Towards efficient and unbiased causal effect estimation from observational data

Debo Cheng
Master of Computer Science

A thesis submitted for the degree of
PhD in Computer and Information Science



University of
South Australia

UniSA STEM
University of South Australia

April 2021

Contents

List of Figures **v**

List of Tables **ix**

1 Introduction **1**

1.1 Background 1

1.2 Contributions 8

1.3 Thesis outline 10

2 Literature review **12**

2.1 Covariate adjustment for Causal Effect Estimation 12

2.2 Adjustment set identification given a causal graph 16

2.3 Adjustment set identification with local structure knowledge 18

2.4 Adjustment set identification without causal structure given 19

2.5 Sufficient dimension reduction for ACE estimation 23

2.6 Causal structure learning from data 25

2.6.1 Constraint-based structure learning methods 25

2.6.2 Score-based structure learning methods 27

2.6.3 Hybrid structure learning methods 27

2.6.4 Local causal structure learning methods 28

2.7 Summary 29

3 Preliminaries **30**

3.1 Graphical causal models 30

3.1.1 Causal DAG 30

3.1.2 Causal MAG 33
3.2 Covariate adjustment based on graphical causal modelling . 37
3.2.1 Covariate adjustment for causal effect estimation . . 37
3.2.2 Graphical criteria for identifying adjustment sets . . 38
3.3 Matching methods for causal effect estimation 41
3.3.1 Potential outcome framework 41
3.3.2 Matching methods 42
3.4 Summary 44

4 Sufficient dimension reduction for average causal effect estimation 45
4.1 Motivation 45
4.2 The theorem for adjustment set identification with SDR . . 47
4.2.1 Problem setting 47
4.2.2 Sufficient condition for identifying an adjustment set using SDR 47
4.3 The CESD algorithm 49
4.3.1 Adjustment set identification using SDR 49
4.3.2 NNM using the discovered adjustment set 51
4.4 Experiments 52
4.4.1 Experiments on the two semisynthetic real-world datasets 54
4.4.1.1 IHDP 54
4.4.1.2 Twins 55
4.4.2 Evaluation with three real-world datasets 56
4.4.2.1 Jobs 56
4.4.2.2 Cattaneo2 58
4.4.2.3 Right Heart Catheterization 59
4.4.3 The quality of matching 61
4.5 Conclusion 62

5 Local causal search for efficient causal effect estimation from data with latent variables 64
5.1 Motivation 64
5.2 Local search for an adjustment set 65

5.2.1 The Theorem 65
5.2.2 Causal effect estimation 67
5.2.3 Removing insignificant adjustment variables 67
5.2.4 The DICE algorithm 69
5.3 Experiments 70
5.3.1 The quality of local structure learning 70
5.3.2 Evaluating DICE with real-world datasets 72
5.3.3 Efficiency evaluation of DICE 76
5.4 Conclusion 78

6 Local causal search for efficient and unique causal effect estimation 79
6.1 Motivation 79
6.2 The theorems for local causal search 81
6.2.1 Problem setting 81
6.2.2 The theorems for local search 82
6.3 The CE^2LS algorithm 87
6.4 Experiments 90
6.4.1 Comparison methods and metric 90
6.4.2 Experiments on synthetic datasets 93
6.4.3 Experiments on synthetic data generated using a benchmark Bayesian network 96
6.4.4 Experiments with real-world datasets 97
6.4.5 Time efficiency 98
6.4.6 The completeness of CE^2LS 100
6.5 Conclusion 103

7 Local causal search for causal effect estimation without the pretreatment variable assumption 104
7.1 Motivation 104
7.2 The theorems for adjustment set discovery 105
7.2.1 Problem setting 105
7.2.2 The theorems 107
7.3 The proposed two algorithms 110

7.4 Experiments 113
7.4.1 Experiments with synthetic data 114
7.4.2 Experiments on two semi-synthetic real-world datasets 117
7.4.3 Experiments on two real-world datasets 118
7.4.4 Efficiency evaluation 120
7.4.5 Sensitive Analysis 121
7.5 Conclusion 123

8 Conclusion and Future Directions **124**
8.1 Summary of contributions 124
8.2 Future directions 127

References **129**

List of Figures

1.1 An overview of causal effect estimation from observational data. The objective of the thesis is to estimate average causal effects (ACE) from observational data based on graphical causal modelling. 2

1.2 An example of Simpson's paradox. The picture was taken from [1]. There exists a positive correlation between dosage and recovery in the whole dataset, but in each subgroup of the data (males and females), there exists a negative correlation between dosage and recovery. 3

1.3 Graphical representation to illustrate the confounding bias when estimating the causal effect of drug dosage W on recovery Y. 4

1.4 An example showing the reason for uncertainty in causal effect estimation. 6

1.5 An example to illustrate M-bias. W, Y and C are observed variables, and U_1 and U_2 are latent variables. Adjusting for set $\{C\}$ yields M-bias [2, 3], although C is correlated to both W and Y. 7

2.1 An overview of causal inference with observational data based on graphical causal models (with reference to the sections in this chapter discussing the related work), including causal structure learning and adjustment set identification for unbiased causal effect estimation. 16

LIST OF FIGURES

3.1 (a) The directed path $V_3 \to V_1 \to V_2$ together with the edge $V_2 \to V_3$ form a directed cycle, (b) the directed path $V_3 \to V_1 \to V_2$ together with the edge $V_2 \leftrightarrow V_3$ form an almost directed cycle. 35

3.2 Two possible configurations in MAGs and PAGs for a visible edge $V_i \to V_j$ [4]. Nodes V_k and V_j must be non-adjacent. . . 37

4.1 A graphical representation of sufficient dimension reduction, $W \perp\!\!\!\perp \mathbf{Q}|\mathbf{Z}$ holds [5], where $\mathbf{X}$ is decomposed into $(\mathbf{Z}, \mathbf{Q})$. . . 48

4.2 Estimated causal effects on the IHDP dataset w.r.t. a 95% confidence interval. The red line represents the ground truth ACE. 55

4.3 Estimated causal effects on the Twins dataset w.r.t. 95% confidence interval. The red line represent the group truth ACE. 57

4.4 Estimated causal effects on the Jobs dataset. The two dotted lines denote the empirical estimated interval with a 95% confident interval. 58

4.5 Estimated causal effects on the Cattaneo2 dataset w.r.t. 95% confident interval. The two dotted lines represent the empirical estimated interval (-250g, -200g). 60

4.6 Estimated causal effects on the RHC dataset w.r.t. 95% confident interval. 61

4.7 The distributions of propensity scores (left four) and reduced covariate dimensions, grouped by the treatment variable W (red for the control group and blue for the treated group) on the Cattaneo2 dataset. 62

5.1 (a) A DAG with two latent variables U_1 and U_2. (b) The corresponding MAG of the DAG. 71

5.2 The precision, recall and F-score of FCI & RFCI & PC-Select algorithms on two groups of synthetic datasets w.r.t. the number of variables and samples. The qualities of parent discovery by the three algorithms are very consistent. but the time efficiencies of the algorithms are quite different, as shown in Fig. 5.4 72

5.3 Histograms of causal effects in ASCETs for Jobs, IHDP and twins. The red line denotes the ground truth, and the bin size is $(\max(Y)/100)$. 76

5.4 The runtime on two groups of synthetic datasets w.r.t. the number of variables and samples. 77

5.5 The runtime on ten BNs. Note: LVIDA did not return results in two hours on BARLEY, HAILFINDER, HEPAR II, WIN95 and ANDES. 77

6.1 A counterexample in which Corollary 3 fails to select an adjustment set by local search. 86

6.2 (a) is a causal DAG for generating synthetic datasets [6], and (b) represents its corresponding MAG when X_2, X_4, X_6 is unobserved. 93

6.3 (a) A causal DAG for synthetic dataset generation [7] and its corresponding MAG in (b) where variables X_7, X_{11}, X_{12} are removed. 93

6.4 REs on synthetic datasets with increasing sample size ($\log_{10}$ on the $Y-$axis). The top figure shows the results of Group I datasets, and the bottom figure shows the results of Group II datasets. As the sample size increases, the relative errors of CE^2LS and EHS decrease significantly. The performances of LVIDA-FCI and LVIDA-rFCI are worse than CE^2LS. . . 95

6.5 The INSURANCE Bayesian network. The red node denotes the treatment variable W, the blue node represents the outcome variable Y, and the causal path between them is shown in green. The crossed variables are the latent variables in the datasets with latent variables. 96
6.6 REs of causal effects estimated by different methods with the datasets generated based on INSURANCE. EHS, LVIDA-FCI, and LVIDA-rFCI are not shown since they did not return results within two hours. CE^2LS outperforms six other adjustment criteria. 97
6.7 Running time in seconds (s) ($\log_{10}$ on the $Y-$axis) of CE^2LS, EHS, LVIDA-FCI and LVIDA-rFCI with the two groups of synthetic datasets described in Section 6.4.2. The top figure shows the results of Group I datasets, and the bottom figure shows the results of Group II datasets. CE^2LS is faster than the others on all datasets. 99

7.1 An example of a PAG $\mathcal{G}$, used in Example 7.1. 109
7.2 The estimated causal effects of all methods in terms of a 95% confident interval. The two dotted lines denote the empirical estimated interval (-250g, -200g). 119
7.3 The estimated causal effects of all methods in terms of a 95% confident interval. There is no true causal effect or empirical estimates on the RHC dataset. 120
7.4 Runtime in seconds of different algorithms. 121
7.5 The boxplots of the estimated causal effects under different COSO Q on three synthetic datasets, i.e., INSURANCE, MILDEW and BARLEY. Choosing different COSO Q, the estimated causal effects are close to the true causal effect on all three synthetic datasets. 122

List of Tables

2.1 A summary of acronyms used in the thesis. 13

2.2 A summary of the data-driven causal effect estimation methods. 25

4.1 Summary of the real-world datasets. 53

4.2 Experimental results on IHDP. 54

4.3 Experimental results on Twins. 56

4.4 Experimental results on Jobs. 57

4.5 Experimental results on Cattaneo2 and RHC. 59

5.1 An example ASCET. 68

5.2 The final ASCET. 68

5.3 Results on Jobs. Bias (%) is the relative error comparing to the ground-truth causal effect. The estimation with the lowest bias is highlighted in each category. 73

5.4 Results on IHDP w.r.t. CE and bias. 74

5.5 Results on Twins w.r.t. CE and bias. 75

6.1 The estimated ACEs and RE with the IHDP and Twins datasets. The lowest RE in each dataset is boldfaced. CE^2LS outperforms all comparisons on both datasets. 98

6.2 Experimental results on the completeness of adjustment set discovery using Corollary 3 and CE^2LS. In the table, the expression A/B (e.g., 99/99) denotes the number of trials (out of 100) when an adjustment set is identified using Corollary 3 (A) and CE^2LS (B) respectively. n is the number of nodes in an MAG, l indicates the number of latent variables, and d denotes the average degree of the generated DAGs used to construct MAGs. '-' denotes that the adjustment sets cannot be found correctly because the density of the constructed MAG is high (the density of a MAG is more than 0.5). . . . 102

7.1 The details of five standard benchmark BNs. 114
7.2 The variables W, Y and the latent variables on the five synthetic datasets. 115
7.3 Estimating ACE on synthetic datasets. (a, b) next to a dataset name denotes the number of nodes (a) and arcs (b). 116
7.4 Estimating ACE on IHDP and Twins. 117
7.5 Estimating ACE on Cattaneo2 and RHC. 119

Abstract

One of the important tasks of causal inference is to estimate the causal effects of the treatment (intervention) on the outcome of interest. Randomized control trials (RCTs) are considered the gold standard for evaluating causal effects, but they are often infeasible because of practical and ethical constraints. Estimating causal effects from nonexperimental (observational) data has gained increasing attention due to easier access to observational data. However, confounding bias is one of the main obstacles in causal effect estimation from observational data because a confounder has an effect on both the treatment and outcome variables that may significantly bias the estimated causal effects.

Covariate adjustment is a widely accepted approach to reducing confounding bias in causal effect estimation from observational data. Recently, several criteria have been proposed based on graphical causal models for determining an appropriate set of adjustment variables. These criteria require the knowledge of a causal graph representing the causal mechanism underlying the data, but in most applications, the causal graph is unknown. Thus, data-driven methods are in demand for finding adjustment sets directly from data. However, the data-driven approach faces four main challenges: data insufficiency, uncertainty in adjustment set identification, latent variables, and low efficiency.

In this thesis, we address the four challenges and make the following contributions.

First, to address the challenge of data insufficiency, we prove that a large set of variables can be reduced to a representative

set that is much smaller (low dimensional) by sufficient dimension reduction. Based on this, we develop a data-driven algorithm, CESD, which uses a supervised kernel dimension reduction method to estimate causal effects from observational data.

Second, to address the challenges of latent variables and low estimation efficiency, we develop a local search theorem to discover a superset of the adjustment set for causal effect estimation from data with latent variables under the pretreatment variable assumption. Then, based on the theorem, we propose an efficient algorithm, DICE, to obtain a bound estimation of causal effects from data with latent variables.

Third, we present theorems to support local search to find proper adjustment sets based on conditional independence tests under the pretreatment variable assumption. We propose a local search algorithm, CE^2LS, that is fast and can estimate causal effects unbiasedly and uniquely from data with latent variables. The CE^2LS algorithm addresses the three challenges of uncertainty in adjustment set identification, latent variables, and low efficiency.

Finally, we relax the pretreatment variable assumption and develop two theorems to ensure that it is sufficient and necessary to conduct conditional independence tests to find proper adjustment sets for unbiased causal effect estimation. Based on the theorems, we propose two novel algorithms, DAVS-Q and DAVS, to discover unique and unbiased causal effects from data with latent variables.

We evaluated the developed algorithms on synthetic and real-world benchmark datasets. The experimental results demonstrated the good performance of the developed algorithms.

In summary, this thesis addressed the main challenges confronting data-driven causal effect estimation by developing theorems and

algorithms for effective and fast search for proper adjustment sets for unbiased and unique causal effect estimation from observational data. The developed algorithms significantly improved the performance of causal effect estimation from observational data and provide better assistance with decision-making in real-world applications.

Declaration

This thesis presents work carried out by myself and does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; to the best of my knowledge it does not contain any materials previously published or written by another person except where due reference is made in the text; and all substantive contributions by other to the work presented, including jointly authored publications, are clearly acknowledged.

Name: Debo Cheng

Signed: 

Date: April 21, 2021

Acknowledgements

I would like to express my deep gratitude to my supervisors, Professor Jiyoung Li, Associate Professor Lin Liu, Associate Professor Jixue Liu and Dr Kui Yu. They always give me a lot of support, guidance and encouragement throughout my doctoral journey. It has been an honour to have the opportunity to study under the supervision of them. Their wisdom and academic style have benefited me for my academic career.

First and foremost, I express my gratitude to my principal supervisor, Professor Jiyoung Li. From him, I learned how to identify problems, solve the problems and present results. His persistent encouragement, great patience and valuable suggestions have helped me through the journey of my doctoral degree. Thanks Jiuyong for bringing me to the research area of causal inference, from which I have enjoyed very much and learned a lot. His kind guidance and advice has helped my thesis to made solid contributions to the research in causal inference.

I particularly appreciate my co-supervisor, Associate Professor Lin Liu. She is a great collaborator and an enthusiastic coach. I benefit so much from her knowledge, expertise and insightful ideas. Moreover, she gives me insightful suggestions and comments for all my research manuscripts. A special thank extends to my co-supervisor, Associate Professor Jixue Liu, whose profound knowledge about data structure and his valuable comments on my works has made my journey easier. I also owe my

gratitude to my co-supervisor, Dr Kui Yu, as I have benefited from our inspiring discussions and collaborations.

I thank all my colleagues in the Data Analytics Group at the University of South Australia (UniSA). In the last few years, we have had many good memories and moments. I particularly thank Associate Professor Thuc Duy Le, Dr Selasi Kwashie, Dr Michael Bewong, Sha Lu and Xiaomei Li for their company and valuable discussions during my P.h.D study.

I reserve my special gratitude to my parents, my brother for their perpetual love, support and understanding, and my fiancée, Wenying Huang, for her consistent encouragement and love.

My research work at UniSA would not have been possible without the generous financial support from China Scholarship Council (CSC), Australian Research Council (ARC) Discovery Project Top-up Scholarship and UniSA International Research Tuition Scholarship (IRTS).

List of publications

1. **Debo Cheng**, Jiuyong Li, Lin Liu, Jixue Liu, Kui Yu and Thuc Duy Le. Causal query in observational data with hidden variables. Pages: 2551-2558, *ECAI2020*.
2. **Debo Cheng**, Jiuyong Li, Lin Liu, Jiji Zhang, Thuc Duy Le and Jixue Liu. Local search for efficient causal effect estimation. Submitted to *Neurips2021*, Under Review.
3. **Debo Cheng**, Jiuyong Li, Lin Liu, Kui Yu, Thuc Duy Le and Jixue Liu. Towards precise causal effect estimation from data with hidden variables. Submitted to *IEEE Transactions on Neural Networks and Learning Systems*, Under Second Round Review.
4. **Debo Cheng**, Jiuyong Li, Lin Liu, Jixue Liu, Thuc Duy Le and Kui Yu. Sufficient dimension reduction for average causal effect estimation. Submitted to *Data Mining and Knowledge Discovery*, Accepted.
5. Zhenlong Xu, Jixue Liu, **Debo Cheng**, Jiuyong Li, Lin Liu and Ke Wang: Assessing the fairness of classifiers with collider bias. arXiv preprint arXiv: *2010.03933*, 2020.

DEDICATION

to my family and fiancée.

Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance).

- Albert Einstein (1953)

Chapter 1

Introduction

1.1 Background

Causal inference investigates cause-and-effect relations between variables. Such relations entail the mechanisms of a system under the influence of interventions [8, 9]. Therefore, causality has been widely studied and utilized in a variety of fields, such as epidemiology [10, 11], econometrics [12, 13, 14], and social science [15, 16], and philosophy [17, 18]. For example, medical researchers may want to evaluate the influence of a new drug on a disease, and government policy makers are interested in estimating the effect of education on the incomes of individuals.

A randomized controlled trial (RCT) is regarded as the *gold standard* for drawing causal conclusions [19, 20]. In an RCT, participants are randomly assigned to a treated group and a control group. Under random assignment, the two groups of participants have similar characteristics. For instance, to test the effectiveness of a new medicine, RCT participants are randomly selected to take the medicine (i.e., assigned to the treated group) or receive placebos (i.e., assigned to the control group). Then, the effect of the medicine is assessed by comparing the outcomes of participants in the treated group and the outcomes of participants in the control group.

Unfortunately, RCTs are often expensive or even infeasible to conduct due to practical and ethical limitations. For example, it is infeasible to manipulate each of the over 20,000 human genes to discover which genes

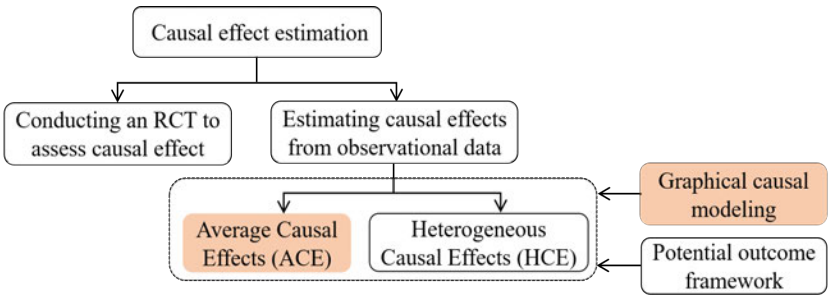


Figure 1.1: An overview of causal effect estimation from observational data. The objective of the thesis is to estimate average causal effects (ACE) from observational data based on graphical causal modelling.

are the causes of a disease, and it is unethical to force participants to smoke to determine the effect of smoking on lung cancer.

As an alternative strategy, causal inference from observational data has attracted increasing attention [21, 22, 23, 24]. Many methods have been proposed to investigate cause-and-effect relations between variables from observational data [13, 18, 25, 26]. Generally, causal inference from observational data contains two tasks, recovering the underlying causal structure over the measured variables [18, 27, 28, 29] and quantifying the strengths of causal relationships (i.e., the causal effects) between the variables [30, 31, 32, 33]. Many methods have been proposed to recover causal structures from observational data [25, 34, 35, 36], and we will briefly discuss them in Chapter 2. In this thesis, we focus on the second task, i.e., quantifying the strengths of causal relationships or causal effect estimation from observational data, as shown in the highlighted part of Fig. 1.1.

Causal effect estimation from observational data is a fundamental problem in causal inference [37, 38]. Many methods have been developed to estimate average causal effects (ACE) [13, 22, 23] and heterogeneous causal effects (HCE) [39, 40, 41]. ACE measures the difference in the average outcome in the treated group and the average outcome in the control group. ACE is the most common starting point for analysing any effect assessment. Therefore, it is very important to study how to estimate ACE unbiasedly from data. In contrast, HCE assesses the causal effects at the subgroup

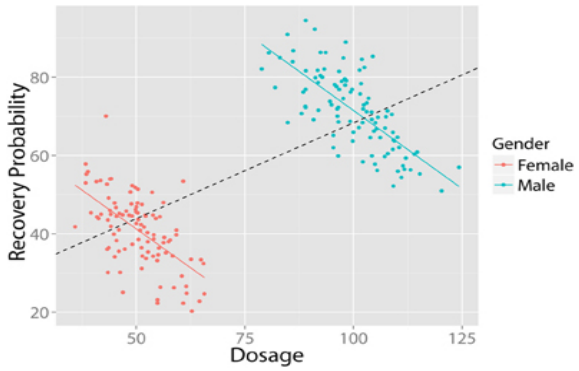


Figure 1.2: An example of Simpson's paradox. The picture was taken from [1]. There exists a positive correlation between dosage and recovery in the whole dataset, but in each subgroup of the data (males and females), there exists a negative correlation between dosage and recovery.

level [26, 42]. One popular method for addressing HCE is to divide the observational data into subgroups (e.g., by gender or state) and then quantify the ACE in each subgroup. We focus on estimating ACE, as shown in the highlighted area in Fig. 1.1.

There are two major paradigms for causal inference from observational data: graphical causal modelling [43] and the potential outcome framework [32]. The former establishes a causal graph with variables, and the causal graph represents the causal mechanisms of the data (i.e., causal relationships between variables), while the latter estimates the missing potential outcomes and then calculates the causal effects. It is worth noting that both graphical causal modelling and the potential outcome framework can be used to estimate causal effects from observational data. Comparisons of the two frameworks can be found in [9, 44, 45].

An overview of causal effect estimation from data is presented in Fig. 1.1, and we focus on employing graphical causal models for quantifying causal effects from observational data.

Confounding bias is the main challenge in causal effect estimation from observational data [46]. In observational data, the investigators have no control over treatments and individuals, and they record the data based on their observations. The measured variables affect both the treatment

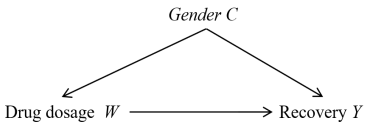


Figure 1.3: Graphical representation to illustrate the confounding bias when estimating the causal effect of drug dosage W on recovery Y.

and outcome in observational data, and hence, removing the effect of these variables is significant for determining the treatment effectiveness. We take Simpson's paradox [47] as an example to show confounding bias. The dosage of a drug W is positively correlated with patient recovery Y in the whole dataset but negatively correlated with Y in the sub-datasets for the male and female groups. A visualisation of the example is given in Fig. 1.2 [1]. The relationships between the three variables can be represented in the causal graph, as shown in Fig. 1.3. The path $W \leftarrow C \rightarrow Y$ results in a confounding bias when estimating causal effects of W on Y without considering C.

Covariate adjustment is a principled method for removing confounding bias by adjusting for a covariate set in causal effect estimation from observational data [48, 49, 50]. However, it is challenging to determine which variables are appropriate to be adjusted for, especially with datasets containing a large number of variables. For instance, some methods [51, 52, 53] adjust for as many variables as possible, as long as these variables are not affected by the treatment variable W and outcome Y. However, in certain circumstances, such an inclusive adjustment leads to incorrect conclusions [3, 54]. Therefore, determining a proper adjustment set (or deconfounding set) is crucial for achieving unbiased causal effect estimation from observational data.

Graphical causal modelling provides a theoretical foundation for determining a proper adjustment set. When a causal DAG (directed acyclic graph) or MAG (maximal ancestral graph) representing the causal mechanism is given, the back-door criterion [9], the adjustment criterion [55], the generalized back-door criterion [48] or the generalized adjustment cri-

terion [49] can be used to determine proper adjustment sets for estimating causal effects from observational data. We take the causal graph in Fig. 1.3 as an example to illustrate how to choose a proper adjustment set by using the back-door criterion. From the causal graph, set $\{C\}$ satisfies the back-door criterion since node C is sufficient to block all back-door paths from W to Y (i.e., all paths having an edge into W), i.e., $W \leftarrow C \rightarrow Y$. Thus, $\{C\}$ can be chosen as a proper adjustment set for estimating causal effects of W on Y from observational data.

When users have data but no causal graph knowledge, estimating causal effects from the data is very challenging. In most applications, the causal graph is unknown such that it is difficult to find a proper adjustment set for estimating causal effects unbiasedly from observational data. Thus, data-driven methods are in demand for finding adjustment sets directly from data. There are four main challenges in the existing data-driven methods: (**I**) data insufficiency, (**II**) uncertainty in adjustment set identification, (**III**) latent variables, and (**IV**) low efficiency.

Data insufficiency results from large-sized adjustment sets in nonparametric causal effect estimation [56, 57]. When the number of variables in the adjustment set is large, a very large dataset is needed for reliable estimation. For example, the number of samples required for causal effect estimation with the widely used matching approach is exponential to the number of variables in the adjustment set to achieve reliable causal effect estimation [57, 58]. Under the assumptions that all observed variables are pretreatment variables and the data are absent of latent variables, sufficient dimension reduction techniques are utilized to address data insufficiency [59, 60, 61]. The application of sufficient dimension reduction allows researchers to directly propose purely data-driven methods for estimating causal effects from observational data. However, these sufficient dimension reduction strategies have not yet been theoretically linked with covariate adjustment.

The uncertainty in causal effect estimation is from the non-determinism in learning causal graphs from observational data. From data, only a

Markov equivalence class[1] of the underlying causal graph can be recovered. Based on these different learned causal graphs (although they are in the same Markov equivalence class), different adjustment sets may be discovered, and different causal effects are estimated using the identified adjustment sets. Thus, existing data-driven causal effect estimation methods often return a multiset of possible causal effects, i.e., a bound estimation of causal effects, instead of a unique estimation [62, 63].

Referring to Fig. 1.4, where Fig. 1.4 (a) presents the underlying DAG[2]. From the observational data, we can learn a CPDAG (completed partially directed acyclic graph)[3] as shown in Fig. 1.4 (b), which results in two Markov equivalent DAGs, as shown in Fig. 1.4 (c). To estimate the causal effect of W on Y, based on DAG (1) in Fig. 1.4 (c), the adjustment set is $\{C\}$, and based on DAG (2) in Fig. 1.4 (c), and the adjustment set is an empty set. Thus, we obtain two possible causal effects of W on Y and are not sure which one is true. This is a bound estimation of causal effects. In a complex dataset, the number of possible causal effects can be large.

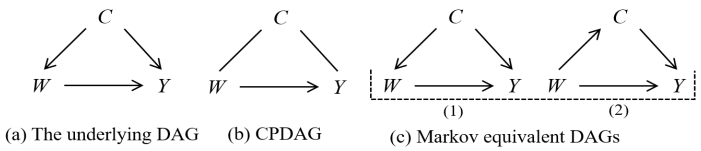


Figure 1.4: An example showing the reason for uncertainty in causal effect estimation.

Adjustment for a variable may result in bias when some important factors are not observed or unmeasured (latent variables) [2, 64]. Conditioning on a variable that is correlated to both W and Y potentially introduces so-called collider bias or M-bias in causal effect estimation [3, 54]. An example of M-bias is given in Fig. 1.5. A correct estimation of the causal effect of

[1]Two causal graphs are said to be Markov equivalent when they encode the same conditional independence relationships between variables [18].

[2]A DAG is a directed graph with no directed cycles [9].

[3]A CPDAG is a graph with directed and undirected edges for representing an equivalence class of DAGs [18].

W on Y is produced when adjusting on no variable, but a biased causal effect estimation is produced when adjusting for $\{C\}$ since the back-door path $W \leftarrow U_1 \rightarrow C \leftarrow U_2 \rightarrow Y$ is open when conditioning on C. The causal graphs in Fig. 1.4 (a) and Fig. 1.5 are indistinguishable from the data since these two causal graphs entail the same set of independencies between the observed variables [65, 66].

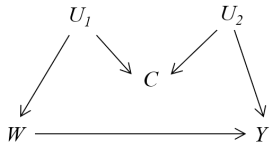


Figure 1.5: An example to illustrate M-bias. W, Y and C are observed variables, and U_1 and U_2 are latent variables. Adjusting for set $\{C\}$ yields M-bias [2, 3], although C is correlated to both W and Y.

Current data-driven causal effect estimation methods are inefficient since they employ global causal structure learning [36, 67]. For example, a widely used data-driven causal effect estimation method, IDA (intervention when the DAG is absent) [62], is used for estimating causal effects directly from data without latent variables. IDA is inefficient because it first needs to learn a CPDAG by a global causal structure learning algorithm (e.g., the PC algorithm [18]) from data and then determine adjustment sets from each DAG in the Markov equivalence class represented by the CPDAG. Learning CPDAG in data is fundamentally NP-complete [68], and existing algorithms do not scale with the number of variables well [18].

The existing data-driven causal effect estimation methods are affected by the four challenges described above. Thus, how to find the proper adjustment sets efficiently and effectively remains unsolved when the causal graph is unknown. This thesis addresses the four challenges by developing a set of theorems and novel algorithms based on graphical causal modelling for estimating causal effects directly from observational data.

The first piece of research work in the thesis provides a theoretical dimension reduction link with adjustment set discovery such that the reduced

set is an adjustment set for covariate adjustment. We focus on studying sufficient dimension reduction [69, 70] for causal effect estimation from data. The significance of this work is that when the size of the adjustment set is reduced, the quantity of data required for the estimation is reduced accordingly, as discussed previously. This work helps solve the first challenge **(I)**, i.e., data insufficiency.

The second piece of work studies local search to discover proper adjustment sets from data with latent variables. Compared to the search for adjustment sets based on global causal structure learning, local causal structure learning can greatly reduce the time complexity of the search for adjustment sets [36] and hence significantly improve the efficiency of causal effect estimation. To the best of our knowledge, there is no literature studying local search to identify proper adjustment sets for causal effect estimation from data with latent variables. This work addresses the third challenge **(III)** and fourth challenge **(IV)** described above.

The third piece of work further improves the performance of local search to determine the proper adjustment sets for a more precise estimation of the causal effect from data with latent variables. In many real-world applications, investigators know some causes (or prior knowledge) of the treatment, and we use such knowledge to improve the quality of causal effect estimation in data. This work addresses challenges (**II**), (**III**), and (**IV**).

The last piece of work in the thesis relaxes the pretreatment variable assumption and obtains *unique* and *unbiased* causal effect estimation from data with latent variables. The work addresses challenges (**II**) and (**III**).

1.2 Contributions

It is challenging to determine which variable should be adjusted for from observational data. The existing data-driven algorithms suffer from four challenges described previously. The problem of finding appropriate adjustment variables is still unsolved when the causal graph is unknown. In this

thesis, we aim to establish several theorems based on graphical causal models to determine the appropriate set of adjustment variables for estimating causal effects from observational data.

Overall, the goal of this thesis is to develop data-driven algorithms to determine proper adjustment sets for estimating causal effects from observational data. The contributions of this thesis are summarized as follows.

- We propose a theorem to show that the adjustment set obtained by sufficient dimension reduction (SDR) is sufficient for controlling confounding bias in causal effect estimation from data without latent variables. To the best of our knowledge, this conclusion has not been reported before. Based on the theorem, we propose a data-driven algorithm *CESD*, causal effect estimator by using sufficient dimension reduction. CESD utilizes kernel dimension reduction [5] to obtain an adjustment set that captures the conditional independence on covariance operators using the *reproducing kernel Hilbert spaces* [71, 72]. The experimental results on two semi-synthetic and three real-world datasets demonstrate the effectiveness of CESD. The experiments also demonstrate that the quality of matching by CESD is better than the quality of matching based on the propensity score. The material in this chapter comes from the literature [73].

- We establish a local search theorem for finding a superset of an adjustment set. The theorem ensures an efficient data-driven algorithm without assuming causal sufficiency in data. To the best of our knowledge, this is the first theorem to support such a local search for estimating causal effects from data with latent variables. Supported by the theorem, we develop an efficient data-driven algorithm, DICE, data-driven causal effect estimation. Experiments on three real-world datasets demonstrate that DICE can outperform state-of-the-art methods. The material in this chapter comes from the literature [74].

- We identify a practical problem setting where unique and unbiased causal effect estimation can be achieved by local search from data with

latent variables. We propose theorems to support the local search for unique and unbiased causal effect estimations from data with latent variables. Based on the proposed theorems, we develop a data-driven algorithm, CE^2LS (causal effect estimation by local search). CE^2LS utilizes the *apriori* approach to reduce the search space of minimal adjustment sets to further improve the search efficiency. CE^2LS is the first practical method based on a local search for finding minimal adjustment sets from data with latent variables. Experiments on both synthetic and real-world datasets demonstrate the high effectiveness and time efficiency of CE^2LS compared to state-of-the-art data-driven methods for causal effect estimation from observational data with latent variables. This chapter is based on the work published in [75].

- To relax the assumption of pretreatment variables in Chapter 6, we develop theorems to support a global search for unique and unbiased causal effect estimation from data with latent variables. Based on the developed theorems, we propose two algorithms, DAVS (data-driven adjustment variable selection) and DAVS-Q (Q is a COSO variable as defined in Definition 7.1), for obtaining unique and unbiased estimation of causal effect from data with latent variables. To achieve a faster search for minimal adjustment sets, both algorithms utilize the *apriori* strategy to reduce the search space of minimal adjustment sets. Experiments show that the two algorithms provide more accurate causal effect estimation than the comparisons that deal with latent variables. This chapter is based on the work published in [76].

1.3 Thesis outline

The rest of the thesis is organised as follows.

- Chapter 2 reviews the literature related to the work in this thesis. We first provide an overview of causal effect estimation by covariate adjustment. Then, we review covariate selection criteria/methods based

on graphical causal models for estimating causal effects from data. In this chapter, we also present a brief overview of causal structure learning algorithms.

