \( e_D = \text{Unif}(-2,1) \) and \( E = 1/\sinh(D) + e_E \) with \( e_E = T(df = 5) \). Notice that all the errors added to the system are non-Gaussian, which satisfy the model requirements of LiNGAM. The ’Exp’ represents exponential distribution with the rate of generation ’r’, ’Unif’ represents uniform distribution and ’T’ represents the student’s t distribution with their respective degrees of freedom by ’df’. Here 500 data points are sampled for each of the nodes and error sets and then applied to algorithm 1 to find the causal directions. The whole process is fully randomized and the consistency of the model is observed for 10 simulations.
Table 3.1 shows the results for synthetic dataset-1 of dimension $(5 \times 5)$ and each node is compared with every other node in the simulation process. Any value in Table 3.1 represents the direction from row variable to column variable. Note that the results shown in the matrix indicate a value 0 if there is no direction and 1 for the existence of direction. Although the DIRECTION matrix records the number of directions found in each case, the algorithm is designed to visit a pair only once in the simulation process to accelerate the evaluation. Table 3.1 provides the results for the most number of simulations out of 10. From 10 simulations, in 6 of the simulations a direction from $\{B \rightarrow C\}$ is found, while 40% results show the direction from $\{C \rightarrow B\}$ and the most probable outcome of $\{B \rightarrow C\}$ is selected.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The following observations are crucial for causal constructions:

1. Node $A$ is completely independent of all other nodes, or none of the nodes have causal directions towards $A$. So node $A$ is a parent in the causal model.

2. Node $E$ can be achieved from any other node, and its dependability on every other node makes it the last node on the causal levels.

3. The directions between $B$ and $C$ are quite uncertain, while they both have a common parent $A$. And both $B$ and $C$ are at same causal level.

4. Assuming that node $D$ only depends on nodes $B$ and $C$, which make it easier to remove the extra direction from $A$ to $D$.

5. Node $E$, being the least independent node of all the others, has causal directions from every node high in the causal level. To simplify the causal model construction, it is assumed that there is only one direction $\{D \rightarrow E\}$ in the kinship relations.

Fig. 3.2 shows the original causal model for the dataset-1 and the constructed causal model using results generated from the simulations.
3.5 Experiments, Results and Comparison Test

Notice that the extra direction \( \{B \rightarrow C\} \), which is only true for 60% of the results is due to the common parent \( A \). And, notice that the change in parent datasets in successive simulations affects the relation in descendant nodes as discussed in the above points. The extra direction in the simulated model does not affect the criteria of DAG: the causal orderings and kinship levels also remain the same for the simulated model when compared with the original causal model.

### 3.5.1.2 Synthetic Dataset-2

A causal structure with 7 nodes having nonlinear relations is generated. The nodes are ordered as \{B, C, A, D, F, G, E\} with the causal directions \{C \leftarrow B \rightarrow D\}, \{B \rightarrow A\}, \{C \rightarrow F \leftarrow A\}, \{A \rightarrow G \leftarrow D\}, \{F \rightarrow E \leftarrow G\}. The nonlinear causal relations are generated through the following processes, \( B = \text{Exp}(r = 1) \), \( C = \text{tan}(B)/2 + e_C \) with \( e_C = \text{Unif}(-2, 2) \), \( A = 1/\sinh(B) + e_A \) with \( e_A = T(d_f = 3) \), \( D = B / (\sin(B) - \cos(B)) + e_d \) with \( e_D = \text{Unif}(-1, 2) \), \( F = \text{tan}(C)/2 + \sin(A) + e_F \) with \( e_F = T(d_f = 5) \), \( G = 1/\sqrt{|A|} + (1 - D)/D + e_G \) with \( e_G = \text{Unif}(-1, 1) \), \( E = \sin(F/2)/(1 - \cos(F/2)) + (1 - 7 * \text{tan}(G/2)) + e_E \) with \( e_E = \text{Exp}(r = 10) \). For these 800 instances of nodes, errors are generated and then the method is applied to find the causal directions. Table 3.2 provides results for Dataset-2.

Table 3.2 shows the causal directions in Dataset-2 found from the simulation. To arrange the nodes in a causal model the order of dependency of each node is followed and are sorted in order from highly independent to least independent. The generated causal model is given in Fig. 3.3 with the original causal model for Dataset-2.
Table 3.2 DIRECTION matrix for Dataset-2

<table>
<thead>
<tr>
<th>Nodes</th>
<th>B</th>
<th>C</th>
<th>A</th>
<th>D</th>
<th>F</th>
<th>G</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

To generate the simulated causal model as shown in Fig. 3.3, the directions $D \rightarrow C$ and $A \rightarrow E$ are discarded. However, the addition of these nodes to the simulated model does not affect the acyclic nature of the causal model and none of the system assumptions are violated. The deletion of these nodes only makes the graph easier to represent in 2-dimensions with their respective causal levels.

To test the ability of the proposed model the next two experiments are carried out with symmetric and non-symmetric mixtures of non-Gaussian distributions. The basic non-Gaussian distributions used are (1) Exponential- $Exp$, (2) Double Exponential- $DExp$, (3) Log Normal- $LNorm$, (4) Student’s $t$- $T$, and (5) Uniform- $Unif$. The dependable nodes are generated by adding a mixture of non-Gaussian distribution of symmetric or non-symmetric types to make it a complex mixture model.
3.5.1.3 Synthetic Dataset-3

The Dataset-3 is a symmetric mixture of non-Gaussian distributions, which validates the flexibility to handle different kinds of datasets without losing the effectiveness of the model. The dependent nodes are generated using a symmetric mixture of arbitrary non-Gaussian errors. For symmetric mixture model a 6 node causal model with node set \{A,B,D,C,E,F\} is generated. Causal directions in the generated model are \{B \leftarrow A \rightarrow C\}, \{A \rightarrow D\}, \{B \rightarrow E \leftarrow D\} and \{D \rightarrow F \leftarrow C\}. The nodes are generated through following process, \(A = DExp(r = 0.15)\), \(B = (1-A)/A^3 + e_B\) with \(e_B = [LNorm(\mu = 0.53, \sigma = 1.37), T(df = 3)]\), \(C = \tan(A) - e_C\) with \(e_C = [T(df = 3), Unif(-3,3)]\), \(D = \cos(1/A)/(\sin(A))^2 - e_D\) with \(e_D = [LNorm(\mu = 0.53, \sigma = 1.37), Unif(-3,3)]\), \(E = 1/\cos(D) + 2/\sin(B) + e_E\) with \(e_E = [Exp(r = 0.35), Unif(-3,3)]\), \(F = D/(5(D - 1)) + C/\cos(C/2) - e_F\) with \(e_F = [Exp(r = 0.35), T(df = 3)]\). The directions found from simulation using the proposed method are shown in Table 3.3.

Table 3.3 DIRECTION matrix for Dataset-3

<table>
<thead>
<tr>
<th>Nodes</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Using the directions obtained from Table 3.3, the simulated model is compared to the original model and is shown in Fig.3.4. Notice that some nodes are removed from the Fig. 3.4 to represent all significant directions.

3.5.1.4 Synthetic Dataset-4

This dataset is generated using non-symmetric mixtures of non-Gaussian distributions. The causal model with 8 nodes as \{A,B,C,D,E,F,G,H\} with causal directions \{C \leftarrow A \rightarrow D\}, \{D \leftarrow B \rightarrow E\}, \{C \rightarrow F \leftarrow D\}, \{D \rightarrow G \leftarrow E\} and \{F \rightarrow H \leftarrow G\} is selected to test the performance of the proposed method. The nodes are generated through the following process, \(A = DExp(r = 1.5)\), \(B = LNorm(\mu = -1.53, \sigma = 3)\), \(C = 2/\tanh(A) + e_C\) with \(e_C = [T(df = 3), T(df = 5)]\), \(D = A^2/(1 - A) + 2/\sin(B) - e_D\) with \(e_D = [LNorm(\mu = -1.53, \sigma = 3), Unif(-3,3)]\), \(E = 2/\tan(B/2) - e_E\) with \(e_E = [Exp(r = 0.75), Unif(-3,3)]\), \(F = \cos(1/A)/(\sin(A))^2 - e_F\) with \(e_F = [Exp(r = 0.75), T(df = 3)]\), \(G = \cos(1/A)/(\sin(A))^2 - e_G\) with \(e_G = [Exp(r = 0.75), T(df = 3)]\), and \(H = 1/\cos(D) + 2/\sin(B) + e_H\) with \(e_H = [Exp(r = 0.75), Unif(-3,3)]\).
3.5 Experiments, Results and Comparison Test

Fig. 3.4 Causal models for Original and Simulated DAGs for Dataset-3

\[ F = 1/\sin(C) - 1/\cosh(1-D) - e_F \text{ with } e_F = \left[ T(df = 5), \text{Unif}(-3, 3) \right], G = \sin(D)/\cos(E) - e_G \text{ with } e_G = \left[ \text{Exp}(r = 0.75), T(df = 3) \right] \text{ and } H = E/\tanh(E) - 1/\tan(F) - e_H \text{ with } e_H = \left[ \text{Exp}(r = 0.75), T(df = 5) \right]. \] The directions found in Dataset-4 using algorithm 1 are provided in the following Table 3.4.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

It may be noticed that all the errors in the Dataset-4 are non-symmetric mixtures of non-Gaussian distributions as given above. The causal model has two independent parent nodes generating 4 levels of kinship relations. Fig. 3.5 shows the results for the simulated model with its original causal structure.

The proposed method is found to work well for different types of datasets and results are discussed and verified in tables and figures given above for synthetic datasets. In the section below, the performance of ALiNGAM for real world data is provided.
3.5 Experiments, Results and Comparison Test

3.5.2 Real World Data Test

The practical use of the proposed method is as yet unclear until it is tested for real world cases. For this purpose, two of the datasets are selected in which the variable relations follow the general conception of pre-existing facts or known ideas. This helps to track the behaviour of the proposed model as well as to validate the estimated results.

3.5.2.1 Abalone Data

First, Abalone data of Asuncion and Newman (2007) is used, available at UCI repository for this real world test. The test criteria are to check that given the measurements of the shell, whether the age of the fish can be predicted in an easier way or not. The long process of finding out the age is to count the number of rings in the shell of the fish. The dependency is to be found by drawing the causal model for the Abalone data. For use, the column of ‘Sex’ are removed from the original data, which only contains the binary values. The variables are \{Length \(- L\), Diameter \(- D\), Height \(- H\), WholeWeight \(- WW\), ShuckedWeight \(- SW\), VisceraWeight \(- VW\), ShellWeight \(- ShW\), Rings \(- R\)\}. Table 3.5 provides the directions in Abalone Data.

From Table 3.5 it is clear that Shell Length- \( L \) depends on Diameter- \( D \), Height- \( H \) and Whole Weight- \( WW \). Notice that the age of Abalone depends on all other measurements taken from the Shell and the directions show that ‘Measurements of the Shell can predict the age of Abalone’.
Table 3.5 DIRECTION matrix for Abalone Data

<table>
<thead>
<tr>
<th>Nodes</th>
<th>L</th>
<th>D</th>
<th>H</th>
<th>WW</th>
<th>SW</th>
<th>VW</th>
<th>ShW</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>WW</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SW</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>VW</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ShW</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

3.5.2.2 Energy Efficiency Data

For the second test, the Energy Efficiency Dataset is used from UCI repository of Asuncion and Newman (2007) which contains 10 variables of which 8 are input nodes and the last 2 are output/response nodes, which are to be predicted. The Input nodes are as follows \{X1\}- Relative Compactness, \{X2\}- Surface Area, \{X3\}- Wall Area, \{X4\}- Roof Area, \{X5\}- Overall Height, \{X6\}- Orientation, \{X7\}- Glazing Area, \{X8\}- Glazing Area Distribution. Output nodes are \{y1\}- Heating Load and \{y2\}- Cooling Load. The test is performed to see whether or not the Heating and Cooling loads are affected by the total area of the Room’. Table 3.6 provides the causal relations in Energy Efficiency dataset.

Table 3.6 DIRECTION matrix for Energy Efficient Data

<table>
<thead>
<tr>
<th>Nodes</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>y1</th>
<th>y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>X6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>X7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>y1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>y2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

It is evident that the Heating and Cooling loads only depend on the total area of a room. The causal relations shown in Table 3.6 validate that ‘Heating Load- \(y1\) and Cooling Load- \(y2\) are get affected by Surface Area- \(X2\), Wall Area- \(X3\) and Roof Area- \(X4\)’. The combined
area of Surface Area- X2, Wall Area- X3 and Roof Area- X4 provide the total area of the room. So, the causal directions are found to be correct from the estimations.

The proposed model works well for real world datasets and results which validate the model are shown in above tables and figures. The assumptions hold correct and causal directions are found to be true for different kinds of test sets. In the next section, a comparison of the proposed method is provided with LiNGAM and DLiNGAM methods.

3.5.3 Comparison Test

The comparison test is performed using the original LiNGAM method by (Shimizu et al., 2006) and the later one of the methods developed based on LiNGAM called DLiNGAM (Shimizu et al., 2011). Both methods provide the facility to estimate connection strengths, error and causal ordering of the nodes. But for exact estimation of the causal directions using these methods, manual searching is required to find the connection strength values from the estimated strength matrix. For comparison, only the correct and wrong are selected for causal direction estimation for each of the methods. Table 3.7 provides the comparison of the Altered-LiNGAM method with LiNGAM and DLiNGAM methods, where CORR and, WRON represent correct and wrong directions respectively. The same synthetic datasets which are used in Sections 3.5.1.1, 3.5.1.2, 3.5.1.3 and 3.5.1.4 are used for this testing of the proposed model.

Table 3.7 Comparison for causal direction estimations in LiNGAM, DLiNGAM, ALiNGAM

<table>
<thead>
<tr>
<th>Methods</th>
<th>LiNGAM</th>
<th>DLiNGAM</th>
<th>ALiNGAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data/Directions</td>
<td>CORR</td>
<td>WRON</td>
<td>CORR</td>
</tr>
<tr>
<td>Dataset 1</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Dataset 3</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Dataset 4</td>
<td>7</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

The above results in Table 3.7 show that the proposed Altered-LiNGAM performed very well for all datasets and most of the time performed a lot better than the previously proposed methods. Notice that all the above methods use the bivariate analysis system, so these methods are better suited for our comparative analysis. The correct directions and wrong directions are estimated strictly from the originally assumed number of edges in these simulated cases. The extra set of directions is those which are estimated by the methods but do not affect the acyclic nature of the model, whereas the wrong directions are those which
are estimated just opposite from what is assumed in case of simulated models. The following points summarize the findings from Table 3.7:

- For methods LiNGAM and DLiNGAM, the directions are estimated by counting the connection strength values from the estimated matrices, whereas the proposed ALiNGAM directly produces direction matrix.

- The total number of estimated directed edges in LiNGAM and DLiNGAM methods is \( \binom{n}{2} \), where \( n \) is the number of nodes. In the case of ALiNGAM, it only estimates the probable causal directions with less than \( \binom{n}{2} \) number of directed edges.

- Methods LiNGAM and DLiNGAM estimated lesser number of correct directions and produced more extra directions, whereas ALiNGAM produced the most number of correct directions and a few extra sets of directions.

- The proposed method (ALiNGAM) did not estimate wrong directions, but in the case of LiNGAM and DLiNGAM, they produced wrong directions as well as missing directions.

In all the above comparisons, the DLiNGAM method produced poor results as compared to any of the two other methods. The results produced by the original LiNGAM are very comparable to those of the ALiNGAM method. Notice that any set of extra directions produced in the model are counted as false positives. For a mixture model of non-Gaussian datasets the proposed method (ALiNGAM) is performed better than both LiNGAM and DLiNGAM methods.

In computational complexity the algorithmic complexity of ALiNGAM is discussed in comparison with both LiNGAM and DLiNGAM methods. The Table 3.8 provides the run time complexity of all the three methods, evaluated using the simulated datasets of 3.5.1.1,3.5.1.2,3.5.1.3,3.5.1.4. Although all the three methods are of same class, but the estimated results produced in final stage are very different.

Table 3.8 provides the runtime complexity for LiNGAM, DLiNGAM and the proposed ALiNGAM methods for simulated datasets. It also provide the number of nodes used for directions estimation in different datasets and the time taken for computation. From the results shown in Table 3.8, it is evident that the proposed method ALiNGAM is highly efficient and computationally faster than both the methods LiNGAM and DLiNGAM. While the method LiNGAM has a better time complexity than DLiNGAM method, but definitely lags behind the proposed ALiNGAM method. The method DLiNGAM has a higher time complexity than both the other methods in comparison.
3.5 Experiments, Results and Comparison Test

Table 3.8 Runtime complexity of LiNGAM, DLiNGAM and ALiNGAM for estimating the causal directions in Simulated Datasets

<table>
<thead>
<tr>
<th>Methods</th>
<th>Time taken in direction estimation in seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dataset-1</td>
</tr>
<tr>
<td></td>
<td>Nodes=5</td>
</tr>
<tr>
<td>LiNGAM</td>
<td>0.188 sec.</td>
</tr>
<tr>
<td>DLiNGAM</td>
<td>3.266 sec.</td>
</tr>
<tr>
<td>ALiNGAM</td>
<td>0.012 sec.</td>
</tr>
</tbody>
</table>

Fig. 3.6 shows the performance of ALiNGAM in comparison to the LiNGAM and DLiNGAM methods, on the estimation of the correct number of directed edges in simulated datasets.

![Performance of Methods based on Correctly Estimated Directions](image)

Fig. 3.6 Performance of LiNGAM, DLiNGAM and ALiNGAM based on the number of correct directions estimated in simulated datasets

Fig. 3.6 provides the performance of ALiNGAM in comparison with LiNGAM and DLiNGAM for the estimation of correct directions in the simulated datasets. It is evident that, ALiNGAM has a high performance score when estimating correct directions, whereas LiNGAM is comparable in this scenario, but produced poor results. An error plot is provided in Fig. 3.7, to show the errors in estimation for all the three methods using simulated datasets.
Fig. 3.7 shows the errors produced by LiNGAM, DLiNGAM and ALiNGAM in direction estimations using datasets 3.5.1.1, 3.5.1.2, 3.5.1.3 and 3.5.1.4. The error percentage is calculated using the number of missing and wrong directions estimated by different methods and the number of edges assumed in the simulation models (as shown in Table 3.5). Noticeably, the proposed ALiNGAM produced no missing or wrong directions in the estimation, while both LiNGAM and DLiNGAM produced many wrong and missing directed edges in simulated models. All the above synthetic, real world data tests and comparison tests validate the performance of the proposed ALiNGAM model.

### 3.6 Discussion and Conclusion

The ALiNGAM or Altered-LiNGAM is a proposed modified method using the basic structure of the LiNGAM model, but the analytic and working principles are completely different from the original method. The ALiNGAM uses the least square method to estimate the parameters of the model and successfully reduces the noises in the system using the probable causal direction criteria. While using a linear bivariate model, the structure of ALiNGAM is fully capable of analyzing nonlinear and noisy models. The Probable Causal Directions work as specific condition for analyzing nonlinear models.
The performance of ALiNGAM is duly verified for synthetic and real world cases as provided in Sections 3.5.1 and 3.5.2. The results show that the system can estimate all possible causal directions in the model and the directed acyclic condition of LiNGAM method is always preserved in all scenarios. The estimated extra directions do not change the primary assumptions of the system. In the comparison test, the method ALiNGAM completely outperformed the original LiNGAM method and the DLiNGAM method which is a modified version of LiNGAM method. Most of the time, the proposed method ALiNGAM produced good results in estimating causal directions and produced very few extra sets of directions, in comparison to both LiNGAM and DLiNGAM methods (Section 3.5.3) where both produced a large number of missing and wrong directions. In both Figs. 3.6 and 3.7, the performance of ALiNGAM was shown, for the estimation of correct directions and error rates in comparison with LiNGAM and DLiNGAM. The proposed method has a high accuracy rate in estimating correct directions and does not produce any missing or wrong edges in simulated studies. The proposed bivariate model has a lower algorithmic complexity which makes it efficient while computing over the larger datasets. The results in Table 3.8 proves the claim that the proposed ALiNGAM has a very low computational complexity. Using the graphical structure construction technique of node arrangement (from most independent nodes at the top to the most dependent nodes towards the bottom), ALiNGAM can be used to provide causal levels using their kinship relations. The method is capable of estimating causal directions in large datasets and the randomized process of estimation also ensures that the produced causal directions do not change with the change in the input order of feature set. While the primary assumption of the system does not favour cyclic analysis, the method retains this by using probable causal direction condition in the estimation process.
Chapter 4

A Multivariate Additive Noise Model for Complete Causal Discovery

4.1 Introduction

The dependency on different factors requires studying the causal reasoning in real world processes. The simplest structures defined by Directed Acyclic Graphs (DAGs) are primarily analyzed using Structural Equation Models (SEMs) (Bollen, 1989). Later a more complex system Bayesian Network (BN) (Pearl, 2000; Spirtes et al., 1993) was designed to provide structural construction conditions using probability criteria for dependent features. Bayesian analysis only provides the functionality to construct DAGs using Bayesian low, graph d-separation, Markov Equivalent Classes, v-structures, Markov Blanket and Markov Chain Monte Carlo (MCMC) simulation. These assumptions lead to a model where nodal relations are defined only using the directed edges and weights of connection strengths to show the amount of information passed on. By definition, the works carried out on causal analysis only establish the model in the form of directed acyclic graph with weighted values on the edge to find the causal model. But none of these inform the quality of causal inference.

Most causal models such as Geiger and Heckerman (1994); Pearl (2000); Spirtes et al. (1993) use Gaussian assumption of the data to provide causal reasoning using Markov Equivalent Classes. Non-Gaussian properties of data can solve the problem for better approximation over real world cases, and so the model of Linear Non-Gaussian Acyclic Model (LiNGAM) was proposed by Shimizu et al. (2005) for causal discovery. The LiNGAM works with Independent Component Analysis (ICA) to estimate causal ordering by measuring the effects of different components independently. If it is known how one feature affects the other, then
the causal direction can be found. The successive arrangement of causal directions, starting from the most independent node to the most dependent node, provides causal ordering. But in LiNGAM, they first estimate the order and then from the order, they find the causal directions, which is an incorrect process. In addition, the later developments over LiNGAM like Shimizu et al. (2006, 2011) also get it wrong.

The nonlinear causal discovery using the additive noise model by Hoyer et al. (2009) estimates causal directions in two variables using nonlinear functional over joint probability densities of independent variable, added with arbitrary noise. The bivariate model is capable of finding directions by comparing two variables at a time. The problem with the bivariate model is in the accuracy of estimation, when a node or feature is dependent on multiple features, that is if multiple nodes have causal directions towards a single node then the model estimates wrong directions. Causal discovery for the continuous case was given by Peters et al. (2014) using the additive noise model. This model relies on the independence of residuals while observed nodes are regressed on each other. The dependency is measured using the Hilbert-Schmidt independence criterion by taking minus log over the functions of the feature set. Although the method claims to detect all possible edges in the model, the results show that method fails mostly for multivariate data with large node sets.

All the proposed papers on bivariate analysis fail to hold on to the graph d-separation condition. The d-separation uses triplet node sets (a set of 3 nodes at least) to observe the flow of information between parent and child nodes by blocking connected paths. Utilizing a bivariate model d-separation criteria cannot be verified, which primarily defines the causal model. So, an effective and easier method for multivariate analysis is proposed in this chapter, to find the complete causal structure using a multivariate additive noise model. It shows that the proposed method is fully capable of constructing multi-nodal causal structures for complete causal analysis. The insufficiency of the bivariate model which leads to the development of MANM is discussed, to provide the necessity of the proposition.

The conditional independence for causal inference is not one of the best tools for the estimation of independent features in the observational set. While the findings of causal inference are about qualitative analysis, the use of conditional independence/dependence merely provides information on quantitative measures. As a solution to this, the causal independence is introduced which better is suited for causal inference and analysis. The major difference between conditional independence and causal independence is discussed with a mathematical explanation to validate the claim.

Until now, the proposed methods emphasize the computation of directed edges from connection strength values. The connection strength is a quantification of the information
that flows from a parent node to child node. On the other hand, the causal inference is all about retrieving the influence of the information which causes it. Also, the issue of finding the goodness of causal inference has not been addressed or pointed out yet. Only finding the values for connection strengths which explicitly contribute towards the measurement for the amount of information passed on, does not conclude anything regarding the quality of inference. In this chapter for the first time, the new concept of the causal influence is introduced, which provides the measure for the goodness of causal inference. The value of causal influence is crucial to detect the causal direction based on its influence on the child nodes. The maximized Causal Influence Factor (CIF) values are used for the successful discovery of causal directions in the feature set when the substructures are d-separable v-structures.

While most of the proposed methods do not scope for the analysis of mixture models of Gaussian and non-Gaussian distributions, the proposed MANM can handle these datasets with ease. To show the capability of MANM method, the complexity of the mixture models was increased by making them both correlated, uncorrelated and mixed cases of both in simulation tests. The proposed model (MANM) is compared with Independent Component Analysis based Linear Non-Gaussian Acyclic Model (ICA-LiNGAM) (Shimizu et al., 2006), Greedy DAG Search (GDS) (Chickering, 2002) and Regression with Sub-sequent Independent Test (RESIT) (Peters et al., 2014) over simulated models and the test results are provided in Section 3.5.3. The choice of these three methods is due to their major contributions and wide usability in causal inference studies. Also, these all provide maximum inference on any dataset of interest and can handle large node sets.