- Chapter 3 introduces the background and definitions involved in this thesis, including the definitions of graphical causal modelling frequently used in our contributions, covariate adjustment based on graphical causal modelling, and matching methods.
- Chapter 4 presents a theoretical link of sufficient dimension reduction with the back-door criterion and the CESD algorithm.
- Chapter 5 presents a local search theorem and the DICE algorithm.
- Chapter 6 presents local search theorems for determining proper adjustment sets based on conditional independence tests and the corresponding data-driven algorithm, CE^2LS.
- In Chapter 7, the assumption of pretreatment variables in Chapter 6 is relaxed. We present theorems to support a global search for precise causal effect estimation and the two data-driven algorithms ($DAVS$ and $DAVS$-Q).
- In Chapter 8, we conclude the thesis and discuss future research.

Chapter 2

Literature review

In this chapter, we review the literature that is related to the work in this thesis. First, we provide an overview of the key concepts and research of covariate adjustment for causal effect estimation (a.k.a. treatment effect estimation), followed by the related work on causal effect estimation based on graphical causal modelling. Then, we present a brief introduction to the algorithms for the causal structure learning from observational data.

2.1 Covariate adjustment for Causal Effect Estimation

Estimating treatment effects from observational data is a critical task of causal inference and has become a key technique supporting decision making in various domains [32, 40, 77, 78, 79]. Confounding bias caused by confounders or confounding variables is a major obstacle to estimating treatment effects from observational data. Covariate adjustment, which adjusts for or controls a proper set of covariates (known as an adjustment set), is the main method for removing confounding bias.

Confounders versus covariates. Let us first clarify the two terms: confounders and covariates in causal inference literature. A common cause of treatment W and outcome Y is often considered as a confounder [64, 80], whereas covariates are the factors which are associated with W, or Y or

Table 2.1: A summary of acronyms used in the thesis.

DAG	Directed acyclic graph
CPDAG	Completed partially directed acyclic graph
MAG	Maximal ancestral graph
PAG	Partial ancestral graph
SDR	Sufficient dimension reduction
ACE	Average causal effect
ACT	Average causal effect on treated individuals
HCE	Heterogeneous causal effect
CATE	Conditional average treatment effect

both [6, 54, 66]. Hence a confounder must be a covariate, but a covariate does not have to be a confounder. A proper adjustment set is a set of covariates when conditioning on them is sufficient to remove confounding bias in causal effect estimation using observational data [9, 64].

Causal effects. From observational data, there are two types of causal effects are often estimated, average causal effect (ACE) [13, 22, 23] and heterogeneous causal effect (HCE) [26, 39, 40, 41, 81]. ACE, also known as average treatment effect, refers to the causal effect of treatment W on outcome Y in the whole population on average, i.e., the difference between the average outcome in the treated group and the average outcome in the control group. Estimating ACE is a general method for understanding whether an action (or treatment) should be taken to the population before making a decision or prediction. HCE, also known as conditional average treatment effect (CATE), is the average causal effect in a subpopulation [40]. This thesis is focused on adjustment set identification for ACE estimation.

There are two major causal modelling frameworks, the potential outcome framework [10, 32, 82], and graphical causal models [45, 46, 83] for representing and inferring causal knowledge, including estimating causal effects from data. In the following, we briefly review the two frameworks, focusing on their support for causal effect estimation.

Potential outcome framework. Under the potential outcome framework, given a treatment W, for an individual i, there are two potential

outcomes, $Y_i(1)$ when i receives the treatment and $Y_i(0)$ when i does not receive the treatment. The individual level causal effect is defined as $\delta_i = Y_i(1) - Y_i(0)$. For each individual, only one of the potential outcomes can be observed at a time since each individual can only receive either treatment or control, but not both [22, 32, 51, 58, 84]. The ACE in a population is defined as $\mathbb{E}[\delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$.

Many methods based on the potential outcome framework have been developed to estimate causal effects from observational data. These methods require the unconfoundedness assumption (Definition 3.2, which is given in Chapter 3). In observational data, the distributions between the treated group and the control group are imbalanced since there is the presence of confounders. To remove confounding bias, the most common method is to adjust for differences in measured covariates to make the distributions of the treated group and control group as similar as possible. There are four main approaches to achieve this purpose, including balance score [32, 85, 86], matching methods [22, 58, 87, 88], re-weighting methods [38, 89, 90, 91, 92] and representation-based learning methods [39, 81, 93].

Graphical causal modelling. A graph consists of a set of nodes and a set of edges (The definition of a graph is presented in Chapter 3) [94]. In general, a graph is called a graphical causal model (a.k.a. causal graph) if the "parent" node (i.e., if there exists an edge $W \to Y$ in the graph, then W is a parent node of Y) in the graph is interpreted as "direct causes". In the causal graph, edges represent direct causal relationships between variables. Hence, the causal graph provides a causal interpretation as well.

Graphical causal models are regarded as almost synonymous with DAGs [95]. In this thesis, the graphical causal model is a general term for these four types of graphs: DAG, CPDAG, MAG and PAG.

Graphical causal models are popular languages for representing the causal relationships between variables since they can entail the assumptions of the data-generating process [18, 46]. Such models reflect the mechanism of the measured joint probability distribution [95]. Currently, the research on graphical causal models includes estimating the causal effects (as shown in

the top box of Fig. 2.1) and learning the structure of causal relationships between variables (as shown in the bottom box of Fig. 2.1) [27, 65].

Potential outcome framework and graphical causal models are complementary with different strengths and are suitable for different questions in causal inference [14]. The graphical causal model entails the key assumptions of the data-generating process and is powerful for illustrating conditional independence statements for the covariates involved [57, 96]. The potential outcome framework concerns treatment assignment mechanism [32], which has connected well with experimental design, and estimates causal effect modelling randomized controlled experiments [14]. The connection between a graphical causal model and a graph consisting of potential outcomes is discussed in detail by Pearl [9, 45]. For the relationships of the two frameworks, we refer the readers to the works of [9, 14, 97].

In this chapter, we review the literature on how to determine adjustment sets based on graphical causal models when confounders are not given (i.e., the red dashed box in Fig. 2.1). Graphical causal models have become a basis for identifying adjustment sets since it provides theoretical insights for inferring which variables result in potential bias [48, 49, 50, 55]. We divide the adjustment set identification methods based on graphical causal models into four categories, causal graph given, local causal structure given, no causal structure given, and SDR (refer to the third row in Fig. 2.1). Once we achieve a proper adjustment set, the unbiased causal effect can be estimated by adjusting for such a set.

The work in this thesis is related to the research shown in the red dashed box in Fig. 2.1, i.e., determining proper adjustment sets based on graphical causal modelling for covariate adjustment, to achieve an unbiased estimation of ACE from observational data. In the next section, we will review in detail the research on adjustment set identification based on graphical causal modelling.

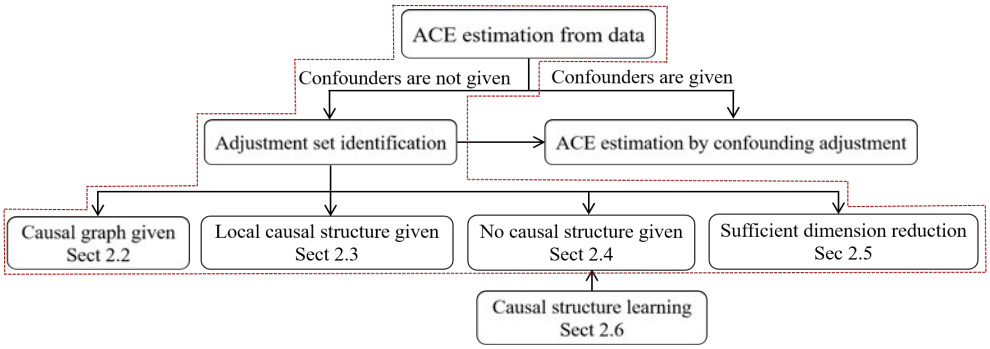


Figure 2.1: An overview of causal inference with observational data based on graphical causal models (with reference to the sections in this chapter discussing the related work), including causal structure learning and adjustment set identification for unbiased causal effect estimation.

2.2 Adjustment set identification given a causal graph

In the last few decades, the research on causal graphical modelling has advanced significantly. One of the major research directions in graphical causal modelling has been the development of the criteria for identifying adjustment sets from a given causal graph. In this section, we review these criteria. We first review the criteria given a causal graph that assumes no latent variables, and then we review the criteria given a causal graph that deals with latent variables. Note that the following provides an overview and less formal discussion of the criteria, and a more formal description of these criteria is given in Chapter 3.

When there are no latent variables, a DAG is often used to represent the causal relationships between variables [94]. When a DAG is given, Pearl's *back-door criterion* [9] is the best-known graphical criteria identifying a proper adjustment set from a given DAG. In a DAG, a back-door path from one node (variable), e.g., the treatment variable W, to another node, e.g., the outcome variable Y, is a path that has an edge pointing to the starting node of the path, e.g., W in the example. Then, with respect to a pair of nodes, e.g., W and Y, a set of nodes (variables) satisfies the back-

door criterion if the set blocks all back-door paths from W to Y. In this case, this set of variables is a proper adjustment set and can be used for covariate adjustment for estimating the ACE of W on Y. For example, in Fig. 1.3, the set $\{C\}$ blocks all back-door paths from W to Y and satisfies the back-door criterion.

The back-door criterion has become the main principle for identifying adjustment sets when a DAG is given. However, the back-door criterion is incomplete w.r.t. covariate adjustment [55], i.e., for some DAGs, the back-door criterion fails to find an adjustment set for yielding unbiased estimation. Thus Shpister et al. [55] proposed the first sufficient and necessary graphical adjustment criterion for covariate adjustment when a DAG is given. The adjustment criterion requires that (1) the descendant of W which lies on a proper causal path [1] from W to Y is forbidden to be included in any adjustment set, and (2) the adjustment set blocks all non-causal paths from W to Y. For example, a DAG in Fig. 1.5, there is a causal path from U_1 to W, and a non-causal path from U_1 to W, i.e., $U_1 \to C \leftarrow U_2 \to Y \leftarrow W$. For the DAG in Fig. 1.5, $\{C\}$, $\{U_2\}$, or $\{Y\}$ satisfies the adjustment criterion w.r.t. the pair (U_1, W). The adjustment criterion implies that adjusting for all ancestors of W and Y that do not lie on causal paths from W to Y is able to remove all confounding bias.

When a system contains latent variables, ancestral graphs are used to represent the causal relationships between observed variables. Ancestral graphs are an extension of DAGs (the formal definition is refer to Chapter 3) [98, 99]. In an ancestral graph, the bi-directed edge $\leftrightarrow$ is used to represent the existence of latent variables. In this thesis, we use the MAG, which is one of the most commonly used ancestral graphs for representing the causal relationships between measured variables [98]. For determining proper adjustment sets from a MAG, Maathuis and Colombo [48] generalized Pearl's back-door criterion to MAG, i.e., generalized back-door criterion. In a MAG, a back-door path from the treatment variable W to the outcome variable Y, is a path if it does not have a visible edge out of W (the

[1] A path from W to Y that forms exclusively of directed arrows out of W is called a causal path, and all other types of paths are called non-causal paths.

visible edge in a MAG is defined in Definition 3.12). A set $\mathbf{Z}$ satisfies generalized back-door criteria w.r.t. (W, Y) if (1) $\mathbf{Z}$ does not contain descendants of W, and (2) $\mathbf{Z}$ blocks all back-door paths from W to Y. The main difference between the generalized back-door criterion and Pearl's back-door criterion is the back-door path, i.e., $W \rightarrow V \cdots \rightarrow Y$ is a back-door path in a MAG if $W \rightarrow V$ is invisible. With respect to covariate adjustment, the generalized back-door criterion is not complete [49].

Van der Zander et al. [100] developed a sound and complete graphical adjustment criterion (a.k.a. constructive back-door criterion) for determining minimal adjustment sets when the MAG is given. The constructive back-door criterion reduces the problem of determining adjustment sets to the problem of searching for m-separating sets (The definition of m-separation is given in Definition 3.7). However, the conclusion assumed the soundness and completeness of the adjustment criterion for DAGs provided in [55]. For providing a generalized graphical adjustment criterion for a MAG, Perković et al. [49] presented A sound and complete criterion for a MAG, i.e., generalized adjustment criterion. The generalized adjustment criterion clarifies the conditions of proper adjustment sets when the MAG is given. The formal definition can be found in Chapter 3.

2.3 Adjustment set identification with local structure knowledge

In practice, the complete causal diagram is rarely known [7, 84]. Therefore, some works have studied how to select adjustment sets when knowledge is available about the local causal structure around the treatment variable W and/or the outcome variable Y. This type of method often makes the assumption that either all causes of W and/or Y are known.

When local structures around W and Y are given, and under the assumptions described above, the following three strategies are presented in the literature to select adjustment variables for confounding control. The first strategy is to include variables that are the direct causes of Y, i.e., risk

factors of Y are sufficient to eliminate confounding bias [101]. This strategy only requires knowledge of the causes of Y. The second strategy, suggested by Glymour et al., is to adjust for all common causes of W and Y [102]. The set of all common causes of W and Y can be regarded as a confounder set, which is a minimal adjustment set for removing confounding bias. The third strategy, proposed by VanderWeele and Shpitser [54] is a disjunctive causal criterion that suggests adjusting for either causes of W or causes of Y or both. All three criteria require knowing the local causal structure around W and/or Y.

2.4 Adjustment set identification without causal structure given

In real-world applications, causal graphs or local causal structures are rarely available. Therefore, we have to resort to causal structure learning (reviewed in section 2.6) to learn the causal structure from data and then apply the graphical criteria to identify the adjustment set based on the learned causal structure. Such a data-driven approach is non-trivial, as from data, only a Markov equivalence class of causal graphs can be learned, and the situation is complicated by other issues such as latent variables and high-dimensional data. Despite the difficulties, in recent years, some data-driven methods have been developed for adjustment set identification and causal effect estimation, although so far, only a handful of such methods exist. In the following, we will review these methods.

IDA (intervention when the DAG is absent) [62] is the pioneering work on data-driven causal effect estimation. IDA assumes that there are no latent variables in a dataset. It first learns a Markov equivalence class of DAGs (i.e., CPDAG) from data using the PC algorithm [18]. Then, for each DAG, to identify all possible adjustment sets for estimating all possible causal effects from data, the IDA output is a bound estimation of causal effects, as different DAGs in the equivalence class give different adjustment sets, resulting in a multi-set of estimates of the causal effect. Nandy et

al. [103] proposed generalized IDA to handle multi-treatments (joint interventions) and an outcome, referred to as Joint-IDA. Joint-IDA relies on the assumption that only the parent sets of treatment nodes (OPIN) are given. It also needs to recover a CPDAG of the underlying causal DAG from the data, and then possible joint parent sets of the treatment nodes are extracted from the learned CPDAG. Finally, a semi-local algorithm is developed to obtain jointly valid parent sets from the learned CPDAG and produces a multi-set of joint causal effects. Following the IDA frameworks, Fang and He [104] developed an IDA method with background knowledge, referred to as NIDA. The authors proposed a locally valid orientation rule in a maximal PDAG (partially directed acyclic graph) which is usually utilized to represent background knowledge. Based on the obtained maximal PDAG, the authors developed a fully local method to estimate all possible causal effects. NIDA can be regarded as a constrained version of the IDA algorithm since it requires background knowledge.

Another line of work requires unconfoundedness (Definition 3.2) and explores data-driven methods for covariate adjustment. De luna et al. [57] proposed minimal adjustment set selection criteria (MASS) when all covariates can be utilized as a valid adjustment set, under the pretreatment variable assumption. MASS criteria are sufficient under the unconfoundedness assumption. MASS criteria provide four minimal adjustment sets from the set of pretreatment variables, i.e., all causes of W, all causes of Y, all causes of W excluding causes of Y and all causes of Y excluding causes of W. When the knowledge of the underlying causal graph is not available for the use of the MASS criteria, causal structure learning methods are necessary for aiding the process of covariate selection. Hence, Häggström [7] proposed utilizing estimated causal graphs to select covariates as proper adjustment sets. The author developed a data-driven covariate selection algorithm (*CovSel* for short) to estimate ACE from data. CovSel employs a causal structure learning method, MMHC (max-min hill-climbing, introduced in Subsection 2.6.3) algorithm [105], to recover a causal graph from data, and then, in conjunction with the MASS criteria to select the adjustment sets that satisfy the MASS criteria for uniquely estimating the causal

effect of W on Y from data. However, MMHC cannot guarantee the correctness of the recovered underlying DAG since MMHC is based on a greedy steepest-ascent search for orienting the edges in the candidate graph, which is not a sound search method for global causal structure learning [36].

For datasets with latent variables, several methods have been developed to identify adjustment sets to obtain bound estimation or unique estimation of causal effects. CE-SAT and LVIDA are the methods that output bound estimation of causal effects. CE-SAT [67] uses logic representation and SAT-based inference to estimate causal effects from data with latent variables. CE-SAT can only deal with very small datasets and is a bound estimation of causal effects. It firstly learns a Markov equivalence class of MAGs (i.e., PAG) from data using the FCI algorithm [18]. Then it employs the d-separation (Definition 3.5) constraints of the PAG into a logical representation by using the ASP-solver [106] to obtain graphs from the PAG. For each graph in the Markov equivalent class, CE-SAT utilizes do-calculate [9, 107] to identify the valid causal effects. The CE-SAT is complete w.r.t. the identifiability. But the time complexity of CE-SAT consists of two parts, learning a PAG and ASP-solver, and both factors are very time-consuming, especially for a large number of variables. Thus, CE-SAT is impractical for high-dimensional data due to the fact that it needs to perform a complete search for solutions [108]. LVIDA (latent variable IDA) [63] employs a constraint-based algorithm, FCI (fast causal inference) [18] or RFCI (really fact causal inference) [35], to recover a Markov equivalence class of MAGs (i.e., a PAG) from data with latent variables. Then for each MAG (enumerated from the learned PAG), LVIDA obtains all possible generalized back-door sets and calculates all possible causal effects conditioning on the obtained all possible generalized back-door sets. Hence, it can only produce a bound estimation of causal effects and return a multi-set of estimated causal effects. Moreover, the performance of LVIDA is contingent on the correctness of the learned PAG. Learning a PAG from data and enumerating all Markov equivalent MAGs from the learned PAG contribute to the time complexity of LVIDA, and the latter significantly affects the efficiency

of LVIDA. Hence LVIDA is not scalable to large datasets with dozens or more variables.

The EHS and GAC methods can produce a unique estimation of causal effects from data with latent variables. Entner et al. [66] proposed a data-driven algorithm (i.e., EHS algorithm for the authors Entner, Hoyer, and Spirtes) to search for adjustment sets directly from data with latent variables. The EHS algorithm employs conditional independence tests to determine an adjustment set for estimating causal effects of W on Y from observational data with latent variables. However, EHS is very inefficient since it conducts an exhaustive search over all the combinations of variables for the conditional independence tests. Moreover, the adjustment sets identified by EHS can be large, containing redundant variables, which can result in inaccurate estimation of causal effects, especially when the number of samples in a dataset is not large enough. The GAC (a data-driven method based on the generalized adjustment criterion) method [49] can estimate a unique causal effect from data with latent variables. GAC needs to recover a PAG from data and obtain proper adjustment sets from the recovered PAG according to the generalized adjustment criterion. However, in practice, due to insufficient data and/or unreliable statistical tests, the PAG learned from the data may have errors in edge orientation and thus may not correctly represent the underlying MAG. That is, the performance of GAC depends on the accuracy of the underlying PAG search. If an inconsistent PAG is applied, the causal effect estimation will be biased. An inconsistent PAG learned from data results in low GAC performance as the experiments shown in Section 7.4.1 of Chapter 7.

To summarize, we reviewed the prior works based on graphical causal modelling, but the performances of these methods suffer from low efficiency since they rely on global causal structure learning or global search strategy in causal effect estimation. Another important issue in causal effect estimation is latent variables. Effectively and efficiently estimating causal effect from data with latent variables remains an open problem.

2.5 Sufficient dimension reduction for ACE estimation

When estimating causal effects from observational data, a large-sized adjustment set easily causes data insufficiency. If an adjustment set has a large number of variables, a very large dataset is needed for obtaining a reliable estimation. To mitigate the problem of data insufficiency caused by large-sized adjustment sets, SDR has been studied in estimating causal effects from data [60, 90, 109, 110]. The application of SDR in causal effect estimation requires some common assumptions, i.e., the absence of latent variables and all measured variables are pretreatment variables [1]. In the following, we first give a summary of SDR and then review several important works based on SDR in causal effect estimations from observational data.

SDR learns a low-dimensional set of the original covariates which explains the response (target) variable of interest, such as the variable W, as effectively as possible [69, 70, 111, 112]. Many SDR methods have been proposed to address various problems caused by the high dimensionality (a.k.a. curse of dimensionality) and form a family of methods, including sliced inverse regression (SIR) [113], estimating the central solution space [114] and kernel dimension reduction [111]. The central task of SDR is to learn an effective subspace for transforming the space of the original variables into a low-dimensional set. SDR can be formulated as conditional independence, i.e., $W \perp\!\!\!\perp \mathbf{X} | \mathbf{\Psi}^T \mathbf{X}$ in which $\mathbf{X}$ denotes the set of original variables and $\mathbf{\Psi}$ is projection matrix [70]. The detailed introduction of SDR is presented in Chapter 4.

SDR has been used in causal effect estimation for developing semiparametric estimators in finite samples [90, 109, 110]. Luo et al. [109] proposed a matching method (MSDR) for estimating causal effects based on SDR. MSDR applies SDR twice in a dataset to construct two low-dimensional sets, one is for the sub-dataset of the treatment samples, and the other is for the

[1] Pretreatment variables are variables not affected by the treatment and the outcome.

sub-dataset of the control samples. Then the learned two low-dimensional sets are used as the adjustment sets (a.k.a. balance score in causal inference [32]) for matching. When the number of samples in a dataset is small, dividing it into two sub-datasets will reduce the performance of the SDR method [112]. Thus, MSDR is not suitable for small datasets. Ghosh et al. [110] proposed a novel estimator based on SDR (named Shrinkage) for average causal effect estimation from data. Shrinkage considers the efficient semiparametric SDR methods in the model of estimating propensity score and the model of predicting the outcome. Then shrinkage takes the estimated propensity scores and outcomes into a classical inverse probability weighting estimator. When both models are inconsistent, the performance of shrinkage is poor. Liu et al. [90] proposed a semiparametric approach (SDRE for short) for estimating ACE from data. SDRE adopted SDR to learn the propensity score weights and then use the inverse probability weighted estimator to estimate ACE from data. The disadvantage of SDRE may be prone to propensity score model misidentification by SDR. Ma et al. [60] proposed a robust and efficient estimator (SSDRE for short) for estimating ACE from data. SSDRE utilized a sparse SDR method to estimate the propensity score and recover the outcome model simultaneously.

However, the above-reviewed methods based on SDR theoretically lack the back-door criterion, i.e., covariate adjustment. That is, in these cases, we do not know whether the applied SDR can theoretically control for confounding bias.

Table 2.2 provides a summary of the above-reviewed data-driven methods. We group the methods according to whether the method can handle latent variables, and whether it is a bound estimation of causal effects.

Overall, estimating causal effects from observational data is an important and challenging task, especially for data with latent variables. There are not many data-driven methods for causal effect estimation based on graphical causal models, and they are either bound estimates or low time efficiency for estimating ACE from data. In this dissertation, efficient data-driven algorithms are developed to determine adjustment sets based on

Table 2.2: A summary of the data-driven causal effect estimation methods.

Estimate	Unique	Latent	Graph based
IDA[62]	No	No	Yes
Joint-IDA [103]	No	No	Yes
NIDA [104]	No	No	Yes
CE-SAT [67]	No	Yes	Yes
LVIDA[63]	No	Yes	Yes
CovSel [7, 57]	Yes	No	Yes
GAC[49]	Yes	Yes	Yes
EHS[66]	Yes	Yes	Yes
MSDR [109]	Yes	No	No
Shrinkage [110]	Yes	No	No
SDRE [90]	Yes	No	No
SSDRE [60]	Yes	No	No

graphical causal models for estimating causal effects from observational data.

2.6 Causal structure learning from data

Causal structure learning algorithms play an important role in estimating causal effects from observational data since the underlying causal graph is rarely available and has to be learned from data. A brief of the main structure learning algorithms, including constraint-based algorithms [36], score-based methods [115, 116] and hybrid methods [105, 117] is presented.

2.6.1 Constraint-based structure learning methods

Constraint-based methods learn the causal structure by performing independence and dependence constraints to shrink the candidate graphs that may have generated the data [18]. Conditional independence tests [118] play a crucial role in constraint-based structure learning since constraint-based methods rely on conditional independence tests on a given dataset to discover the underlying causal structure.

The SGS (Spirtes-Glymour-Scheines) algorithm is the first constraint-based algorithm developed by Spirtes et al. [18]. The central idea of SGS is to delete edges one by one from a complete undirected graph based on a series of conditional independence tests. However, SGS is prohibitively inefficient since each pair of variables needs to use the elements in the power set of the remaining variables as the conditioning sets. The time complexity of SGS is exponential to the number of variables. SGS can only deal with data containing a small number of variables.

The PC (Peter-Clark) algorithm is an advanced version of the SGS algorithm, designed by Peter Spirtes and Clark Glymour [18] to reduce the time complexity of SGS. The PC algorithm also starts with a complete undirected graph and then obtains the causal structure by deleting edges when the pair of variables are conditionally independent given a conditioning set. The major difference between SGS and PC is that PC conducts the conditional independence tests in a certain order to reduce the overall number of tests.

SGS and PC algorithms require that all variables are observed, so when a system involves latent variables, both algorithms may produce the wrong structures. Inductive causation with latent variables (IC) [119] can learn the causal structure from data with latent variables [9]. It extends the definition of DAG to represent the causal structure with the existence of latent variables by using bi-directed edges. The bi-directed edge between two variables denotes that there exists a latent variable between the two variables.

Inspired by the idea of the IC algorithm, FCI [18] and RFCI [35] were proposed to learn the underlying causal structure from data with latent variables. Both algorithms recover a Markov equivalence class PAG of a MAG from data. The initial skeleton learning of both FCI and RFCI is the same as the PC algorithm, and both algorithms perform additional conditional independence tests to determine the latent variables in the learned graph.

2.6.2 Score-based structure learning methods

In addition to constraint-based causal structure learning, there are two other types of methods for causal structure learning, score-based and hybrid methods. Score-based methods attempt to maximize a scoring function to obtain the causal graph from data. Low scores should be given to the graphs which contain the incorrect edges or orientations. In detail, a search space is defined for all possible causal graphs. Then, a scoring function is utilized to assess the candidate graphs. Finally, the candidate causal graph is selected if the causal graph best fits the data.

Many scoring functions have been developed to evaluate the candidates [105, 120], such as the Bayesian information scoring (BIC) [121], K2 scoring function [120], the BDeu scoring function [122], and the BDe scoring function. The BIC score is the most widely used metric that maximizes the likelihood of measured variables by regularization of the number of parameters and the sample size. However, learning the causal structure by maximal score is NP-complete [68]. Therefore, many methods propose to achieve a local optimum through heuristics, such as greedy equivalence search (GES) [123] and its extension fast GES [124].

Recently, some works based on deep learning have been developed to learn the global causal structure from data [116, 125, 126, 127]. These methods belong to score-based methods and have advantages from representation learning [128, 129]. Additive noise models can sometimes uniquely recover the causal DAG from data with some strong assumptions [130, 131, 132, 133]. We do not consider additive noise models since their assumptions are quite different from what we consider in this thesis and belong to another framework. Because they are not directly related to this thesis, we do not review them here and refer the readers to the survey [134].

2.6.3 Hybrid structure learning methods

Hybrid methods adopt the strategies of both constraint-based and score-based methods to learn the underlying causal structure from data. Most hybrid methods first perform a constraint-based method to reduce the search

space and then employ a scoring function to pick up a candidate who is the best fit for the data. The max-min hill-climbing (MMHC) algorithm [105] is the most representative hybrid method, which combines the ideas of local causal structure learning (constraint-based) and score-based methods. In detail, MMHC first recovers an undirected structure, then a greedy-based Bayesian scoring with the hill-climbing search is used to orient the undirected edges for recovering the causal graph. MMHC is one of the best effective causal structure learning methods [7].

2.6.4 Local causal structure learning methods

Although many works have been proposed to learn a global causal structure from data, the efficiency of learning causal structure is still an open problem in causal inference. Instead of learning a global causal structure, a few works have proposed to learn a local causal structure [25, 34, 36].

Local causal structure learning methods discover the local causal relationships of a fixed target variable (denoted as Y) in a causal DAG, i.e., a part of the causal DAG. Generally speaking, local causal structure refers to the parents and children set of Y in the causal DAG, denoted as $PC(Y)$. PC-Select [34] can learn the causal structure from high-dimensional data, which is a simple version of the PC algorithm. The main idea of PC-Select is the same as the PC algorithm, i.e., PC-Select performs conditional independence tests to detect the conditional dependence between a pair of nodes. Instead of recovering a global causal structure, PC-Select discovers $PC(Y)$. In this dissertation, PC-Select (a.k.a. PC-Simple) [34] is employed to discover the local causal structure from data.

Another local causal structure is the Markov blanket of Y (denoted as $MB(Y)$), which is a minimum size of the conditional set, i.e., conditioning on $MB(Y)$, all other variables become independent of Y [36]. In a causal DAG, $MB(Y)$ contains $PC(Y)$ and the spouses of Y (two variables are spouses if they share a common child node). Many works have been developed to learn the local causal structure, such as incremental association Markov blanket (IAMB) [135], max-min parents and children (MMPC) for

discovering Markov blanket (MMMB) [136], HITON-MB [25], and iterative parent and children for determining Markov blanket (IPC-MB) [36] . These methods attempt to discover $MB(Y)$.

2.7 Summary

In this chapter, we reviewed the works related to the work in this thesis. We focus on reviewing principles and methods based on graphical causal modelling in causal effect estimation. In addition, we also reviewed some main structure learning algorithms, some of which are directly applied in this work.

SDR has been used to address the large-sized adjustment set in causal effect estimation from data. The existing methods focus on using SDR to obtain a low-dimensional set for building models on either the treatment variable W, or the response variable Y, or both. However, a theoretical link between SDR and covariate adjustment is lacking.

In many applications, the causal graph is rarely known and local causal knowledge is not available. Recovering a Markov equivalent class of the underlying causal graph is needed for causal effect estimation, and based on the recovered causal graph represented in the Markov equivalent class, different adjustment sets can be determined. Additionally, the existing methods suffer from low time efficiency since they rely on global causal structure learning or global search strategy.

Motivated by these challenges, in this thesis, a set of theorems based on graphical causal modelling is proposed to support finding small-sized adjustment sets. Based on the proposed theorems, data-driven algorithms are developed to estimate causal effects from observational data.

Chapter 3

Preliminaries

The objective of the thesis is to identify proper adjustment sets for unbiased causal effect estimation from observational data under the framework of graphical causal modelling. In this chapter, first, the background and definitions of graphical causal modelling are introduced. Then, we introduce covariate adjustment based on graphical causal modelling. Finally, we provide a brief introduction to the matching methods for estimating causal effects using identified adjustment sets.

3.1 Graphical causal models

In this section, we introduce two major types of causal graphs used in graphical causal modelling: causal DAG and causal MAG.

Throughout the thesis, we use an upper-case letter to represent a variable, a lower-case letter to indicate the value of a variable, and a boldfaced upper-case letter to denote a set of variables. Moreover, we use a letter in calligraphic font to denote a graph.

3.1.1 Causal DAG

When there are no latent variables, a causal DAG is often used to encode the causal relationships between variables.

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a graph, which consists of a set of nodes $\mathbf{V} = \{V_1, \ldots, V_p\}$ representing the variables in a problem domain and a set of edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ representing the relationships between the variables. $\mathcal{G}$ is a directed graph if each edge in $\mathbf{E}$ is directed or has an orientation, e.g., $V_i \to V_j$ is a direction pointing from V_i to V_j. A DAG, i.e., acyclic directed graph, is a directed graph without cycles (i.e., no path starting from a node V and then pointing back to V). A causal DAG is a DAG in which each edge represents a direct causal relationship. For example, the edge $V_i \to V_j$ in a causal DAG indicates that V_i is a direct cause of V_j, and V_j is a direct effect of V_i.

In the following, we describe in detail some of the terms or concepts related to a (causal) DAG.

In a graph, a *path* π between V_i and V_j is a sequence of edges that connects a sequence of distinct nodes, denoted as $\pi = \langle V_i, \ldots, V_j \rangle$, where V_i and V_j are end points of the path. The number of edges included in a path is called the length of the path. In a DAG, a *directed path* from V_i to V_j is a path that starts from V_i and all edges on the path are directed towards V_j, i.e., $V_i \to \cdots \to V_j$. In a causal DAG, a direct path from V_i to V_j is named a *causal path* from V_i to V_j, and non-directed paths are non-causal paths.

In a DAG, if we have $V_i \to V_j$, then V_i is a parent of V_j, and V_j is a child of V_i. Two nodes are adjacent if there is an edge between them. If there exists a directed path from V_i to V_j, then V_i is an *ancestor* of V_j, and V_j is a *descendant* of V_i. In a DAG, an intermediate node V_k is a collider or a V-structure is formed at V_k if there exists $V_i \to V_k \leftarrow V_j$, and the two nodes V_i and V_j are married to each other since they share a child node V_k. $Pa(V), Ch(V), Adj(V), An(V), De(V)$ and $Sp'(V)$ denote the sets of all parents, children, adjacent nodes, ancestors, descendants and spouses of V in a DAG, respectively.

Now, we introduce a central concept used in graphical causal modelling, conditional independence among variables [9].

Definition 3.1 (Conditional independence)**.** *Two variables V_i and V_j are conditionally independent given another set of variables $\mathbf{Z}$, denoted as $V_i \perp\!\!\!\perp$*

$V_j \mid \mathbf{Z}$ *if and only if* $prob(V_i = v_i, V_j = v_j \mid \mathbf{Z} = \mathbf{z}) = prob(V_i = v_i \mid \mathbf{Z} = \mathbf{z}) * prob(V_j = v_j \mid \mathbf{Z} = \mathbf{z})$ *for all values of* V_i, V_j, *and* $\mathbf{Z}$, *with* $prob(\mathbf{Z} = \mathbf{z}) > 0$.

In graphical causal modelling with DAGs, the following conditions or assumptions are often involved.

Definition 3.2 (Markov condition [137]). *Given a DAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *and the joint probability distribution of* $\mathbf{V}$. *We say that* $prob(\mathbf{V})$ *satisfies the Markov condition with respect to* $\mathcal{G}$ *if, for* $\forall V_i \in \mathbf{V}$, V_i *is conditionally independent of all non-descendants of* V_i, *given the set* $Pa(V_i)$ *in* $\mathcal{G}$.

When the Markov condition is satisfied, the joint probability distribution $prob(\mathbf{V})$ can be factorised as $prob(\mathbf{V}) = \prod_i prob(V_i|Pa(V_i))$.