The main goal of causal analysis is to find the directionally connected acyclic nodal structures which can be used for complete causal explanation. While constructing such models, many problems arise, due to the wrongly estimated directions which result in missing directions, and mostly due to overestimated extra sets of directions. These issues have been studied in comparison with ICA-LiNGAM, GDS, and RESIT, to find the effectiveness and accuracy of MANM while constructing causal models from the estimated causal directions. Results for model under-fitting, reverse-fitting and over-fitting are calculated using mathematical formulas and graphs are provided to show the performances of MANM in comparison with other methods.

Starting the discussion by defining the multivariate additive noise model in Section 4.2, and the identifiability of the model towards complete causal estimations is discussed in Section 4.3. Section 4.4 provides the tools for estimation using the proposed method and experimental results for simulated causal structures, and real world tests are provided.
in Section 4.5. The performance of the proposed method is verified with ICA-LiNGAM, GDS and RESIT, for their model constructibility and accuracy in Section 4.5.4. A concise discussion of the work and future perspective are given in Section 4.6.

### 4.2 Multivariate Additive Noise Model: MANM

The proposed multivariate additive noise model (MANM) depends on the imposed system assumptions to produce a complete causal model in the form of Directed Acyclic Graphs (DAGs). The assumptions are as follows for MANM: Consider the multi-nodal structure $G$ as a DAG with $n$ number of nodes represented by $\{X_i\}$ where $(i = 1, \ldots, n)$ and each $X_i$ is a matrix of $(m \times 1)$. All the nodes in $G$ are arranged in order from the most causal independent node at the top of the graph to the least causal independent ones towards the bottom. So, the edge directions in the DAG are from the top towards the bottom, and none of the nodes observed later than the earlier ones have directed edge towards any earlier node. The above assumptions produces a causal model where the graph starts from the parent node and goes down to descendant nodes while following a **top-bottom** graphical construction.

**Causal Independence:** Any causal direction $\{x_i \rightarrow x_j\}$ can be described using the causal order $O(x_i) > O(x_j)$. Consider a child node $x_h$ and its parent node set $Pa(x_h) = \{x_i, x_j, x_k, x_l\}$, where parents are of the following order $\{O(i) > O(j) > O(k) > O(l)\}$. Here $x_h$ can only be influenced by the parent set $Pa(x_h)$ through the information transferred from the parent node. The direction $\{Pa(x_h) \rightarrow x_h\}$ causes $Pa(x_h)$ to become causally independent of $x_h$. Every node in $Pa(x_h)$ has directed edges towards $x_h$, but the converse is not true. In the parent set node $x_i$ is most causal independent and $x_l$ is least causal independent (as from the order set).

While observing datasets, the variables are represented as nodes in the causal model with multivariate relations. Assume that in a multivariate causal model a dependent node is at least connected with two conditionally independent nodes like a $v$-structure and consider the case of $\{x_i \rightarrow x_k \leftarrow x_j\}$. The causal model for the parent nodes $\{x_i, x_j\}$ generating $x_k$ can be given as a linear combination of functions of parent nodes to the multiple of their respective connection strengths and added external independent noise as follows:

$$x_k = f(Pa(x_k).S_{tk}, n_k) = f(x_i.S_{ik}) + f(x_j.S_{jk}) + n_k(i \neq j \neq k), \quad (4.1)$$

where $S_{ik}$ and $S_{jk}$ are connection strengths for directions $\{x_i \rightarrow x_k\}$ and $\{x_j \rightarrow x_k\}$ respectively and $n_k$ is the additive noise for $x_k$. All the noises added in MANM are independent of each
other and are continuous random variables of both Gaussian and non-Gaussian distributions. For the multivariate linear model eq. 4.1 can be written as:

\[ x_k = S_{ik}x_i + S_{jk}x_j + n_k (i \neq j \neq k), \]

which provides the scope for the study of LiNGAM methods by Hyvärinen and Smith (2013); Shimizu et al. (2006, 2011) as special cases where noises are only non-Gaussian. In equation 4.2, if any of the connection strength values from \( \{ S_{ik}, S_{jk} \} \) are zero then only one direction will be found for the non-zero connection strength producing a bivariate model. Therefore, the models of Hoyer et al. (2009) and Peters et al. (2011, 2014) can be studied under special cases of MANM. If both the connection strength values are zero then node \( x_k \) will be observed as in the form of pure independent additive noise or as an exogenous variable in the causal model. However, the aim is to minimize the noise factors in eq. 4.1 such that the child nodes can be observed as the maximized causal influence of parent nodes. In the next Section, the identifiability of the proposed model is discussed.

### 4.3 Causal Identification using MANM

It has to make sure that MANM can discover all the causal directions in a data set and then whether or not the model can hold the acyclic condition while constructing the causal models. The conditional probabilities is checked first using graph d-separation, Markov Condition, Causal Markov Condition and Markov equivalent classes.

**Causal Independence vs. Conditional Independence:** The primary difference between causal independence and conditional independence is that where the first one is significant for qualitative analysis, the latter one only provides the quantitative information. This can be understood in this simple example: consider the node set \{x,y,z\}, where the causal independence is defined from the causal ordering \( O(x) > O(y) > O(z) \). Now in case of bivariate analysis the conditional probabilities for \{x,y\} can be given as,

\[ (x \perp \! \! \! \perp y) = (y \perp \! \! \! \! \perp x) \]
\[ P(x|y) = P(y|x), \]

\[ (4.3) \]
which provides quantitative information. But in the case of causal inference the relation is as following:

\[
P(x|y) \Rightarrow (y \rightarrow x) \Rightarrow y \text{ is causal independent of } x, \\
P(y|x) \Rightarrow (x \rightarrow y) \Rightarrow x \text{ is causal independent of } y.
\] (4.4)

This is a qualitative information and useful for causal inference. So, conditional independence does not necessarily show the causal independence, but the converse is always true.

The graph **d-separable** condition (Pearl, 1998) provides the necessary criteria to block the flow of information while observing a causal dependent node as and in the form of other causal independent nodes in a graph. Also, the d-separable condition provides the basic infrastructure for causal inference in the smaller sub-structural level. Considering **d-separable** for the case \( \{X \leftrightarrow Z \leftrightarrow Y\} \), all the followings are equivalent:

1. \( X \rightarrow Z \rightarrow Y, X \leftarrow Z \rightarrow Y \) is **d-separated** by \( Z \) when \( Z \) is not observed.
2. \( X \rightarrow Z \leftarrow Y \) is **d-separated** by \( Z \) when \( Z \) or its other descendant nodes is observed.

In point (2), by conditioning on node \( Z \) the nodes \( X \) & \( Y \) become mutually dependent while previously both were mutually independent. Only this type of causal directions or \( V \)-structures is suitable for \( d \)-separation and can be used as sub-structures for causal discovery.

In case of bivariate analysis, the criteria of \( d \)-separable are missing and not implemented for structural learning. It requires at least 3 sets of nodes forming a \( V \)-structure for this kind of structural learning. So, bivariate models fail to provide the correct causal directions in multivariate causal structures.

Consider the DAG \( G \) with nodes \( \{x_i, x_j, x_k, x_l\} \) with edge directions \( \{x_j \leftarrow x_i \rightarrow x_k\} \) and \( \{x_j \rightarrow x_l \leftarrow x_k\} \). Using MANM, the complete causal structure is needed to be discovered for this graph \( G \). Using \( d \)-separation, graph \( G \) can be represented by eq. 4.1 as,

\[
(i)x_j = f(x_i.S_{ij}) + f(x_k.S_{kj} = 0) + n_j = f(x_i.S_{ij}) + n_j, \\
(ii)x_k = f(x_i.S_{ik}) + f(n_j.S_{jk} = 0) + n_k = f(x_i.S_{ik}) + n_k, \\
(iii)x_l = f(x_j.S_{jl}) + f(x_k.S_{kl}) + n_l.
\] (4.5)

In eq. 4.5, for \( (i) \) directions are \( \{x_i \rightarrow x_j \leftarrow x_k\} \) and for \( (ii) \) directions are \( \{x_i \rightarrow x_k \leftarrow x_j\} \). As there is no direction from \( \{x_k \rightarrow x_j\} \) and \( \{x_j \rightarrow x_k\} \), the connection strengths \( \{S_{jk}, S_{kj}\} \) become zero. Notice that using the \( d \)-separable condition any three nodes can be written as MANM as given in eq. 4.1.
Applying the Markov Condition (MC) on the set of nodes \( \{x_i, x_j, x_k, x_l\} \) in \( G \), the joint distribution factorization can be done in such a way that the conditional probabilities are ordered from the earlier to the later ones. In this case for causal relations, the Causal Markov Condition (CMC) says that given the condition on parents of \( x_i \) i.e. \( \text{Pa}(x_i) \), the variable \( x_i \) is independent of all the other variables which are not direct causes of \( x_i \) in the graph \( G \) i.e. if \( X \rightarrow Y \) then \( \text{Pr}(X|Y, \text{Pa}(X)) = \text{Pr}(X|\text{Pa}(X)) \). Using the MC and CMC, the joint probabilities can be studied for the system given in eq. 4.5. Eq. 4.5 can be represented as follows using MC and CMC:

\[
\begin{align*}
\text{Pr}(\text{Pa}(x_j), x_j) &= \text{Pr}(x_i).\text{Pr}(x_j|x_i) \\
\text{Pr}(\text{Pa}(x_k), x_k) &= \text{Pr}(x_i).\text{Pr}(x_k|x_i) \\
\text{Pr}(\text{Pa}(x_l), x_l) &= \text{Pr}(x_j).\text{Pr}(x_k|x_j,x_k).
\end{align*}
\] (4.6)

Estimation over the eq. 4.6 needs proposal distribution to approximate over the probability density functions (PDFs). To overcome this issue, the linear model can be used as given in eq. 4.2 rather using eq. 4.1. But this will not be an exact approximation, considering the facts that a linear model is tried for approximating a nonlinear model. In this case for probability density function, approximations are required for mean=\( \mu \) and variance=\( \sigma^2 \) of the distribution, and this can be done by using a projection of the nonlinear PDF as in eq. 4.1 on to a linear function as in eq. 4.2. Note that the mean and standard deviation remains the same while using the projection.

Given a set of data \( \{X, Y, Z\} \) each of size \( (m \times 1) \), the causal relations in these are to be found as shown in Fig. 4.1a. Using the graph d-separable as explained above three Markov Equivalent models can be constructed as shown in Figs. 4.1b,4.1c, and 4.1d. Markov Equivalent models are the graphs with the same vertices and edges irrespective of their edge directions.

The aim is to find the causal directions i.e. \( \{Y \rightarrow X \leftarrow Z\} \), \( \{X \rightarrow Y \leftarrow Z\} \) and \( \{X \rightarrow Z \leftarrow Y\} \) whether or not the method is able to detect the additional/extra directions \( \{Y \leftrightarrow Z\} \), \( \{X \leftrightarrow Z\} \), and \( \{X \leftrightarrow Y\} \) respectively as shown in Figs. 4.1b,4.1c and 4.1d. Notice that these extra sets of directions are detected due to the condition of d-separation upon \( v \)-structures. These three Markov equivalent models for detection of causal directions can be represented using eq. 4.2 as:

\[
\begin{align*}
X &= S_{YX}Y + S_{ZX}Z + n_X, \quad Y = S_{XY}X + S_{ZY}Z + n_Y, \quad \& \\
Z &= S_{XZ}X + S_{YZ}Y + n_Z.
\end{align*}
\] (4.7)
4.4 Estimations over MANM

Fig. 4.1 All possible causal relations in \{X, Y, Z\} as shown in (a) and three different Markov equivalent models shown by (b), (c) and (d).

From the comparison between the three causal models using eq. 4.7, all the possible directions or causal influences can be successfully estimated in the substructures. Notice that the additional directions produced in the model do not change the acyclicity of the model. The estimations for causal directions using parameter acceptance criteria and causal influence factor values are discussed in the next section.

4.4 Estimations over MANM

To estimate causal directions for the models given in Figure 1.1, the parameter acceptance criteria is needed. Consider the model \{X \rightarrow Y \leftarrow Z\} from Fig. 4.1c. This can be represented using equation 4.1 by \( Y = f(X,S_{XY}) + f(Z.S_{ZY}) + n_Y \) and as a projection of eq. 4.1 by \( Y = S_{XY}.X + S_{ZY}.Z + n_Y \) using eq. 4.2. The projection will be the exact estimation of the PDF only when the following two parameter acceptance criteria are true,

\[
(I) F_{XY}^1 = \frac{S_{XY}.E(X)}{E(Y)} > \frac{E(n_Y)}{E(Y)} \quad \text{and} \quad (II) F_{ZY}^2 = \frac{S_{ZY}.E(Z)}{E(Y)} > \frac{E(n_Y)}{E(Y)} \tag{4.8}
\]

In eq. 4.8, the values \( F^1, F^2 \) are the Causal Influence Factors (CIFs), where \( F_{XY}^1 \) show the direction \( \{X \rightarrow Y\} \) and \( F_{ZY}^2 \) for \( \{Z \rightarrow Y\} \). The values \( E(X), E(Y), E(Z), E(n_Y) \) are expectations of \( X, Y, Z \) and \( n_Y \) respectively. Any successive update of causal direction in these cases will only be implemented if new values of \( (F_1, F_2) \) are greater than the old values. As the noises are minimized using acceptance criteria, at the same time the CIF values are also maximized. So, the process only estimates those directions which are independent of noise factors and influenced by parent nodes only.
The causal directions in \( \{X, Y, Z\} \) can be estimated using two methods.