Definition 3.3 (Causal sufficiency [18]). *Given a dataset* $\mathbf{D}$ *with a set of random variables* $\mathbf{V}$. $\mathbf{D}$ *satisfies causal sufficiency if for every pair of measured variables in* $\mathbf{V}$, *all their common causes are also measured and in* $\mathbf{V}$.

The assumption of causal sufficiency requires that all common causes of the measured variables in the underlying data-generating mechanism are also measured.

Definition 3.4 (Faithfulness [18]). *A DAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *is faithful to a joint probability distribution* $prob(\mathbf{V})$ *if and only if every independence present in* $prob(\mathbf{V})$ *is entailed by* $\mathcal{G}$ *and the Markov condition. A joint distribution* $prob(\mathbf{V})$ *is faithful to a DAG,* $\mathbf{E}$*) if and only if there exists a DAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, *which is faithful to the joint distribution* $prob(\mathbf{V})$.

The Markov condition and the faithfulness assumption establish a close relationship between a causal DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and the joint probability distribution $prob(\mathbf{V})$ such that we can read all the independence relationships in $prob(\mathbf{V})$ from $\mathcal{G}$.

A set of DAGs that encode the same set of conditional independence relationships is known as a Markov equivalence class. A CPDAG [27, 138] can represent a Markov equivalence class, and a CPDAG can be learned

from data by using the structure learning algorithms, e.g., the PC algorithm described in Chapter 2.

d-separation is a graphical criterion that enables the identification of conditional independence between variables entailed in a DAG [9].

Definition 3.5 (d-separation)**.** *A path π in a DAG is said to be d-separated (or blocked) by a set of nodes $\mathbf{Z}$ if and only if*

1. *π contains a chain $V_i \to V_k \to V_j$ or a fork $V_i \leftarrow V_k \to V_j$ such that the middle node V_k is in $\mathbf{Z}$, or*

2. *π contains a collider V_k such that V_k is not in $\mathbf{Z}$ and no descendant of V_k is in $\mathbf{Z}$.*

A set $\mathbf{Z}$ is said to d-separate V_i from V_j if and only if $\mathbf{Z}$ blocks every path between V_i and V_j.

3.1.2 Causal MAG

The assumption of causal sufficiency is often unwarranted in most applications since there is not a closed world and latent variables exist [98, 99]. *Ancestral graphs* have been developed to represent the underlying causal mechanism of a system where latent variables and selection variables exist [98]. We assume that there are no selection variables since selection variables usually block causal effect identification by covariate adjustment alone [49, 100, 139].

In this thesis, we employ a type of ancestral graph, i.e., MAG, for adjustment set identification. Therefore, in the following, we introduce the definitions and concepts related to ancestral graphs and MAG. Since ancestral graphs are one type of (directed) mixed graph, we start our introduction with mixed graphs, following the definitions and notations in [4, 49].

Edges in a graph $\mathcal{G}$ can be *directed* $\to$, *bidirected* $\leftrightarrow$, *non-directed* $\circ\!\!-\!\!\circ$ or *partially directed* $\circ\!\!\to$. We use "*" to denote an arbitrary edge mark. For example, $*\!\!\to$ can represent $\to$, $\leftrightarrow$, or $\circ\!\!\to$. A *mixed graph* [4, 49] may contain directed edges $\to$ and bidirected edges $\leftrightarrow$. A *partial mixed graph*

may include any type of edge. Therefore, a mixed graph is a special type of partial mixed graph.

In a partial mixed graph, a *possibly directed path* from V_i to V_j is a path from V_i to V_j that does not contain an arrowhead pointing in the direction of V_i. In a partial mixed graph, a *possibly causal path* from V_i to V_j is a possibly directed path from V_i to V_j. If a path from V_i to V_j is not a possibly causal path, then it is called a *non-causal path* between V_i and V_j.

In a partial mixed graph, if there exists a *possibly directed path* from V_i to V_j, V_i is a *possible ancestor* of V_j, and V_j is a *possible descendant* of V_i. In a partial mixed graph, if $V_i \leftrightarrow V_j$, V_i and V_j are married. Note that this is different from the concept of spouses in a DAG. Two nodes in a partially mixed graph are adjacent if there is an edge (of any type) between them. We use $PossAn(V)$, $PossDe(V)$, $Sp(V)$ and $Adj(V)$ to denote the sets of all possible ancestors, possible descendants, spouses and adjacent nodes of V, respectively. "Possible" here means uncertain, e.g., if there exists $V_i \circ\!\!\rightarrow V_k \rightarrow V_j$, then we have $V_i \in PossAn(V_j)$ and $V_j \in PossDe(V_i)$ since there exists the edge $V_i \circ\!\!\rightarrow V_k$.

Unlike that in a DAG, in a partial mixed graph, a node V_i is a collider if there exists $V_k *\!\!\rightarrow V_i \leftarrow\!\!* V_l$ in the partial mixed graph. A *collider path* is a path on which every non-endpoint node is a collider. A path π with a length of one is a *trivial collider path*. A subpath $< V_i, V_j, V_k >$ is an (un)*shielded triple* if V_i and V_k are (not) adjacent. A path π is *unshielded* if all successive triples on π are unshielded [49]. A node V_j is called a *definite non-collider* on π if there exists at least an edge out of V_j on π or $V_i \not*\!\!\rightarrow V_j \leftarrow\!\!\not* V_k$ is an unshielded triple [4]. A node on a path π is of *definite status* if it is a collider or a definite non-collider on π. A collider on a path π is referred to as a *definite collider* since it is always of definite status. A path π is of *definite status* if every non-endpoint node on π is of definite status [49].

In a partial mixed graph, as illustrated in Fig. 3.1 (a), a directed cycle is a path that starts and ends at the same node, and all edges on the path are directed edges pointing in the same direction, whereas an *almost directed cycle* consists of a directed path from V_i to V_j together with $V_i \leftrightarrow V_j$, as illustrated Fig. 3.1 (b). Now we present the definition of an ancestral graph.

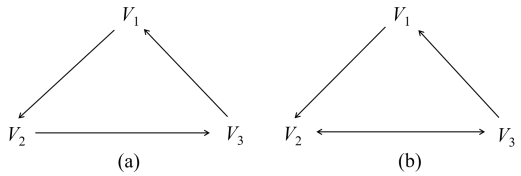


Figure 3.1: (a) The directed path $V_3 \to V_1 \to V_2$ together with the edge $V_2 \to V_3$ form a directed cycle, (b) the directed path $V_3 \to V_1 \to V_2$ together with the edge $V_2 \leftrightarrow V_3$ form an almost directed cycle.

Definition 3.6 (Ancestral graph [98])**.** *A mixed graph $\mathcal{G}$ is an ancestral graph if*

1. *$\mathcal{G}$ does not contain directed cycles or almost directed cycles, and*
2. *when there exists an edge $V_i \leftrightarrow V_j$ in $\mathcal{G}$, then $\mathcal{G}$ must not contain the directed edges $V_i \to V_j$ or $V_i \leftarrow V_j$.*

Richardson and Spirtes [98] extended d-separation w.r.t. a DAG to the following defined m-separation w.r.t. an ancestral graph.

Definition 3.7 (m-separation [35, 98])**.** *In an ancestral graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, a path π between V_i and V_j is said to be m-separated by a set of nodes $\mathbf{Z} \subseteq \mathbf{V} \setminus \{V_i, V_j\}$ (possibly $\varnothing$) if*

1. *π contains a subpath $\langle X_l, X_k, X_s \rangle$ such that the middle node X_k is a non-collider on π and $X_k \in \mathbf{Z}$;*
2. *π contains $X_l \,{*\!\!\to}\, X_k \leftarrow {*\!\!}X_s$ such that $X_k \notin \mathbf{Z}$ and no descendant of X_k is in $\mathbf{Z}$.*

Two nodes V_i and V_j are said to be m-connected by $\mathbf{Z}$ in $\mathcal{G}$ if V_i and V_j are not m-separated by $\mathbf{Z}$.

Based on the definition of m-separation and ancestral graphs, we have the following definition of an MAG.

Definition 3.8 (MAG [98])**.** *An ancestral graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is referred to as a MAG if for every pair of non-adjacent nodes V_i and V_j in $\mathcal{G}$, there exists $\mathbf{Z} \subseteq \mathbf{V} \backslash \{V_i, V_j\}$ such that $\mathbf{Z}$ m-separates V_i and V_j.*

From a DAG with a set of observed variables $\mathbf{O}$ and a set of unobserved variables $\mathbf{U}$, following the construction rule specified in [4], one can construct an MAG with nodes $\mathbf{O}$ such that all the conditional independence relationships among the observed variables entailed by the DAG are entailed by the MAG and vice versa, and the ancestral relationships in the DAG are maintained in the MAG.

If two MAGs represent the same set of conditional independencies among measured variables, they are called *Markov equivalent*, and formally, we have the following definition.

Definition 3.9 (Markov equivalent MAGs [140]). *Two MAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ with the same nodes are said to be* Markov equivalent*, denoted $\mathcal{G}_1 \sim \mathcal{G}_2$, if for all triple nodes A, B, C, A and B are m-separated by C in $\mathcal{G}_1$ if and only if A and B are m-separated by C in $\mathcal{G}_2$.*

The set of all MAGs that encode the same set of conditional independence forms a *Markov equivalence class* [18, 140]. A set of Markov equivalent MAGs can be represented uniquely by a PAG.

Definition 3.10 (PAG [4]). *Let $[\mathcal{G}]$ be the Markov equivalence class of an arbitrary MAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. The PAG $\mathcal{P} = (\mathbf{V}, \mathbf{E})$ for $[\mathcal{G}]$ is a partial mixed graph such that*

1. *$\mathcal{P}$ has the same adjacencies as $\mathcal{G}$ does,*
2. *an arrowhead mark (same as the tail mark) is in $\mathcal{P}$ if and only if it is shared by all MAGs in $[\mathcal{G}]$.*

Following [35], we define the possible d-separation set, denoted as $pds(W, Y, \mathcal{G})$.

Definition 3.11 ($pds(W, Y, \mathcal{G})$ [35]). *Given a PAG or MAG $\mathcal{G}$ and a pair of variables (W, Y) in $\mathcal{G}$, $X \in pds(W, Y, \mathcal{G})$ if and only if there is a path π between V and W in $\mathcal{G}$ such that for every subpath $< V_l, V_k, V_h >$ on π, either V_k is a collider or $< V_l, V_k, V_h >$ is a triangle, i.e., each pair of nodes in the triple are adjacent.*

We distinguish two types of directed edges in an MAG/PAG by the following visibility definition.

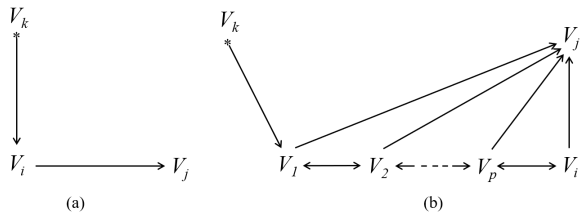


Figure 3.2: Two possible configurations in MAGs and PAGs for a visible edge $V_i \to V_j$ [4]. Nodes V_k and V_j must be non-adjacent.

Definition 3.12 (Visibility [4])**.** *Given a PAG or an MAG $\mathcal{G}$, a directed edge $V_i \to V_j$ is visible if there is a node V_k not adjacent to V_j, such that either there is an edge between V_k and V_i that is in V_i or there is a collider path between V_k and V_i that is in V_i and every node on this path is a parent of V_j. Otherwise, $V_i \to V_j$ is said to be invisible.*

It is worth noting that as a DAG is a special type of MAG (when there are no latent variables), the above definition of visibility also applies to a DAG. Fig. 3.2 shows two possible graphical configurations that can result in a visible edge $V_i \to V_j$ in an MAG or a PAG.

3.2 Covariate adjustment based on graphical causal modelling

In this section, we introduce the definition of covariate adjustment for causal effect estimation based on graphical causal modelling.

3.2.1 Covariate adjustment for causal effect estimation

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a graph, and $\mathbf{V} = \{W, Y\} \cup \mathbf{X}$, where W is a binary treatment variable ($W = w$ ($w \in \{0, 1\}$)), where $w = 0$ indicates receiving treatment, $w = 1$ represents receiving the control, Y denotes the outcome of interest, and $\mathbf{X}$ is the set of all other measured variables. We want to query the total average causal effects of W on Y in $\mathcal{G}$, as defined below.

Definition 3.13 (Average causal effects [46]). *The average causal effects of W on Y is defined as*

$$ACE(W,Y) = \mathbb{E}(y|do\ (w = 1)) - \mathbb{E}(y|do\ (w = 0)) \tag{3.1}$$

where $\mathbb{E}$ denotes the expectation operations, and $do(\cdot)$ is the do-operator [9] indicating the manipulation of W by setting W to a value $w \in \{0, 1\}$.

With a proper adjustment set $\mathbf{Z}$, unbiased estimation of $ACE(W,Y)$ can be calculated from data directly. If the random variables are discrete analogues, the average causal effect can be calculated [54] by

$$ACE = \sum_z [\mathbb{E}(Y \mid w = 1, \mathbf{Z} = z) - \mathbb{E}(Y \mid w = 0, \mathbf{Z} = z)]prob(\mathbf{Z} = z) \tag{3.2}$$

Adjustment amounts to dividing the data into groups that are homogeneous relative to adjustment set $\mathbf{Z}$ [9]. That is, if we find a proper adjustment set $\mathbf{Z}$, the causal effect of W on Y can be calculated unbiasedly. Furthermore, the average causal effect on the treated individuals can be estimated by

$$ACT = \sum_z [\mathbb{E}(Y \mid w = 1, \mathbf{Z} = z)]prob(\mathbf{Z} = z) \tag{3.3}$$

As a result, identifying a proper adjustment set $\mathbf{Z}$ is crucial for the computation of causal effects. In the next section, we introduce graphical criteria to determine $\mathbf{Z}$ based on a causal graph.

3.2.2 Graphical criteria for identifying adjustment sets

Several graphical criteria have been developed to determine proper adjustment sets based on a causal graph such that the causal effect of W on Y can be calculated unbiasedly by Eq. (3.2). When there are no latent variables and the causal DAG is known, the *back-door criterion* [9] can be used to identify an adjustment set $\mathbf{Z}$.

Definition 3.14 (Back-door criterion). *Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, w.r.t. an ordered pair of variables (W, Y), a set of variables $\mathbf{Z} \subseteq \mathbf{V} \setminus \{W, Y\}$ is said to satisfy the back-door criterion in a given DAG if*

1. $\mathbf{Z}$ *does not contain a descendant node of* W*;*

2. $\mathbf{Z}$ *blocks every back-door path between* W *and* Y *(i.e., the paths between* W *and* Y*, which contain a directed edge pointing into* W*).*

Once $\mathbf{Z}$ is found, the causal effects of W on Y can be estimated by Eq. (3.2) unbiasedly.

In many real-world applications, some variables may be unobserved or unmeasured, meaning that causal sufficiency is not satisfied. Maathuis and Colombo proposed a generalized back-door criterion for identifying adjustment sets from DAGs, CPDAGs, MAGs, and PAGs. In the following, we first introduce the generalized back-door path and then the generalized back-door criterion.

Definition 3.15 (Generalized back-door path [48])**.** *Given a DAG, CPDAG, MAG or PAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$*, for an ordered pair of variables* (W, Y) *in* $\mathbf{V}$*, a path in* $\mathcal{G}$ *is a generalized back-door path from* W *to* Y *if it does not have a visible edge out of* W*.*

Definition 3.16 (Generalized back-door criterion [48])**.** *Given a DAG, CPDAG, MAG, or PAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$*, w.r.t. an ordered pair of variables* (W, Y) *in* $\mathbf{V}$*, a set of variables* $\mathbf{Z} \subseteq \mathbf{V} \setminus \{W, Y\}$ *satisfies the generalized back-door criterion and is known as a generalized back-door set relative to* (W, Y) *in* $\mathcal{G}$ *if*

1. $\mathbf{Z}$ *does not contain any possible descendants of* W *in* $\mathcal{G}$*, and*

2. $\mathbf{Z}$ *blocks every definite status back-door path from* W *to* Y *in* $\mathcal{G}$*.*

The generalized back-door criterion is sufficient but not necessary to identify an adjustment set from a given DAG, CPDAG, MAG, or PAG $\mathcal{G}$. Perković et al. [49] extended the results in [48, 55, 100] and developed a generalized adjustment criterion. Before introducing the generalized adjustment criterion, in the following, we first introduce two important concepts, amenability and forbidden set, regarding a causal graph.

The causal effect cannot be estimated by covariate adjustment in some graphs (such as DAG, CPDAG, MAG, or PAG) since causal ambiguities

exist in the system [100]. The amenability requirement defined below is used to guarantee that the causal effect can be estimated by covariate adjustment.

Definition 3.17 (Amenability [100])**.** *Given a DAG, a CPDAG, an MAG or a PAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$*, where* $\mathbf{V} = \{W, Y\} \cup \mathbf{X}$*.* $\mathcal{G}$ *is adjustment amenable if every proper possibly directed path from* W *to* Y *in* $\mathcal{G}$ *starts with a visible edge.*

A node on any causal path from W to Y in $\mathcal{G}$ should not be included in an adjustment set since such a node would block the causal path from W to Y [49]. Furthermore, including a descendant of W may open a non-causal path from W to Y, such as node A on path $W \to \cdots \to A \leftarrow \cdots \to Y$. This leads to the definition of the forbidden set, as follows, for defining the generalized adjustment criterion.

Definition 3.18 (Forbidden set; $Forb(W, Y, \mathcal{G})$ [49])**.** *Given a DAG, a CPDAG, an MAG or a PAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$*, w.r.t. an ordered pair of nodes* (W, Y) *in* $\mathbf{V}$*, a forbidden set is defined as* $Forb(W, Y, \mathcal{G}) = \{X \in \mathbf{V} : X \in PossDe(W, \mathcal{G})$*, and* X *lies on a possible directed path from* W *to* Y *in* $\mathcal{G}\}$*.*

Definition 3.19 (Generalized adjustment criterion [49])**.** *Given a DAG, a CPDAG, an MAG or a PAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$*, w.r.t. an ordered pair of nodes* (W, Y) *in* $\mathbf{V}$*, a set of variables* $\mathbf{Z}$ *satisfies the generalized adjustment criterion if*

1. $\mathcal{G}$ *is adjustment amenable w.r.t.* (W, Y)*,*
2. $\mathbf{Z} \cap Forb(W, Y, \mathcal{G}) = \emptyset$*, and*
3. *all definite status non-causal paths between* W *and* Y *are blocked (i.e., m-separated) by* $\mathbf{Z}$ *in* $\mathcal{G}$*.*

The above introduced graphical criteria have established a systematic theoretical basis for identifying valid adjustment sets in a given causal graph. However, in practice, as a causal graph is rarely known, data-driven adjustment set identification is still an open question.

3.3 Matching methods for causal effect estimation

Eq. (3.2) provides a conceptual definition for causal effect estimation by adjustment. When an adjustment set is identified, matching methods developed in the potential outcome framework can be used for causal effect estimation more effectively. In our work, we use existing matching methods for causal effect estimation after identifying an adjustment set. Thus, we provide a brief introduction to the matching methods in this section. Since matching methods are based on the potential outcome framework, we first present the background information about the potential outcome model.

3.3.1 Potential outcome framework

With respect to a binary treatment W, for each sample (individual) i, there are two potential outcomes, $Y_i(1)$ when receiving the treatment $(W = 1)$ and $Y_i(0)$ when not receiving the treatment $(W = 0)$. For the same individual, only one of the potential outcomes can be observed, and the other one is counterfactual [51, 141]. We use $Y_i \in \mathbb{R}$ to denote the observed outcome of sample i, and we have $Y_i = W * Y_i(1) + (1 - W) * Y_i(0)$. The counterfactual outcome for individual i is denoted as Y_i^*. We omit the subscript i when the meaning is clear.

The fundamental challenge is that Y_i^* is unmeasured/unobserved for each individual i [51, 86]. In many situations, it is reasonable to assume that applying a treatment to one individual does not affect the outcome of another individual, i.e., make the following stable unit treatment value assumption (SUTVA) [32, 142].

Assumption 3.1 (SUTVA)**.** *The potential outcomes for any individual do not vary with the treatments assigned to other individuals. For each individual, there are non-different forms or versions of each treatment level that lead to different potential outcomes.*

The SUTVA implies that (1) the independence of individuals, that is, the treatment assignment of one individual, does not interfere with another, and (2) for each individual, there are no latent forms of treatments.

Another important assumption of the potential outcome model is unconfoundedness regarding the relationship between the treatment assignment mechanism and the potential outcome.

Assumption 3.2 (Unconfoundedness)**.** *The potential outcomes* $(Y(1), Y(0))$ *are conditionally independent of the treatment variable* W *given a set of covariates* $\mathbf{X}$*, i.e.,* $(Y(1), Y(0)) \perp\!\!\!\perp W|\mathbf{X}$*.*

The assumption of unconfoundedness means that there is "no latent confounder" in the system, and the assumption of causal sufficiency is satisfied. Additionally, $\mathbf{X}$ is a proper adjustment set for unbiased estimation of the causal effect of W on Y. Moreover, the unconfoundedness assumption and the following overlap assumption together form the so-called strong ignorability assumption [22, 32].

Assumption 3.3 (Overlap)**.** *Every sample has a non-zero probability of receiving treatment when conditioned on* $\mathbf{X}$*, i.e.,* $0 < prob(W = 1|\mathbf{X}) < 1$*.*

In the potential outcome framework, strong ignorability cannot be tested on data since it involves unobserved outcomes.

3.3.2 Matching methods

Matching is one of the most popular approaches to estimating causal effects from observational data [58]. There are two key components in a matching method, a distance measure and a matching method. We first introduce the measure of distance or "closeness" utilized in matching.

Distance measure. For a matching method, the distance measure is used to assess the similarity between two individuals (i and j) w.r.t. the adjustment set $\mathbf{Z}$. Four distances are often used in matching methods, including exact distance, Mahalanobis distance, propensity score, and linear

propensity score. The exact distance between i and j, i.e., $Dist(\mathbf{z}_i, \mathbf{z}_j)$ is defined as:

$$Dist(\mathbf{z}_i, \mathbf{z}_j) = \begin{cases} 0 & if \mathbf{z}_i = \mathbf{z}_j, \\ \infty & otherwise. \end{cases} \tag{3.4}$$

The Mahalanobis distance is used to measure the distance of each pair $(\mathbf{z}_i, \mathbf{z}_j)$ over the space of the adjustment set $\mathbf{Z}$ and is defined as follows.

$$Dist(\mathbf{z}_i, \mathbf{z}_j) = \{(\mathbf{z}_i - \mathbf{z}_j)^T \hat{\Sigma}_{\mathbf{z}}^{-1} (\mathbf{z}_i - \mathbf{z}_j)\}^{\frac{1}{2}} \tag{3.5}$$

$\mathbf{z}_i$ and $\mathbf{z}_j$ are the value vectors of the adjustment set $\mathbf{Z}$ of the i-th and j-th individuals. For ACE, Σ is the variance-covariance matrix of $\mathbf{Z}$ in both groups; for ACT, Σ is the variance-covariance matrix of $\mathbf{Z}$ in the full control group.

A propensity score is defined as the probability of an individual being assigned to the treatment conditioning on a set of measured adjustments. The true propensity score is unknown in observational data and must be calculated [22]. The estimated propensity score can be interpreted as the conditional probability of the treatment W given the set of adjustment sets $\mathbf{Z}$:

$$e(\mathbf{z}) = prob(W = 1 \mid \mathbf{Z} = \mathbf{z}) \tag{3.6}$$

where $e(\mathbf{z})$ is often calculated by utilizing logistic regression [143] and generalized boosted models (GBM) [144]. The distance by propensity score is defined as

$$Dist(\mathbf{z}_i, \mathbf{z}_j) = |e(\mathbf{z}_i) - e(\mathbf{z}_j)| \tag{3.7}$$

The linear propensity score is defined as

$$Dist(\mathbf{z}_i, \mathbf{z}_j) = |logit(e(\mathbf{z}_i)) - logit(e(\mathbf{z}_j))| \tag{3.8}$$

Matching methods. Nearest neighbour matching (NNM) is one of the most popular matching strategies [82]. For each individual i, NNM selects the most similar individual in the opposite treatment group. Thus,

we can treat the outcome of the matched individual as the missing potential (counterfactual) outcome of i.

Other matching strategies are available to improve the performance of NNM, such as optimal matching [145], ratio matching [85], and with or without replacement according to the size of the treated group and the size of the control group [146].

3.4 Summary

In this chapter, we first introduced graphical causal modelling, including causal DAG and causal MAG. Then, we presented covariate adjustment for causal effect estimation based on graphical causal modelling and introduced the graphical criteria for identifying adjustment sets. Identifying proper adjustment sets based on graphical causal modelling is an essential task in our work, and these graphical criteria are used in Chapters 4, 5, 6 and 7. Finally, we also presented matching methods from causal effect estimation from data.

Chapter 4

Sufficient dimension reduction for average causal effect estimation

In this chapter, we study the problem of determining a proper adjustment set by learning a reduced set by SDR from the original variable space. The work in this chapter addresses the challenge of data insufficiency caused by large adjustment sets. We provide a theoretical conclusion based on graphical causal modelling; i.e., the reduced set learned by SDR is sufficient to eliminate confounding bias under the assumptions of strong ignorability (i.e., assumptions 3.2 and 3.3 in Chapter 3) and pretreatment variables. We present a data-driven causal effect estimator based on a specific SDR method.

4.1 Motivation

Covariate adjustment is a primary method for removing confounding bias when estimating ACE from observational data. When data contain pretreatment variables, it is a common practice to use all pretreatment variables as the adjustment (or deconfounding) set [82]. However, the large size of the adjustment set can significantly affect the performance of a causal effect estimator since including irrelevant variables in an adjustment set

is suboptimal for estimating causal effects from data [57]. Note that an adjustment set with over 10 variables will be large when using Eq. (3.2) for adjustment since the stratification produces more than 1,000 subgroups. Thus, a small-sized adjustment set is preferred for nonparametric estimation of causal effect [6, 56, 57].

SDR techniques have been utilized to create a small set of variables in a transformed space for estimating ACE from data [89, 110]. As reviewed in Section 2.5, several methods have been developed to learn the reduced set for ACE estimation [60, 90, 110]. However, it is not clear whether the reduced set learned by SDR guarantees unbiased causal effect estimation by covariate adjustment, i.e., and there is no theoretical connection between SDR and covariate adjustment.

To obtain a small-size adjustment set, in this chapter, we prove that the reduced set learned by SDR is sufficient to eliminate confounding bias under the assumptions of strong ignorability (i.e., assumptions 3.2 and 3.3 in Chapter 3) and pretreatment variables. Based on this, we developed a novel causal effect estimator, CESD (the Causal effect estimator by using sufficient dimension reduction). CESD utilizes kernel dimension reduction (KDR) [5], an SDR method, to learn an adjustment set from data. Then, CESD employs nearest neighbour matching (NNM) [56, 82] based on the learned adjustment set for ACE estimation.

To the best of our knowledge, this is the first work to prove that the reduced covariate set by using SDR is a proper adjustment set. This result opens the door for developing new methods of causal effect estimation with a large number of variables. In the rest of this chapter, we present the developed theorem and the proposed CESD algorithm. We also present experiments for evaluating CESD and discuss the results.

The chapter is based on the paper [73] "Sufficient Dimension Reduction for Average Causal Effect Estimation", which was accepted by the journal of Data Mining and Knowledge Discovery.

4.2 The theorem for adjustment set identification with SDR

In this section, we first introduce the problem setting. After that, we present the theorem to support *SDR* for adjustment set identification. That is, the reduced covariate set $\mathbf{Z} = \mathbf{\Psi}^T\mathbf{X}$ by SDR is sufficient to remove confounding bias in ACE estimation.

4.2.1 Problem setting

We assume that (1) the covariate set $\mathbf{X}$ contains the pretreatment variables and (2) the *strong ignorability* holds (i.e., both assumptions 3.2 and 3.3 hold). The strong ignorability assumption means that there is "*no latent confounder*" in the system, i.e., and the assumption of causal sufficiency is satisfied. The problem definition of this chapter is given as follows.

Problem 4.1. *We convert the problem of determining an adjustment set from the original covariate space* $\mathbf{X}$ *to the problem of learning a low-dimensional reduced set* $\mathbf{Z}$ *from* $\mathbf{X}$ *such that* $W \perp\!\!\!\perp Y \mid \mathbf{Z}$ *is in the manipulated DAG* $\mathcal{G}_{\underline{W}}$.

where $\mathcal{G}_{\underline{W}}$ denotes a manipulated graph obtained by deleting all arrows out of W from G. When the adjustment set $\mathbf{Z}$ is learned, the causal effect of W on Y can be estimated unbiasedly by adjusting for $\mathbf{Z}$ as in Eq. (3.2) (or Eq. (3.3)).

4.2.2 Sufficient condition for identifying an adjustment set using SDR

Let us consider the treatment assignment W as a binary classification problem, i.e., the probability density function of W given the value of $\mathbf{X}$ is $prob_{W|\mathbf{X}}(w|\mathbf{x})$. SDR attempts to search for a projection $\mathbf{\Psi} \in \mathbb{R}^{p\times r}$, where $r < p$, such that

$$W \perp\!\!\!\perp \mathbf{X}|\mathbf{\Psi}^T\mathbf{X} \tag{4.1}$$

where $\mathbf{\Psi}^T\mathbf{X}$ is the orthogonal projection of $\mathbf{X}$ onto the column subspace of $\mathbf{\Psi}$, and the column subspace of $\mathbf{\Psi}$ refers to the *dimension reduction space*

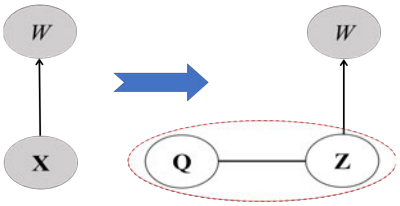


Figure 4.1: A graphical representation of sufficient dimension reduction, $W \perp\!\!\!\perp \mathbf{Q}|\mathbf{Z}$ holds [5], where $\mathbf{X}$ is decomposed into $(\mathbf{Z}, \mathbf{Q})$.

(DRS) [69, 70]. The primary interest is the *central* DRS since it has a well-known invariance property [69, 70].

Definition 4.1 (Central DRS [69])**.** *The column space of* $\mathbf{\Psi}$ *is a central DRS if the column space of* $\mathbf{\Psi}$ *is a DRS with the smallest possible dimension* r*.*

Identifying a projection $\mathbf{\Psi}$ makes Eq. (4.1) equivalent to searching for a projection $\mathbf{\Psi}$, which makes W and $(\mathbf{I} - \mathbf{\Psi}^T)\mathbf{X}$ conditionally independent given $\mathbf{\Psi}^T\mathbf{X}$. That is, $\mathbf{X}$ can be decomposed into $(\mathbf{Z} = \mathbf{\Psi}^T\mathbf{X}, \mathbf{Q} = (\mathbf{I} - \mathbf{\Psi}^T)\mathbf{X})$, $\mathbf{Z}$ is associated with W and $\mathbf{Q}$ is independent of W given $\mathbf{Z}$, i.e., $W \perp\!\!\!\perp \mathbf{Q}|\mathbf{Z}$ (See Fig. 4.1).

Now, we provide the following theorem to ensure that finding an adjustment set from $\mathbf{X}$ can be converted to the problem of learning the reduced covariate set $\mathbf{Z}$ by SDR.

Theorem 4.1. *Let* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *be the underlying causal DAG, in which* $\mathbf{V}$ *contains the treatment variable* W*, the outcome variable* Y*, and a set of all pretreatment variables* $\mathbf{X}$*. If there exists a central DRS (the column space of* $\mathbf{\Psi}$*) such that* $W \perp\!\!\!\perp \mathbf{Q} \mid \mathbf{Z}$*, where* $\mathbf{Z} = \mathbf{\Psi}^T\mathbf{X}$ *and* $\mathbf{Q} = (\mathbf{I} - \mathbf{\Psi}^T)\mathbf{X}$*, then* $\mathbf{Z}$ *is a proper adjustment set for estimating the causal effect of* W *on* Y *unbiasedly.*

Proof. Under the pretreatment assumption of $\mathbf{X}$, there is no descendant node of W in $\mathbf{X}$, and there is a directed edge $W \to Y$ in the DAG $\mathcal{G}$. In the manipulated DAG $\mathcal{G}_{\underline{W}}$, there are only back-door paths between W and Y. Hence, all back-door paths between W and Y are blocked by the set of $\mathbf{X}$, i.e., $W \perp\!\!\!\perp Y|\mathbf{X}$ holds in $\mathcal{G}_{\underline{W}}$.

We now prove that if the column space of $\mathbf{\Psi}$ is a central DRS, then $\mathbf{Z}$ is sufficient to block all back-door paths between W and Y. We have $\mathbf{Z} = \mathbf{\Psi}^T\mathbf{X}$ and $\mathbf{Z}$ satisfies $W \perp\!\!\!\perp \mathbf{Q}|\mathbf{Z}$ by the property of SDR since the column space of $\mathbf{\Psi}$ is a central DRS. We have $W \perp\!\!\!\perp Y|\mathbf{X}$ in $\mathcal{G}_{\underline{W}}$ and $\mathbf{X} = (\mathbf{Z}, \mathbf{Q})$; then, $W \perp\!\!\!\perp Y|(\mathbf{Z}, \mathbf{Q})$ holds in $\mathcal{G}_{\underline{W}}$. As $W \perp\!\!\!\perp \mathbf{Q}|\mathbf{Z}$ is satisfied, $W \perp\!\!\!\perp (Y, \mathbf{Q})|\mathbf{Z}$ holds in $\mathcal{G}_{\underline{W}}$ by the contraction property of conditional independence. According to the decomposition property of conditional independence, $W \perp\!\!\!\perp (Y, \mathbf{Q})|\mathbf{Z}$ is sufficient to infer $W \perp\!\!\!\perp Y|\mathbf{Z}$ in $\mathcal{G}_{\underline{W}}$. That is, conditioning on $\mathbf{Z}$ appears to block the flow of information from W to Y in $\mathcal{G}_{\underline{W}}$, i.e., $\mathbf{Z}$ is sufficient to block all back-door paths from W to Y in $\mathcal{G}$. Therefore, $\mathbf{Z}$ satisfies the back-door criterion and is a proper adjustment set. □

Theorem 4.1 has shown that the reduced covariate set $\mathbf{Z}$ by SDR is sufficient for removing confounding bias when estimating the causal effects of W on Y from data. Furthermore, Theorem 4.1 guarantees the soundness of any SDR method for estimating causal effects from data. In the next section, we propose a sound data-driven method based on Theorem 4.1 for estimating causal effects from data.

4.3 The CESD algorithm

In this section, we introduce a specific method for identifying the adjustment set by using an SDR method (i.e., kernel dimension reduction [111]) and then present the proposed CESD algorithm.

4.3.1 Adjustment set identification using SDR

We present a kernel-based SDR method to identify an adjustment set from data. We utilize the cross-covariance operators on *reproducing kernel Hilbert space* (RKHS) [71] $\mathcal{H}$ to establish an objective function for dimensionality reduction. By default, we use the *Gaussian kernel*, i.e., $k(x_i, x_j) = exp(-\frac{\|x_i - x_j\|^2}{2\delta^2})$, where δ is the kernel width.

Given two RKHS, $(\mathcal{H}_1, k_1)$ and $(\mathcal{H}_2, k_2)$, which are over the measured spaces $(\Omega_1, \mathfrak{B}_1)$ and $(\Omega_2, \mathfrak{B}_2)$, with the positive measurable kernels k_1 and

k_2. For the pair of W and the set of $\mathbf{X}$ on $\Omega_1 \times \Omega_2$, the cross-covariance operator from $\mathcal{H}_1$ to $\mathcal{H}_2$ is defined by the relation:

$$< g, \Sigma_{W\mathbf{X}} f >_{\mathcal{H}_2} = \mathbb{E}_{\mathbf{X}W}[f(\mathbf{X})g(W)] - \mathbb{E}_{\mathbf{X}}[f(\mathbf{X})]\mathbb{E}_W[g(W)] \tag{4.2}$$

for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$. Hence, the conditional covariance operator $\Sigma_{WW|\mathbf{Z}}$ on $\mathcal{H}_1$ can be defined as follows.

$$\Sigma_{WW|\mathbf{Z}} := \Sigma_{WW} - \Sigma_{W\mathbf{Z}}\Sigma_{\mathbf{ZZ}}^{-1}\Sigma_{\mathbf{Z}W} \tag{4.3}$$

Theorem 7 in [5] shows that $\Sigma_{WW|\mathbf{Z}} \geq \Sigma_{WW|\mathbf{X}}$ for any $\mathbf{Z}$, and $\Sigma_{WW|\mathbf{X}} - \Sigma_{WW|\mathbf{Z}} = 0 \Leftrightarrow W \perp\!\!\!\perp \mathbf{Q}|\mathbf{Z}$. That is, minimizing $\hat{\Sigma}_{WW|\mathbf{Z}}$ is the most informative direction for searching the optimal $\mathbf{Z}$. Hence, searching for a set of the reduced covariates $\mathbf{Z}$ such that $W \perp\!\!\!\perp \mathbf{Q}|\mathbf{Z}$ holds is equivalent to optimizing the minimized conditional covariance operator $\hat{\Sigma}_{WW|\mathbf{Z}}$. The determinant of $\hat{\Sigma}_{WW|\mathbf{Z}}$ can be written as follows.

$$\det \hat{\Sigma}_{WW|\mathbf{Z}} = \frac{\det \hat{\Sigma}_{[W\mathbf{Z}][W\mathbf{Z}]}}{\det \hat{\Sigma}_{\mathbf{ZZ}}} \tag{4.4}$$

where $\hat{\Sigma}_{[W\mathbf{Z}][W\mathbf{Z}]} = \begin{pmatrix} \hat{\Sigma}_{WW} & \hat{\Sigma}_{W\mathbf{Z}} \\ \hat{\Sigma}_{\mathbf{Z}W} & \hat{\Sigma}_{\mathbf{ZZ}} \end{pmatrix} = \begin{pmatrix} (\hat{K}_W + \epsilon \mathbf{I}_n)^2 & \hat{K}_W \hat{K}_{\mathbf{Z}} \\ \hat{K}_{\mathbf{Z}} \hat{K}_W & (\hat{K}_{\mathbf{Z}} + \epsilon \mathbf{I}_n)^2 \end{pmatrix}$, where ϵ is a positive regularization parameter. $\hat{K}_W$ and $\hat{K}_{\mathbf{Z}}$ are the centralized *Gram matrices* defined as follows.

$$\hat{K}_W = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T) G_W (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T) \tag{4.5}$$

$$\hat{K}_{\mathbf{Z}} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T) G_{\mathbf{Z}} (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T) \tag{4.6}$$

where $(G_W)_{i,j} = k(w_i, w_j)$, $(G_{\mathbf{Z}})_{i,j} = k(\mathbf{Z}_i, \mathbf{Z}_j)$ and $\mathbf{1}_n = (1, \ldots, 1)^T$ is a vector with all elements equal to 1.

To solve Eq. (4.4), gradient descent with line search can be used. The matrix of parameters is updated iteratively by

$$\mathbf{\Psi}^{t+1} = \mathbf{\Psi}^t - \beta \frac{\partial \log \det \hat{\Sigma}_{WW|\mathbf{Z}}}{\partial \mathbf{\Psi}} = \mathbf{\Psi}^t - \beta Tr[\hat{\Sigma}_{WW|\mathbf{Z}}^{-1} \frac{\partial \hat{\Sigma}_{WW|\mathbf{Z}}}{\partial \mathbf{\Psi}}] \tag{4.7}$$

where the trace norm in Eq. (4.7) can be rewritten as $2\epsilon Tr[\hat{\Sigma}^{-1}_{WW|\mathbf{Z}}\hat{K}_W(\hat{K}_\mathbf{Z}+\epsilon\mathbf{I}_n)^{-1}\frac{\partial\hat{K}_\mathbf{Z}}{\partial\mathbf{\Psi}}(\hat{K}_\mathbf{Z}+\epsilon\mathbf{I}_n)^{-2}\hat{K}_\mathbf{Z}\hat{K}_W]$, and β is optimized through golden section search [5]. All of these matrices in Eq. (4.7) can be obtained directly based on Eq. (4.5) and Eq. (4.6). Therefore, the problem of identifying an adjustment set $\mathbf{Z}$ can be achieved by optimizing Eq. (4.4).

4.3.2 NNM using the discovered adjustment set

Given the learned adjustment set $\mathbf{Z}$, the next step is to infer the counterfactual outcome denoted as $Y_i^*(w_i)$. *Nearest neighbour matching* (NNM) is a well-known method for such inference [56, 82]. With NNM, the unobserved (a.k.a. counterfactual) outcome of an individual i is imputed by the observed outcome of an individual j who has similar covariates (the values of $\mathbf{Z}$) in the opposite treatment group. We use the Mahalanobis distance to measure the distance of each pair (z_i, z_j) in Eq. (3.5). This strategy of NNM is utilized in the Adadie-Imbens estimator [56] and is defined as follows.

$$Y_i^*(w_i) = Y_k(1-w_i); \quad k = \underset{j\in\mathbf{D}_{(1-w_i)}}{\arg\min}\, Dist(\mathbf{z}_i, \mathbf{z}_j) \tag{4.8}$$

where $\mathbf{D}_{(1-w_i)}$ is the dataset with the treatment of $1-w_i$.

The complete procedure of CESD is listed in Algorithm 1. The first line calculates $\hat{K}_W$ and $\hat{K}_\mathbf{Z}$, where the kernel width δ is set to 5 for CESD. The Line 2 is to calculate $\det\hat{\Sigma}_{WW|\mathbf{Z}}$. Lines 3 to 6 aim to obtain the central DRS $\mathbf{\Psi}$ by optimizing Eq. (4.4). Line 7 is to obtain the adjustment set $\mathbf{Z}$. Line 8 calculates the Mahalanobis distance of each pair of individuals over the space of $\mathbf{Z}$. Line 9 imputes the counterfactual outcome $Y_i^*(w_i)$ for each factual outcome. Line 10 calculates the causal effects by Eq. (3.2) or Eq. (3.3) according to the different requirements.

Theorem 4.2 (The soundness of CESD algorithm)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be the underlying causal DAG, in which $\mathbf{V}$ contains the treatment variable W, the outcome variable Y, and a set of all pretreatment variables $\mathbf{X}$. CESD is sufficient to infer the causal effect of W on Y unbiasedly.*

Algorithm 1 The causal effect estimator by using sufficient dimension reduction (CESD)

Input: Dataset $\mathbf{D}$ with W, Y and pretreatment variables $\mathbf{X}$. The parameters ϵ, δ, the maximum number of iterations Ite and the dimension of the reduced covariates r.

Output: Causal effect

1: Compute Eq. (4.5) and Eq. (4.6)
2: Calculate $\det \hat{\Sigma}_{WW|\mathbf{Z}}$ by Eq. (4.4)
3: $t = 1$
4: **while** $t \leq Ite$ or $\mid \mathbf{\Psi}^{t+1} - \mathbf{\Psi}^{t} \mid \geq \epsilon$ **do**
5: $\quad \mathbf{\Psi}^{t+1} = \mathbf{\Psi}^{t} - \beta Tr[\hat{\Sigma}^{-1}_{WW|\mathbf{Z}} \frac{\partial \hat{\Sigma}_{WW|\mathbf{Z}}}{\partial \mathbf{\Psi}}]$
6: **end while**
7: Compute $\mathbf{Z} = \mathbf{\Psi}^{T}\mathbf{X}$
8: Compute $Dist(z_i, z_j)$ over $\mathbf{Z}$ via Eq. (3.5).
9: Impute $Y_i^*(w_i)$ via Eq. (4.8).
10: Calculate the causal effect of W on Y.
11: **return** Causal effect

Proof. The column space of $\mathbf{\Psi}$ is optimal by minimizing Eq. (4.4) because $\Sigma_{WW|\mathbf{Z}} \geq \Sigma_{WW|\mathbf{X}}$ for any $\mathbf{Z}$, and $\Sigma_{WW|\mathbf{X}} - \Sigma_{WW|\mathbf{Z}} = 0 \Leftrightarrow W \perp\!\!\!\perp \mathbf{Q}|\mathbf{Z}$. In other words, $\mathbf{Z} = \mathbf{\Psi}^{T}\mathbf{X}$ contains the same information as $\mathbf{X}$ relative to W. By Theorem 4.1, $\mathbf{Z}$ is a proper adjustment set. Therefore, the causal effect of W on Y is correctly and unbiasedly identified by CESD. □

Complexity analysis: Three parts contribute to the time complexity of CESD. The calculation of $\det \hat{\Sigma}_{WW|\mathbf{Z}}$ is matrix multiplication, which has a time complexity of $\mathbf{O}(np^2)$. Solving Eq. (4.4) requires a linear search, i.e., $\mathbf{O}(np)$ where n is the sample size. The calculation of NNM is $\mathbf{O}(nr^2)$. Therefore, the time complexity of CESD is $\mathbf{O}(np^2)$.

4.4 Experiments

Evaluating a causal effect estimator is very challenging since we rarely have the ground truth of causal effects on real-world datasets. Following the ex-

Table 4.1: Summary of the real-world datasets.

Name	#treated	#control	#samples	#variables
IHDP	139	608	747	24
Twins	3,275	1,546	4,821	40
Jobs	297	2,915	3,212	9
Cattaneo2	864	3,778	4,642	20
RHC	1,103	1,604	2,707	72

isting literature, we evaluate CESD on five datasets, including two semisynthetic real-world datasets, IHDP [37] and Twins [147], and three real-world applications, Job training [15], Cattaneo2 [110] and RHC [148]. A brief description of the five datasets is shown in Table 4.1.

The developed CESD method consists of kernel dimension reduction and NNM, which are implemented by the $\mathbb{R}$ packages *KDRcpp*[1] and *Matching* [149], respectively. To evaluate the performance of CESD, we compare it against the state-of-the-art causal effect estimators including **MDM**: Mahalanobis distance matching [21]; **PSM**: propensity score matching with logistic regression [82]; **CBPS**[2]: covariate balancing propensity score [150]; **PAW**: the set of causes of W with PSM [7]; **PAY**: the set of causes of Y with PSM [7]; **CausalForest**[3]: random forest regression for estimating causal effect [151]; **Shrinkage**[4]: inverse probability weighting estimator based on SDR for average causal effect estimation [110] and **MSDR**: matching using SDR [109].

Parameter settings. For CausalForest, we set the number of trees to 200. For Shrinkage, MSDR, and CESD, the dimension of the reduced covariates r is set to 2. For CESD, the parameters ϵ, δ and Ite are set to 0.0001, 5 and 20, respectively, for all datasets.

Evaluation metrics. We evaluate the performance of all algorithms using standard deviation (SD), the root-mean-square error (RMSE) and the

[1] `https://github.com/aschmu/KDRcpp`

[2] `https://cran.r-project.org/web/packages/CBPS/index.html`

[3] `https://cran.r-project.org/web/packages/grf/index.html`

[4] `https://www.stat4reg.se/software/sdrcausal`

Table 4.2: Experimental results on IHDP.

Methods	ACE	SD	RMSE	Bias (%)
MDM	3.9394	1.8085	0.4187	9.61%
PSM	3.9406	2.3966	0.4175	9.58%
CBPS	4.0732	2.5376	0.2849	6.54%
PAW	3.0977	1.4850	1.2604	28.92%
PAY	2.6755	1.7595	1.6826	38.61%
CausalForest	3.5021	8.2676	0.8561	19.64%
Shrinkage	4.8637	0.3049	0.5056	11.60%
MSDR	4.9508	1.7604	0.5927	13.60%
CESD	3.9563	1.7958	0.4018	9.22%

estimation bias (%) (relative error) when the ground truth is available. In this paper, we visualize the estimated causal effects and their confidence intervals with a confidence level of 95%.

4.4.1 Experiments on the two semisynthetic real-world datasets

4.4.1.1 IHDP

The IHDP dataset is observational data from a randomized trial based on the Infant Health and Development Programme (IHDP), which investigated the effects of high-quality intensive care on low-birth-weight and premature infants [37]. The indicator variable, representing with/without high-quality intensive care, is used as a treatment variable. IHDP consists of 747 samples with 24 pretreatment variables, among which 608 are control units (samples), and 139 are treated units. The simulated outcomes are generated by using setting "A" in the $\mathbb{R}$ package *npci*[1], and the ground truth of the causal effect, i.e., 4.36 is obtained by the noiseless outcome according to the same procedures suggested by Hill [37].

[1] `https://github.com/vdorie/npci`

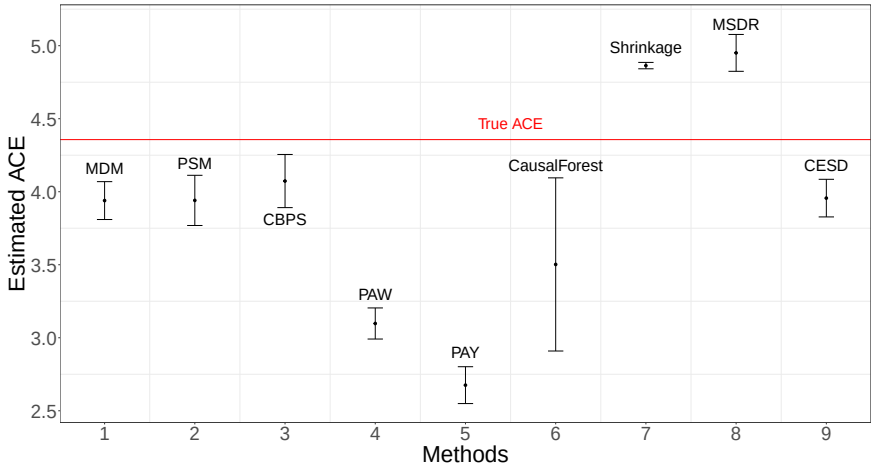


Figure 4.2: Estimated causal effects on the IHDP dataset w.r.t. a 95% confidence interval. The red line represents the ground truth ACE.

The experimental results of all estimators are listed in Table 4.2 and display the estimated causal effects and their confidence intervals with a confidence level of 95% in Fig. 4.2. According to the experimental results, these algorithms can be divided into two groups: Group I includes methods whose estimates are close to the ground truth, i.e., MDM, PSM, CBPS, Shrinkage, MSDR, and CESD; the remaining methods (including PAW, PAY, and CausalForest) are in Group II. The performance of CESD is competitive with the methods in Group I and better than the methods in Group II. CBPS performs very similarly to CESD when considering the confidence interval. CBPS performs well since it directly optimizes the balance of covariates between the treatment and control groups.

4.4.1.2 Twins

The Twins dataset is collected from twin births in the USA between 1989 and 1991, with infants having birth weights less than 2,000 g [152]. We remove samples with missing values from the original dataset and have 4,821 twin pairs left with 40 covariates. The weight of an infant is considered the treatment variable: W=1 for a baby who is heavier in the twin pair; W=0 otherwise. The mortality after one year is the outcome. The ground

Table 4.3: Experimental results on Twins.

Methods	ACE	SD	RMSE	Bias (%)
MDM	-0.0160	0.0110	0.0089	35.83%
PSM	-0.0267	0.0121	0.0018	7.08%
CBPS	-0.0266	0.0121	0.0017	6.90%
PAW	-0.0301	0.0118	0.0052	21.03%
PAY	-0.0240	0.0106	0.0009	3.77%
CausalForest	-0.0174	0.0122	0.0075	30.14%
Shrinkage	-0.0178	0.0563	0.0070	28.42%
MSDR	-0.0225	0.0117	0.0024	9.70%
CESD	-0.0268	0.0113	0.0019	7.63%

truth causal effect is -0.025. To simulate an observational study, we follow Louizos et al.'s suggestion [147] to randomly select one of the two twins as the observed infant and hide the other by applying the setting: $W_i|\mathbf{x}_i \sim Bern(sigmoid(\beta^T\mathbf{x} + \varepsilon))$, where $\mathbf{x}$ denotes the 40 other covariates, and $\beta^T \sim \mathcal{U}((-0.1, 0.1)^{40\times 1})$ and $\varepsilon \sim \mathcal{N}(0, 0.1)$.

The experimental results of all methods are presented in Table 4.3 and visualized in Fig. 4.3. In the figure, the performance of the estimators can be divided into Group I, including PSM, CBPS, PAY, MSDR, and CESD, whose results are close to the true ACE, and Group II, including the remaining methods. We see that the performance of CESD is competitive with the methods in Group I and better than the methods in Group II.

4.4.2 Evaluation with three real-world datasets

4.4.2.1 Jobs

The Job Training dataset (or Jobs) is a widely used benchmark dataset in causal inference, which consists of the original LaLonde dataset (297 treated samples and 425 control samples) [15] and the Panel Study of Income Dynamics (PSID) observational group (2,490 control samples) [150]. There are 9 covariates, including age, schooling in years; indicators for black and Hispanic; marital status; school degree; previous earnings in 1974, 1975;

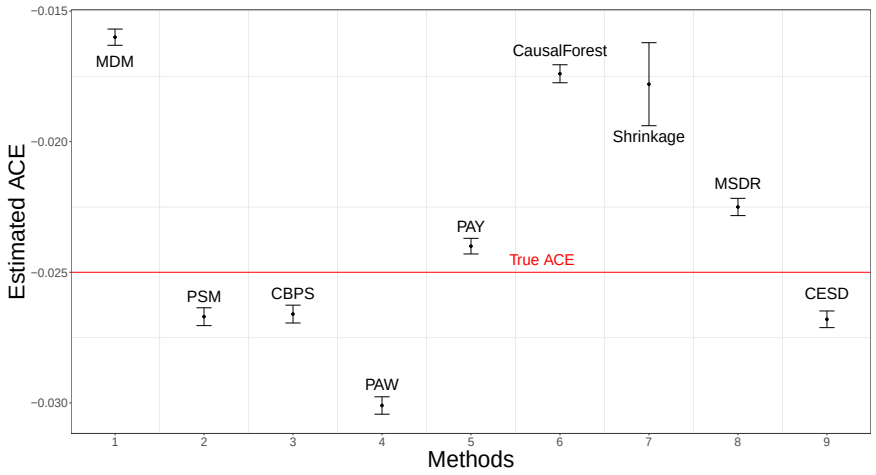


Figure 4.3: Estimated causal effects on the Twins dataset w.r.t. 95% confidence interval. The red line represent the group truth ACE.

and whether the 1974 earnings variable is missing. The job training status, i.e., with/without job training, is defined as the treatment variable W. The earnings in 1978 are defined as the outcome variable Y. Because the dataset contains records of people taking part in the training only, as in [15], we estimate the ACT using CESD and all comparing methods against the ground truth ACT, which is $886 with a standard error of $448 [150].

Table 4.4: Experimental results on Jobs.

Methods	ACT	SD	RMSE	Bias (%)
Ground truth	886	448	N/A	N/A
MDM	-593.84	612.19	292.16	167.02%
PSM	-516.62	608.63	1,402.22	158.26%
CBPS	423.30	1,295.20	462.70	52.00%
PAW	-867.37	626.66	1,753.37	197.90%
PAY	-671.33	598.24	1,557.34	175.77%
CausalForest	-5,755.78	2,989.81	6,641.78	749.64%
Shrinkage	-45.33	-344.55	931.33	105.12%
MSDR	-122.68	755.70	1,008.68	113.85%
CESD	756.00	540.13	129.61	14.63%

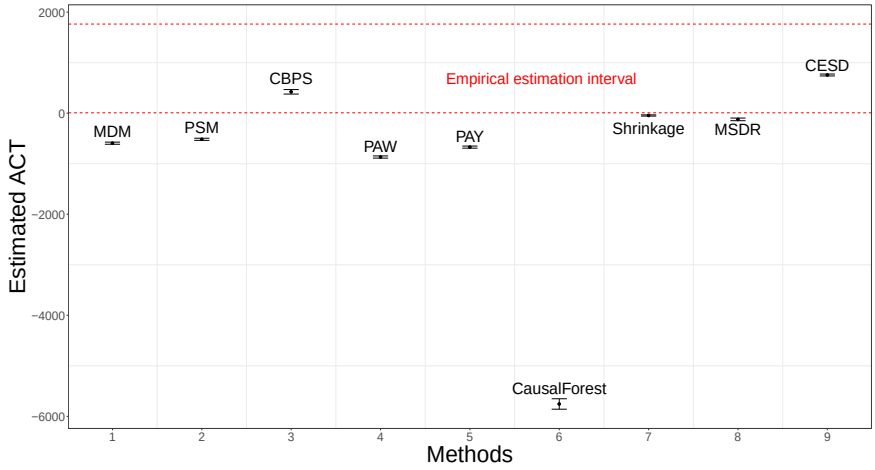


Figure 4.4: Estimated causal effects on the Jobs dataset. The two dotted lines denote the empirical estimated interval with a 95% confident interval.

We summarize the results of all methods in Table 4.4 and draw them in Fig. 4.4. In the figure, we see that CBPS and CESD are in Group I, where the methods' estimates fall within the empirical estimation interval, and other methods are not in the interval, and hence, in Group II. CESD achieves completive results with CBPS. Moreover, the estimates by Group II methods lead to the opposite conclusion, i.e., employees who participate in job training receive smaller incomes than employees who do not participate in job training.

4.4.2.2 Cattaneo2

The dataset [153] is usually used to study the ACE of maternal smoking status during pregnancy (W) on babies' birth weight (in grams)[1]. Cattaneo2 consists of birth weights of 4,642 singleton births in Pennsylvania, USA [152, 153]. Cattaneo2 contains 864 smoking mothers (W=1) and 3,778 non-smoking mothers (W=0). The dataset contains several covariates: mother's age, mother's marital status, an indicator for the previous infant where the newborn died, mother's race, mother's education, father's

[1] `http://www.stata-press.com/data/r13/cattaneo2.dta`

Table 4.5: Experimental results on Cattaneo2 and RHC.

Methods	Cattaneo2		RHC	
	ACE	SD	ACE	SD
MDM	-159.26	25.61	0.0491	0.0243
PSM	-188.47	33.23	0.0199	0.0384
CBPS	-189.16	32.60	0.0209	0.0422
PAW	-152.99	31.75	0.0209	0.0306
PAY	-157.15	20.92	0.0319	0.0205
CausalForest	-260.11	59.30	0.0256	0.0345
Shrinkage	-285.36	19.31	-0.0215	0.6224
MSDR	-181.80	23.94	0.0387	0.0234
CESD	-221.47	28.60	0.0196	0.0207

education, number of prenatal care visits, months since last birth, an indicator of firstborn infant and indicator of alcohol consumption during pregnancy. The authors [152] found a strong negative effect of maternal smoking on the weights of babies about $200g$ to $250g$ lighter for a baby with a mother smoking during pregnancy by statistical analysis on all covariates.

All results on this dataset are shown in Table 4.5 and displayed in Fig. 4.5. The range of the estimated causal effects of smoking on babies' birth weight is -285.36g to -152g. From Fig. 4.5, we see that only the estimated ACE by CESD falls within the empirically estimated interval (-250g, -200g), i.e., the estimated effects by CESD are consistent with the original study [152].

4.4.2.3 Right Heart Catheterization

Right heart catheterization (RHC) is a real-world dataset from an observational study regarding a diagnostic procedure for the management of critically ill patients [154]. The RHC dataset can be downloaded from the $\mathbb{R}$ package *Hmisc*[1]. RHC contains information on hospitalized adult patients from five medical centres in the USA. These hospitalized adult pa-

[1] `https://CRAN.R-project.org/package=Hmisc`

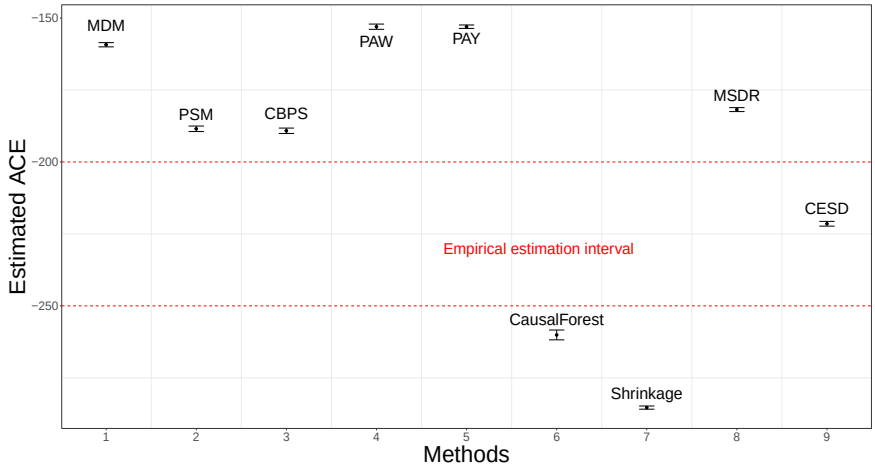


Figure 4.5: Estimated causal effects on the Cattaneo2 dataset w.r.t. 95% confident interval. The two dotted lines represent the empirical estimated interval (-250g, -200g).

tients participated in the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). Treatment W indicates whether a patient received an RHC within 24 hours of admission. The outcome Y is whether a patient died at any time up to 180 days after admission. The original RHC dataset has 5,735 samples with 73 covariates. We preprocess the original data, as suggested by Loh et al. [155], and the final data contain 2,707 samples with 72 covariates.

The experimental results on this dataset are represented in Table 4.5 and drawn in Fig. 4.6. In the experimental results, the CESD results are consistent with those of PSM, CBPS, PAW, PAY, and CausalForest. The estimated causal effects by the methods indicate that applying RHC leads to higher mortality within 180 days than not applying RHC.

In summary, based on all experimental results with the five datasets, we can conclude that CESD achieves estimates that are close to true or empirically estimated causal effect values in all datasets and are consistently in the high-performing method group across all the datasets, and CESD is the only method that is in the high-performing group across all five datasets. The closest high-performing method is CBPS from a widely used $\mathbb{R}$ package *CBPS* in causal effect estimation. All these results demonstrate the

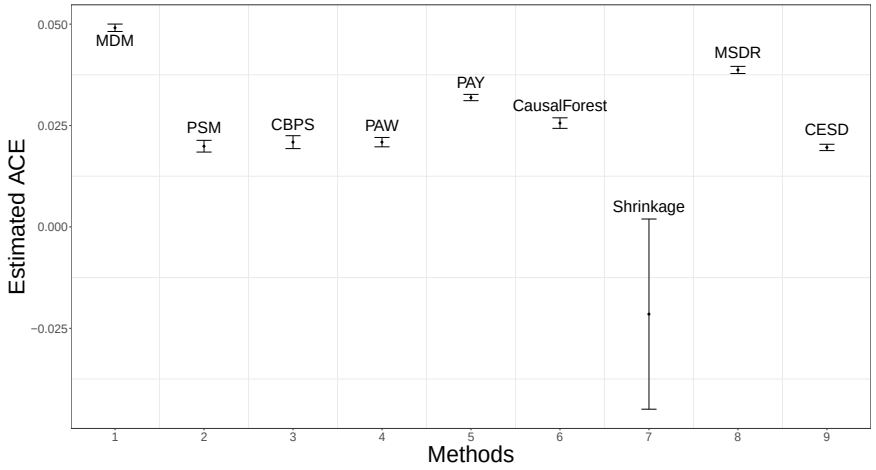


Figure 4.6: Estimated causal effects on the RHC dataset w.r.t. 95% confident interval.

robustness of the proposed method.

4.4.3 The quality of matching

To investigate the better performance of CESD compared with the other methods, we dig into the Cattaneo2 dataset to see the matching process, which is crucial for causal effect estimation methods that are based on propensity score and estimate propensity score using all covariates (PSM), factors of W only (PAW), factors of Y only (PAY), or balanced propensity score between treatment and control groups (CBPS). These methods all reduce the covariate set to a one-dimensional propensity score for matching. We show the distributions of the estimated propensity scores in the treated and control groups in Fig. 4.7 (left four subfigures). We see the obvious mismatch of propensity score distributions in the two groups, which leads to the loss of power in matching. Mismatch results in a large variance in the estimated causal effects [58]. When the covariate set is reduced to two dimensions by CESD, we can see that the distributions of each reduced dimension in the treated and control groups (the curves in the right panel in Fig. 4.7) largely overlap. The overlap improves matching in causal effect

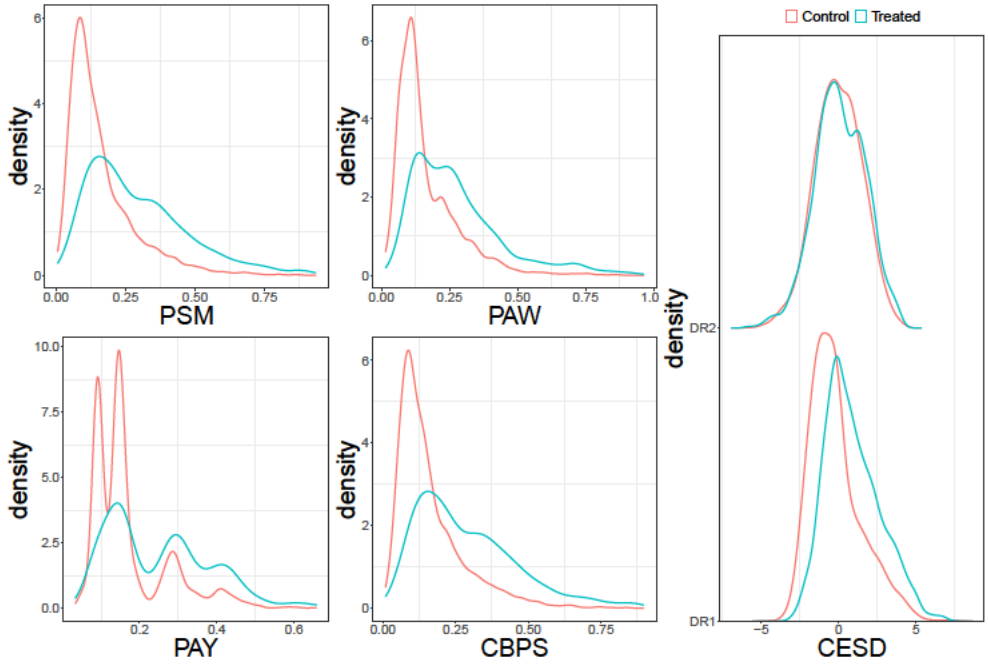


Figure 4.7: The distributions of propensity scores (left four) and reduced covariate dimensions, grouped by the treatment variable W (red for the control group and blue for the treated group) on the Cattaneo2 dataset.

estimation, and this provides an explanation for the good performance of CESD.

4.5 Conclusion

Using all pretreatment variables as the adjustment set is a common practice for causal effect estimation. However, it results in low data efficiency since adjustment based on a large number of variables may leave only a very small number of data points in each stratum. Reducing the variables to a single scalar, e.g., propensity score, may lead to large bias since irrelevant variables affect the quality of the single scalar. In this chapter, we proposed using SDR to transform the original set of pretreatment variables to a small-sized reduced set w.r.t. the treatment variable W for causal effect estimation.

We provided theoretical support for average causal effect estimations

through SDR by proving that the reduced set learned by SDR is sufficient to remove confounding bias under the assumptions of pretreatment variables and strong ignorability. This is the first theoretical conclusion that the reduced set by using SDR is a proper adjustment set.

With the support of the proposed theorem, we developed a data-driven algorithm, CESD, to estimate causal effects from observational data. CESD consists of kernel dimension reduction and nearest neighbour matching. Experimental results on two semisynthetic real-world datasets and three real-world datasets demonstrated that CESD performs consistently very well in all five datasets in comparison with the state-of-the-art methods, implying the potential usefulness in causal effect estimation in different areas. The experiments also investigated whether the quality of matching using the adjustment sets identified by CESD is better than the quality of matching based on the propensity score.

A limitation of CESD is that it does not deal with latent variables. In the next three chapters, we relax the restriction of the absence of latent variables and develop data-driven methods to identify adjustment sets from observational data with latent variables.

Chapter 5

Local causal search for efficient causal effect estimation from data with latent variables

In Chapter 4, we developed a theorem to support the use of SDR for finding proper adjustment sets and proposed an SDR-based algorithm, CESD, to estimate causal effects from observational data. However, CESD requires the assumption of no latent confounding variables, which is not always held in real-world applications. In this chapter, we present our first method (out of four) for identifying the adjustment set from data with latent variables. The method in this chapter addresses the challenges of latent variables and efficiency.

5.1 Motivation

It is challenging to obtain an unbiased estimation of causal effects from observational data with latent variables. As illustrated in Fig. 1.5, adjusting for variable C leads to M-bias, where C is related to both W and Y. From the data, it is difficult to distinguish the causal graphs in Fig. 1.4 (a) and Fig. 1.5 since they represent the same set of conditional independencies between measured variables. Therefore, the methods developed with the causal sufficiency assumption do not work on data with latent variables [3].

Additionally, the existing methods are very inefficient for estimating causal effects from data with latent variables. The main reason is that the underlying causal graph is rarely known in real-world applications. The commonly used approach is to recover a causal graph from data [49]. However, from the data, only a Markov equivalence class of causal graphs can be learned instead of a unique causal graph. Moreover, learning a global structure from data is NP-complete [68]. For example, LVIDA [63] first needs to learn a PAG of the underlying MAG from data. Then, LVIDA enumerates all valid MAGs represented in the learned PAG. Finally, for each MAG, all possible adjustment sets are determined for estimating all possible causal effects. All three steps result in the low efficiency of LVIDA. Hence, it is impossible for LVIDA to scale the number of variables.