1. Use the maximum likelihood estimation (MLE) method to find parameters when the system is represented by eq. 4.7. After parameter estimation, using acceptance criteria given in eq. 4.8 the causal influence factors can be found.

2. For linear models as shown in eq. 4.7 and for acceptance criterion given in eq. 4.8, the convex optimization method can be used to obtain the CIF values for estimation of causal directions.

Only choose the causal directions which have the highest CIF values i.e. high \( F_1 \) and \( F_2 \) values. For the set \( \{X, Y, Z\} \) there will be six causal influence factors, and the matrix CIF can be given as

\[
CIF = \begin{bmatrix}
F_{yx}^1 & F_{zx}^2 \\
F_{xy}^1 & F_{zy}^2 \\
F_{xz}^1 & F_{yz}^2
\end{bmatrix}.
\] (4.9)

For an input dataset of size \((m \times n)\), construct matrix \( IF \) of size \((n \times n)\) to store the updated CIF values. Any value \( IF[i, j] \) in the \( IF \) matrix shows the CIF value for \( \{i \rightarrow j\} \) and depending on these the causal models are constructed following the graph top-bottom construction method. Notice that for causal model construction always select those directions which have CIF values greater than 0.75, such that the influence will be significant with a less probability of wrong direction detection of 0.25. Algorithm 2 is provided below for implementation of MANM using all the criteria explained above: In the next section, the experimental results are provided using MANM.

**Algorithm 2 Multivariate Additive Noise Model**

1. **Input**: Matrix of size \((m \times n)\)
2. Generate \( Comb \) of size \((3 \times \binom{n}{2})\) the combination of all three tuples from \(n\) variables.
3. Construct matrix \( IF \) of size \((n \times n)\)
4. for \( i = 1 \) to \(^3\binom{n}{3}\) do
5. Solve for each triplets of \( Comb[i, \cdot] \) using mle or optimization for eq. 4.7 for the constraints given in eq. 4.8.
6. Find the matrix \( CIF \).
7. Find \( \{M_1, M_2, M_3\} \) from \( M_1 = \max(F_{yx}^1, F_{yx}^1) \), \( M_2 = \max(F_{zx}^2, F_{xz}^1) \), \( M_3 = \max(F_{zy}^2, F_{yz}^2) \).
8. Update \( IF \) depending on \( \{M_1, M_2, M_3\} \).
9. end for
10. **return** \( IF \)
4.5 Experiments

The significance of MANM for generating highly accurate causal directions is tested in following experiments. For experimental purposes, different types and sizes of simulated and real world datasets are used. Simulated models are generated with both linear and nonlinear causal relations using mixture models of Gaussian and non-Gaussian distribution types. For real world test, the datasets are selected from the different fields of study.

4.5.1 Simulated Model Results

The test using simulated models contains 5 causal models. The first one is a 5-node linear correlated non-Gaussian model and the second one is a 9-node nonlinear causal model where noises are sampled from Gaussian and non-Gaussian distributions. A mixture model test set of 6 and 8 nodes are generated with linear and nonlinear relations, where samples are drawn from the mixture of Gaussian and non-Gaussian distributions. The last causal model is a 13-node structure generated using both linear, nonlinear functionals with the mixture of both Gaussian and non-Gaussian distributions. Dataset 5 has both correlated and uncorrelated relations in the node sets, while all the other sets are generated with specific correlations (i.e. either +, - or 0 correlations). For data generation, the used non-Gaussian distributions are (1) Uniform distribution- $U(min, max)$, (2) Log Normal distribution- $LN(\mu, \sigma)$; where $\mu$ and $\sigma$ represents the mean and standard deviations respectively, (3) Exponential distribution- $Exp(r)$; where $r$ is the rate of generation determines the density function, (4) Double Exponential distribution- $DExp(r)$ and (5) Student’s $t$-distribution- $T(df)$; where $df$ is the degrees of freedom for $t$-distribution, and (6) The normal distribution- $N(\mu, \sigma)$.

Simulated models with the indicated number of nodes sets are generated using specific parent structures, such that the causal relations can be verified after the estimation of causal directions. Without having any knowledge on generated datasets it will be difficult to validate the results for any method. When in use, the constructed causal relation does not affect the process of estimation, however remains effective while needed to be compared with the original causal model.

The choice for correlated and uncorrelated datasets is to provide the experimental proof that

$$(\pm, 0)\text{Correlation} \not\Rightarrow \text{Causality}.$$

Results are provided for the following datasets in correlated and uncorrelated conditions. All the sets generated for this purpose are selectively small, as generating nonlinear models
with many node sets and with specific correlation structure is a fairly difficult process. For an estimation of causal directions, each of the simulation tests was performed 10 times to validate the outcome of the method.

### 4.5.1.1 Dataset 1 (Linear correlated non-Gaussian Model)

The generated set is a 5-node linear correlated causal model of linear non-Gaussian distributions. The nodes are \{A, B, E, C, D\} with causal directions \{B \leftarrow A \rightarrow C\}, \{B \rightarrow D \leftarrow C\} and \{A \rightarrow E \leftarrow C\}. The linear non-Gaussian model is generated through the following process:

\[ A = \text{Exp}(1.7) - 1, \quad B = 2A - n_B \quad \text{with} \quad n_B = U(-1, 2), \quad C = 3.5A - n_C \quad \text{with} \quad n_C = N(0, 1.5), \]

\[ D = -B - 0.2C + n_D \quad \text{with} \quad n_D = T(3) \quad \text{and} \quad E = -A - 0.5C + n_E \quad \text{with} \quad n_E = T(5). \]

For dataset 1, both Gaussian and non-Gaussian noises are added into the model. For this, 500 data points were samples for each node and noise set and algorithm 2 is used to find CIF values for model constructions. Table 4.1 below shows the results for \textit{IF} matrix showing the existence of direction in any two set of nodes from rows to column variables.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.963</td>
<td>0.958</td>
<td>0.710</td>
<td>0.966</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.872</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0.596</td>
<td>0</td>
<td>0.931</td>
<td>0.945</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.443</td>
<td>0</td>
</tr>
</tbody>
</table>

Using the maximized CIF values given in Table 4.1, the following observations are concluded for construction of causal model:

- Node \(A\) is the most causal independent node. So, \(A\) can be represented as a parent node or starting node at the top of the causal model. Also, node \(A\) affects all other nodes, which needs a better arrangement for causal construction.

- Node \(B\) and \(C\) are the second most causal independent nodes, while \(D\) and \(E\) are the most causal dependent ones in the model.

- The directions \(\{D \rightarrow E\}\) and \(\{E \rightarrow B\}\) are rejected as their CIF scores are very low in comparison to other nodes. For \(\{C \rightarrow B\}\) the chances are 50%, which may exist but not very strong.
Considering above points and our original generation process, Fig. 4.2 provides a comparison of the original and simulated causal models.

![Original Causal Model](image1)
![Simulated Causal Model](image2)

Fig. 4.2 Causal Model for Dataset 1

To generate a very similar model to the originally assumed causal model, the direction \( \{C \rightarrow B\} \) and \( \{A \rightarrow D\} \) are not considered. The parents in the causal model have an influence over the child nodes and even the correlations in variables are very high due to linearity in causal relations.

**4.5.1.2 Dataset 2 (Non-Linear Uncorrelated non-Gaussian Model)**

The nonlinear uncorrelated non-Gaussian model is a 9-node causal model with node set \( \{A,C,D,E,B,F,G,I,H\} \) and the causal directions are \( \{C \leftarrow A \rightarrow D\}, \{C \rightarrow E\}, \{C \rightarrow B \leftarrow D\}, \{D \rightarrow F\}, \{E \rightarrow G \leftarrow B\}, \{B \rightarrow I \leftarrow F\} \) and \( \{G \rightarrow H \leftarrow I\} \). The nodes are generated using nonlinear functions as,

\[
A = \text{Exp}(1.3), \ C = \cosh(A) + n_C \text{ with } n_C = T(2),
D = \tanh(A)/2 + n_D \text{ with } n_D = U(-2, 2), \ E = C/\sin(C) + n_E \text{ with } n_E = T(3), \ B = \sin(C) - \cos(D) - n_B \text{ with } n_B = N(0, 1.5) + U(-3, 3), \ F = D^2/\cos(D) + n_F \text{ with } n_F = U(-1, 1) + N(-1.5, 2), \ G = E + 1/\sin(B) + n_G \text{ with } n_G = [\text{Exp}(3) + U(-1, 1)], \ I = \sin(B) - \sin(F) - n_I \text{ with } n_I = U(-3, 3) \text{ and } H = G/\cos(G) + I/\tan(I) + n_H \text{ with } n_H = U(-2, 3).
\]

For model generation, 1000 samples were taken for nodes and noises. Also, some of the additive noises in the model are linear combinations of Gaussian and non-Gaussian distributions as shown above. The causal directions estimated using MANM from maximized CIF values are provided in Table 4.2.
Table 4.2 Maximized CIF values for Dataset 2.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>A</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>B</th>
<th>F</th>
<th>G</th>
<th>I</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.98</td>
<td>0.99</td>
<td>0.81</td>
<td>0.71</td>
<td>0.56</td>
<td>0.73</td>
<td>0.43</td>
<td>0.64</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
<td>0.98</td>
<td>0.24</td>
<td>0.66</td>
<td>0</td>
<td>0.15</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0.98</td>
<td>0</td>
<td>0.49</td>
<td>0.96</td>
<td>0.88</td>
<td>0.37</td>
<td>0.31</td>
<td>0.70</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.32</td>
<td>0.84</td>
<td>0.52</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0.37</td>
<td>0</td>
<td>0</td>
<td>0.95</td>
<td>0.79</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.83</td>
<td>0</td>
<td>0.12</td>
<td>0.89</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In Table 4.2 estimated results are quite good, considering the fact that features are totally uncorrelated and functions are carefully selected to provide a real-world data similarity. Fig. 4.3 shows the constructed causal model with originally assumed model.

Simulated causal model for dataset 2 only produces two extra directions from \( \{D \rightarrow C\} \) and \( \{F \rightarrow B\} \) in comparison with the original model and all the other extra sets of directions are excluded while constructing the model. Owing to randomness in the data generation or function estimations and common parent relations, the extra sets of directions are observed in the simulated model. Notice that the addition of extra directions in the original causal model does not change the acyclicity of the simulated causal model.

In the next two datasets, the results for mixture model estimations are provided. Both symmetric and non-symmetric non-Gaussian mixtures are considered (Bach and Jordan,
2002), to construct causal models with nonlinear relations. Also, the correlations in the features were taken into account which affect the computation while using regression analysis (Hoyer et al., 2009; Peters et al., 2011, 2014).

### 4.5.1.3 Dataset 3 (Non-Linear Uncorrelated Symmetric Non-Gaussian Mixture Model)

Dataset 2 is a 6-node causal model with nonlinear relations and feature sets are generated using non-Gaussian distributions whereas noises are the symmetric mixture of non-Gaussian distributions. So, the resultant model becomes a nonlinear symmetric non-Gaussian mixture model. The node set as they ordered in the model is \{P, Q, S, R, T, V\} with causal directions \{Q ← P → R\}, \{Q → S ← P\}, \{Q → T ← R\} and \{T → V ← R\}. The nodes are derived using the following nonlinear processes:

- \(P = LN(0.37, 1.7)\)
- \(Q = P^2/(1-P) - n_Q\) where \(n_Q = [\text{Exp}(0.52), T(3)]\)
- \(R = \text{tanh}(P) - n_R\) where \(n_R = [T(5), U(-3, 3)]\)
- \(S = 3\sin(P) - \tan(Q) + n_S\) where \(n_S = [T(3), T(5)]\)
- \(T = \sin(Q) + \sin(R) - n_T\) where \(n_T = [\text{DExp}(0.83), T(3)]\)
- \(V = \tan(T) + R/\sinh(R) + n_V\) where \(n_V = [\text{Exp}(0.52), T(5)]\)

Here 600 data instances were sampled to generate each feature set with additive noises. Note that all the feature sets are carefully drawn to make them uncorrelated. Using the CIF values, the causal directions are successfully derived in the simulated model. Fig. 4.4 shows the simulated model with one extra direction \{R → S\} and the originally assumed causal model.