To address these challenges of latent variables and the low efficiency caused by global structure learning, in this chapter, we develop a practical solution for efficient causal effect estimation from data with latent variables. Under the pretreatment variable assumption, our solution finds a superset of adjustment sets for unbiased causal effect estimation based on local causal structure learning considering latent variables. To achieve this goal, we develop a theorem to support the search for a superset of proper adjustment sets in the local causal structure around the treatment and outcome variables. Based on the theorem, we developed DICE (data-driven causal effect estimation), an effective and efficient algorithm for adjustment set discovery based on local causal structure and considering latent variables in data.

The chapter is based on the paper [74] "Causal query in observational data with hidden variables", which was published at the 24th European Conference on Artificial Intelligence, ECAI2020, Pages: 2,551-2,558.

5.2 Local search for an adjustment set

5.2.1 The Theorem

Let $\mathbf{D}$ be a dataset containing a binary treatment variable W, an outcome variable Y that is binary or numerical, and $\mathbf{X}$ be a set of all other variables

of any type. We assume that the dataset is generated from an underlying causal MAG $\mathcal{M}$, which is adjustment amenable to the ordered pair (W, Y). We also assume that all the variables in $\mathbf{X}$ are measured before applying the treatment and observing Y, indicating that variables in $\mathbf{X}$ are all non-descendants of W or Y in $\mathcal{M}$. The pretreatment variable assumption is realistic, as it reflects what normally happens in practice. That is, all the other variables are often measured before the treatment is observed and the outcome is observed. This assumption is commonly used in [37, 54, 57, 150].

In the proposed Theorem 5.1 below, $\mathcal{M}_{\underline{W}}$ denotes the manipulated $\mathcal{M}$ from which $W \to Y$ has been removed, and $\mathcal{M}_{\underline{W}}$ is still an MAG by the closure property of an MAG [98].

Theorem 5.1 (Adjustment set discovery via local search)**.** *Given a causal MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ that is adjustment amenable to the ordered pair (W, Y), there exists at least an adjustment set $\mathbf{Z} \subseteq \mathbf{X}$ for (W, Y) such that $\mathbf{Z}$ is a subset of $Adj(W \cup Y)$, where $Adj(W \cup Y)$ is the shorthand of $\{Adj(W) \setminus \{Y\}\} \cup \{Adj(Y) \setminus \{W\}\}$ in $\mathcal{M}_{\underline{W}}$.*

Proof. The requirement that $\mathcal{M}$ is adjustment amenable relative to the ordered pair (W, Y) ensures that the edge $W \to Y$ is visible and that the causal effect of W on Y is identifiable. There are two types of paths between W and Y, causal paths and back-door paths. In our problem setting, pretreatment variables $\mathbf{X}$ are non-descendants of W and Y, and hence, $\mathcal{M}_{\underline{W}}$ does not contain any causal path between W and Y.

To prove that an adjustment set $\mathbf{Z}$ in the MAG $\mathcal{M}$ is a subset of $Adj(W \cup Y)$, we show that the set $Pa(Y) \setminus \{W\}$ blocks all back-door paths from W to Y. Let π be any back-door path between W and Y, and let X denote the node adjacent to Y on π. Note that X is not W; for the assumption, $W \to Y$ is visible and so is not a back-door path. There are two cases as follows.

1. The edge between X and Y is $X \to Y$. In this case, X is a non-collider on π and is in the set $Pa(Y) \setminus \{W\}$, so π is blocked by $Pa(Y) \setminus \{W\}$.

2. The edge between X and Y is not $X \to Y$. Then, it must be a bidirected edge: $X \leftrightarrow Y$; otherwise, X is a descendant of Y and, hence, a descendant of W, contradicting the assumption. This means that X is not

an ancestor of Y and, hence, is also not an ancestor of W in the MAG $\mathcal{M}$. Then, on π, there must be a collider. Let C be the collider closest to X on π. Then, C must be a descendant of X (possibly X itself), which means that C does not have a descendant in $Pa(Y) \setminus \{W\}$; hence, π is blocked by $Pa(Y) \setminus \{W\}$.

Therefore, given an MAG $\mathcal{M}$, $Pa(Y) \setminus \{W\}$, which is a subset of $Adj(W \cup Y)$, is an adjustment set for the ordered pair (W, Y). □

5.2.2 Causal effect estimation

Based on the developed theorem, we can first employ a local causal structure discovery algorithm to obtain $Adj(W \cup Y)$ from data; then, our answer to a causal query is an adjustment set-causal effect table, or ASCET, where each row contains a candidate adjustment set (a subset of $Adj(W \cup Y)$) and the causal effect estimated corresponding to the candidate adjustment set.

Example. Assume $Adj(W \cup Y) = \{X_1, X_2, X_3\}$. Table 5.1 shows an example ASCET, where $\surd$ denotes that a variable is in a candidate adjustment set and $\times$ otherwise.

5.2.3 Removing insignificant adjustment variables

We reduce the size of ASCET by removing variables that do not significantly affect causal effect estimation. To test whether a variable significantly affects causal effect estimation, we compare the causal effect estimated by including the variable with the causal effect estimated by excluding the variable in all cases. If the average difference is smaller than a threshold, the variable is insignificant. More precisely, we have the following definition.

Table 5.1: An example ASCET.

X_1	X_2	X_3	CE
×	×	×	0.4
√	×	×	0.3
×	√	×	0.4
×	×	√	0.5
√	√	×	0.4
√	×	√	0.1
×	√	√	0.5
√	√	√	0.1

Table 5.2: The final ASCET.

X_1	X_3	CE
×	×	0.4
×	×	0.3
×	√	0.5
√	√	0.1

Definition 5.1. *For each $X \in Adj(W \cup Y)$, let $\mathbf{Z} \subseteq Adj(W \cup Y)$ be a candidate adjustment set containing X and $\mathbf{Z}' = \mathbf{Z} \setminus \{X\}$, and the sensitivity of X is defined as:*

$$Sen(X) = \frac{1}{\rho} \sum_{\mathbf{Z}} |CE_{\mathbf{Z}} - CE_{\mathbf{Z}'}| \tag{5.1}$$

where ρ is the size of the power set of $Adj(W \cup Y) \setminus \{X\}$ and $CE_{\mathbf{Z}}$ and $CE_{\mathbf{Z}'}$ are average causal effects of W on Y with the adjustment sets $\mathbf{Z}$ and $\mathbf{Z}'$, respectively.

The sensitivity of variable X reflects its impact on the estimated causal effect of W on Y. When X has low sensitivity, no significant error is introduced if we exclude it from $Adj(W \cup Y)$. Therefore, X is removed from $Adj(W \cup Y)$, and the size of ASCET is reduced.

Example. We use the example in Table 5.1 to illustrate how we reduce the size of ASCET. We use X_1 as an example, $\rho = 4$ as the size of the power set of $\{X_2, X_3\}$ is 4, and then $Sen(X_1) = \frac{1}{4}(|CE_{\{X_1\}} - CE_{\phi}| + |CE_{\{X_1,X_2\}} - CE_{\{X_2\}}| + |CE_{\{X_1,X_3\}} - CE_{\{X_3\}}| + |CE_{\{X_1,X_2,X_3\}} - CE_{\{X_2,X_3\}}|) = \frac{1}{4}(0.1 + 0 + 0.4 + 0.4) = 0.25$. In the same way, $Sen(X_2) = 0.025$ and $Sen(X_3) = 0.175$. Hence, X_2 has a low sensitivity such that we can exclude it from $Adj(W \cup Y)$ and the corresponding rows in ASCET can be removed. The final result is presented as Table 5.2.

5.2.4 The DICE algorithm

Algorithm 2 presents the proposed DICE algorithm. DICE algorithm contains two parts. Part 1 (lines 1 to 13) is for finding $Adj(W \cup Y)$. First, a local causal structure learning algorithm (such as PC-Select [34] or HITON-PC [36]) is used to find the parents of W and Y. In this work, we use PC-Select. Given its high accuracy, PC-Select is a local version of PC [18], an algorithm for learning a global causal structure. Then, we check whether $Adj(W) = Adj(W) \setminus \{Y\}$ is empty to determine whether the causal effect of W on Y is identifiable from the data. If $Adj(W)$ is empty, no adjustment set can be found, and DICE returns an empty ASCET and terminates.

Part 2 of Algorithm 2 (lines 7 to 16) first estimates each possible causal effect when adjusting each subset of $Adj(W \cup Y)$. We use propensity score matching (PSM) [22] to estimate causal effects, where the propensity score is calculated by *glm* and matching is performed by *Match* in the $\mathbb{R}$ packages *stats* and *Matching* [156], respectively. The adjustment sets and the corresponding causal effects estimated are added to the ASCET. Next, DICE calculates the sensitivity of each variable in $Adj(W \cup Y)$, and if a variable's sensitivity is below the given threshold, DICE removes all the adjustment sets containing the variable and the corresponding causal effects of W on Y ($CE(W, Y)$ for short) from the ASCET.

Time-complexity analysis. PC-Select and PSM contribute the most to the time complexity of DICE. In the worst case, the complexity of PC-Select is $\mathbf{O}(p2^{p-1}n)$ [34], where p is the number of variables and n is the number of samples. In our problem, the worst case is when all variables of $\mathbf{X}$ are causes of W and/or Y, which, however, rarely occurs. In most cases, PC-Select can handle thousands of variables [34]. The time complexity of PSM is $\mathbf{O}(p^2)$. Our experiment in Section 5.3 shows that DICE scales with both p and n well.

Algorithm 2 The data-dreven causal effect estimation algorithm (DICE)

Input: Dataset $\mathbf{D}$ with a treatment variable W, the set of pretreatment variables $\mathbf{X}$, an outcome variable Y and a sensitivity threshold τ.

Output: The ASCET.

1: Call a local causal structure learning algorithm to find $Adj(W) \setminus \{Y\}$ and $Adj(Y) \setminus \{W\}$ from $\mathbf{D}$, respectively.
2: $ASCET \leftarrow \varnothing$
3: **if** $Adj(W)$ is empty **then**
4: Adjustment set could not be found.
5: **return** $ASCET$
6: **else**
7: **for** each subset $\mathbf{Z} \subseteq Adj(W \cup Y)$ **do**
8: calculate $CE(W, Y)$ by adjusting $\mathbf{Z}$.
9: add pair $(\mathbf{Z}, CE(W, Y))$ to ASCET.
10: **end for**
11: For each $X \in Adj(W \cup Y)$ calculate $Sen(X)$.
12: **if** $Sen(X) < \tau$ **then**
13: remove from ASCET all the rows containing X.
14: **end if**
15: **return** $ASCET$
16: **end if**

5.3 Experiments

5.3.1 The quality of local structure learning

Before evaluating DICE, we assess the quality of the local structure learning algorithm, i.e., PC-Select used in this work versus global structure learning algorithms. The synthetic datasets are generated according to the underlying causal DAG in Fig. 5.1 (a) with 8 pretreatment variables, 2 latent variables, W and Y by following the procedure in [7]. The underlying causal MAG is presented in Fig. 5.1 (b), i.e., U_1 and U_2 are removed from these generated datasets.

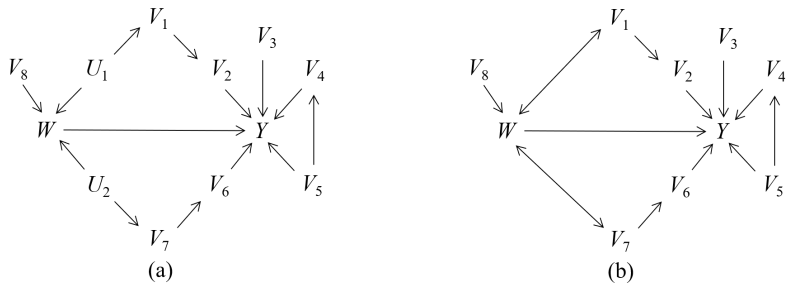


Figure 5.1: (a) A DAG with two latent variables U_1 and U_2. (b) The corresponding MAG of the DAG.

We generate two groups of datasets to evaluate the impact of sample size and the number of variables on the quality of the local structure learning algorithm. The first group of datasets includes 10 datasets, each containing 5 k samples for 10 variables generated from Fig. 5.1 (b), plus 0, 10, 20, ..., and 90 random variables. The second group of datasets includes 20 datasets with 10 fixed variables of the MAG but with varying sample sizes of 5 k, 10 k, ..., 95 k, and 100 k. These datasets with latent variables are used to evaluate the quality of the local structure learning algorithm.

We use two global structure learning algorithms, FCI [18] and RFCI [35], to learn from each of the generated datasets a PAG, which represents a Markov equivalence class of MAGs encoding the same dependence/independence relations in the data. Then, we extract $Adj(W)$ and $Adj(Y)$ from the learned PAG. For PC-Select, we apply it twice to a dataset, one having W as the target and the other having Y as the target to learn $Adj(W)$ and $Adj(Y)$. The implementations of FCI, RFCI, and PC-Select are from the $\mathbb{R}$ package *pcalg* [157], with the default parameter settings and a significance level (α) of 0.05.

Results. We perform the experiments 10 times on each of the 30 datasets. We draw the mean results (precision, recall, and F-score) of the three algorithms in Fig. 5.2. The local structure learning algorithm PC-Selects achieves similar performance as the global structure learning algorithm in $Adj(W)$ and $Adj(Y)$, indicating that adjustment set discovery in data by local search is reliable.

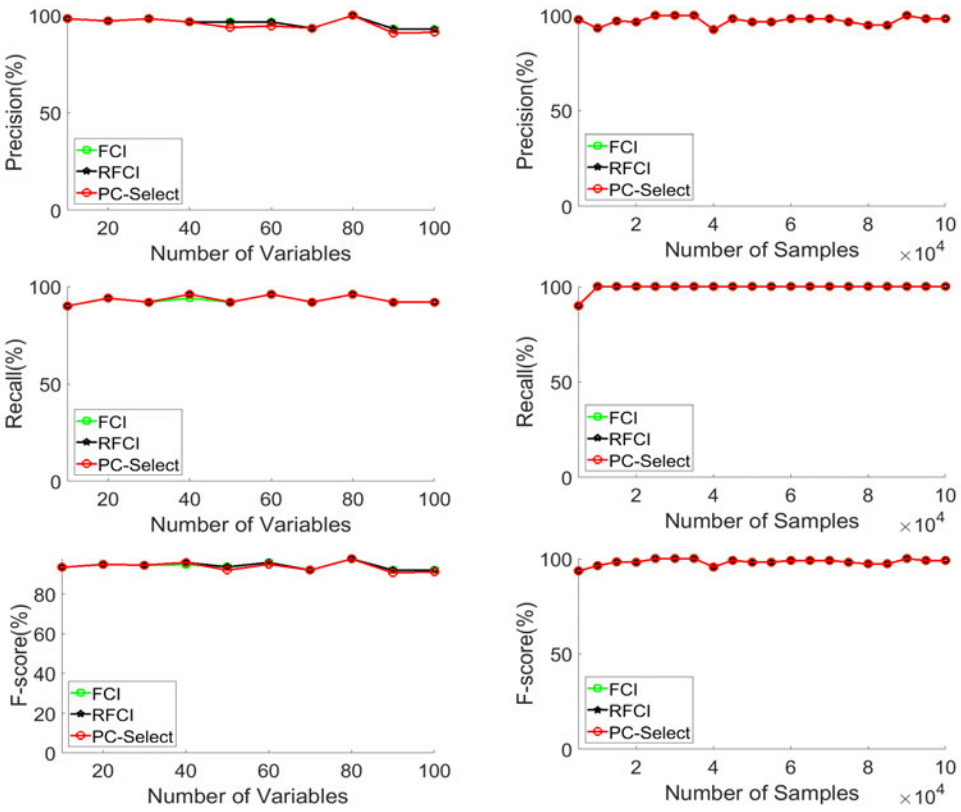


Figure 5.2: The precision, recall and F-score of FCI & RFCI & PC-Select algorithms on two groups of synthetic datasets w.r.t. the number of variables and samples. The qualities of parent discovery by the three algorithms are very consistent. but the time efficiencies of the algorithms are quite different, as shown in Fig. 5.4

5.3.2 Evaluating DICE with real-world datasets

We evaluate the effectiveness of DICE on three real-world datasets: Job Training (Jobs) [15], IHDP [37] and Twins [152]. A detailed description of these three datasets can be found in Section 4.4 of Chapter 4.

The job data contain records of people taking part in the training only. As in [15], we estimate the average causal effect on treated (ACT) for DICE and all comparisons against the ground truth of ACT, which is $886 [150]. For both IHDP and Twins, we estimate the average causal effect (ACE) for DICE and all comparisons against the ground truth of ACE, which are 4.38

Table 5.3: Results on Jobs. Bias (%) is the relative error comparing to the ground-truth causal effect. The estimation with the lowest bias is highlighted in each category.

Method	ACT	Bias (%)	Remarks
LVIDA	-6,645.1	827.4%	The best of 4 estimates
DICE	**834.7**	**5.8**%	The best of 128 estimates
DICE	1,745.5	97.0%	The most probable estimate
PSM	-947.6	201.0%	
CBPS	423.3	52.0%	
CFR	**742.0**	**16.0%**	
LASSO	-475.3	153.6%	
BART	-245.2	127.7%	
TMLE	-1,901.4	314.6%	
CF	-4,438.4	600.9%	The best of 30 forests

and -0.025, respectively.

There are two major objectives for the experiments.

First, we show that the true adjustment set can be found, as shown in Theorems 5.1, and therefore, the smallest bias of DICE will be very small. We compare DICE with a data-driven causal effect estimation method, LVIDA [63][1], which uses FCI to learn a PAG and searches for an adjustment set in each MAG enumerated from the PAG. LVIDA returns a set of causal effects, one for the adjustment set discovered from each MAG. In the experiment, we report the best estimates for both methods to show the quality of adjustment sets found.

Second, we show that a causal effect estimated by DICE is comparative with the best estimates by methods using domain knowledge. Some statistical and machine learning methods are available for causal effect estimation, with the assumptions of a known covariate set and unfoundedness. The typical methods include **PSM**, propensity score matching with logistic regression [22]; **CBPS**, covariate balancing propensity score [150]; **CFRNet**,

[1] https://github.com/dmalinsk/lv-ida

Table 5.4: Results on IHDP w.r.t. CE and bias.

Method	ACE	Bias(%)	Remarks
LVIDA	3.97	9.2%	The best of 300 estimates
DICE	**4.41**	**0.7**%	The best of 16 estimates
DICE	**4.48**	**2.5**%	The most probable estimate
PSM	4.95	13.2%	
CBPS	4.56	4.3%	
CFR	4.86	11.0%	
LASSO	5.06	15.7%	
BART	4.69	7.3%	
TMLE	4.86	11.2%	
CF	4.25	2.9%	The best of 30 forests

a deep learning framework for counterfactual regression with a theoretical error bound (named CFR) [39]; **LASSO**, linear regression with the regularization ℓ_1-norm to predict the factual outcome [158]; **BART**, Bayesian additive regression tree [159], a nonlinear model that has been applied for counterfactual inference [37]; **TMLE**, targeted maximum likelihood estimation, a doubly robust method [160]; and **CausalForest**, random forest regression to estimate causal effect [151] (**CF**). In the experiments, we compare DICE with all the abovementioned methods.

The three datasets favour the above methods since the unconfoundedness assumption is satisfied. Since the datasets were obtained from well-designed observational studies, the covariates were chosen by domain experts.

DICE returns ASCET, the set of adjustment sets and causal effect pairs. When reporting causal effects estimated by DICE, we use the most likely value in the ASCET after removing insignificant variables with the sensitivity threshold of $\tau = 0.1$. We group the causal effect in the ASCET by using the bin size of $(\max(Y)/100)$. The average causal effect in the most frequent bin is used as an estimated causal effect.

The parameter setting of DICE is that the *match* function of an estimate

Table 5.5: Results on Twins w.r.t. CE and bias.

Method	ACE	Bias (%)	Remarks
LVIDA	-0.007	72%	The best of 36 estimates
DICE	**-0.027**	**8.0**%	The best of 64 estimates
DICE	-0.039	59.2%	The most probable estimate
PSM	-0.018	56.7%	
CBPS	-0.010	59.4%	
CFR	-0.011	56.3%	
LASSO	-0.016	36.2%	
BART	-0.010	41.8%	
TMLE	-0.016	32.9%	
CF	**-0.017**	**32.6%**	The best of 30 forests

is set to "ATT" for Jobs and "ATE" for IHDP and Twins, as in the prior work in [37, 39]. Other parameters are set as the default. For CF, the parameter *num.trees* is set from 10 to 300 with an increment of 10, and the best result is reported. For PSM and CBPS, all variables are included in the adjustment variable set.

The experimental results are shown in Tables 5.3, 5.4, 5.5, and Fig. 5.3. From the results, we make the following conclusions.

First, DICE finds the correct adjustment set through local search. Compared with the other methods, the lowest biases of DICE are consistently the smallest in the three datasets and are significantly smaller than the lowest biases of LVIDA, which searches adjustment sets by the global search.

Second, causal effects estimated by DICE are comparable with those estimated by other methods that use domain knowledge. For the IHDP dataset, DICE achieves the smallest bias. For the Jobs dataset, DICE is ranked 3rd based on the smallest biases. For the Twin dataset, the largest absolute bias is only 0.015, and hence, we consider that the estimates by all methods are similar and good.

To show that those comparison methods do not work in a dataset when the unconfoundedness assumption is not satisfied, we generated datasets

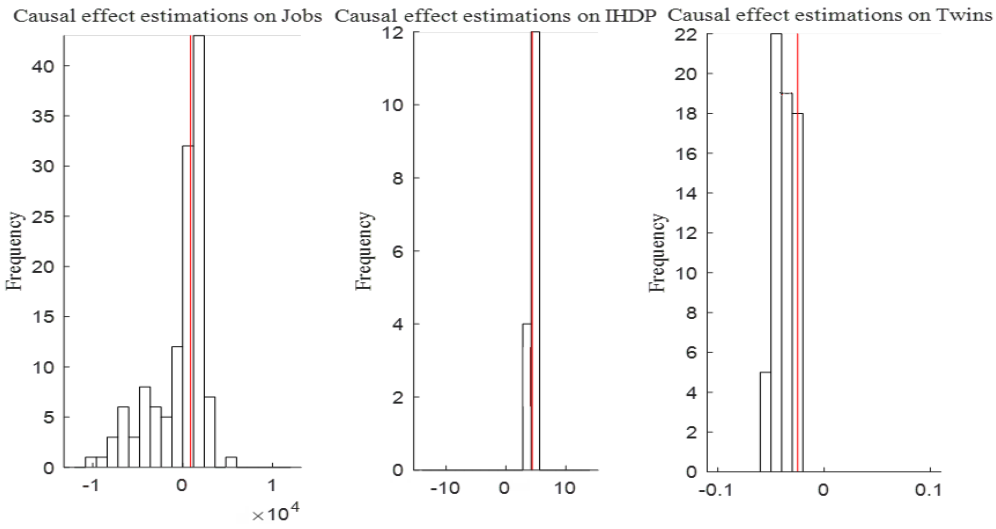


Figure 5.3: Histograms of causal effects in ASCETs for Jobs, IHDP and twins. The red line denotes the ground truth, and the bin size is $(\max(Y)/100)$.

with M-bias [3]. The performance of some methods deteriorates greatly. Implementations of other methods do not work with datasets with more than 200 k records. The experimental results are shown in Chapter 6.

5.3.3 Efficiency evaluation of DICE

As PC-Select is a major contributor to the complexity of DICE, in this section, we first evaluate the efficiency of PC-Select and then the overall efficiency of DICE. The computations are performed on a PC with 2.6 GHz Intel Core i7 and 16 GB of memory.

Efficiency evaluation of the structure learning algorithms. We record the running time of the structure learning algorithms on the synthetic datasets introduced in Section 5.3.1 and draw the mean running time in Fig. 5.4. PC-Select is faster than FCI and RFCI, and compared to FCI and RFCI, PC-Select scales with the number of variables and samples very well.

Experiments on ten standard benchmark BNs. To evaluate the efficiency of DICE, we use the datasets generated from the ten benchmark

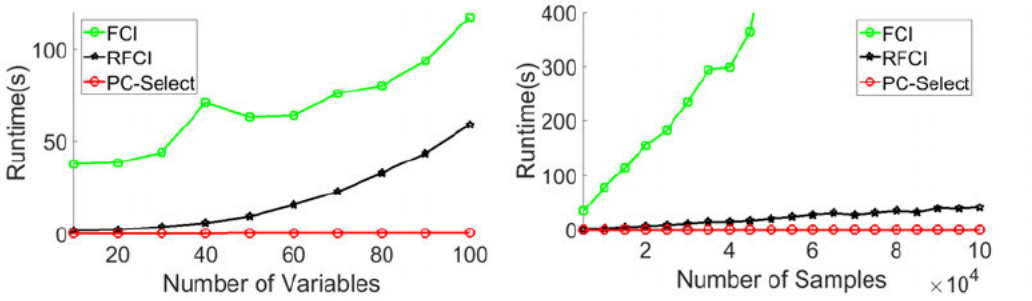


Figure 5.4: The runtime on two groups of synthetic datasets w.r.t. the number of variables and samples.

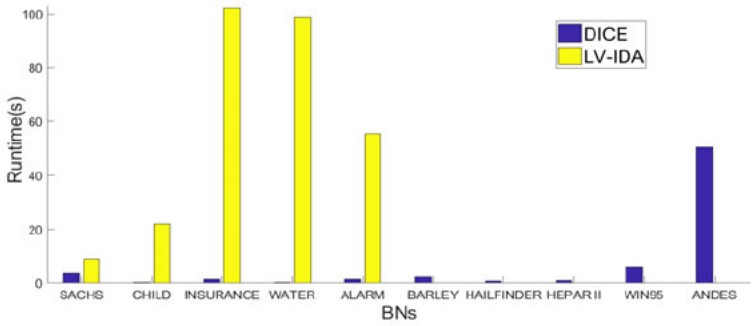


Figure 5.5: The runtime on ten BNs. Note: LVIDA did not return results in two hours on BARLEY, HAILFINDER, HEPAR II, WIN95 and ANDES.

BNs from the BN repository[1]: SACHS, CHILD, INSURANCE, WATER, ALARM, BARLEY, HAILFINDER, HEPAR II, WIN95 and ANDES. With each BN, we choose a variable without child nodes as the outcome variable and then select one of its parents without child nodes (except the outcome) as the treatment variable. We generate ten synthetic datasets from the ten BNs with 5,000 samples each using the $\mathbb{R}$ package *bnlearn* [161]. Then, we hide 5% variables (only one variable is latent for the small BNs SACHS and CHILD), which lie on a noncausal path between the treatment and outcome variables. The same parameter settings are used for DICE and LVIDA as in Subsection 5.3.2.

Results. As shown in Fig. 5.5, DICE is faster than LVIDA across all 10 datasets. In particular, LVIDA did not return results for the 5 datasets

[1] http://www.bnlearn.com/bnrepository/

generated using the larger BNs within two hours, while DICE was completed in seconds (at most 50+ seconds) for all datasets.

5.4 Conclusion

The existing methods for adjustment set identification from data with latent variables either use global search or exhaustive local search, both leading to the low efficiency of these algorithms.

In this chapter, we developed a theorem to support the use of efficient local causal structure discovery by assuring the correctness of the result of causal effect estimation obtained with local search. Based on the theorem, we developed DICE to query data for causal effects and the confounders potentially impacting the causal effects. The results returned by DICE provide decision makers not only valuable information on the effect of changing one variable in their systems but also awareness of the other factors that can influence the effect.

Experiments with synthetic and real-world data demonstrated the effectiveness and efficiency of DICE. It was shown that DICE produces better causal effect estimation than LVIDA and works on datasets where LVIDA fails. The causal effects estimated by DICE are as good as the state-of-the-art methods that use domain knowledge.

Chapter 6

Local causal search for efficient and unique causal effect estimation

In Chapter 5, we discussed the local search for estimating causal effects from data with latent variables. The developed DICE algorithm provides a scalable method for answering causal queries from data with latent variables. However, the DICE algorithm, like most other data-driven methods, only provides a bound estimation of causal effects. In most applications, investigators are interested in a unique or deterministic causal effect estimation. Hence, in this chapter, we study how to obtain unique and unbiased causal effect estimations from data with latent variables.

6.1 Motivation

Most data-driven methods that are based on graphical causal modelling can only provide a bound estimation of causal effects when the causal effect is identifiable. The reason is that we can only learn a Markov equivalence class of causal structures from data, instead of a unique causal structure [18, 48], and we may not know which structure in the equivalence class is the true causal structure. Consequently, the causal effect estimation based on the equivalence class of causal structures is a bound estimation of causal effects,

i.e., a multi-set of all possible causal effects, which is from the uncertainty in adjustment set identification. Each causal effect is estimated by adjusting for a possible adjustment set that is identified from a causal structure represented in the equivalence class.

In a complex dataset, the number of returned possible adjustment sets can be very large, which makes the queried answer less useful. Therefore, it is necessary and meaningful to propose methods to obtain a small number of adjustment sets. In this chapter, we consider identifying *the minimal adjustment sets* (Definition 6.3) by local search from data with latent variables. It is worth noting that identifying minimal adjustment sets are important for the application of causal effect estimation in practice because large-sized adjustment sets can lead to poor estimator performance and make the queried answer less useful [56, 57, 66]. For example, a biologist has only a limited budget to attempt a few experiments, and she cannot try many possible adjustment sets. An advertisement company can try a limited number of A/B tests to make a promotion decision.

It is desirable to obtain unbiased and unique causal effect estimation by adjusting for minimal adjustment sets identified from the local causal structure when estimating causal effects from data with latent variables. The current methods handling data with latent variables all produce bound estimations, except for the EHS method [66]. However, EHS is very inefficient due to the exhaustive local search the method conducts. Therefore, in this chapter, we develop new theorems based on the EHS condition (as described in Theorem 6.1) to support the local search for adjustment sets when the causal effect is identifiable and develop an efficient algorithm based on the new theorems to achieve unique and unbiased causal effect estimation from data with latent variables.

However, it is impossible to estimate causal effects from data uniquely and unbiasedly without additional constraints. Thus, in this chapter, we identify a practical problem setting (details in Section 6.2) under which a causal effect can be estimated from data uniquely. Then, we develop new local search theorems with the setting to support local search for proper adjustment sets from data with latent variables. Based on the theorems,

we propose a new data-driven algorithm, CE^2LS (causal effect estimation by local search), to find *minimal* adjustment sets for *unique* and *unbiased* causal effect estimation from data with latent variables. CE^2LS is fast, as it conducts a local search for adjustment sets and employs an efficient *pattern mining strategy*. To the best of our knowledge, CE^2LS is the first practical method based on the local search for finding minimal adjustment sets from data with latent variables.

The chapter is based on the paper [75] " Local search for causal effect estimation", which was submitted to the Thirty-fifth Conference on Neural Information Processing Systems, Neurips2021 and is currently under review.

6.2 The theorems for local causal search

In this section, we first introduce our problem setting. After that, we present the local search theorems and corollaries for determining proper adjustment sets.

6.2.1 Problem setting

The EHS condition [66] can infer proper adjustment sets based on the conditional independence and dependence tests directly from data with latent variables. The EHS condition searches for proper adjustment sets can be described as the following theorem.

Theorem 6.1 (EHS condition [66])**.** *Given a dataset* $\mathbf{D}$ *with treatment variable* W*, outcome variable* Y *and a set of pretreatment variables* $\mathbf{X}$*. If there exists a variable* $S \in \mathbf{X}$ *and a set* $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$ *such that (1)* $S \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ *and* $S \perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{W\}$*, then* $\mathbf{Z}$ *is an adjustment set.*

The EHS condition fully supports a data-driven approach (i.e., EHS algorithm) to determine an adjustment set directly from data with latent variables. However, the EHS algorithm [66] is very inefficient due to the global search required to search for S and $\mathbf{Z}$ in $\mathbf{X}$.

We now define the problem to be solved in this chapter and then develop the theorems for local search.

Definition 6.1 (Problem Definition)**.** *Given a dataset* $\mathbf{D}$ *generated from an underlying MAG* $\mathcal{M} = (\mathbf{V}, \mathbf{E})$*, where* $\mathbf{V} = \{W, Y\} \cup \mathbf{X}$*, and* $W \to Y$ *is in* $\mathcal{M}$ *as a visible edge. We identify a proper adjustment set* $\mathbf{Z} \subseteq \mathbf{X}$ *w.r.t.* (W, Y) *in the local structures around* Y *and* W*,* $Pa(Y) \setminus \{W\}$*,* $Adj(Y) \setminus \{W\}$ *or* $\{Adj(W) \cup Adj(Y)\} \setminus \{W, Y\}$ *based on the conditional independence tests in the EHS condition (Theorem 6.1).*

Assuming data are generated from an *underlying* MAG is fundamentally different from assuming a known MAG from which adjustment sets are identified. The underlying MAG is unknown to us, and we identify adjustment sets from data directly. Additionally, the pretreatment variable assumption is widely accepted in practice [37, 151] and used by existing methods [54, 57, 66].

6.2.2 The theorems for local search

In our problem setting (i.e., data generation is governed by an underlying MAG containing $W \to Y$ as a visible edge, and the pretreatment variable assumption holds). Based on the generalized adjustment criterion (Definition 3.19), we have the following corollary.

Corollary 1. *In our problem setting, in the underlying MAG* $\mathcal{M}$*, if* $\exists \mathbf{Z} \in \mathbf{V} \setminus \{W, Y\}$*, which blocks all definite status non-causal paths from* W *to* Y*,* $\mathbf{Z}$ *is a proper adjustment set, i.e.,* $ACE(W, Y)$ *can be estimated unbiasedly by adjusting for* $\mathbf{Z}$ *as in Eq. (3.2).*

Proof. There exists a visible edge $W \to Y$, so $\mathcal{M}$ is adjustment amenable relative to (W, Y) based on the definition of amenability [100]. Under the pretreatment variable assumption, the *forbidden set* is $\emptyset$ because for $\forall X \in \mathbf{X}$, $X \notin De(W)$ in $\mathcal{M}$. Hence, according to the generalized adjustment criterion 3.19, if $\exists \mathbf{Z} \in \mathbf{V} \setminus \{W, Y\}$ satisfies the third condition of the generalized adjustment criterion, then $\mathbf{Z}$ is an adjustment set w.r.t. (W, Y). Therefore, the causal effect of W on Y can be estimated unbiasedly by adjusting for $\mathbf{Z}$ as in Eq. (3.2). □

The third condition of the generalized adjustment criterion 3.19 requires that all definite status non-causal paths between W and Y are blocked by $\mathbf{Z}$. Obviously, under the assumptions that $\mathbf{X}$ includes the pretreatment variables only and there exists a visible edge $W \to Y$, then other paths from W to Y are definite status non-causal paths. Thus, in this chapter, according to Definition 3.15, a definite status non-causal path from W to Y is a generalized back-door path from W to Y.

Searching for $\mathbf{Z}$ in the space of $\mathbf{X}$ is infeasible when the size of $\mathbf{X}$ is large, hence, the following theorem for local search.