![Fig. 4.4 Causal Model for Dataset 3](image_url)

Notice that only those CIF values are considered which have at least a 75% chance of affecting other nodes. The arrangement of the simulated model is done by shorting all nodes from most causal independent to least causal independent ones and using a graphical tree arrangement technique from top-to-bottom.
4.5 Experiments

4.5.1.4 Dataset 4 (Non-Linear Correlated Non-Symmetric Non-Gaussian Mixture Model)

To test the efficiency and identifiability of the proposed method perhaps the best model is a nonlinear correlated non-symmetric mixture of non-Gaussian distributions. The generated causal structure is a 8-node DAG with node set \( \{ V_1, V_2, V_4, V_3, V_5, V_7, V_6, V_8 \} \) and with directed edges as \( \{ V_2 \leftarrow V_1 \rightarrow V_3 \}, \{ V_1 \rightarrow V_4 \}, \{ V_2 \rightarrow V_5 \leftarrow V_4 \}, \{ V_4 \rightarrow V_7 \leftarrow V_5 \}, \{ V_4 \rightarrow V_6 \leftarrow V_3 \} \) and \( \{ V_5 \rightarrow V_8 \leftarrow V_6 \} \). The DAG is generated through the following process,

- \( V_1 = DExp(0.83), \quad V_2 = 2tanh(V_1) - n_{V_2} \) where \( n_{V_2} = [T(3), U(-3, 3)] \),
- \( V_3 = 2*exp(V_1) - n_{V_3} \) with \( n_{V_3} = [LN(0.37, 1.7), T(5)] \),
- \( V_4 = (2 - V_1)/\sqrt{|V_4|} + n_{V_4} \) with \( n_{V_4} = [U(-3, 3), LN(0.37, 1.7)] \),
- \( V_5 = V_2^2 - V_4/(2cos(V_4)) - n_{V_5} \) where \( n_{V_5} = [T(5), U(-3, 3)] \),
- \( V_6 = V_3/cos(V_3) + 1/cos(V_4) - n_{V_6} \) with \( n_{V_6} = [Exp(0.52), T(5)] \),
- \( V_7 = tan(V_4) - cos(1/V_5) - n_{V_7} \) where \( n_{V_7} = [U(-3, 3), Exp(0.52)] \) and \( V_8 = sin(V_5) - (V_6/5 - 1) - n_{V_8} \) where \( n_{V_8} = [DExp(0.83), T(5)] \).

For the dataset 3, 800 samples are drawn to generate the above model and using algorithm 2 the IF matrix is estimated to construct the causal model. Due to similarity in data generating functions in \( V_6 \) and \( V_7 \), the direction \( \{ V_6 \rightarrow V_7 \} \) is produced.

The simulated causal model is given in Fig. 4.5 with the original DAG. Notice that the direction \( \{ V_6 \rightarrow V_7 \} \) does not change the acyclic nature of the simulated causal model.
4.5 Experiments

4.5.1.5 Dataset 5 (Correlated, Uncorrelated mixed datasets of both Gaussian, Non-Gaussian Distributions of Linear and Non-Linear Functions)

This data set is generated from a 13-node causal structure, where child nodes are both linear and nonlinear functions of a more complicated distribution sets with a mixture of both symmetric, non-symmetric Gaussian and non-Gaussian distributions. Also, the causal nodes have both correlated and uncorrelated relations among them, which makes it more complicated to trace the parent and child structures. The 13-node set and causal directions are \( \{D \leftarrow A \rightarrow E \leftarrow B \rightarrow F \leftarrow C \rightarrow G\}, \{D \rightarrow H \leftarrow E \rightarrow I\}, \{F \rightarrow J \leftarrow G\}, \{H \rightarrow K \leftarrow I \rightarrow L \leftarrow J\} \) and \( \{K \rightarrow M \leftarrow L\} \). The data set is generated through the following process:

- \( A = \text{Exp}(0.52) \)
- \( B = T(5) \)
- \( C = \text{LN}(0.15, 2.3) \)
- \( D = A + n_D \) with \( n_D = N(0.5, 1.5) \)
- \( E = \sin(A) + 3B/(5-B) + n_E \) with \( n_E = \{T(3), U(-4,2)\} \)
- \( F = \cos(B) - \cos(C) + n_F \) with \( n_F = T(3) \)
- \( G = C + n_G \) with \( n_G = N(-0.3, 2.3) \)
- \( H = \sin(D) + \tan(E) - n_H \) with \( n_H = \{\text{Exp}(0.52), \text{LN}(-0.37, 1.7)\} \)
- \( I = 3/E + n_I \) with \( n_I = T(5) \)
- \( J = F - G + n_J \) with \( n_J = \{\text{DEExp}(1.35), U(-3,2)\} \)
- \( K = H/\cos(H) + 2\sin(I) + n_K \) with \( n_K = T(3) \)
- \( L = 1/I - 2J - n_L \) with \( n_L = N(-1.3, 2.7) \) and \( M = (2 - K)/K + L + n + M \) with \( n_M = U(-4,3) \).

For this data set 1500 points are sampled and 10 simulations are performed with a randomly ordered input of node sets to conform the estimated causal directions. Fig. 4.6 shows the simulated model generated from MANM with the originally assumed causal model.

In Fig. 4.6, MANM estimated one missing node \( \{B \rightarrow F\} \) and estimated three extra sets of directions \( \{A \rightarrow B, C \rightarrow D\} \) and \( \{I \rightarrow K\} \). For model construction and representation, the direction \( \{C \rightarrow D\} \) is removed.
The method MANM can analyze wide sets of simulated data cases as discussed above. In the next section, the performance of MANM is tested for real world cases.

### 4.5.2 Real World Test Results

The ability of the proposed MANM was tested fully for different real world cases. A broad range of datasets from various fields of studies were used, which are available at UCI repository (Asuncion and Newman, 2007).

#### 4.5.2.1 Forest Fires Data

The first real world test set is the forest fires dataset from UCI repository (Asuncion and Newman, 2007), which contains measured weather conditions for different aspects and the areas affected by fire in the northeast region of Portugal. The data is helpful for the prediction of forest fires but the task is to find out which environmental factors affect forest fires by representing them in a causal structure. Only the major environmental features like (1) Temperature- $Temp$, (2) Relative Humidity- $RH$, (3) Wind Speed- $Wind$, (4) Rain Fall- $Rain$ and (5) Area Burnt- $Area$ were selected from the forest fire data set. Using the CIF values, the causal directions in forest fire data are estimated and generated causal model is shown in Fig. 4.7. For this all CIF value greater than 0.5 are considered to make sure that any feature responsible for forest fires must be observed.

![Fig. 4.7 Causal relations in natural factors for Forest Fires Data.](image)

From Fig. 4.7 it is observed that relative humidity has a lower tendency to affect forest fires while temperature and wind speed have a higher tendency to cause a forest fire. Furthermore, the wind controls the temperature, rain, and spread of fire in the forest.
4.5 Experiments

4.5.2.2 Abalone Data

The Abalone data (Asuncion and Newman, 2007) is sampled to prove the methodology to calculate the age of the fish from different measurements of the shell. The time-consuming process counts the number of rings in the shell of the fish and the addition of value 1.5 with the number of rings in order to calculate the age of the fish. The acquisition is that "shell measurements can predict the age of the fish or the number of rings in the shell" and to prove the above point the causal directions are to be estimated. The causal relations in variables Length- $L$, Diameter- $D$, Height- $H$, Whole Weight- $WW$, Shucked Weight- $SW$, Viscera Weight- $VW$, Shell Weight- $ShW$, and the number of Rings- $R$ are observed using MANM. The results are shown in Table 4.3.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>$L$</th>
<th>$D$</th>
<th>$H$</th>
<th>$WW$</th>
<th>$SW$</th>
<th>$VW$</th>
<th>$ShW$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>0</td>
<td>0.97</td>
<td>0.88</td>
<td>0.75</td>
<td>0.74</td>
<td>0.77</td>
<td>0.67</td>
<td>0</td>
</tr>
<tr>
<td>$D$</td>
<td>0</td>
<td>0</td>
<td>0.875</td>
<td>0.774</td>
<td>0.74</td>
<td>0.776</td>
<td>0.682</td>
<td>0</td>
</tr>
<tr>
<td>$H$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.721</td>
<td>0.833</td>
<td>0.753</td>
<td>0.613</td>
<td>0</td>
</tr>
<tr>
<td>$WW$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.972</td>
<td>0.953</td>
<td>0.82</td>
<td>0.881</td>
</tr>
<tr>
<td>$SW$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$VW$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.834</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$ShW$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.667</td>
<td>0.775</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$R$</td>
<td>0</td>
<td>0.025</td>
<td>0.260</td>
<td>0</td>
<td>0.251</td>
<td>0.217</td>
<td>0.399</td>
<td>0</td>
</tr>
</tbody>
</table>

From the maximized CIF values produced in Table 4.3, it can be concluded that the 'age of the fish can be predicted from the shell measurements in particular by measuring the diameter, height, Shucked Weight, Viscera Weight and Shell Weight of the fish shell'. The results show that $R$ affects $\{D, H, SW, VW, ShW\}$, so the values of $\{D, H, SW, VW, ShW\}$ are higher in aged fishes only or shells with the higher number of rings.

4.5.2.3 Boston Housing Data

The Boston Housing data (Asuncion and Newman, 2007) is a 14-variable continuous set with only one binary variable, featuring housing values in the suburbs of Boston. All the variables are used to check some of the general known facts usually believed to be true, such as 'the houses with better access to highways have a higher price range, industrial and high populated areas have high Nitric Oxide concentrations, residential areas with higher facilities have higher tax values'. The following variables are observed from the
Boston Housing data set, (1) ZN- residential land zones over 25000 sq.ft., (2) IND- non-retail business acres, (3) CR- Charles River (= 1 if tract bounds river; 0 otherwise), (4) NOX- nitric oxides concentration, (5) RM- average number of rooms per dwelling, (6) AGE- owner-occupied units built prior to 1940, (7) RAD- accessibility to radial highways, (8) TAX- full-value property-tax rate per $10,000, (9) PTR- pupil-teacher ratio by town and (10) LSP: % lower status of the population. Table 4.4 shows the causal relations in the considered features for the Boston Housing data.

Table 4.4 Observed CIF values for causal relations in selected variables of Boston Housing Data.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>ZN</th>
<th>IND</th>
<th>CR</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>RAD</th>
<th>TAX</th>
<th>PTR</th>
<th>LSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZN</td>
<td>0</td>
<td>0</td>
<td>0.47</td>
<td>0.81</td>
<td>0</td>
<td>0</td>
<td>0.82</td>
<td>0</td>
<td>0.96</td>
<td>0</td>
</tr>
<tr>
<td>IND</td>
<td>0.56</td>
<td>0</td>
<td>0.65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.92</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NOX</td>
<td>0.79</td>
<td>0.79</td>
<td>0.83</td>
<td>0</td>
<td>0.91</td>
<td>0.76</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RM</td>
<td>0.71</td>
<td>0.63</td>
<td>0.75</td>
<td>0.43</td>
<td>0</td>
<td>0.73</td>
<td>0.56</td>
<td>0.93</td>
<td>0</td>
<td>0.98</td>
</tr>
<tr>
<td>AGE</td>
<td>0.67</td>
<td>0.97</td>
<td>0.82</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RAD</td>
<td>0</td>
<td>0</td>
<td>0.46</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TAX</td>
<td>0.54</td>
<td>0.95</td>
<td>0.53</td>
<td>0.76</td>
<td>0</td>
<td>0.94</td>
<td>0.87</td>
<td>0</td>
<td>0</td>
<td>0.50</td>
</tr>
<tr>
<td>PTR</td>
<td>0</td>
<td>0.87</td>
<td>0.61</td>
<td>0.55</td>
<td>0.76</td>
<td>0.97</td>
<td>0.74</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LSP</td>
<td>1</td>
<td>0.98</td>
<td>0.70</td>
<td>0.96</td>
<td>0</td>
<td>0.59</td>
<td>0.85</td>
<td>0</td>
<td>0.77</td>
<td>0</td>
</tr>
</tbody>
</table>

Using the maximum CIF values greater than to 0.8 provided in Table 4.4, the residential land zone ZN affects the features NOX, RAD, PTR. All the regions such as ZN, IND, NOX, AGE, TAX, PTR and LSP have access to highways (RAD) and this is true in real world cases. From CIF values, it is clear that CR has not affected any of the features on the data set, whereas if any variable shows the relation with CR, it indicates that the river is bounded on that region (e.g. industrial areas have bounded river). Notice that industrial areas (IND) and high population zones (RM, AGE, PTR) have higher nitric oxide concentrations (NOX) or they show a positive relation with NOX. Also, industrial areas (IND) and high population zones (RM, AGE, PTR, LSP) have high TAX values for properties and even have higher valued facilities. So, the known assumptions are found to be correct with the estimated causal relations in the Boston housing data set.
4.5 Experiments

4.5.3 Comparison Test

To compare the performance of MANM with methods proposed previously, the methods ICA-LiNGAM, GDS and RESIT were selected. All the considered methods provide continuous data inference structure and provide causal inference and structural construction facilities.