Theorem 6.2 (Adjustment set in local structure)**.** *In our problem setting,* $Pa(Y) \setminus \{W\}$ *is an adjustment set w.r.t.* (W, Y).

Proof. To prove Theorem 6.2, we show that the three conditions of the generalized adjustment criterion (Definition 3.19) are satisfied by $Pa(Y) \setminus \{W\}$. According to Corollary 1, the first two conditions are satisfied. Therefore, in the following, we prove the third condition, i.e., $Pa(Y) \setminus \{W\}$ blocks all generalized back-door paths from W to Y.

Let π be an arbitrary generalized back-door path from W to Y. Under the problem setting, π must be in the form of $W \leftarrow * \ldots X * \rightarrow Y$, where $X \in \mathbf{X}$. There are two cases to consider.

1. π ends with $X \to Y$, i.e., $X \in Pa(Y) \setminus \{W\}$. In this case, X is a non-collider on π, so π is blocked by X. Moreover, for any other parent of Y in $Pa(Y) \setminus \{W\}$, denoted as P, if it is on π, adding P to the current blocking set (i.e., $\{X\}$) will not open π; thus, $Pa(Y) \setminus \{W\}$ blocks π.

2. π ends with $X \leftrightarrow Y$, i.e., $X \in Sp(Y) \setminus \{W\}$. This means that X is not an ancestor of Y and there must be a collider on π; otherwise, there is a directed path from X to W. Then, because $\mathcal{M}$ contains $W \to Y$, X is an ancestor of Y, which contradicts the previous statement that X is not an ancestor of Y. Furthermore, let C be the collider closest to X on π. There are two possible cases for C. (1) $C = X$, then we have $* \rightarrow C \leftrightarrow Y$. In this case, $\forall X' \in Pa(Y) \setminus \{W\}$, X' must not be in $De(C)$; otherwise, $C \in An(Y)$, which contradicts the fact that X (i.e., C) is not an ancestor of Y. As a result, adding $Pa(Y) \setminus \{W\}$ to the current blocking set (i.e., empty set) will not open π. (2) $C \neq X$ and C must be a descendant of

X, i.e., $*\!\!\rightarrow C \leftarrow \cdots \leftarrow X \leftrightarrow Y$. In this case, for the same reason as in case (1), $\forall X' \in Pa(Y) \setminus \{W\}$, X' must not be in $De(C)$. Hence, adding $Pa(Y) \setminus \{W\}$ to the current blocking set (i.e., empty set) will not open π. Therefore, $Pa(Y) \setminus \{W\}$ blocks π.

Thus, we can conclude that $Pa(Y) \setminus \{W\}$ blocks all generalized backdoor paths from W to Y, and $Pa(Y) \setminus \{W\}$ must be a proper adjustment set relative to (W, Y). □

Theorem 6.2 helps restrict the search of an adjustment set to the local structure around Y. However, from the data, we can only find $Adj(Y)$ instead of $Pa(Y)$. Therefore, we have the following corollary, which uses $Adj(Y)$ instead of $Pa(Y)$.

Corollary 2. *In our problem setting, assume that there exists a cause Q of W and $Q \notin Adj(Y)$. A proper adjustment set $\mathbf{Z}$ w.r.t. (W, Y) can be identified in $Adj(Y) \setminus \{W\}$ using the conditional independence test $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$.*

Proof. $Pa(Y) \setminus \{W\}$ is a proper adjustment set according to Theorem 6.2. Since $Pa(Y) \setminus \{W\} \subseteq Adj(Y) \setminus \{W\}$, there must exist at least one proper adjustment set $\mathbf{Z} \subseteq Adj(Y) \setminus \{W\}$. Because $Q \notin Pa(Y) \setminus \{W\}$ and $Q \not\perp\!\!\!\perp W$, a set $\mathbf{Z} \subseteq \mathbf{X} \setminus \{Q\}$ satisfies $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$; then, $\mathbf{Z}$ is an adjustment set by the EHS search in Theorem 6.1. Thus, an adjustment set $\mathbf{Z} \subseteq Adj(Y) \setminus \{W\}$ can also be identified using $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$. □

Corollary 2 enables us to use the EHS condition to find an adjustment set locally, but it may not be practical either because finding a cause of W from data is difficult, as we may not be able to distinguish a cause from a spouse. Therefore, we relax the requirement of a cause of W to be a COSO (cause or spouse of the treatment only) variable as defined below.

Definition 6.2 (COSO)**.** *In our problem setting, a variable $S \in \mathbf{V} \setminus \{W, Y\}$ is a COSO variable w.r.t. (W, Y) if and only if S is adjacent to W and not adjacent to Y in $\mathcal{M}$. That is, $S \in Adj(W) \setminus (Adj(Y) \cup \{Y\})$.*

From the above definition, we see that it is easier to identify a COSO variable from data. Given a COSO variable and under our problem setting, first, we can state the EHS condition as the following theorem.

Theorem 6.3 (Adjustment set identification given a COSO variable)**.** *In our problem setting, assume that there exists a COSO variable* S*,* $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$ *is a proper adjustment set w.r.t.* (W, Y) *if* $S \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$ *holds.*

Proof. There are two types of COSO variables, i.e., either $S = Q$ or $S \in Sp(W)$. For $S = Q$, Corollary 2 proved that a proper adjustment set $\mathbf{Z} \subseteq Adj(Y) \setminus \{W\}$ can be identified by $Q \perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{W\}$. Thus, in the following proof, we consider the second type, i.e., $S \in Sp(W)$.

The first two conditions of the generalized adjustment criterion are satisfied by Corollary 1, so if $\mathbf{Z}$ satisfies the third condition of the generalized adjustment criterion, then $\mathbf{Z}$ is an adjustment set. In other words, we need to prove that if $S \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$ holds, then $\mathbf{Z}$ blocks all definite status non-causal paths from W to Y, i.e., all generalized back-door paths from W to Y. We prove this using contradiction. Assume that a generalized back-door path from W to Y, π, is m-connecting given $\mathbf{Z}$, i.e., π is not blocked by $\mathbf{Z}$. There are the following two cases of π (depending on whether π includes S): (1) π contains S and thus π is of the form $W \leftrightarrow S \ldots *\!\!\rightarrow Y$, then S is m-connected to Y given $\mathbf{Z} \cup \{W\}$, i.e., $S \not\perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$ and this contradicts $S \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$. (2) π does not contain S. Then, the concatenation of the edge $S \leftrightarrow W$ and π produce a path of the form $S \leftrightarrow W \leftarrow\!\!* X \ldots *\!\!\rightarrow Y$, where X is the adjacent node of W on π. This path is an m-connecting path from S to Y given $\mathbf{Z} \cup \{W\}$ because S is m-connected to X given W (note that W is a collider) and W is m-connected to Y given $\mathbf{Z}$. This means $S \not\perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$, which contradicts $S \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$. Thus, if $S \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$ holds, $\mathbf{Z}$ blocks all generalized back-door paths from W to Y. Therefore, all three conditions of the generalized adjustment criterion are satisfied, and $\mathbf{Z}$ is a proper adjustment set w.r.t. (W, Y). □

Note that Theorem 6.3 still requires a global search in $\mathbf{X}$, so for local search, we present the following corollary.

Corollary 3. *In our problem setting, assume that there exists a COSO variable* S*,* $\mathbf{Z} \subseteq Adj(Y) \setminus \{W\}$ *is a proper adjustment set w.r.t.* (W, Y) *if* $S \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$ *holds.*

Proof. $Pa(Y) \setminus \{W\}$ is a proper adjustment set according to Theorem 6.2. Since $Pa(Y) \setminus \{W\} \subseteq Adj(Y) \setminus \{W\}$, there must exist at least one proper

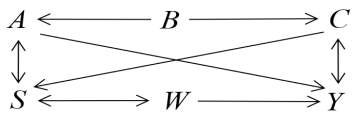


Figure 6.1: A counterexample in which Corollary 3 fails to select an adjustment set by local search.

adjustment set $\mathbf{Z} \subseteq Adj(Y) \setminus \{W\}$. From Theorem 6.3, a set $\mathbf{Z}$ satisfies $S \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$, and then $\mathbf{Z}$ is a proper adjustment set. Hence, a set $\mathbf{Z} \subseteq Adj(Y) \setminus \{W\}$ satisfies $S \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$; then, $\mathbf{Z}$ is a proper adjustment set relative to (W, Y). □

Corollary 3 provides a sound solution for finding an adjustment set locally. However, it may not ensure the same level of completeness as the EHS condition (i.e., Theorem 6.3 in our problem setting and given a COSO variable), in the sense that there may exist an adjustment set that satisfies the EHS condition but does not reside in the local structure searched. We provide an example to show this.

Example 6.1. *Refer to the MAG in Fig. 6.1, denoted as $\mathcal{M}$, which contains a pair of variables (W, Y), and $\mathbf{X} = \{S, A, B, C\}$ are all pretreatment variables. Note that S is a COSO variable and $W \to Y$ is visible in $\mathcal{M}$.*

For every $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$, $W \not\perp\!\!\!\perp Y \mid \mathbf{Z}$, and $S \not\perp\!\!\!\perp Y \mid \mathbf{Z}$. Note that $Adj^{\prime}(W \cup Y) = \{S, A, C\}$ and for every $\mathbf{Z} \subseteq Adj^{\prime}(W \cup Y) \setminus \{S\}$, we have $S \not\perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{W\}$ since adjusting for $\{A, C\}$ only leaves the path $S \leftrightarrow A \leftarrow B \to C \leftrightarrow Y$ open in $\mathcal{M}$.

However, the simulations in Section 6.4.6 indicate that the chance of failing to find an adjustment set using Corollary 3 is small ($< 10\%$).

To mitigate the completeness issue with Corollary 3 without sacrificing much efficiency, in our proposed CE^2LS algorithm, we extend the search space to $Adj(W \cup Y) \setminus \{W, Y\}$, where $Adj(W \cup Y)$ is a short form of $Adj(W) \cup Adj(Y)$. Our experiments have shown that the extended search space improves completeness. However, in $Adj(W \cup Y) \setminus \{W, Y\}$, there might be multiple adjustment sets, and randomly using one of the adjustment sets

can cause random fluctuations in causal effect estimation [162]. Therefore, CE^2LS first uses each of the adjustment sets found to obtain an estimation of the causal effect and then averages over all the estimated causal effects as its output.

Furthermore, a large adjustment set requires more data samples for reliable statistical tests [57], so we only aim to find all minimal adjustment sets in $Adj(W \cup Y) \setminus \{W, Y\}$.

Definition 6.3 (A minimal adjustment set)**.** **Z** *is a minimal adjustment set if no proper subset of* **Z** *is an adjustment set.*

6.3 The CE^2LS algorithm

In this section, we present the details of the CE^2LS algorithm. To achieve high efficiency, in addition to local search, CE^2LS employs a frequent pattern mining approach (the *apriori* approach) [163] to conduct a level-wise search in $Adj(W \cup Y) \setminus \{W, Y\}$ to discover minimal adjustment. The level-wise search is based on the fact that a superset of a minimal adjustment set is not a minimal adjustment set and hence satisfies the upwards closure property as the frequent patterns.

As shown in Algorithm 3, CE^2LS first (lines 1 to 3) searches for $Adj'(W \cup Y)$ (i.e., $(Adj(W) \cup Adj(Y)) \setminus \{W, Y\}$) and the set of COSO variables $\mathbf{COSO} = Adj(W) \setminus (Adj(Y) \cup \{Y\})$. Line 1 finds $Adj(W)$ and $Adj(Y)$ based on a local causal structure learning algorithm. There are a few local search algorithms available for discovering the adjacent nodes of W and Y from data, such as *PC. Select* [34], *MMPC* and *HITON-PC* [36]. In our implementation, the *PC. Select* algorithm is employed.

Algorithm 3 Causal effect estimation by local search (CE^2LS)

Input: Dataset $\mathbf{D}$ with the treatment W, pretreatment variables $\mathbf{X}$, the outcome Y.

Output: average $ACE(W, Y)$.

1: call a local structure learning algorithm to search for $Adj(W)$ and $Adj(Y)$ from $\mathbf{D}$.
2: $Adj'(W \cup Y) = Adj(W) \cup Adj(Y) \setminus \{W, Y\}$
3: **COSO** $= Adj(W) \setminus (Adj(Y) \cup \{Y\})$
4: let $\psi = \emptyset$; /*Search for minimal adjustment sets, and calculate $ACE(W, Y)$*/
5: **for** each variable $S \in$ **COSO do**
6: $\quad \mathbf{C}_1 = Adj(W \cup Y) \setminus \{W, Y\} \setminus \{S\}$
7: $\quad k = 1$
8: $\quad$ **while** $\mathbf{C}_k \neq \emptyset$ **do**
9: $\quad\quad$ **for** each element $\mathbf{Z} \in \mathbf{C}_k$ **do**
10: $\quad\quad\quad$ **if** $\mathbf{Z}$ is not in ψ **then**
11: $\quad\quad\quad\quad$ **if** $S \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$ **then**
12: $\quad\quad\quad\quad\quad$ calculate $ACE(W, Y)$ via Eq. (3.2) and remove $\mathbf{Z}$ from $\mathbf{C}_k$
13: $\quad\quad\quad\quad\quad$ $\psi \leftarrow \cup(\mathbf{Z}, ACE(W, Y))$
14: $\quad\quad\quad\quad$ **end if**
15: $\quad\quad\quad$ **end if**
16: $\quad\quad$ **end for**
17: $\quad\quad$ $k = k + 1$
18: $\quad\quad$ $\mathbf{C}_k = Candidate.gen(\mathbf{C}_{k-1}, k)$ /*Apriori pruning strategy in Algorithm 4*/
19: $\quad$ **end while**
20: **end for**
21: **return** ψ

In Line 3, the algorithm finds the set of COSO variables $\omega = Adj(W) \setminus (Adj(Y) \cup \{Y\})$. Then, for each COSO variable in ω, Lines 5 to 20 aim to search for the minimal adjustment sets in $Adj'(W \cup Y) \setminus \{S\}$. Line 10 prevents repeated tests for the adjustment sets that are already in ψ using a different COSO variable. The test in Line 11 assesses whether a candidate

Algorithm 4 Candidate adjustment sets generation (*Candidate.gen*)

Input: $\mathbf{C}_{k-1}$; k.

Output: $\mathbf{C}_k$.

1: $\mathbf{C}_k = \emptyset$;
2: Hash the first $(k-2)$ variables of all sets in $\mathbf{C}_{k-1}$
3: **for** each pair $C_i, C_j \subseteq \mathbf{C}_{k-1}$ with the same first $(k-2)$ variables. **do**
4: $C_s = \cup(C_i, C_j)$
5: $Csubsets = subset(C_s, k-1)$ /*Generating all subsets of C_s with the size $k-1$.*/
6: **if** every element of $Csubsets$ is in $\mathbf{C}_{k-1}$ **then**
7: Add C_s to $\mathbf{C}_k$
8: **end if**
9: **end for**
10: **return** $\mathbf{C}_k$

set is an adjustment set according to Corollary 3. In Line 12, the causal effect of W on Y is estimated based on Eq. (3.2) by adjusting for the current minimal adjustment set $\mathbf{Z}$.

The function *Candidate.gen* in Algorithm 4 uses the *Apriori* pruning strategy as in frequent pattern mining [163] for candidate generation. Once a minimal adjustment set $\mathbf{Z}$ is found at level k, none of its supersets are generated as a level l $(l > k)$ candidate minimal adjustment set. We employ a *hash table* to store candidates to save search time (see Line 2 in Algorithm 4). All variables are sorted in lexicon order for fast pruning. The first $k-2$ variables of all sets in $\mathbf{C}_{k-1}$ are hashed for generating new candidates in $\mathbf{C}_k$ (see Line 3). Line 4 combines a pair of sets with the same first $k-2$ variables to create a new candidate with k variables. Lines 5 to 9 prune candidate set $\mathbf{C}_k$ by checking all subsets of a candidate to ensure that the candidate is potentially minimal.

Time complexity analysis. We analyse the time complexity of the CE^2LS algorithm. Assume that the input dataset contains p variables and n samples. The complexity for Line 1 of CE^2LS is determined by the complexity of *PC.select* [34]. The crude complexity bound of *PC.select*

is $max(\mathbf{O}(p * 2^{q_1} * n), \mathbf{O}(p * 2^{q_2} * n))$, where q_1 and q_2 are the sizes of $Adj(W) \setminus \{Y\}$ and $Adj(Y) \setminus \{W\}$, respectively. Let $q = \max(q_1, q_2)$. The complexity for the first part is $\mathbf{O}(p*2^q*n)$. For Lines 5 to 20 of CE^2LS, the number of iterations of the loop is at most p, and the number of iterations of the while loop from Lines 8 to 19 determines the time complexity of the *Apriori* algorithm [163]. Theoretically, its time complexity is $\mathbf{O}(2^l)$, where l denotes the size of $Adj^l(W \cup Y) \setminus \{S\}$. Because of the pruning, l is a small number. We let $l = q$ for the simplicity. In the worst scenario, the time complexity of the second part is $\mathbf{O}(p * 2^q)$. Hence, the overall complexity of CE^2LS is $\mathbf{O}(p * 2^q * n)$, which is largely determined by the number of neighbouring variables around W and Y. In practice, the largest size of the conditional set for conditional independence tests in *PC.select* can be set to a small number, say, k [34, 36]. The size of the largest minimal adjustment set is also set to a small number, and we use k again. The complexity is $\mathbf{O}(p * 2^k * n)$, where k is a small number, typically 2 to 5.

6.4 Experiments

In this section, we evaluate CE^2LS and compare it with the benchmark causal inference methods/criteria on synthetic and real-world datasets. Our goal is to validate the effectiveness of CE^2LS in terms of unbiased causal effect estimation, its time efficiency and completeness. All experiments are performed on a PC with a 2.6 GHz Intel Core i7 processor and 16 GB of memory (RAM).

6.4.1 Comparison methods and metric

In total, nine benchmark methods/criteria are used in the comparison, including the pretreatment adjustment (PRE) [51], the four minimal adjustment set selection (MASS) criteria [57] $\hat{\mathbf{X}}_{\to W}$, $\hat{\mathbf{X}}_{\to Y}$, $\hat{\mathbf{Q}}_{\to W}$ and $\hat{\mathbf{Z}}_{\to Y}$, the disjunctive cause criterion [54] $\hat{\mathbf{X}}_{\to W,Y}$, EHS [66], LVIDA with FCI (LVIDA-FCI), and LVIDA with rFCI (LVIDA-rFCI) [63]. EHS returns a set of unbiased estimates and uses the average as its final output. LVIDA provides a

set of possible estimates and uses the average as its final result. The details of the comparisons are introduced as follows:

1. The pretreatment adjustment (PRE) [51], which adjusts for all pretreatment variables $\mathbf{X}$ for estimating $ACE(W, Y)$.

2. The minimal adjustment set selection (MASS) criteria, which refer to the four criteria used to select the minimal adjustment set from pretreatment variables, are summarized in [57]. The four criteria select all causes of W, all causes of Y, all causes of W excluding causes of Y, or all causes of Y excluding causes of W as the adjustment set. Following the notation in [57], the identified adjustment sets are denoted as $\hat{\mathbf{X}}_{\to W}$, $\hat{\mathbf{X}}_{\to Y}$, $\hat{\mathbf{Q}}_{\to W}$ and $\hat{\mathbf{Z}}_{\to Y}$.

3. The disjunctive cause criterion [54], which considers the union of all causes of W and all causes of Y as the adjustment set, denoted as $\hat{\mathbf{X}}_{\to W,Y} = (\hat{\mathbf{X}}_{\to W} \cup \hat{\mathbf{X}}_{\to Y})$.

4. EHS [66], a data-driven adjustment set identification from data with latent variables based on the EHS condition in Theorem 6.1.

5. LVIDA [63], a causal effect estimation algorithm based on global causal structure learning. The LVIDA method using FCI for causal structure learning is denoted as LVIDA-FCI, and the LVIDA method using rFCI (really fast causal inference, rFCI [35]) is called LVIDA-rFCI.

Calculating $ACE(W, Y)$ by variable adjustment for all methods. Linear regression and logistic regression are used to estimate the average causal effect of W on Y with variable adjustment, i.e., $ACE(W, Y)$ as given in the Eq. (3.2) for continuous and binary outcomes, respectively. With linear regression, the causal effect is calculated as the partial regression coefficient of W. It has been shown that when the adjustment set is correct, the coefficient of W is $ACE(W, Y)$ [63]. For binary outcomes, the marginal causal odds ratio (MCOR), as in Eq. (6.1), is the estimated causal effect of

$ACE(W, Y)$. Logistic regression is employed to obtain $\log(MCOR)$, and this is implemented by an $\mathbb{R}$ package *stdReg* [164].

$$MCOR = \frac{prob(Y = 1, do(W = 1))/prob(Y = 0, do(W = 1))}{prob(Y = 1, do(W = 0))/prob(Y = 0, do(W = 0))} \tag{6.1}$$

For implementing CE^2LS, *PC.select* from the $\mathbb{R}$ package *pcalg* is employed to learn the local causal structure around W and Y, i.e., $Adj(W)$ and $Adj(Y)$. The conditional independence tests are implemented by using *gaussCItest* and *binCItest* from the same $\mathbb{R}$ package *pcalg*. The average value of these $ACE(W, Y)$ estimated by CE^2LS is regarded as the final result.

EHS was originally implemented in MATLAB [1] [66]. We implement the $\mathbb{R}$ version of EHS, as the MATLAB implementation is for the linear Gaussian model only. As with CE^2LS, we utilize *gaussCItest* and *binCItest* to implement the conditional independence tests for EHS. The maximum size of a conditioning set is set to 6 for EHS in all experiments, as it is impractical to use large conditioning sets (although, as per its design, EHS conducts conditional independence tests over all possible conditioning sets). As the output of EHS contains multiple estimated values for $ACE(W, Y)$, each corresponding to a proper adjustment set identified by EHS, we use the average of the multiple causal effect values estimated by EHS as its final result in the comparison.

For LVIDA, we use the implementation by the authors [63][2], and the functions FCI and rFCI are from the $\mathbb{R}$ package *pcalg*. The function *lv.ida* chooses *method="local"* to ensure the fast search of valid adjustment set $\mathbf{Z}$. For the binary outcome, we use *stdReg* to replace the linear regression. As LVIDA is a bound estimation method that outputs a multi-set of causal effect values, each corresponding to a potential (not necessarily proper) adjustment set, we use the average of these estimated values as the result of LVIDA in the comparison.

[1] `https://sites.google.com/site/dorisentner/publications/CovariateSelection`

[2] `https://github.com/dmalinsk/lv-ida`

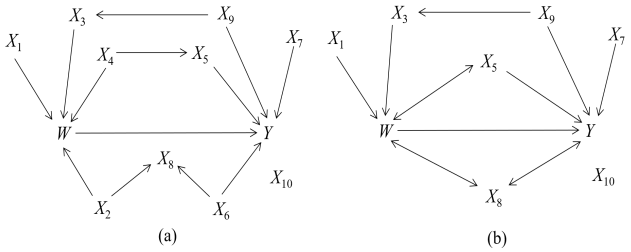


Figure 6.2: (a) is a causal DAG for generating synthetic datasets [6], and (b) represents its corresponding MAG when X_2, X_4, X_6 is unobserved.

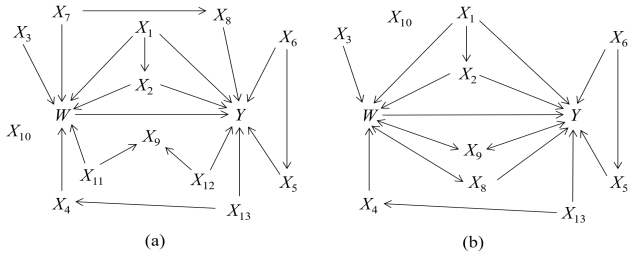


Figure 6.3: (a) A causal DAG for synthetic dataset generation [7] and its corresponding MAG in (b) where variables X_7, X_{11}, X_{12} are removed.

In all experiments, the significance level (α) is set to 0.05 for all algorithms.

Evaluation metric. The relative error (RE, %), which is the absolute error of the estimated causal effect relative to the ground-truth causal effect, is used to evaluate the effectiveness of the methods.

6.4.2 Experiments on synthetic datasets

The DAG in Fig. 6.2 (a) [6] contains 12 variables, and the DAG in Fig. 6.3 (a) [7] includes 15 variables. The DAG in Fig. 6.2 (a) has three back-door paths between W and Y (a sparse structure around W and Y), and the DAG in Fig. 6.3 (a) has seven back-door paths between W and Y (a denser structure around W and Y than the DAG in Fig. 6.2 (a)). The corresponding MAGs are provided in Fig. 6.2 (b) and 6.3 (b), respectively.

Both true MAGs contain an M-structure, which causes the well-known Berkson's paradox [165], also known as the M-bias [3, 78]. An M-structure

contains the path $W \leftarrow U_1 \rightarrow M \leftarrow U_2 \rightarrow Y$, where W, M and Y are observed, but U_1 and U_2 are unobserved. Conditioning on M generates a spurious association between W and Y, i.e., M in an adjustment set results in a biased estimation relative to the causal effect of W on Y. To assess the ability of CE^2LS to deal with latent variables, all synthetic datasets are generated with latent variables and M-bias.

We first generate two groups of synthetic datasets without latent variables based on the two true DAGs in Fig. 6.2 (a) and 6.3 (a). We follow the processes for synthetic data generation in the literature [6] and [7], respectively. Second, we generate synthetic datasets with latent variables by removing several common causes from the above-generated datasets and denote them as Group I datasets and Group II datasets. Each group includes 14 datasets with a range of sample sizes, i.e., 1 k (denotes 1,000), 2 k, 3 k, 4 k, 5 k, 6 k, 7 k, 8 k, 9 k, 10 k, 30 k, 50 k, 100 k, and 150 k samples. Columns of variables X_2, X_4 and X_6 for Group I datasets based on the DAG in Fig. 6.2 (a) and columns of X_7, X_{11} and X_{12} for Group II datasets based on the DAG in Fig. 6.3 (a) are removed. The resulting datasets correspond to MAGs in Fig. 6.2(b) and Fig. 6.3 (b), respectively.

The true causal effect of all synthetic datasets in Group I is 0.5 [6] and of all synthetic datasets in Group II is 2 [7]. To avoid the bias caused by data generation, we repeat the procedure 50 times and generate 50 datasets in each setting. Finally, we conduct experiments on these datasets with latent variables to verify the performance of CE^2LS in estimating causal effects. We report the average of the results of the 50 repeated experiments.

From Fig. 6.4, we make the following observations. (1) The methods that assume causal sufficiency, including PRE, $\hat{\mathbf{X}}_{\rightarrow W}$, $\hat{\mathbf{Q}}_{\rightarrow W}$, $\hat{\mathbf{X}}_{\rightarrow Y}$, $\hat{\mathbf{Z}}_{\rightarrow Y}$ and $\hat{\mathbf{X}}_{\rightarrow W,Y}$, have a large relative error on both group datasets because they do not deal with hidden variables. (2) The performances of CE^2LS and EHS are consistent and improve with the increase in sample size. (3) In most cases, the performances of LVIDA-FCI and LVIDA-rFCI are worse than CE^2LS and EHS since LVIDA-FCI and LVIDA-rFCI rely on the correctness of the learned global causal structure.

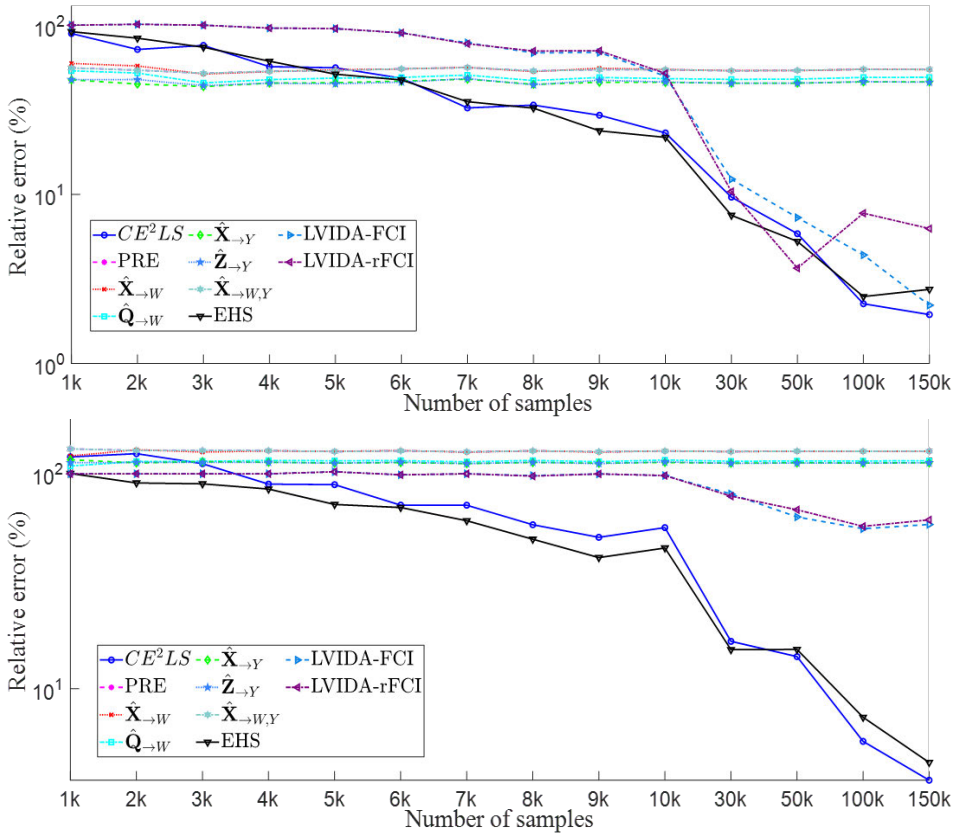


Figure 6.4: REs on synthetic datasets with increasing sample size ($\log_{10}$ on the $Y-$axis). The top figure shows the results of Group I datasets, and the bottom figure shows the results of Group II datasets. As the sample size increases, the relative errors of CE^2LS and EHS decrease significantly. The performances of LVIDA-FCI and LVIDA-rFCI are worse than CE^2LS.

The poor performances of LVIDA-FCI and LVIDA-rFCI on Group II datasets reinforce our argument in the Introduction section that there is a need for causal effect estimation using local causal structure learning. The complexity for optimally learning a global structure increases exponentially with the number of variables. LVIDA-FCI and LVIDA-rFCI read off the adjustment sets from a learned PAG, and both depend heavily on the quality of the learned PAG. This is demonstrated by the (better) performance of EHS and CE^2LS; both determine adjustment sets directly from data, both are based on conditional independence tests, but the former is inefficient

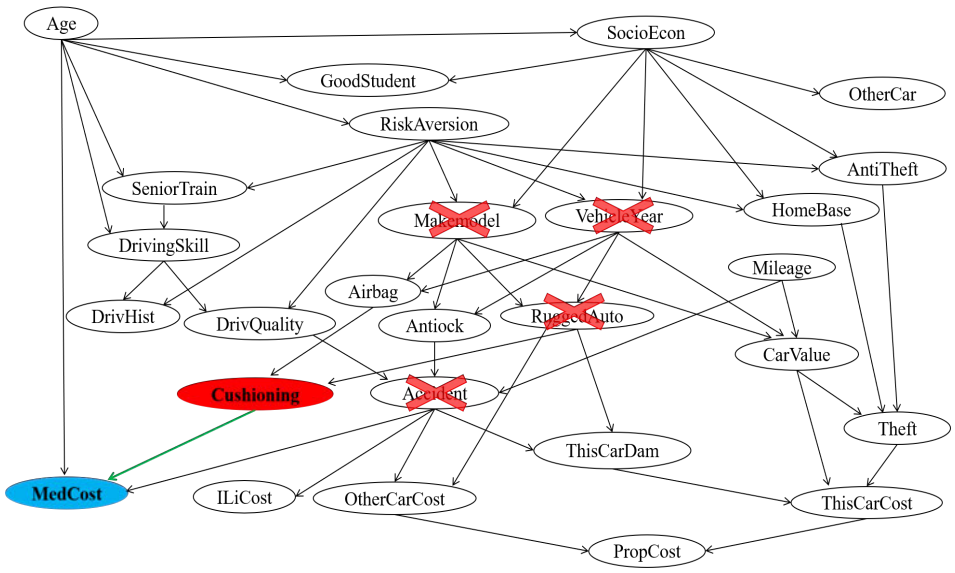


Figure 6.5: The INSURANCE Bayesian network. The red node denotes the treatment variable W, the blue node represents the outcome variable Y, and the causal path between them is shown in green. The crossed variables are the latent variables in the datasets with latent variables.

as a result of global search, and the latter is efficient since local structure learning is used.

6.4.3 Experiments on synthetic data generated using a benchmark Bayesian network

A benchmark Bayesian network (BN) [161], the INSURANCE network, is used to generate data for the experiments in this section. INSURANCE contains 27 nodes and 52 arcs (see Fig. 6.5)[1]. We select the node *Cushioning* as the treatment variable W and the node *MedCost* as the outcome variable Y such that all other variables are pretreatment variables (i.e., non-descendant nodes in the network).

We first generate 7 synthetic datasets with 200 k, 250 k, 300 k, 350 k, 400 k, 450 k, and 500 k samples based on INSURANCE. The ground-truth

[1]The network can be downloaded at `http://www.bnlearn.com/bnrepository/discrete-medium.html`

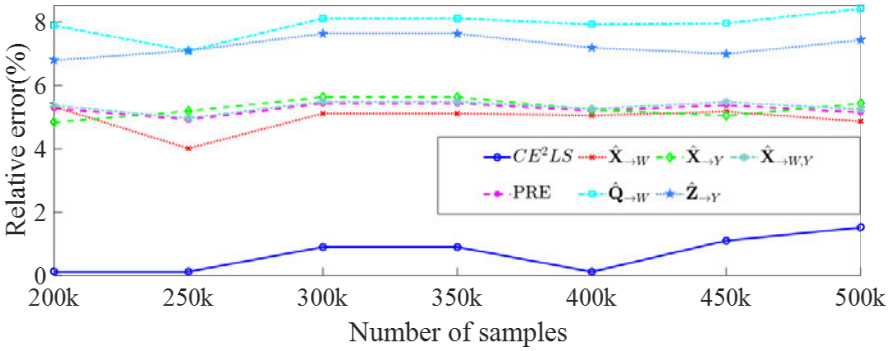


Figure 6.6: REs of causal effects estimated by different methods with the datasets generated based on INSURANCE. EHS, LVIDA-FCI, and LVIDA-rFCI are not shown since they did not return results within two hours. CE^2LS outperforms six other adjustment criteria.

causal effect $ACE(W, Y)$ for each dataset is calculated by adjusting for variables *Airbag* and *RuggedAuto*, which are identified by using the back-door criterion [94].

In this section, we evaluate the performance of CE^2LS on datasets with latent variables. Then, we generate the 7 datasets with latent variables by removing four variables from the above-generated datasets: *VehicleYear*, *MakeModel*, *RuggedAuto* and *Accident*. The resulting causal MAG contains an M-structure $W \leftrightarrow ThisCarDam \leftrightarrow Y$, and hence, all synthetic datasets include M-bias. The 7 datasets with latent variables, EHS, LVIDA-FCI, and LVIDA-rFCI did not return results within two hours due to their low efficiency. Hence, we only compare CE^2LS with the other 6 methods, and the results are shown in Fig. 6.6. CE^2LS outperforms all six methods for causal effect estimation, and this is expected since the six methods require causal sufficiency.

6.4.4 Experiments with real-world datasets

We further evaluate the effectiveness of CE^2LS by performing experiments on two real-world datasets: IHDP [37] and Twins [147]. A detailed introduction of both datasets can be found in Section 4.4 of Chapter 4.

Table 6.1: The estimated ACEs and RE with the IHDP and Twins datasets. The lowest RE in each dataset is boldfaced. CE^2LS outperforms all comparisons on both datasets.

Methods	IHDP		Twins	
	ACE	RE (%)	ACE	RE (%)
CE^2LS	**4.51**	**3.49**%	**-0.02348**	**5.66**%
PRE	3.94	9.61%	-0.01597	35.83%
$\hat{\mathbf{X}}_{\to W}$	3.59	17.71%	-0.02123	14.70%
$\hat{\mathbf{Q}}_{\to W}$	3.69	15.19%	-0.01946	21.84%
$\hat{\mathbf{X}}_{\to Y}$	3.71	14.96%	-0.01326	46.72%
$\hat{\mathbf{Z}}_{\to Y}$	3.70	15.18%	-0.01228	50.65%
$\hat{\mathbf{X}}_{\to W,Y}$	4.13	5.26%	-0.01702	31.63%
EHS	4.05	7.07%	-0.03287	24.27%
LVIDA-rFCI	3.81	14.58%	-0.02127	17.03%

Results on IHDP and Twins. The comparison results of all algorithms on both datasets are listed in Table 6.1. For both datasets, CE^2LS achieved the lowest REs of 3.49% and 5.66%, respectively. Both datasets satisfy causal sufficiency, and all methods are expected to perform well. However, CE^2LS stands out because it finds the minimal adjustment sets and hence utilizes small datasets well. Taking the IHDP dataset as an example, the size of all the minimal adjustment sets found by CE^2LS is 2, whereas the number of covariates used by PRE is 24, and the sizes of $\hat{\mathbf{X}}_{\to W}$, $\hat{Q}_{\to W}$, $\hat{\mathbf{X}}_{\to Y}$, $\hat{Z}_{\to Y}$, $\hat{\mathbf{X}}_{\to W,Y}$ are 7, 4, 8, 3, and 15, respectively. The sizes of the adjustment sets found by EHS are 2 to 6 and 0 to 3 by LVIDA-rFCI. The low-performance LVIDA is also caused by incorrect causal structures learned via global structure learning.

6.4.5 Time efficiency

We compare the time efficiency of CE^2LS with the methods that deal with latent variables, including EHS, LVIDA-FCI and LVIDA-rFCI. All experiments are performed on a PC with 2.6 GHz Intel Core i7 processor and 16

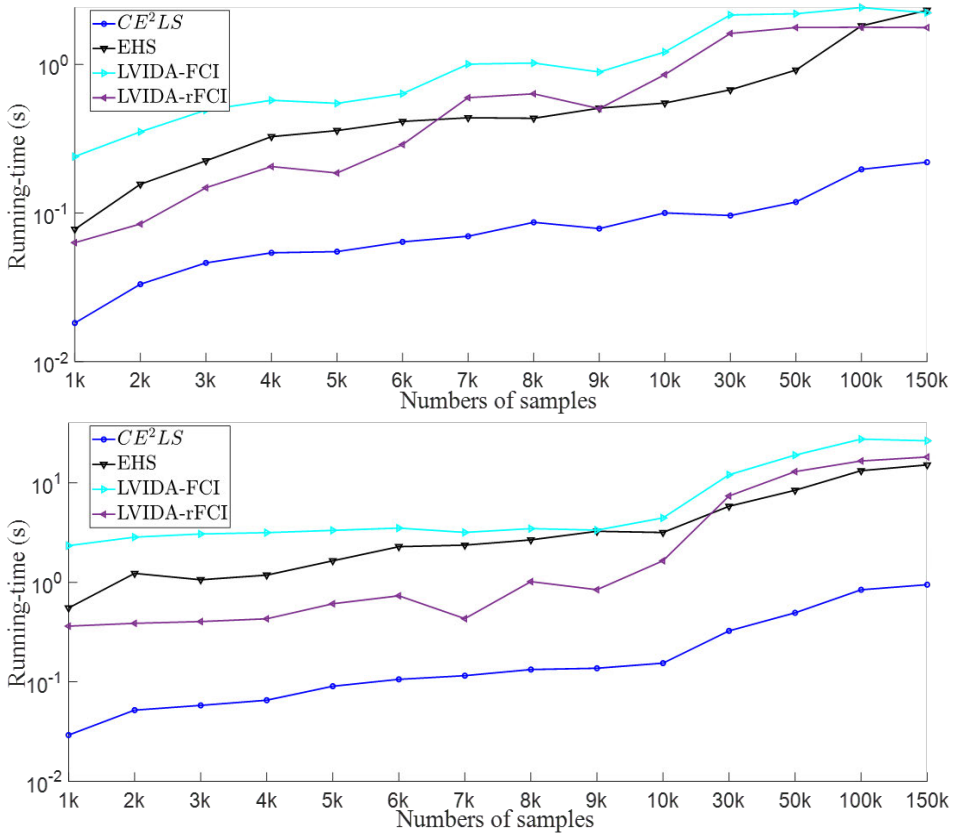


Figure 6.7: Running time in seconds (s) ($\log_{10}$ on the $Y-$axis) of CE^2LS, EHS, LVIDA-FCI and LVIDA-rFCI with the two groups of synthetic datasets described in Section 6.4.2. The top figure shows the results of Group I datasets, and the bottom figure shows the results of Group II datasets. CE^2LS is faster than the others on all datasets.

GB of memory (RAM).

We record the running time of CE^2LS, EHS, LVIDA-FCI and LVIDA-rFCI from the experiments on the two groups of synthetic datasets (as described in Section 6.4.2) and draw the mean running time in Fig. 6.7. We can see that CE^2LS is faster than the others. As described in Section 6.4.3, EHS, LVIDA-FCI and LVIDA-rFCI did not return results within 2 hours in the synthetic datasets of INSURANCE BN, but CE^2LS completed in 5 minutes. Furthermore, the current global structure learning method has a very high time complexity even if some optimized search strategies are

used [36, 74, 166]. Hence, the local search method developed in this paper makes it possible to estimate causal effects unbiasedly from data generated by a complex underlying causal diagram.

6.4.6 The completeness of CE^2LS

As mentioned previously, using Corollary 3 may lead to less complete discovery than using the EHS condition since the local search in Corollary 3 may miss adjustment sets that satisfy the EHS condition but are not included in the local structure searched. In this section, we conduct a simulation to examine the completeness of Corollary 3.

The $\mathbb{R}$ packages *pcalg* [157] and *ggm* [167] are utilized to randomly generate MAG. It is difficult to generate MAG directly. We adopt the strategy of generating DAG first and then construct an MAG by removing some common causes in the generated DAG and adding bidirected edges in the resulting DAG. However, the new graph may not be an ancestral graph since there may exist an almost cycle. Hence, we repeat the procedure to obtain an ancestral graph. To replicate the experiment, the specific functions and steps are described below.

First, we generate a random DAG $\mathcal{G}$ via the function *randomDAG* in the package *pcalg*. There are two parameters, 'n' denoting the number of nodes in $\mathcal{G}$ and '*prob*' represents the probability of connecting a node to another node with higher topological ordering in the $\mathcal{G}$. In our experiments, 'n' is set to 20, 30, 40, and 50. Theoretically, the average degree of the generated DAG (i.e., d) is $(n-1) * prob$. In this experiment, d is set to 2, 4, 6, and 8. We generate the DAG $\mathcal{G}$ with the edge $W \to Y$ and exclude the pair (W, Y), and other nodes are either the ancestors of W, or the ancestors of Y.

Second, we randomly remove the required number of common causes as latent variables from the generated DAG $\mathcal{G}$. The number of latent variables is denoted as 'l'. l is set to 2, 4, 6, 8, 10, and 12. In this step, we only remove a node with more than two children, i.e., common cause. After a node X is removed from the generated DAG $\mathcal{G}$, for each pair in $Ch(X)$,

add a bidirected edge in $\mathcal{G}$, and for each node in $Pa(X)$, add a directed edge to every node in $Ch(X)$. In this way, we obtain a new graph $\mathcal{G}'$. The function *isAG* in the $\mathbb{R}$ package *ggm* is employed to examine whether $\mathcal{G}'$ is an ancestor graph. If $\mathcal{G}'$ is an ancestral graph, then the function *max* in the $\mathbb{R}$ package *ggm* is used to construct an MAG $\mathcal{M}$; otherwise, $\mathcal{G}'$ is discarded and the above procedure is repeated.

Finally, we generate a synthetic dataset that is faithful to the constructed MAG $\mathcal{M}$. In this step, the original generated DAG $\mathcal{G}$ is used to generate a synthetic dataset. The function *rmvDAG* in the package *pcalg* is utilized to generate multivariate data with a dependency structure specified by the DAG $\mathcal{G}$. The sample size is set to 10,000 for all generated data. Then, the variables deleted from the DAG $\mathcal{G}$ are removed from the generated dataset to obtain the synthetic data with latent variables such that the generated dataset is faithful to the constructed MAG $\mathcal{M}$.

In our experiments, for each setting (number of nodes, average node degree and number of deleted common causes), we randomly construct 100 MAGs and generate the corresponding synthetic datasets. We run Corollary 3 (and algorithm CE^2LS) on the datasets to count the number of cases when an adjustment set is identified. When a constructed MAG is too dense (the density is more than 0.5), we do not report its results. The experimental results are reported in Table 6.2.

In Table 6.2, we observe that in most cases, adjustment sets are found by local search using Corollary 3. Since CE^2LS searches in the expanded local space, there are more cases when CE^2LS can find adjustment sets. Overall, the completeness rates of both Corollary 3 and CE^2LS are close to 100% in all experimental situations.

Table 6.2: Experimental results on the completeness of adjustment set discovery using Corollary 3 and CE^2LS. In the table, the expression A/B (e.g., 99/99) denotes the number of trials (out of 100) when an adjustment set is identified using Corollary 3 (A) and CE^2LS (B) respectively. n is the number of nodes in an MAG, l indicates the number of latent variables, and d denotes the average degree of the generated DAGs used to construct MAGs. '-' denotes that the adjustment sets cannot be found correctly because the density of the constructed MAG is high (the density of a MAG is more than 0.5).

n	l	d = 2	d = 4	d = 6	d = 8
20	2	99 / 99	98 / 100	93 / 97	92 / 98
	4	100 / 100	97 / 97	96 / 98	-
	6	98 / 98	97 / 97	94 / 95	-
	8	99 / 99	95 / 96	-	-
	10	97 / 97	95 / 96	-	-
30	2	100 / 100	100 / 100	96 / 98	95 / 97
	4	100 / 100	100 / 100	98 / 99	95 / 98
	6	100 / 100	99 / 99	99 / 100	93 / 93
	8	99 / 99	100 / 100	96 / 98	94 / 97
	10	100 / 100	99 / 100	97 / 98	95 / 96
	12	99 / 99	97 / 97	98 / 99	93 / 93
40	2	99 / 99	100 / 100	99 / 100	98 / 100
	4	100 / 100	100 / 100	98 / 100	96 / 97
	6	100 / 100	99 / 99	100 / 100	96 / 98
	8	100 / 100	100 / 100	99 / 100	96 / 97
	10	100 / 100	99 / 99	100 / 100	95 / 99
	12	98 / 98	100 / 100	98 / 99	95 / 96
50	2	100 / 100	100 / 100	100 / 100	99 / 99
	4	100 / 100	100 / 100	99 / 100	98 / 99
	6	100 / 100	100 / 100	99 / 100	96 / 97
	8	100 / 100	100 / 100	99 / 99	98 / 100
	10	99 / 99	100 / 100	99 / 100	98 / 99
	12	100 / 100	100 / 100	99 / 99	97 / 98

6.5 Conclusion

In this chapter, we identified a practical problem setting where proper adjustment sets can be found via local search in data with latent variables, which enables rapid, unique and unbiased causal effect estimation.

We developed several theorems and corollaries to ensure the soundness of the local causal search. Based on the developed theorems and corollaries, we designed an efficient algorithm, CE^2LS, for finding adjustment sets and estimating causal effects from data with latent variables. A level-wise search strategy was employed to further improve efficiency and to find minimal adjustment sets via local causal structure discovery.

We conducted experiments on synthetic and real-world datasets to evaluate the performance of CE^2LS. The results demonstrated that CE^2LS can obtain more accurate causal effect estimation than the other existing methods, and CE^2LS is much faster than other methods that handle data with latent variables.

Chapter 7

Local causal search for causal effect estimation without the pretreatment variable assumption

In Chapter 6, we studied the local search for unique and unbiased causal effect estimation from data with latent variables. The proposed method in Chapter 6, CE^2LS, can handle large datasets and provides unique and unbiased ACE and exact adjustment sets. However, CE^2LS relies on the assumption of pretreatment variables (and so do most existing methods), but in some cases, information about pretreatment variables is unavailable. In this chapter, we relax the pretreatment variable assumption and develop a new method to find proper adjustment sets to achieve precise causal effect estimation without the pretreatment variable assumption.

7.1 Motivation

When the pretreatment variables assumption does not hold, obtaining a unique and unbiased estimation of a causal effect becomes challenging. The main challenge comes from the descendants of W that exist in the set of

measured variables. The methods developed under the pretreatment variable assumption (including CESD, DICE and CE^2LS developed in this thesis and most existing methods) do not work on this case since adjusting for an adjustment set that contains a descendant of W will potentially produce a biased estimation [43, 49, 78].

In this chapter, we relax the pretreatment variables assumption and develop algorithms for unique and unbiased causal effect estimation from data with latent variables. To obtain an adjustment set that does not contain the descendants of W and remove the uncertainty in adjustment identification, we first identify a practical problem setting, i.e., a COSO variable Q (a cause or spouse of the treatment only variable). A COSO variable Q is a variable $Q \in Adj(W) \setminus \{Y\}$ and $Q \notin Adj(Y) \setminus \{W\}$.

We show that by using a COSO variable, the proper adjustment set will be identified from data with latent variables, and hence, the unbiased causal effect is estimated. We develop the two theorems to support the discovery of the proper adjustment sets. Based on the developed theorems, we propose two data-driven algorithms, DAVS-Q and DAVS, to estimate causal effects uniquely and unbiasedly from data with latent variables.

The chapter is based on the paper [76] "Towards precise causal effect estimation from data with hidden variables", which was submitted to IEEE Transactions on Neural Networks and Learning Systems, and is currently review.

7.2 The theorems for adjustment set discovery

In this section, we first introduce our problem setting and then present the theorems for adjustment set identification.

7.2.1 Problem setting

From observational data, what we learn is an equivalence class of causal graphs. Hence, the problem of causal effect estimation does not have a

unique solution in general. To obtain a unique and unbiased estimation of a causal effect, assumptions are needed. In this chapter, we identify a practical case where the causal effect can be uniquely estimated from data with latent variables. Specifically, we assume that in the equivalence class of the MAGs learned (represented by a PAG), there exists a cause or spouse of the treatment only (COSO) variable, as defined below.

Definition 7.1 (COSO variable)**.** *Let $\mathcal{G}$ be the PAG learned from a dataset with treatment W, outcome Y and the set of other observed variables $\mathbf{X}$. A variable $Q \in \mathbf{X}$ is a COSO variable if $Q \in (Pa(W) \cup Sp(W))$ and $Q \notin (Pa(Y) \cup Sp(Y))$.*

This definition comes from the COSO in Definition 6.2 and adds a constraint compared to Definition 6.2 since there is more than one causal path from W to Y. We assume that there exists at least a COSO variable in the underlying causal MAG. This means that the underlying causal MAG is adjustment amenable to (W, Y), i.e., the unbiased estimate can be identified through covariate adjustment.

A COSO variable is different from an instrumental variable (IV) [168, 169] in the following three aspects. First, an IV is known *a priori* to have a causal effect on W, but a COSO Q may be in $Sp(W)$. Second, the causal effect of an IV on Y should be through W, but a COSO Q may have other causal paths to Y. Third, an IV is normally given, but COSO Q can be testable from data.

A COSO variable can be easily found in many applications. For example, when studying the effect of smoking on lung cancer [18], family influence causes smoking but does not cause lung cancer directly. Family influence is a COSO w.r.t. (smoking, lung cancer). In the study of the effect of job training on income, marriage impacts job training but does not directly affect income [15]. Marriage is a COSO for (job training, income). Compared to the existing data-driven methods for unique and unbiased causal effect estimation, which require that $\mathbf{X}$ contains only pretreatment variables or satisfies the unconfoundedness assumption (as discussed in the chapter 2),

our COSO variable assumption is more practical and enables broader real-world applications.

7.2.2 The theorems

We estimate the causal effect of treatment of interest, W, on Y, the outcome from data with hidden variables. We designed our theorems to identify adjustment sets by conditional independence tests. To prove our theorems, we first propose the lemma below.

Lemma 7.1. *Given an MAG or a PAG $\mathcal{G}$ with a pair of variables (W, Y) and a set of nodes $\mathbf{Z}$ in $\mathcal{G}$, $\mathcal{G}$ is adjustment amenable relative to (W, Y) and $\mathbf{Z} \cap Forb(W, Y, \mathcal{G}) = \emptyset$. If all generalized back-door paths between W and Y are blocked by $\mathbf{Z}$ in $\mathcal{G}$, then $\mathbf{Z}$ blocks all definite status non-causal paths between W and Y in $\mathcal{G}$.*

Proof. All definite status non-causal paths between W and Y in Definition 3.19 are either generalized back-door paths between W and Y or $W \circ\!\!\rightarrow \ldots *\!\!\rightarrow C \leftarrow\!\!* \ldots Y$. The latter is naturally blocked by $\emptyset$ since C is a collider on the path. There are two situations for the latter case: (1) if path π is a form of $W \circ\!\!\rightarrow \ldots \circ\!\!\rightarrow C \leftarrow\!\!* \ldots Y$, then $C \in Forb(W, Y, \mathcal{G})$ such that $C \notin \mathbf{Z}$ because $\mathbf{Z} \cap Forb(W, Y, \mathcal{G}) = \emptyset$, and (2) if path π is the form of $W \circ\!\!\rightarrow \ldots \circ\!\!\rightarrow X_k \leftrightarrow C \leftarrow\!\!* \ldots Y$ and $C \in \mathbf{Z}$, then $X_k \notin \mathbf{Z}$ since $X_k \in Forb(W, Y, \mathcal{G})$.

Therefore, if all generalized back-door paths between W and Y are blocked by the set $\mathbf{Z}$ in $\mathcal{G}$, then $\mathbf{Z}$ satisfies the blocking condition in the generalized adjustment criterion. □

Now, we present the following theorem to describe the condition for searching for an adjustment set in data with hidden variables under our problem setting.

Theorem 7.1. *Given a PAG $\mathcal{G}$ containing a pair of variables (W, Y) and the other variables $\mathbf{X}$, assume that there exists a COSO variable $Q \in \mathbf{X}$. There exists a set of variables $\mathbf{Z} \subseteq \mathbf{X} \setminus (\{Q\} \cup Forb(W, Y, \mathcal{G}))$; then, $\mathbf{Z}$ is an adjustment set if and only if $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$.*

Proof. As Q is a COSO variable, $\mathcal{G}$ must contain an edge $Q{*\!\!\rightarrow}\, W$, i.e., there is an edge from Q into W. As there is at least a causal path from W to Y, there must be an edge out of W. Then, by Definition 3.12, when there is an edge out of W, the edge must be visible. Thus, the PAG $\mathcal{G}$ is adjustment amenable w.r.t. (W, Y) by Definition 3.17, i.e., condition (I) of the generalized adjustment criterion (Definition 3.19) holds.

The condition $\mathbf{Z} \subseteq \mathbf{X} \backslash (\{Q\} \cup Forb(W, Y, \mathcal{G}))$ removes all possible descendants of W in $\mathcal{G}$. Hence, the set $\mathbf{Z}$ satisfies condition (II) of the generalized adjustment criterion (Definition 3.19).

We first prove the sufficient condition, i.e., if $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$, then $\mathbf{Z}$ blocks all definite status non-causal paths, i.e., condition (III) of the generalized adjustment criterion (Definition 3.19) is satisfied, and thus $\mathbf{Z}$ is an adjustment set given that conditions (I) and (II) of the generalized adjustment criterion have been proved above. We prove this using contradiction. Assume that a generalized back-door path π from W to Y is not blocked by $\mathbf{Z}$. π must be in the form of $W {\leftarrow\!\!*} \ldots {*\!\!\rightarrow}\, Y$ according to Definition 3.15. Moreover, edges $Q{*\!\!\rightarrow}\, W$ and π collide at W (i.e., W is a collider), so Q is m-connected to Y given $\mathbf{Z} \cup \{W\}$. This violates $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$. Therefore, all generalized back-door paths are blocked by $\mathbf{Z}$, and thus condition (III) of the generalized adjustment criterion holds according to Lemma 7.1.

Next, we prove the necessary condition, i.e., if $\mathbf{Z} \subseteq \mathbf{X} \backslash (\{Q\} \cup Forb(W, Y, \mathcal{G}))$ is an adjustment set, then $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$. When $\mathbf{Z}$ is an adjustment set, then $\mathbf{Z}$ blocks all definite status non-causal paths from W to Y by condition (III) of the generalized adjustment criterion (Definition 3.19). Hence, by Lemma 7.1, $\mathbf{Z}$ blocks all generalized back-door paths between W and Y. Moreover, the paths from Q to Y, which are the form of $Q{*\!\!\rightarrow}\, W$, and the causal paths between W and Y (as they must be in the form of $Q{*\!\!\rightarrow}\, W \rightarrow \cdots \rightarrow Y$) are blocked by W. Thus, Q and Y are m-separated given $\mathbf{Z} \cup \{W\}$ in PAG $\mathcal{G}$, i.e., $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$. □

To reduce the search space for $\mathbf{Z}$ by Theorem 7.1, we develop the following theorem.

Theorem 7.2. *Given a PAG $\mathcal{G}$ containing a pair of variables (W, Y) and the other variables $\mathbf{X}$, assume that there exists a COSO variable $Q \in \mathbf{X}$.*

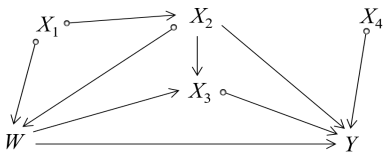


Figure 7.1: An example of a PAG $\mathcal{G}$, used in Example 7.1.

There exists a set of variables $\mathbf{Z} \subseteq pds(W, Y, \mathcal{G}) \setminus (\{Q\} \cup Forb(W, Y, \mathcal{G}))$*; then,* $\mathbf{Z}$ *is an adjustment set if and only if* $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup \{W\}$.

Proof. $pds(W, Y, \mathcal{G})$ is a subset of $\mathbf{X}$, and from [63], $pds(W, Y, \mathcal{G})$ contains all nodes that lie on the generalized back-door paths between W and Y. Hence, the search for a valid adjustment set can be performed only in $pds(W, Y, \mathcal{G}) \setminus (\{Q\} \cup Forb(W, Y, \mathcal{G}))$ instead of in $\mathbf{X} \setminus (\{Q\} \cup Forb(W, Y, \mathcal{G}))$. Then, the proof follows from the proof of Theorem 7.1. □

The following example illustrates Theorems 7.1 and 7.2.

Example 7.1. *Refer to the PAG in Fig. 7.1, denoted as* $\mathcal{G}$*, which contains a pair of variables* (W, Y) *and has* X_1 *as a COSO variable. From* $\mathcal{G}$*,* $Forb(W, Y, \mathcal{G}) = \{X_3, Y\}$ *and* $pds(W, Y, \mathcal{G}) = \{X_1, X_2, X_3, Y\}$.

Considering Theorem 7.1, as $\mathbf{X} \setminus (\{Q\} \cup Forb(W, Y, \mathcal{G})) = \{X_2, X_4\}$*, and* $\mathbf{Z}$ *can be a subset of* $\{X_2, X_4\}$*, i.e.,* $\{X_2\}$*,* $\{X_4\}$ *or* $\{X_2, X_4\}$*. From* $\mathcal{G}$*,* $X_1 \perp\!\!\!\perp Y | \{X_2, W\}$ *and* $X_1 \perp\!\!\!\perp Y | \{X_2, X_4, W\}$*, following Theorem 7.1,* $\{X_2\}$ *and* $\{X_2, X_4\}$ *are adjustment sets for unbiased estimation of* $ACE(W, Y)$*. According to* $\mathcal{G}$*, we can see that* X_2 *blocks all four generalized back-door paths from* W *to* Y*:* $W \leftarrow\!\circ X_1 \circ\!\rightarrow X_2 \rightarrow Y$*;* $W \leftarrow\!\circ X_1 \circ\!\rightarrow X_2 \rightarrow X_3 \circ\!\rightarrow Y$*;* $W \leftarrow\!\circ X_2 \rightarrow X_3 \circ\!\rightarrow Y$ *and* $W \leftarrow\!\circ X_2 \rightarrow Y$.

Considering Theorem 7.2, as $pds(W, Y, \mathcal{G}) \setminus (\{Q\} \cup Forb(W, Y, \mathcal{G})) = \{X_2\}$*, the search space for* $\mathbf{Z}$ *reduces to* $\{X_2\}$ *instead of* $\{X_2, X_4\}$ *when following Theorem 7.1.*

Theorem 7.3. *Given a PAG* $\mathcal{G}$ *containing a pair of variables* (W, Y) *and the other variables* $\mathbf{X}$*, assume that there exists a COSO variable* $Q \in \mathbf{X}$*. Let* $\mathcal{Z}$ *be the set of all adjustment sets found following Theorem 7.1 or Theorem 7.2, the estimated value of* $ACE(W, Y)$ *by adjusting* $\mathbf{Z}$ *is the same, i.e., unique* $\forall \mathbf{Z} \in \mathcal{Z}$.

Proof. Any adjustment set $\mathbf{Z}$ found by following Theorem 7.1 or Theorem 7.2 is a proper adjustment set for all MAGs enumerated from the given PAG since all back-door paths from W to Y in all the MAGs are blocked by $\mathbf{Z}$. Hence, the causal effect estimated by adjusting any $\mathbf{Z}$ is unbiased. In other words, the estimated causal effect is always the same when adjusting any $\mathbf{Z}$ found with Theorem 7.1 or Theorem 7.2. □

Theorem 7.3 shows that the estimated causal effect is unique when adjusting on any of the adjustment sets found with Theorem 7.1 or Theorem 7.2. This is a significant advantage of our proposed method over most existing data-driven methods, such as LVIDA [63], which provides a bound or a multi-set of estimated causal effects, and GAC [49], which suffers from errors of the edge orientations in PAG since $W \perp\!\!\!\perp Y \mid \mathbf{Z}$ are not testable from data. It is also worth noting again that Theorems 7.1 and 7.2 only assume the existence of a COSO variable in the data, which is more practical and less restricted than the pretreatment variable or unconfoundedness assumption.

7.3 The proposed two algorithms

In this section, based on Theorem 7.2 and Theorem 7.3, we propose two algorithms, DAVS-Q (Algorithm 5) and DAVS (Algorithm 6), for unique and unbiased causal effect estimation from data with latent variables. From the data, it is desirable to use multiple adjustment sets for robustness. Thus, the *Apriori pruning strategy* [163] is used to search for all minimal adjustment sets from the data. The correctness of the search strategy is guaranteed by the truth, i.e., any superset of a minimal adjustment set must not be a minimal adjustment set.

Algorithm 5 Data-driven adjustment variable selection given Q (DAVS-Q)

Input: Dataset $\mathbf{D}$ with W, Y and $\mathbf{X}$, $Q \in \mathbf{X}$ and PAG $\mathcal{G}$ learned from $\mathbf{D}$.

Output: $ACE(W, Y)$.

Let $ACE(W, Y) = NULL$
if $Q \perp\!\!\!\perp Y | \mathbf{W}$ **then**
  Calculate $ACE(W, Y)$ via Eq. (3.2) given $\mathbf{Z} = \varnothing$
else
  Obtain $pds(W, Y, \mathcal{G})$ and $Forb(W, Y, \mathcal{G})$ from $\mathcal{G}$
  $\mathbf{C}_1 = pds(W, Y, \mathcal{G}) \setminus (\{Q\} \cup Forb(W, Y, \mathcal{G}))$
  $k = 1$
  while $\mathbf{C}_k \neq \varnothing$ **do**
    for each element $\mathbf{Z} \in \mathbf{C}_k$ **do**
      if $\mathbf{Z}$ is not in $\mathbf{\Psi}$ **then**
        if $Q \perp\!\!\!\perp Y | \mathbf{Z} \cup W$ **then**
          Calculate $ACE(W, Y)$ via Eq. (3.2) given $\mathbf{Z}$ and remove $\mathbf{Z}$ from $\mathbf{C}_k$
          $\mathbf{\Psi} \leftarrow \cup(\mathbf{Z}, ACE(W, Y))$
        end if
      end if
    end for
    $k = k + 1$
    /*Apriori pruning strategy in Algorithm 4*/
    $\mathbf{C}_k = Candidate.gen(\mathbf{C}_{k-1}, k)$
  end while
end if
return $ACE(W, Y)$ in $\mathbf{\Psi}$.

Both algorithms take as input a dataset and a PAG learned from the dataset using a structure learning algorithm at the choice of users. DAVS-Q requires a given COSO variable, while DAVS finds COSO variables from data. In many applications, a COSO variable is known based on domain knowledge, and DAVS-Q is the better choice. When a COSO variable is unknown, users can employ DAVS, which finds the set of candidate COSO variables, $\mathbf{Q}$. For each variable $Q \in \mathbf{Q}$, DAVS calls DAVS-Q to estimate

Algorithm 6 Data-driven adjustment variable selection without given Q (DAVS)

Input: Dataset $\mathbf{D}$ with W, Y and $\mathbf{X}$, PAG $\mathcal{G}$ learned from $\mathbf{D}$.

Output: $ACE(W, Y)$.

1: Let $ACE(W, Y) = NULL$
2: Find $Pa(W), Sp(W)$, $Pa(Y)$ and $Sp(Y)$ from $\mathcal{G}$
3: $\mathbf{Q} = (Pa(W) \cup Sp(W)) \setminus (Pa(Y) \cup Sp(Y))$
4: **for** each $Q \in \mathbf{Q}$ **do**
5: $\quad ACE(W, Y) =$ DAVS-Q $(\mathbf{D}, Q, \mathcal{G})$
6: **end for**
7: **return** $ACE(W, Y)$

$ACE(W, Y)$.

Therefore, the two algorithms take the same search strategy, *apriori pruning strategy* [163], and estimate the causal effect using this found adjustment set (see lines 7-18 of Algorithm 5). Furthermore, the algorithms use a bottom-up search (from single variables to multiple variables) for efficient utilization of data in conditional independence tests. Note that an empty set is also a legitimate adjustment set if it satisfies Theorem 7.2. This situation is checked and dealt with in Lines 2-3 of Algorithm 5. The pseudocode of *Apriori pruning strategy* is provided in Algorithm 4, which is used in Line 17 to generate level k candidate adjustment sets.

In our implementation, the input PAG $\mathcal{G}$ is learned using rFCI [35] implemented in the $\mathbb{R}$ package *pcalg* [157]. The conditional independence test tool is implemented by *GaussianCItest* and *binCItest* in *pcalg* for Gaussian and binary datasets, respectively. Eq. (3.2), the calculation of $ACE(W, Y)$ is implemented by *lm* in the $\mathbb{R}$ packages *stats* [62, 63] and *stdGlm* in the $\mathbb{R}$ packages *stdReg* [6] for continuous and binary outcomes Y, respectively.

Time complexity analysis. The time complexity of DAVS-Q and DAVS is largely determined by the time taken by the rFCI algorithm [35] for learning PAG $\mathcal{G}$ from data $\mathbf{D}$. Lines 7-18 in DAVS-Q (Algorithm 5) can be time consuming, but the size of $\mathbf{Z}$ is often small in practice, usually containing 2 to 5 variables. When the size of $\mathbf{Z}$ is large, a very large dataset

is needed for reliable conditional independence tests. Hence, the size of $\mathbf{Z}$ is normally returned as small for reliable tests. With DAVS, the size of $\mathbf{Q}$ is normally small. The complexity of Algorithm 4 is $\mathbf{O}(2^l)$, where l denotes the size of $pds(W, Y, \mathcal{G}) \setminus (\{Q\} \cup Forb(W, Y, \mathcal{G}))$, which is lower than the complexity of the rFCI algorithm [35]. Furthermore, the complexity of DAVS is linear with $|\mathbf{Q}|$. Therefore, the overall complexity of DAVS is also determined by the rFCI algorithm.

7.4 Experiments

The experimental goals are to evaluate the performance of the proposed DAVS-Q and DAVS algorithms in terms of the data with latent variables. We evaluate the two algorithms on five synthetic datasets with latent variables and compare them with the methods that can be used to handle data without the pretreatment variable assumption, including IDA, LVIDA, and GAC (except DICE, CE-SAT and EHS, since EHS and DICE require the pretreatment variable assumption, and CE-SAT has a high time complexity). To verify whether the proposed algorithms can be used in real applications, we perform experiments on two semi-synthetic and two real-world datasets and compare them with all comparisons in Section 6.4.1, plus four popular average causal effect estimation methods, including a classic propensity score matching method (PSM) [51], covariate balancing propensity score (CBPS) [150], CFRNet [39] and causal forest (CF) [151].

Implementation. We use the IDA implementation in the $\mathbb{R}$ package *pcalg* [157]. The GAC is implemented by the $\mathbb{R}$ packages *pcalg* and *Dagitty* [170]. DICE is implemented by the functions *pc.select*, *GaussianCItest* and *binCItest* in the $\mathbb{R}$ package *pcalg*. The classic PSM [51] is implemented by *glm* in the $\mathbb{R}$ package *stats* and *matching* in the $\mathbb{R}$ package *Matching* [149]. CBPS is from the $\mathbb{R}$ package *CBPS*. The implementation of CFRNet is from the authors' code[1]. CF is implemented in the $\mathbb{R}$ package *grf*.

[1] `https://github.com/clinicalml/cfrnet`

Table 7.1: The details of five standard benchmark BNs.