Independent Component Analysis based Linear Non-Gaussian Acyclic Model (ICA-LiNGAM): The method ICA-LiNGAM was proposed by Shimizu et al. (2006) to estimate Structural Equation Models (SEMs) using the non-Gaussian, continuous datasets. ICA-LiNGAM produces triangular matrices for connection strength values and an estimated causal ordering. The detection of possible causal directions needs the manual user to find the directions from the triangular strength matrix. The operations include the independent component analysis based decomposition method which estimates the strength matrix from the given input matrices of observed variables. The method has a better performance for causal models with node sets less than or equal to 8.

Greedy DAG Search (GDS): This algorithm is based on greedy optimization and a modified version of Greedy Equivalence Search algorithm (GES). The GDS proposed by Chickering (2002) is a score-based method to find interventional Markov equivalence classes of a DAG, based on samples of observational data. The process maximizes the score by adding and removing directed edges in the observed sets until the edge operations do not improve the score further. The edge operations remain the same for all equivalent classes for a particular pair of nodes and both bi-directed and undirected edges have the same score. The method works in three stages: first, adding the edge in between the observed nodes; second, removing the edge from the nodes; and the third, reverting the edge in the nodes. In each step, the score is improved otherwise the method returns to the previous DAG.

Regression with sub-sequent Independent Test (RESIT): This method RESIT by Peters et al. (2014) is based on the additive noise model (ANM). RESIT provides causal estimations using regression technique with a sub-sequent independence test for each observed bivariates through the Hilbert-Schmidt Independence Criterion (HSIC). The method uses linear bivariate models for regression analysis, and the independence test uses covariance and cross covariance relations to find the independent variables in the observed sets. The resultant matrix of ones and zeros provides the results for causal directions, where 1 shows an estimated direction.
MANM is compared with these three methods using the simulated datasets given in Sections 4.5.1.1, 4.5.1.2, 4.5.1.3, 4.5.1.4 and 4.5.1.5. For each method, the tests are performed to estimate the causal structures, and from the estimated causal models the number of correct, wrong, extra, and missing sets of directions are estimated. If in a comparison test a method estimates the exact directions as considered for parent set, then it is counted as a correct direction. Whenever the methods produced a reversed direction in comparison to the original model then it is a wrong direction and wrong directions can result in a cyclic model. Whenever a method estimates a direction which primarily is not available in the original causal model and does not produce cycles in the estimated model, then these are counted as an extra set of directions. If an edge is primarily present in the assumed causal structure but found to be absent after estimation, then it is counted as a missing node.

Table 4.5 Comparison between ICA-LiNGAM, GDS, RESIT & MANM using Simulated Data Sets.

<table>
<thead>
<tr>
<th>Simulated Data</th>
<th>Directions</th>
<th>ICA-LiNGAM</th>
<th>GDS</th>
<th>RESIT</th>
<th>MANM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set 1</td>
<td>Correct</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Samples= 500</td>
<td>Wrong</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nodes=5, Edges=6</td>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Extra</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Data Set 2</td>
<td>Correct</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Samples= 1000</td>
<td>Wrong</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Nodes=9, Edges=12</td>
<td>Missing</td>
<td>0</td>
<td>8</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Extra</td>
<td>24</td>
<td>2</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>Data Set 3</td>
<td>Correct</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Samples= 600</td>
<td>Wrong</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Nodes=6, Edges=8</td>
<td>Missing</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Extra</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Data Set 4</td>
<td>Correct</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>Samples= 800</td>
<td>Wrong</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Nodes=8, Edges=11</td>
<td>Missing</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Extra</td>
<td>17</td>
<td>2</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>Data Set 5</td>
<td>Correct</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Samples= 1500</td>
<td>Wrong</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Nodes=13, Edges=16</td>
<td>Missing</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Extra</td>
<td>62</td>
<td>6</td>
<td>28</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4.5, provides the results for simulated datasets using the methods ICA-LiNGAM, GDS, RESIT, MANM and the number of correct, wrong, missing and extra sets of directions in the estimated causal models respectively. For the datasets used in the comparison test, the
number of nodes and edges which are primarily assumed for the original causal model were provided in the respective sets.

From Table 4.5, it is observed that MANM outperformed ICA-LiNGAM, GDS and RESIT in comparison tests. The methods ICA-LiNGAM, GDS and RESIT produced good results for linear models and gave a poor performance for nonlinear mixture models (correlated and uncorrelated). A performance test is given in the next section for all four methods used in the comparison test for the same number of simulated data cases.

In comparison, all three methods are very different in their working principles and solving techniques. They all have different time complexities for solving different feature sets and produce different results. In comparison to ICA-LiNGAM, GDS and RESIT which only estimate direction sets, connection strength values, and errors, the proposed method MANM estimates and provides all the required information necessary to construct the complete causal model such as directions in features, acyclic edge sets, connection strength values, causal influence values and error sets. So technically MANM gave a higher time complexity in comparison to all other methods. Table 3.4 below provides a comparative time complexity of ICA-LiNGAM, GDS, RESIT and MANM for the different node sets with the correct number of estimated directed edge in comparison with the assumed edges sets.

Table 4.6 Time complexity for methods ICA-LiNGAM, GDS, RESIT & MANM for solving different node sets and correctly estimated directed edge sets with respect to the assumed edge sets

<table>
<thead>
<tr>
<th>Number of Nodes &amp; Edges Assumed</th>
<th>Time taken &amp; Correctly Estimated Directed Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICA-LiNGAM</td>
</tr>
<tr>
<td>Nodes= 5 Edges= 6</td>
<td>0.30 sec.</td>
</tr>
<tr>
<td>Nodes= 6 Edges= 8</td>
<td>0.39 sec.</td>
</tr>
<tr>
<td>Nodes= 8 Edges= 11</td>
<td>1.47 sec.</td>
</tr>
<tr>
<td>Nodes= 9 Edges= 12</td>
<td>0.48 sec.</td>
</tr>
<tr>
<td>Nodes= 13 Edges= 16</td>
<td>0.43 sec.</td>
</tr>
</tbody>
</table>

Table 4.6 shows the time taken in seconds and the number of correct directed edges computed by the methods ICA-LiNGAM, GDS, RESIT and MANM to estimate different nodes sets with assumed number of directed edges. While both RESIT and MANM have
similar complexities to that of GDS and ICA-LiNGAM, MANM estimates the most number of correct edges of all in comparison. As MANM extracts most and all relevant information from datasets in comparison to any other method, it has a higher complexity. But the methods ICA-LiNGAM and GDS with a lower complexity do not provide better estimation results. MANM provides better estimation with very comparable time complexity in comparison to other methods.

4.5.4 Performance Test

For a performance test, the results from comparison test are used to find the performance of the methods ICA-LiNGAM, GDS, RESIT and MANM. The performance of the methods depends on how well the estimated causal directions can be used to construct the causal model. So, three new scoring methods are defined depending on the fitting of the estimated causal structures (in comparison to originally assumed model) to find the performance of the methods. The scores are generated using the number of correct, wrong, extra and missing directions estimated by the methods. The next section finds the estimation errors accrued in the process of causal constructions from the performance score.

**Model Under Fitting:** Whenever a method estimates less nodes than the original model, it becomes difficult to construct a connected directed acyclic graph. This type of problem may arise due to the number of missing nodes in the estimated model. This leads to a model under fitting problem when constructing a causal model from the estimated causal directions. Model under fitting has a negative (-) impact on the system and in extreme case for all missing nodes, the methods have a negative performance score. A mathematical representation of model under fitting is given as:

\[
\text{Model Under Fitting Score} = \frac{CR - MS}{EA},
\]

where \(CR, MS\) and \(EA\) show the number of correct directions, missing directions and edges assumed. Fig. 4.8 shows the performance of ICA-LiNGAM, GDS, RESIT and MANM estimated using the results of comparison tests shown in Table 4.5 and eq. 4.10 using simulated datasets.

From Fig. 4.8, it is observed that MANM have a zero model under fitting error in their estimated causal directions. But GDS and RESIT have a large model under-fitting errors,
4.5 Experiments

The Straight Line at 1 shows the Perfect Score with no under fitting error

**Model Under Fitting**: The Straight Line at 1 shows the Perfect Score with no under fitting error. Specifically, GDS has a poor performance in comparison to the other three methods. ICA-LiNGAM has produced comparatively small under fitting error.

**Model Reverse Fitting**: Owing to the estimation of wrong directions the constructed causal model acts as a reverse model from the original causal model. This type of error leads to the wrong detection of causal effects and has the same effect as the model under-fitting. So, for the performance test, these effects are cancelled from the positively estimated correct directions and signed as a negative (-) performance score. Both missing, wrong directions have a negative effect on causal construction and the performance of model fitting. Using the values of wrong directions in place of missing directions in eq. 4.10,

$$\text{Model Reverse Fitting Score} = \frac{CR - WR}{EA},$$  \hspace{1cm} (4.11)

where $CR$, $WR$ and $EA$ show the number of correct directions, wrong directions and edges assumed. Fig. 4.9 shows the performance results for the Model Reverse Fitting score in comparison to ICA-LiNGAM, GDS, RESIT and MANM.

Fig. 4.9 shows that MANM does not produce model reverse fitting errors, whereas ICA-LiNGAM, GDS and RESIT suffer from model reverse fitting errors to a great extent.
Fig. 4.9 Performance for Model Reverse Fitting for ICA-LiNGAM, GDS, RESIT and MANM over Simulated Datasets.

**Model Over Fitting**: Most of the time methods estimate a large number of extra directions in the estimation process which do not necessarily produce the cycles in the constructed model but complicate the construction process. Due to so many extra numbers of causal directions, it has become difficult to arrange them in a causal model. Also, this leads to the issue of an 'elimination' problem, where the common effects caused due to parents in a causal model have to be eliminated. However, with so many extra directions, elimination becomes difficult. This type of error results in model over fitting in the construction process. Using eq. 4.10 for the model over fitting by replacing positive extra directions in place of negative missing values, the resultant is a positive score (> 1). This can be represented as,

\[ \text{Model Over Fitting Score} = \frac{CR + EX}{EA}, \]  

where \( CR, EX \) and \( EA \) show the number of correct directions, extra directions and edges assumed. The performance for model over fitting for ICA-LiNGAM, GDS, RESIT and MANM on simulated datasets is provided in Fig. 4.10.

While MANM has produced very few extra edges, it has a very consistent and a very low model over fitting error. At the same time ICA-LiNGAM, GDS and RESIT have a large model over fitting errors. In particular, ICA-LiNGAM has a high model over-fitting error score in comparison to the other three methods.
From all the performance results shown in Fig. 4.8, 4.9 and 4.10, it is evident that MANM has better performance scores and has a better causal model constructibility than ICA-LiNGAM, GDS and RESIT.

4.6 Conclusion

In this chapter a Multivariate Additive Noise Model (MANM) has been proposed for complete causal discovery. The model assumptions are well imposed, and it is shown that structural consistency is completely achievable, where a complete causal model is totally identifiable. The proposed MANM can easily estimate multivariate datasets using d-separable V-structures, which work as sub-structural analysis. New terminologies of causal independence and causal influence are introduced for qualitative analysis on causal models. The score of Causal Influence Factor (CIF) provides the measure for causal influence and is useful for causal direction detection. Estimation for model fitting issues in causal construction have been provided, which verify the performance of the considered method towards causal construction.

In simulated and real world results, the novelty and constructibility of the MANM towards the complete causal model were verified without violating any of the criteria for DAG. This
multivariate model also provides scope for a bivariate analysis of discrete and continuous data types under special cases. The real world cases are studied using a variety of datasets. Model validations are also discussed and found to be correct with the prior facts/knowledge. The causal structure remains acyclic even after the method estimates additional directions in the features. In comparison, it was shown that the proposed method not only performed better than ICA-LiNGAM, GDS and RESIT, but also did well for different types of datasets where these methods failed. In the compared complexity test, MANM produced better results while estimating the most number of directed edges in a competitive time. In the performance test, the results were quite good, as MANM gave better causal construction ability with no estimated wrong or missing edges in model fitting cases. Noticeably, MANM gave low error in the model over fitting case and significantly produces no model under fitting or reverse fitting errors. The proposed method gives a better performance for mixed distribution data types and higher accuracy rate for correct causal direction detection. It is worth mentioning that the model has not been tested for cyclic structures as primary assumptions only provide scope for acyclic causal identification.
Chapter 5

Inscribing New Rules in Causality: Resolving the Divergence in Theory and Practical Implication

5.1 Missing Sufficiency for Causal Discovery

The model of causal analysis given by Pearl (2000), provides definitions, and propositions necessary for basic assumptions. Failing to hold on to these fundamental criteria makes any estimation process insufficient, although the sufficiency for complete causal discovery is not provided or addressed in any of the published works. All the assumptions made in theories do not add up when implemented in practice. Such facts can be understood by the following points.

- While the flow of the information or connection strength values inform about the amount of force transmitted from the source towards receiver, they do not explicitly explain how good the impact is.

- There is no such measure available to define and address this impact which is crucial for cause-effect relation.

- In practical cases, not having any measure of this impact leads to a situation where it is very hard to distinguish between a perfect and a bad hit.

- In causal language, this impact is the causal influence which is escalated because of connection strength or the information passed on from parent to the child.
5.2 The Quantitative Analysis of Causality

It may not be clear yet, what exactly the measure of causal influence means, and how useful it is for causal discovery. In the following sections, causal influence and its contribution to qualitative causal analysis in comparison with quantitative analysis techniques are elaborated.

Next, for the construction of the causal model, the concept of causal level is discussed. Causal levels are a cumulative assembly of sub-structural independent causal triplets/ d-separables. These d-separable models are detected using causal influence values and causal independence conditions. Also, the qualitative difference in causal independence and conditional independence is discussed in a mathematical explanation with suitable examples. This discussion is divided into two parts, one with quantitative analysis and the second as qualitative analysis.

5.2 The Quantitative Analysis of Causality

For a long time, what all the researchers have provided is the quantitative analysis of the causality (without any concern for it), although the only purpose of causal inference is about qualitative analysis. The first step in any analysis is to find a suitable model for the system. The next section presents models which are available or used for causal inference.

5.2.1 Model Selection

The first choice for graphical analysis is Structural Equation Models (SEMs) proposed by Bollen (1989). The linearity of SEM makes it easier to implement for path estimations. But path finding is not what is considered for causality: if this were true then all the other models such as linear or nonlinear ones would do the same. A more common and simple functional relation that can explain both linear and nonlinear cases can be given as:

$$ n_j = f(n_i, e_j), \quad (5.1) $$

where the direction/path \( \{n_i \rightarrow n_j\} \) is estimated, with the system containing a noise \( e_j \). While the particular representation for linear and nonlinear cases will be significantly different for the above bivariate model, the purpose remains the same: to estimate the parameters in the model. This type of model is better suited for nonlinear causal models and more efficient than the linear models.
In the case of linear system, a bivariate model for estimation of causal direction in between \((n_i, n_j)\) can be given as follows:

\[
    n_j = c_{ij} n_i + e_j (i \neq j).
\]  

(5.2)

where \(c_{ij}, (i \neq j)\) is the connection strength for the directed edge \(\{n_i \rightarrow n_j\}\) and \(e_j\) is a noise in the system. For any graph with \(m\) number of nodes, there will be \(\binom{m}{2}\) number of path coefficients and \(m\) number of noises for estimation. In every model, the nodes are arranged in a definite order where none of the later nodes have directed edges towards the earlier observed ones. This implies every prior node has a directed path to reach any of the later child nodes. This ordering from the parent node to successive descendant nodes can follow an ascending or descending order depending on the causal influence factors.

**Example:** Consider a case where a particularly prominent factor is significantly improved through the information passed from parent to child nodes. This can be understood in the case of medical observations where the primary host of a particular disease may not be severely affected by it. But as it spreads from primary host to secondary and then to others, the case may become more infectious with worsening behaviours. In this case, the influence becomes more effective from the parent stage to the descendant stages, which will favour an ascending ordering for the node arrangements.

The problem is reduced where the only requirement is to find the significant paths responsible for causal evaluation.

The above bivariate model was used for causal analysis by Hyvärinen and Smith (2013); Peters et al. (2011, 2014); Shimizu et al. (2006, 2005, 2011); Spirtes et al. (1993) with different parameter estimation methods. But the problem arises whether the bivariate model is sufficient to analyze causal inference. This question is explained in the following section. Before that, some of the propositions and definitions useful for causal inference are discussed.

**Graph d-Separable:** The condition of d-separable makes it easier to analyze any graph by observing or blocking the flow of information. For the structure \(\{x \leftarrow y \rightarrow z\}\) the following d-separable conditions are equivalent:

i. The structures \(\{x \rightarrow y \rightarrow z\}\) and \(\{x \leftarrow y \leftarrow z\}\) are d-separable when \(y\) is not observed.

ii. In \(\{x \leftarrow y \rightarrow z\}\), the graph is d-separated for unobserved \(y\).

iii. The graph \(\{x \rightarrow y \leftarrow z\}\) is d-separated whenever \(y\) or the descendant of \(y\) is observed.
Both (i) and (ii) have the same conditional independence where \((x \perp z|y)\). But in (iii), the observation of \(y\) makes the nodes \((x, z)\) become mutually dependent while before both are mutually independent of each other. The graph (iii) is of a common effect type referred to as V-structure and it is the only directed graph which can be used to solve the causal inference in sub-structural levels. The causal analysis from the point of probabilistic inference requires conditional probabilities of child nodes with respect to their parent nodes.

**The Markov Condition:** Consider a graph \(G\) with vertex set \(\{V_1, V_2, \ldots, V_n\}\) with the probability distribution \(P\). Applying the Markov Condition on the graph \(G\), the joint distribution factorization of conditional probabilities can be represented in an order from prior to later ones. A mathematical representation can be given as:

\[
P(V_1, V_2, \ldots, V_n) = \prod_i P(V_i|Pa(V_i)),
\]

(5.3)

where \(Pa(V_i)\) represent the parent of \(V_i\).

**Causal Markov Condition:** Consider the above graph \(G\) and the subset of the vertex set \(\{v \in V|v \subseteq V\}\) and let \(v = \{x \rightarrow y \leftarrow z\}\). Then the Causal Markov Condition says that the subset \(v\) is independent of all the other variables in the set \(V\) which are not the direct causes. Then for the subset \(v\), for any set \(\{S \subseteq V|S \not\rightarrow v\}\) the Causal Markov Condition can be written as:

\[
P(y|S, Pa(y)) = P(y|Pa(y)) = P(y|x, z).
\]

(5.4)

These definitions are the foundations for later causal development and analysis. How the bivariate models fail to hold on to these basic requirements is discussed in following sections.

**Markov Equivalent Classes:** The sub-structural analysis through d-separable V-structures requires a minimal arrangement to learn the feature relation. In this regard, the Markov equivalent classes are very helpful which offer the criteria for the minimal arrangement of V-structures. It says the connected causal models with the same number of node sets are Markov equivalent in their respective classes if they have the same connected edge sets irrespective of the edge directions in them. This is explained in Fig. 4.1 where Fig. 4.1a represent a completely connected causal model with all possible causal directions and Figs. 4.1b, 4.1c and 4.1d show the possible Markov equivalent classes for the Fig. 4.1a.
5.2 The Quantitative Analysis of Causality

The \(d\)-separable V-structures which are Markov equivalent are shown in Figs. 4.1b, 4.1c and 4.1d for the causal model in Fig. 4.1a with node set \(\{X, Y, Z\}\). The paper by He and Geng (2008) provides a better understanding for the use of Markov equivalent classes on causal inference using graphical sub-structural analysis.

Notice that by comparing the causal relations in Markov equivalent classes the correct directions can be found, however for this comparison, the causal influence values are to be estimated. Then by selecting the highest causal influence, the causal directions can be confirmed. This process is explained in the qualitative analysis section of the causal inference.

5.2.2 Insufficiency of Bivariate Model

The primary bivariate model is analyzed using probability and conditional probability for causal relations. The solution of conditional independence given using probability measures is definitely a quantitative case. Consider the below bivariate model:

\[
x_j = x_i + e_j. \quad (5.5)
\]

Here the causal direction \(\{x_i \rightarrow x_j\}\) is estimated, where the order of node is given as \(O(i) > O(j)\) and \(e_j\) is the noise/additive noise in the system for \(x_j\).

The probability representation of the exact case \(\{x_i \rightarrow x_j\}\) using the causal Markov condition can be given as

\[
P(x_j | Pa(x_j)) = P(x_j | x_i). \quad (5.6)
\]

This turns out to be the conditional probabilities of child node for observed parent sets. The question arises whether the conditional probability helps for causal identification. The representation of conditional probability is discussed in the below.

**Observation from Conditional Probability:** The main criterion for checking conditional probability is to find out whether two nodes which are observed on condition are dependent or independent. The probability value inform whether these two observed nodes are conditionally dependent or independent. The conditional dependence case in between the node \(\{x, y\}\) can be represented as:

\[
(x \not\perp \not\!
\!
\perp y) = (y \not\perp \not\!
\!
\perp x),
\]

\[
P(x|y) = P(y|x). \quad (5.7)
\]
Any conditional independence case of a node \( \{w \in V | w \rightarrow y \} \), has a representation of

\[
(w \perp \perp y) = (y \perp \perp w),
\]

\[
P(w|x) = P(w), P(x|w) = P(x).
\]

(5.8)

It is evident that the effect of conditional analysis for causal calculation does not make any sense. Both conditional dependence or independence in the bivariate case make both the nodes dependent or independent at the same time. The example below explains this in a simple case study.

**Example:** The conditional dependence or independence problem in the bivariate case can be better understood from this example. Consider the case of a DNA test for mother, father and child. The DNA samples A, B and C are unlabelled, but in real case A is the father, B is mother and C is the child. Without knowing which is a mother and a child’s DNA, let us assume that there is 50% similarity in B and C’s DNAs test results. This also means there is a 50% difference/dissimilarity in both DNAs. Can it be confirmed which one is a parent and which one is a child from this result? The answer is ‘No’. To resolve this problem, the father’s DNA sample is required. Comparing the DNA samples in A with C, may result in a 45% similarity. Then without any doubt, one can conclude that A and B are parents and C is the child, where one obvious case is that DNAs of A and B are completely dissimilar. But even then from similarity test, it is hard to confirm which one is the father and which one is the mother for the multivariate case. And for bivariate, it is even more complicated to distinguish between mother, father and child from similarity tests. These case scenarios for the test can be changed and the results will remain the same. In no case for two DNA comparisons, the parent can be identified.

It is clear that for the bivariate case, the parent and child features cannot be concluded from conditional dependence or independence test results. Hopefully, these help to clarify the claim that bivariate models are insufficient for causal analysis. The previous statement need to be corrected, because until now, what the conditional independence is trying to do is to find out the dependence or independence in two features for the bivariate case. So, it is nowhere nearly discussing the causal inference in all the above cases, till the conditional probability is used.

**Unused d-separation Criteria:** Let us add one more point to this claim. If it is noticed in the discussion that, the definition of graph d-separation is not applied anywhere in the case of the bivariate model. To use d-separation condition at least three nodes are needed, but this
is impossible in case of bivariate models. This is the point where one primary proposition for causal analysis has just been violated. How is that true? Arguing that after analyzing the structure for the bivariate model and after getting all the possible causal directions, the graph d-separation still can be used for causal construction. That is not possible, because d-separation is a condition which enables us to identify the parent and child nodes in the process of separation. What it means is that, if in the primary stage while causal directions are solved, if it did not include the d-separation (or not analyzing 3 variables at a time for directions) then, the directions estimated are not the causal directions (as in the case of bivariate models). So, our method of estimation and model should use the d-separation criteria whenever the directions are estimated. This is the turning point where bivariate models seem to be using the fundamental conditions wrongly and are found to be insufficient.

For this case, Fig. 4.1 can be followed which provides the d-separable V-structures necessary for sub-structural learning using Markov equivalent classes. But while using a bivariate model, the d-separable V-structures cannot be found for sub-structural analysis.

The goal is to find the qualitative analysis for causal inference, however, this analysis is stuck with quantitative results of conditional probability. The question follows how that can be. This means how the conditional probability is a quantitative analysis. The answer can be found in the following section.

5.3 Causal Inference: A Qualitative Analysis

Much have been discussed about the problem of bivariate models for causal analysis, explaining why they are insufficient and those leading to the quantitative inference in the above sections. So, what exactly does qualitative analysis means, how is it different from the conditional independence or dependence cases and how to identify it?

Let us begin the discussion with bivariate models. If the bivariate model is not useful, then consider the model given below.

5.3.1 A Multivariate Additive Noise Model

It is clear that insufficient bivariate models lead to the development of multivariate models. But how many dependent nodes are needed to be considered or how many parents should be taken into consideration for the estimation of causal directions in parent-child relation? Taking a closer look at the d-separation condition which is not used in the bivariate case,
it takes at least three nodes to find causal directions. That means exactly two parents and one child node are required. In terms of mathematics, two nodes are needed to represent the third node in a linear or nonlinear equation format with an added noise (as additive noise model suggests). Why three, why not four or five or more? The easiest one to analyze for d-separation are three node sets and adding more nodes to it needs more skills and restrictions to find the directions in them. So, it is always easy to go with three nodes which preserve the primary assumptions. This can be represented by the following equation for the case of triplet \( \{x \to y \leftarrow z\} \) as,

\[
y = f(x) + f(y) + e_y. \tag{5.9}
\]

Eq 5.9 can be a linear or nonlinear one, where \( e_y \) represents the additive noise in the system. A linear form of the eq. 5.9 with connection strength and additive noise can be given as:

\[
y = c_{xy}.x + c_{zy}.z + e_y. \tag{5.10}
\]

where \( c_{xy}, c_{zy} \) are the connection strength values for directions \( \{x \to y, z \to y\} \) and \( e_y \) is the additive noise in the system. Analyzing the importance of this equation reveals that the d-separation condition is fully implementable in this case. Also, while estimating for causal directions the equation can be used to impose the d-separation criteria.

The model properties used in case of bivariate models also hold here. So for a node set \( V = \{v_1, v_2, ..., v_n\} \), the order of the nodes can be given as \( \{O(v_1) > O(v_2) > ... > O(v_n)\} \) following a parent to child structural arrangement. But in previous cases, the order of the nodes depends on how they are connected in the graph or more specifically, using the conditional independent order over the conditional dependent ones. As argued before, the conditional probability could not justify the ordering of the nodes in the causal model, so a new type of independence is introduced for causal analysis, called causal independence.

### 5.3.1.1 Causal Independence

Causal independence is a qualitative signifier of the causal direction. While the basics of conditional independence provide the quantitative support, these do not specify the directions in feature sets. The causal independence/dependence are very specific and exclusive to identify the directions in the observed feature sets. Causal independence also signifies the order of a node in the causal structure. While most causal independent nodes are arranged at the top of the structure, the least independent ones are arranged in descending order in the causal model. Let us start from the conditional dependence case as given for the observed
feature set \( \{ x \to y \leftarrow z \} \) as,

\[
P(y|x) = P(x|y) \\
\Rightarrow (y \not\perp \perp x) = (x \not\perp \perp y)
\]

But in the case of causal dependence, the same conditional dependence has a different meaning, and it can be seen in the following:

\[
P(y|x) \Rightarrow (x \to y) = x \text{ is causal independence of } y \\
\quad = y \text{ is causal dependence on } x,
\]

\[
P(x|y) \Rightarrow (y \to x) = y \text{ is causal independence of } x \\
\quad = x \text{ is causal dependence on } y.
\]

It is evident why the conditional independence/dependence is not very useful for direction detection, whereas the causal independence/dependence exclusively informs about the direction in the feature set. So, all the criteria and primary assumptions defined for causal inference can be used in eq. 5.9 with causal independence criteria to find the causal directions.

Different methods can be used to analyze the eq. 5.9 to find the directions in the triplets. The question is, what parameter values are needed to be estimated for a successful causal inference? Following the structure of causal model, the connection strength values are required to detect the causal directions and the error values added into the model. As it has been discussed, the qualitative analysis of causal model requires causal influence values, let us discuss the role of causal influence.

### 5.3.1.2 Causal Influence

The role of causal influence is to verify the goodness of information passed on from parent node to the child node. The connection strength values show the weight/amount of information transferred, but the potential of that information only can be verified using the causal influence condition. The goodness of information depends on the quality of information passed and the level of noise added into the system. The meaning of this is that, if there is more noise/error added into the system then the influence of transmitted information becomes less effective. Therefore the more the noise the system contains, the less influential the connection strength becomes. But causal influence does not only depend on the information, it also depends on the feature value. So, the parent and its connection strength combined provide the causal influence on the child node.
In Fig. 4.1, 3 minimal d-separable Markov equivalent V-structures were provided, which are useful for the detection of causal directions from the comparison of their causal influence values. By solving eq. 5.10, the values of the unknowns can be estimated easily i.e. \( \{c_{xy}, c_{zy}, e_y\} \) for the case \( x \rightarrow y \leftarrow z \) as shown in Fig. 4.1c. Using eq. 5.10 the causal influence in the features can be defined by,

\[
CI_{xy} = \frac{c_{xy} \cdot E(x)}{E(y)} < \frac{E(e_y)}{E(y)}, \quad CI_{zy} = \frac{c_{zy} \cdot E(z)}{E(y)} < \frac{E(e_y)}{E(y)}.
\] (5.13)

The values \( \{CI_{xy}, CI_{zy}\} \) represent the causal influence/ causal influence factors for direction sets \( \{x \rightarrow y, z \rightarrow y\} \) and value \( E(.) \) represents the expectations of the indicated variables. The values of \( CI \) indicate the goodness of causal inference for the direction set \( \{x \rightarrow y \leftarrow z\} \) and the same process is applied for the other Markov equivalent classes as shown in Figs. 4.1b and 4.1d. The values of \( CI \) can be represented using a percentage range from \([0,100\%]\) or the probability range of \([0,1]\).

The causal influence values indicate the existence of direction in the observed features and how influential/effective the information is for constructing the child node. After examining all possible direction sets for a dataset, these can be arranged using the causal independence condition. The most causally independent node will be the top of the causal structure, where nodes are arranged in descending order, depending on their causal independence.

### 5.3.1.3 Causal Level

The causal structures are arranged using the causal influence values which represent the causal independence of an observed feature. The effective arrangement of nodes can be done in a way such that the graphical arrangement follow a tree like but connected and cyclic/acyclic arrangement in an ascending or a descending order using the causal independence of the node set. This kind of ordering produces the levels where nodes are grouped using their causal independence. Fig. 5.1 provides an example of causal levels in the shown causal structure.

Fig. 5.1 represents a 3-causal levelled model with 7 nodes. In Fig. 5.1, the nodes \( \{V_1, V_3\} \) are the two parent sets and represent the parent causal level. The node set \( \{V_2, V_4, V_5\} \) are the children of parent set and are shown in the child causal level. In the next level, the two nodes \( \{V_6, V_7\} \) which are the grandchildren of parents, are represented in grandchild causal level. This kind of causal level represents the kinship relation in the causal structure. Furthermore, the causal levels can be shown using different relation types as is present in the feature set.
5.4 Conclusion

In this chapter, an in-depth introduction and developments of causal analysis was provided with the solution to multiple unresolved problems. This chapter has discussed the issues of causal inference and difficulties with suitable examples in places to make it relevant and clear.
The contents of the references which are included can best explain the necessity of the newly proposed rules for causal analysis.

In the introduction to the subject, the complexity of the subject and issues leading to the introduction of new methods were discussed. The claim for the difference in the quantitative and qualitative analysis of causal inference were discussed elaborately. With example and discussions, it was shown why the quantitative analysis as used by most in previous methods (bivariate models) has failed to intercept, infer and conclude on causal analysis. The newly designed multivariate models are efficient on using the fundamental criteria for causal inference and provide a better evaluation system than the bivariate models.

The claim for the benefits and usefulness of the newly proposed quantitative analysis on causality is complete and well supported by given mathematical explanations. These explanation also supports the claim that causal independence is more suitable and mathematically sound for causal analysis than conditional independence as previously used in most cases. The proposals of causal influence and causal levels provide an all new dimension to causal analysis with measures to find the inference quality and structural arrangements. Mostly the new rule of causal influence supports the evaluation of the goodness of any causal inference, which is essential to find causal relations in the features. The causal levels which are estimated through causal influence are useful for causal construction and provide more detailed information on kinship relations.
Chapter 6

Conclusion & Future Work

6.1 Conclusion

Causal analysis or causality is a subject of great interest which does not explicitly provide predictability on the matter of concern but provides the technique to observe how things have happened and what the causes are for such occurrences. It does not help one to predict the uncertainties but it helps to be observant/aware of the factors which are capable of triggering these. It helps to find the footprints of incidents which are not observed in time but need to be studied and examined for their effects in time. Causality produces the deeper relations in factors which have critical impacts and information that make more sense in crucial processes.

Although causal studies have not yet found their way into mainstream science, the results that they have provided in earlier stages are quite promising for further analysis. Some recent progress in causal inference shows broad application of the subject in different areas where data mining and data analysis techniques have failed to evaluate the interdependent relationships.

This work is an extension of all the previous works that have been introduced into causality, but also it introduces and addresses some major findings, rules and methods which complete the gaps in causal studies. As Chapter 5 discussed the qualitative properties of the proposed definitions and criteria for complete causal discovery, the same can be drawn as the main findings in this regard. Chapter 5 provides both problem identity and solution in a very conclusive manner which summarizes the motivation of the entire work as represented in this thesis.
6.2 Possibilities of Progress

In the beginning, this thesis introduced causality in Chapters 1 and 2 establish the background, previous methods, and insufficiency of the proposed methods. The unaddressed issues are discussed in all the chapters including Chapters 3, 4, and 5; and these also propose the methods to rectify these issues on which the whole work of this thesis is built. Though Chapter 3 presents just a minor evolution or addition to the existing technique with the change in the estimation process that helps to detect causal direction, yet it is a significant change and its success in estimations is well examined in given simulated and experimental results and in the comparison tests.

The problem in causality study is the complete analysis which it suffers because of the use of bivariate models. Bivariate models also fail to utilize d-separation criteria in causal estimation. A major improvement in this is the proposed multivariate additive noise model, the use of causal independence and causal influence values enable it to be fully capable of detecting complete causal models. In Chapter 4, the shown experimental results provide the capability of the MANM in causal estimation in case of different data types and complicated feature relations. It also provides the issues in causal model construction with different model fitting issues. The process of causal construction from estimated causal directions suffers from the model under fitting, model over fitting and model reverse fitting issues, which are broadly discussed and examined for different cases in comparison with similar methods in the same context.

The importance and necessity of the proposed causal independence, causal influence, and causal levels are broadly discussed in Chapter 5 with suitable examples and mathematical explanations. There is an endeavour to resolve the gap in theoretic causal estimation and practical implication. The new rules support the causal analysis for better and precise evaluation of interdependency in factors and the discovery of significant factors responsible for the cause-effect relations.

6.2 Possibilities of Progress

Though the subject of causality has wide implication and scopes, still it has not been able to attract a broad group of researchers to it owing to the difficulties it holds. Although the issues in the causal analysis have been discussed in every chapter of this thesis, which also can be concluded as a straight indication for future work in this regard. However, this section discusses the possibilities in an explicit manner.
A significant step will be to introduce and implement causality in different fields for better knowledge extraction. The standard techniques are not as effective as causality in extracting functional relations from observations. The utilization of this can be very useful in the fields of economics, psychology, sociology, machine learning, structural analysis and medical observational studies.

Causality is extremely useful for prevention analysis where the sources/factors responsible for certain causes are needed to be estimated. The uncertainty in prediction can not be eliminated completely, but situations can be prevented from getting worse by eliminating or removing the causes which are responsible for such.

The noise or external influences can be fully detected using causal analysis as well as their effect towards the occurrences. It is informative for observational studies where some important features are neglected because of the notion of trivial.

The defined causal influence is not completely developed in a sense that it may not fit with other estimation processes of the same type. So, there is every possibility of introducing new methods to measure and estimate causal influences for different cases into this field of study.

The introduction of causal levels is new and very crucial for structural construction, and representation of causal models. It also can be used for causal classification to represent different features in similar groups for a deeper understanding of their relations.

Until now for simplification, causal models are assumed to be acyclic, which is not true for every relation. The natural phenomena favour the cyclic process for their continuity and recyclable property. For a broader application of this topic, the objective is to find the cyclic relations in reversible systems. A consideration of these possibilities will help to find deeper and better relations in causal models.

As an exploratory data analysis, causality still lags behind in many different aspects which are discussed in this thesis. Also, the process of knowledge discovery using causality has not been simplified very much. So, there is a vast scope for researchers to simplify different model based on estimation processes. Many different scenarios have been considered for datasets in simulations and the proposed methods are tested for extremity, but still a huge amount of work needs to be done in this regard. Data analysts can use the system of causality in their endeavours to find new branches and applications of it.

The prediction from data or pattern analysis is strictly dependent on the continuous behaviour of the samples or the feature relations. If this is true, then causality can be used for predicting future effects by tracing their past relations. The fact must be considered while
prediction is uncertain and can only be achieved if facts are in a repetitive pattern, so causality can be the used as a prediction method by analyzing the cause-effect relation.
References


References


Appendix A

Publications

Journals


Journals Communicated


Index

ALiNGAM, 51
Bayes Theorem, 15
Causal Independence, 69, 100
Causal Influence, 101
Causal Influence Factors, 73
Causal Level, 102
Causal Markov Condition, 46, 72, 96
CCA, 32
COCO, 29
Collider, 16
d-separable, 71, 95
D-separation, 16
Direct-LiNGAM, 43
DirectLiNGAM, 36
GDS, 85
HSIC, 29, 44
HSN, 29
ICA, 28
ICA-LiNGAM, 33, 85
KCI, 30
KERNELICA, 28
KMI, 28, 31
LiNGAM, 43
MANM, 69
Markov Blanket, 22
Markov Condition, 72, 96
Markov Equivalent Classes, 96
Markov Equivalent models, 72
Markov Model, 15
MBS, 26
MCMC, 16
Model Over Fitting Error, 89
Model Reverse Fitting Error, 89
Model Under Fitting Error, 88
Multivariate Additive Noise Model, 99
PCD, 46
PGM, 26
PNL, 35
RESIT, 85
SEM, 11
V-structure, 16, 71