	CHILD	INSURANCE	MILDEW	ALARM	BARLEY
Nodes	20	27	35	37	48
Arcs	25	52	46	46	84

Evaluation metrics. We evaluate the performance of all algorithms using the estimated bias (%), which is the relative error w.r.t. the ground truth. For the dataset without the ground truth of causal effect, we evaluate the proposed algorithms by empirical estimation and visualize the estimated causal effects in terms of 95% confident intervals.

Parameter settings. The same parameter settings are used for DAVS-Q, DAVS, and LVIDA, and the significance level (α) is set to 0.05. The average outputs of DAVS-Q and DAVS are considered the final result. Since IDA, LVIDA, EHS, and DICE each return a multi-set of causal effects, the average is considered the most likely estimation. For EHS, we constrain the size of the conditional set to 6. Otherwise, it cannot produce a result within two hours for the real-world datasets used.

7.4.1 Experiments with synthetic data

We use 5 benchmark Bayesian networks (BNs) from the Bayesian Network Repository[1]: CHILD, INSURANCE, MILDEW, ALARM, and BARLEY to generate synthetic datasets. The numbers of nodes and arcs of the BNs are summarized in Table 7.1. For each BN, we choose a variable with multiple incoming edges as the outcome variable Y and select one of Y's parents as the treatment variable W. We generate 5 synthetic datasets from the 5 BNs with 10,000 samples each by using the $\mathbb{R}$ package *bnlearn* [161]. Then, we hide 5% variables that are on back-door paths from W to Y from each synthetic dataset. For the convenience of reproducing our experiments, we provide the information of W, Y, and the latent variables on the five synthetic datasets in Table 7.2.

[1] `http://www.bnlearn.com/bnrepository/`

Table 7.2: The variables W, Y and the latent variables on the five synthetic datasets.

BNs	W	Y	Latent variables
CHILD	LowerBodyO2	CardiacMixing	LungParench
INSURANCE	RiskAversion	DrivHist	Age and MakeModel
MILDEW	meldug_3	meldug_4	temp_2 and meldug_2
ALARM	VALV	CCHL	PMB and SHNT
BARLEY	dgv1059	protein	saa1id, dgv5980 and ng1ilg

As the pretreatment variable assumption does not hold in these datasets, only methods without assuming pretreatment variables, i.e., IDA, LVIDA, and GAC are used in the comparisons. The causal effects calculated using Eq. (3.2), and the adjustment set identified based on the back-door criterion [9] on DAGs of the complete BNs are the ground truth of the causal effects.

The results in Table 7.3 show that the results of DAVS-Q and DAVS have significantly smaller estimated biases than IDA, LVIDA, and GAC. The synthetic experimental results demonstrated that our two developed algorithms can not only provide unique causal effect estimation but also more accurate estimations of the causal effects. IDA handles data without latent variables, so its poor performance is as expected. LVIDA and GAC have the same performance on all synthetic datasets because they are PAG-based algorithms. Moreover, the adjustment set $\mathbf{Z}$ identified by LVIDA and GAC heavily relies on the correctness of the learned PAG since the adjustment set $\mathbf{Z}$ is determined from the manipulated PAG but not identified directly from the data. Different from LVIDA and GAC, DAVS-Q and DAVS perform the conditional independence test $Q \perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{W\}$ to identify the proper adjustment set $\mathbf{Z}$ directly from the data. This is the main reason that the performances of DAVS-Q and DAVS outperform the performances of LVIDA and GAC.

Synthetic experiments confirmed that the developed two algorithms can handle data well, which violates the assumptions of pretreatment variables and unconfoundedness. This means that the developed two algorithms have

Table 7.3: Estimating ACE on synthetic datasets. (a, b) next to a dataset name denotes the number of nodes (a) and arcs (b).

Methods	CHILD(20, 25)		INSURANCE(27, 52)		MILDEW(35, 46)		ALARM(37, 46)		BARLEY(48, 84)	
	ACE	Bias(%)	ACE	Bias(%)	ACE	Bias(%)	ACE	Bias(%)	ACE	Bias(%)
IDA	-0.4352	75.29%	-0.4976	75.96%	0.3322	75.98%	0.3469	72.33%	-0.2772	73.81%
LVIDA	-0.4352	75.29%	-0.4811	76.75%	0.3435	75.16%	0.0712	94.32%	-0.2602	75.42%
GAC	-0.4353	75.15%	-0.4812	76.75%	0.3433	75.17%	0.2854	77.23%	-0.2460	76.76%
DAVS-Q	**-1.7933**	**1.81**%	**-2.0972**	**2.21**%	**1.3912**	**3.11**%	**1.3698**	**9.27**%	-1.0267	3.02%
DAVS	**-1.7933**	**1.81**%	-2.1787	5.25%	1.4349	3.78%	0.8361	33.31%	**-1.0529**	**0.55**%

Table 7.4: Estimating ACE on IHDP and Twins.

Methods	IHDP		Twins	
	ACE	Bias(%)	ACE	Bias(%)
IDA	3.01	31.03%	-0.02308	7.29%
LVIDA	3.81	12.62%	-0.02127	14.71%
GAC	3.81	12.62%	-0.02127	14.71%
DAVS	**4.00**	**8.25**%	**-0.02403**	**3.46**%
CFRNet	4.15	4.82%	-0.01955	20.25%
DICE	**4.48**	**2.50**%	-0.03768	51.38%
EHS	4.05	7.07%	-0.03287	24.27%
$\hat{\mathbf{X}}_{\to W}$	3.59	17.66%	-0.02123	14.70%
$\hat{\mathbf{Z}}_{\to W}$	3.69	15.37%	-0.01946	21.84%
$\hat{\mathbf{X}}_{\to Y}$	3.71	14.91%	-0.01326	46.72%
$\hat{\mathbf{Z}}_{\to Y}$	3.70	15.14%	-0.01228	50.65%
$\hat{\mathbf{X}}_{\to W,Y}$	4.13	5.27%	-0.01702	31.63%
PSM	3.94	9.63%	-0.01597	35.83%
CBPS	4.07	6.54%	**-0.02660**	**6.90**%
CF	3.50	19.64%	-0.01741	30.14%

a wider range of applications than the estimated methods, which require the assumptions of either pretreatment variables or unconfoundedness, or both.

7.4.2 Experiments on two semi-synthetic real-world datasets

We evaluate the performance of DAVS on two semi-synthetic real-world datasets, i.e., IHDP [37] and Twins [147, 152]. A detailed introduction to both datasets can be found in Section 4.4 of Chapter 4. Because the underlying DAGs (or MAGs) of the real-world datasets are unknown, the COSO variable Q needs to be given by domain experts or knowledge, so DAVS-Q does not evaluate on these datasets.

Results on IHDP. From the results in Table 7.4, we can see that DAVS outperforms IDA, LVIDA, and GAC, which rely on causal structure

learning. DAVS, CFRNet, DICE, EHS, $\hat{\mathbf{X}}_{\rightarrow W,Y}$, PSM, and CBPS have comparable performance. Moreover, DAVS is better than the other left estimators. This means that DAVS can work well on small samples of data.

Results on Twins. The experimental results of all methods are reported in Table 7.4. In Table 7.4, the DAVS performs the best among all comparisons. The performances of IDA, LVIDA, and GAC, which rely on causal structure learning, outperform most other methods (excluding $\hat{\mathbf{X}}_{\rightarrow W}$ and CBPS).

7.4.3 Experiments on two real-world datasets

We also evaluate DAVS on two real-world datasets, i.e., Cattaneo2 [153] and RHC [154]. A detailed description of these two real-world datasets can be found in Section 4.4 of Chapter 4. Since both datasets have no true causal effect, we judge the performance of DAVS according to whether DAVS has achieved performance consistent with other algorithms. The 95% C.I. is employed to evaluate the estimated interval of DAVS.

Results on Cattaneo2. The experimental results of all methods are listed in Table 7.5 and shown in Fig. 7.2. The estimated range of maternal smoking on babies' birth weight is -275.25g to -120.88g. Furthermore, the estimated effects from the IDA, DAVS, DICE, $\hat{\mathbf{X}}_{\rightarrow W}$ and $\hat{\mathbf{Z}}_{\rightarrow Y}$ methods fall within the empirical estimated interval (-250g, -200g), as shown in Fig. 7.2. Other estimated results by the remaining methods deviate from the empirical estimates (including the deep learning-based method, CFRNet, tree-based method, CF). Hence, DAVS has better performance than most of the state-of-the-art estimation methods and is consistent with the empirically estimated effects [152]. Overall, the experimental results provide strong evidence that smoking leads to lower birth weight.

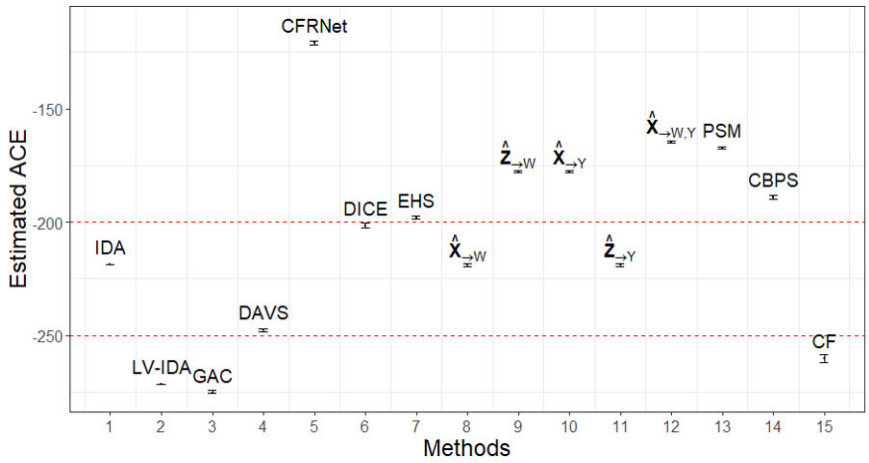


Figure 7.2: The estimated causal effects of all methods in terms of a 95% confident interval. The two dotted lines denote the empirical estimated interval (-250g, -200g).

Table 7.5: Estimating ACE on Cattaneo2 and RHC.

Methods	Cattaneo2		RHC	
	ACE	95% C.I.	ACE	95% C.I.
IDA	-218.82	/	0.06656	/
LVIDA	-271.38	(-271.46, -271.30)	0.04004	(0.03971, 0.04038)
GAC	-275.25	(-275.54, -274.25)	0.04004	(0.03971, 0.04038)
DAVS	-247.62	(-248.33, -246.90)	0.03951	(0.03944, 0.03958)
CFRNet	-120.88	(-121.63, -120.13)	0.01293	(0.01032, 0.01554)
DICE	-201.64	(-202.71, -200.58)	0.02903	(0.02817, 0.02990)
EHS	-198.05	(-198.60, -197.50)	/	/
$\hat{\mathbf{X}}_{\rightarrow W}$	-218.82	(-219.47, -218.18)	0.04069	(0.03992, 0.04147)
$\hat{\mathbf{Z}}_{\rightarrow W}$	-177.56	(-178.05, -177.06)	0.01564	(0.01499, 0.01630)
$\hat{\mathbf{X}}_{\rightarrow Y}$	-177.56	(-178.05, -177.06)	0.02797	(0.02731, 0.02863)
$\hat{\mathbf{Z}}_{\rightarrow Y}$	-218.82	(-219.47, -218.18)	0.01905	(0.01827, 0.01983)
$\hat{\mathbf{X}}_{\rightarrow W,Y}$	-164.68	(-165.20, -164.16)	0.05218	(0.05145, 0.05291)
PSM	-167.22	(-167.74, -166.70)	0.06135	(0.06059, 0.06210)
CBPS	-189.16	(-190.10, -188.22)	0.02093	(0.01934, 0.02250)
CF	-260.11	(-261.82, -258.40)	0.02563	(0.02433, 0.02693)

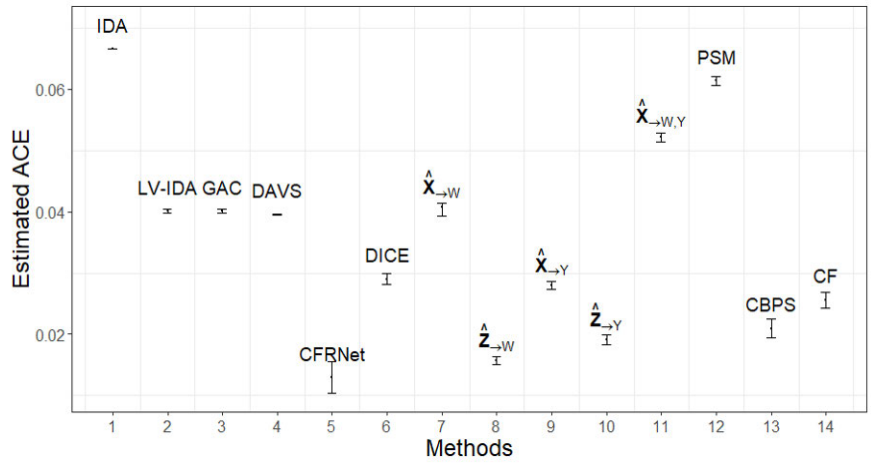


Figure 7.3: The estimated causal effects of all methods in terms of a 95% confident interval. There is no true causal effect or empirical estimates on the RHC dataset.

Results on RHC. The experimental results on the RHC dataset are listed in Table 7.5 and displayed in Fig. 7.3 (excluding EHS since it cannot return results in two hours). According to the results, the estimated ACE range is 0.01293 to 0.06656. The DAVS result is consistent with the estimated results from most of the estimated methods. Overall, the estimated causal effects of all methods (including our developed DAVS method) show that applying RHC results in higher mortality with 180 days than the patient who does not apply RHC.

In summary, according to the experimental results on two semi-synthetic and two real-world datasets, we conclude that DAVS obtains estimates that are approximate to the true or empirically estimated causal effects on four (semi-synthetic) real-world datasets. It also shows that DAVS has very good performance and can be used in real-world applications such as state-of-art comparison methods.

7.4.4 Efficiency evaluation

All computations were conducted on a PC with a *2.6 GHz IntelCore i7* and *16 GB* of memory. The algorithm runtimes on all datasets (synthetic

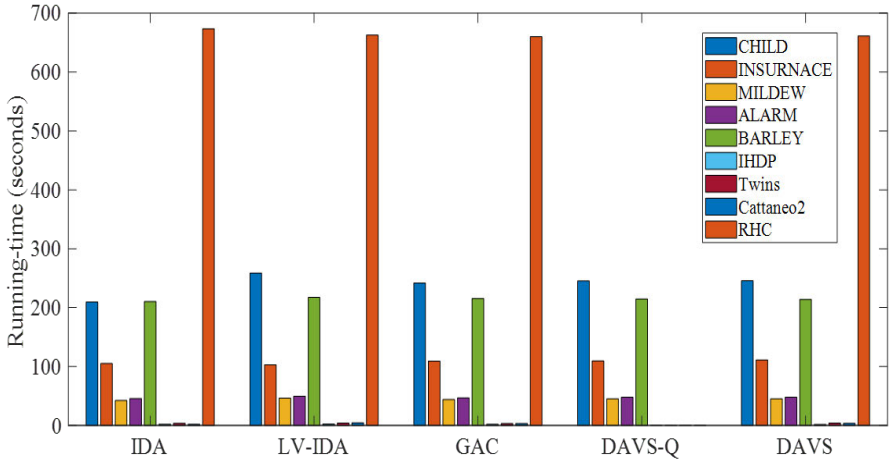


Figure 7.4: Runtime in seconds of different algorithms.

and real-world data) are shown in Fig. 7.4 (we compare DAVS with IDA, LVIDA, and GAC). In Fig. 7.4, IDA is the fastest since learning DAGs from data is faster than learning MAGs. GAC, LVIDA, and our proposed two algorithms have very similar time efficiencies on five synthetic datasets. Furthermore, on four real-world datasets, GAC, LVIDA, and DAVS also have very similar time consumption.

7.4.5 Sensitive Analysis

To investigate the sensitivity of COSO Q on the estimated results of DAVS, we analyse the estimated results from DAVS on the synthetic datasets in Section 7.4.1. According to the true MAGs of ALARM and CHILD, there exists a single COSO variable Q in the synthetic datasets. Their corresponding estimated results are reported in Table 7.3. We visualize the estimated causal effects under different COSO variables Q in Fig. 7.5 on three synthetic datasets, i.e., INSURANCE, MILDEW, and BARLEY. From Fig. 7.5, we can see that these estimated causal effects are very close to the true causal effect under different COSO variables Q on all three synthetic datasets, which further confirms the soundness of Theorem 7.3. In other words, the estimated causal effects are not sensitive to the COSO variable Q.

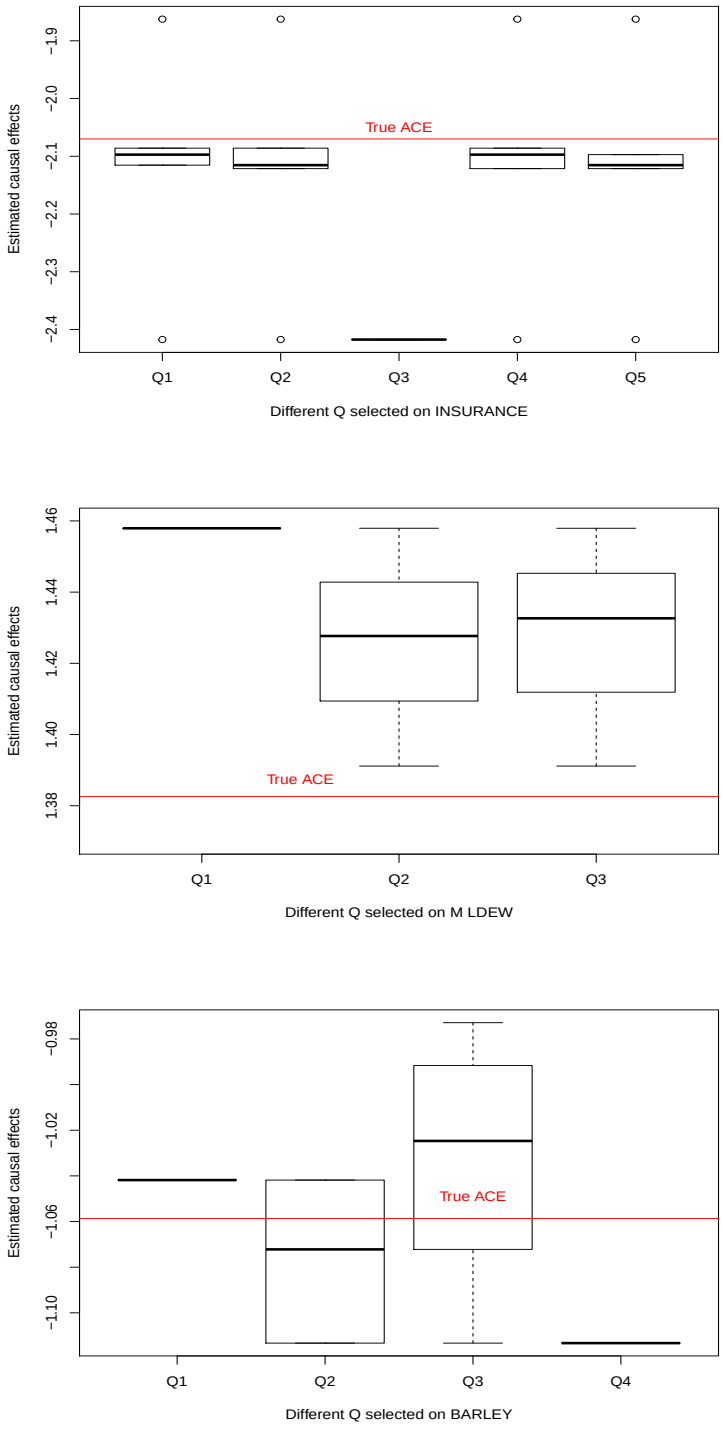


Figure 7.5: The boxplots of the estimated causal effects under different COSO Q on three synthetic datasets, i.e., INSURANCE, MILDEW and BARLEY. Choosing different COSO Q, the estimated causal effects are close to the true causal effect on all three synthetic datasets.

7.5 Conclusion

Previous studies of causal effects estimation are based on an assumption of pretreatment variables, which limits the application scope of data-driven algorithms. When there are multiple causal paths from W to Y, most existing methods and the three algorithms developed in the previous three chapters can address this situation.

In this chapter, we relaxed the pretreatment variable assumption and identified a practical problem setting in which the causal effects of W on Y can be estimated uniquely and unbiasedly from observational data with latent variables. In particular, under our problem setting, we developed two theorems to ensure that it is sufficient and necessary to perform conditional independence tests on data to find a proper adjustment set for unbiased causal effect estimation. We proposed two data-driven algorithms, DAVS-Q and DAVS, to obtain unique and more accurate causal effect estimation from data with latent variables.

Experimental results on synthetic, semi-synthetic and real-world datasets demonstrate that our proposed algorithms achieve better results for data with latent variables compared to the state-of-the-art causal effect estimators. Compared to existing causal effect estimators, our proposed algorithms relax the pretreatment variable assumption and allow latent variables and can provide a unique and unbiased estimation. This significantly improves the practical applications of data-driven causal effect estimation.

Chapter 8

Conclusion and Future Directions

In recent decades, estimating ACE has been an important research direction, as it provides a scientific and effective method for quantifying causal relationships in many real-world applications. Some criteria based on graphical causal modelling have been developed to select proper adjustment sets. Nevertheless, when these criteria are applied to estimate ACE from observational data in real-world applications, there are still gaps, such as data insufficiency, uncertainty in adjustment identification, latent variables and low time efficiency in estimation methods. In this thesis, we bridged some of the gaps between theory and practice while focusing on effectively determining proper adjustment sets by graphical criteria for estimating ACE from data.

8.1 Summary of contributions

We identified four challenges in estimating causal effects from observational data. The first challenge is data insufficiency. One promising method to address data insufficiency is to identify a proper adjustment set by feature space transformation. However, transformation techniques lack a theoretical connection with covariate adjustment. The second challenge is that most of the existing causal effect estimation algorithms based on graphical

causal modelling can only produce a bound estimation of causal effects instead of unique estimations. This uncertainty in estimated causal effects is caused by nondeterminism in adjustment set identification. The third challenge is that bias may be introduced in ACE estimation when there are latent variables in the system. The last challenge is that existing estimation methods have low time efficiency.

In Chapter 4, under the pretreatment variables assumption, we provided a theorem to connect SDR and the back-door criterion, and we proved that a reduced set learned by SDR satisfies the back-door criterion and is sufficient for removing confounding bias in causal effect estimation from data without latent variables. This work addressed the first challenge, i.e., data insufficiency. Based on the proposed theorem, we developed a data-driven algorithm, CESD, to estimate ACE from data. The set of irrelevant representative variables and W are conditionally independent given the learned adjustment set $\mathbf{Z}$ in the space of RKHS. Moreover, the proposed theorem guarantees that many sufficiency dimension reduction methods can be used to estimate ACE from observational data without latent variables. Our work is the first to link SDR with adjustment set identification theoretically. The experimental results demonstrated the effectiveness of the developed algorithm in causal effect estimations.

In most real-world applications, it is inevitable that the data may lack some important factors because there are latent variables in the system that cannot be measured. In this case, algorithms assuming causal sufficiency may draw incorrect conclusions. Additionally, many current algorithms that are designed to handle latent variables are very inefficient. To address these practical issues, i.e., the third and fourth challenges, local search is an intuitive idea to improve efficiency. In Chapter 5, under the pretreatment variable assumption, we propose a local search theorem for identifying proper adjustment sets in data with latent variables. Moreover, we developed an efficient algorithm, DICE, to estimate ACE in data with latent variables. The proposed DICE algorithm significantly improves the efficiency of adjustment set discovery by reducing the original search space to the local structure around W and Y. We evaluated DICE using synthetic

and real-world datasets and compared it with the existing methods. The experimental results showed that DICE can obtain a small bound of causal effects.

Like most existing methods, DICE also provides a bound estimation of causal effects. In many situations, domain knowledge is not available for us to select the most accurate estimation from a set of estimations produced by an algorithm, and thus it is desirable to have algorithms that can obtain unique ACE estimations and deterministic identification of adjustment sets from data with latent variables. Hence, in Chapter 6, based on the local search algorithm developed in Chapter 6 and the EHS condition, we developed a set of theorems to support the data-driven method for deterministic and minimal adjustment sets from data with latent variables under the pretreatment variables assumption. Following the developed theorems, we proposed a data-driven algorithm, CE^2LS, to obtain a unique and unbiased ACE estimation from data with latent variables. We use the *apriori* strategy to search for minimal adjustment sets to further improve efficiency. Experiments on both synthetic and real-world datasets demonstrated high effectiveness and efficiency of CE^2LS.

The methods proposed in Chapter 4, Chapter 5 and Chapter 6 all require the pretreatment variables assumption. Although the pretreatment assumption has been commonly adopted by ACE estimation methods and, in general, is a practical assumption, sometimes, it can restrict the application of ACE estimation. In Chapter 7, we developed theorems to search for proper adjustment sets without the pretreatment variables assumption. We presented two data-driven algorithms, DAVS-Q and DAVS, to obtain unique and unbiased ACE from data with latent variables. The experimental results showed the effectiveness of the two proposed algorithms compared to the state-of-the-art causal effect estimation methods.

In conclusion, we studied theorems based on graphical causal modelling to support efficient and effective algorithms for estimating causal effects from data. In this thesis, we addressed the fundamental and challenging issues encountered by the existing data-driven approach, including data insufficiency, uncertainty in adjustment set identification, latent variables, and

low efficiency, potentially caused by the high dimensionality of data. Therefore, our work provides good insights into adjustment set identification and addresses practical data-driven causal effect estimation for high-dimensional data.

8.2 Future directions

In the following, we identify several potential future directions that can extend the work in this thesis:

- *How can the CESD framework be extended to deal with multiple treatments and multiple outcomes?* In real-world applications, multiple treatments and outcomes are very common [49, 171]. Nabi et al. [171] proposed a semi-parametric causal SDR method for multiple treatments, but their work did not address multiple outcomes. Therefore, developing theorems to support SDR for both multiple treatments and multiple outcomes simultaneously is a very meaningful research direction.

- *How can CE^2LS and $DAVS$ be extended to deal with selection bias?* Removing selection bias from estimation is an important research topic in causal inference. Correa and Bareinboim [172] developed sufficient and necessary graphical conditions for covariate adjustment in latent projection graphs, which allows the presence of selected variables. However, in their work, it was assumed that the causal graph was given. It will be an interesting extension to use the COSO variable proposed in this thesis to design a local search method to remove selection bias from data without assuming a known causal graph.

- *How can the local search idea be used in additive noise models?* For additive noise models, sometimes the causal DAG can be recovered under certain strong assumptions [130, 131, 132, 133]. Hence, joining the local search idea and additive noise model for quick identification of adjustment sets from data will be interesting.

- *How can the local search idea be used in this thesis to help design deep learning causal inference methods?* Recently, machine learning-based algorithms, especially deep learning-based algorithms, have promoted the development of causal inference [39, 93, 147, 173, 174]. Deep learning-based methods estimate causal effects by considering all covariate variables when attempting to eliminate confounding bias, and they do not consider latent variables. Hence, using the local search idea proposed in this thesis for deep learning-based methods to learn a PAG from data with latent variables is a promising direction.

References

[1] Rogier Kievit, Willem Eduard Frankenhuis, Lourens Waldorp, and Denny Borsboom. **Simpson's paradox in psychological science: a practical guide**. *Frontiers in psychology*, **4**:513, 2013. v, 3, 4

[2] Ian Shrier and Robert W Platt. **Reducing bias through directed acyclic graphs**. *BMC medical research methodology*, **8**(1):70, 2008. v, 6, 7

[3] Judea Pearl. **Myth, confusion, and science in causal analysis**. *Tech. Rep. R-348*, 2009. Los Angeles, CA: University of California. v, 4, 6, 7, 64, 76, 93

[4] Jiji Zhang. **Causal reasoning with ancestral graphs**. *Journal of Machine Learning Research*, **9**(Jul):1437–1474, 2008. vi, 33, 34, 36, 37

[5] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. **Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces**. *Journal of Machine Learning Research*, **5**(Jan):73–99, 2004. vi, 9, 46, 48, 50, 51

[6] Janine Witte and Vanessa Didelez. **Covariate selection strategies for causal inference: Classification and comparison**. *Biometrical Journal*, **61**(5):1270–1289, 2019. vii, 13, 46, 93, 94, 112

[7] Jenny Häggström. **Data-driven confounder selection via Markov and Bayesian networks**. *Biometrics*, **74**(2):389–398, 2018. vii, 18, 20, 25, 28, 53, 70, 93, 94

[8] Peter Spirtes. **Introduction to causal inference.** *Journal of Machine Learning Research*, **11**(5), 2010. 1

[9] Judea Pearl. *Causality*. Cambridge university press, 2009. 1, 3, 4, 6, 13, 15, 16, 21, 26, 31, 33, 38, 115

[10] James M Robins and Sander Greenland. **Identifiability and exchangeability for direct and indirect effects**. *Epidemiology*, pages 143–155, 1992. 1, 13

[11] Sander Greenland, Judea Pearl, and James M Robins. **Causal diagrams for epidemiologic research**. *Epidemiology*, pages 37–48, 1999. 1

[12] Manuel Arellano and Olympia Bover. **Another look at the instrumental variable estimation of error-components models**. *Journal of econometrics*, **68**(1):29–51, 1995. 1

[13] Alberto Abadie and Guido W Imbens. **Matching on the estimated propensity score**. *Econometrica*, **84**(2):781–807, 2016. 1, 2, 13

[14] Guido Imbens. **Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics**. Technical report, National Bureau of Economic Research, 2019. 1, 15

[15] Robert J LaLonde. **Evaluating the econometric evaluations of training programs with experimental data**. *The American economic review*, pages 604–620, 1986. 1, 53, 56, 57, 72, 106

[16] Gary King, Robert O Keohane, and Sidney Verba. *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press, 1994. 1

[17] Clark Glymour. **Learning causes: Psychological explanations of causal explanation**. *Minds and machines*, **8**(1):39–60, 1998. 1

[18] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000. 1, 2, 6, 7, 14, 19, 21, 25, 26, 32, 36, 69, 71, 79, 106

[19] Marcia L Meldrum. **A brief history of the randomized controlled trial: From oranges and lemons to the gold standard**. *Hematology/Oncology Clinics*, **14**(4):745–760, 2000. 1

[20] Angus Deaton and Nancy Cartwright. **Understanding and misunderstanding randomized controlled trials**. *Social Science & Medicine*, **210**:2–21, 2018. 1

[21] Donald B Rubin. **Using multivariate matched sampling and regression adjustment to control bias in observational studies**. *Journal of the American Statistical Association*, **74**(366a):318–328, 1979. 2, 53

[22] Paul R Rosenbaum and Donald B Rubin. **The central role of the propensity score in observational studies for causal effects**. *Biometrika*, **70**(1):41–55, 1983. 2, 13, 14, 42, 43, 69, 73

[23] Paul R Rosenbaum et al. *Design of observational studies*, **10**. Springer, 2010. 2, 13

[24] Peter C Austin. **An introduction to propensity score methods for reducing the effects of confounding in observational studies**. *Multivariate behavioral research*, **46**(3):399–424, 2011. 2

[25] Constantin F Aliferis, Ioannis Tsamardinos, and Alexander Statnikov. **HITON: a novel Markov Blanket algorithm for optimal variable selection**. In *AMIA Annual Symposium Proceedings*, **2003**, page 21. American Medical Informatics Association, 2003. 2, 28, 29

[26] Susan Athey and Guido W Imbens. **Machine learning methods for estimating heterogeneous causal effects**. *stat*, **1050**(5), 2015. 2, 3, 13

[27] Christopher Meek. **Causal inference and causal explanation with background knowledge**. In *the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 403–411, 1995. 2, 15, 32

[28] Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeff Ullman. **Scalable techniques for mining causal structures**. *Data Mining and Knowledge Discovery*, **4**(2):163–192, 2000. 2

[29] Dominik Janzing, Patrik O Hoyer, and Bernhard Schölkopf. **Telling cause from effect based on high-dimensional observations**. *arXiv preprint arXiv:0909.4386*, 2009. 2

[30] Andrew Gelman. **Causality and statistical learning**, 2011. 2

[31] Kosuke Imai, Marc Ratkovic, et al. **Estimating treatment effect heterogeneity in randomized program evaluation**. *The Annals of Applied Statistics*, **7**(1):443–470, 2013. 2

[32] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015. 2, 3, 12, 13, 14, 15, 24, 41, 42

[33] Miguel A Hernán and James M Robins. **Causal inference: what if**, 2020. 2

[34] Peter Bühlmann, Markus Kalisch, and Marloes H Maathuis. **Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm**. *Biometrika*, **97**(2):261–278, 2010. 2, 28, 69, 87, 89, 90

[35] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. **Learning high-dimensional directed acyclic graphs with latent and selection variables**. *The Annals of Statistics*, pages 294–321, 2012. 2, 21, 26, 35, 36, 71, 91, 112, 113

[36] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. **Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation**. *Journal of Machine Learning Research*, **11**(Jan):171–234, 2010. 2, 7, 8, 21, 25, 28, 29, 69, 87, 90, 100

[37] Jennifer L Hill. **Bayesian nonparametric modeling for causal inference**. *Journal of Computational and Graphical Statistics*, **20**(1):217–240, 2011. 2, 53, 54, 66, 72, 74, 75, 82, 97, 117

[38] Susan Athey, Guido W Imbens, and Stefan Wager. **Approximate residual balancing: debiased inference of average treatment effects in high dimensions**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**(4):597–623, 2018. 2, 14

[39] Uri Shalit, Fredrik D Johansson, and David Sontag. **Estimating individual treatment effect: generalization bounds and algorithms**. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017. 2, 13, 14, 74, 75, 113, 128

[40] Susan Athey and Guido Imbens. **Recursive partitioning for heterogeneous causal effects**. *Proceedings of the National Academy of Sciences*, **113**(27):7353–7360, 2016. 2, 12, 13

[41] Claudia Shi, David Blei, and Victor Veitch. **Adapting neural networks for the estimation of treatment effects**. In *Advances in neural information processing systems*, pages 2507–2517, 2019. 2, 13

[42] Alexandros Rekkas, Jessica K Paulus, Gowri Raman, John B Wong, Ewout W Steyerberg, Peter R Rijnbeek, David M Kent, and David van Klaveren. **Predictive approaches to heterogeneous treatment effects: a scoping review**. *BMC Medical Research Methodology*, **20**(1):1–12, 2020. 3

[43] Judea Pearl. *An Introduction to Causal Inference*. 2015. 3, 105

[44] David Galles and Judea Pearl. **An axiomatic characterization of causal counterfactuals**. *Foundations of Science*, **3**(1):151–182, 1998. 3

[45] Judea Pearl. **An introduction to causal inference**. *The International Journal of Biostatistics*, **6**(2):1–62, 2010. 3, 13, 15

[46] Judea Pearl et al. **Causal inference in statistics: An overview**. *Statistics surveys*, **3**:96–146, 2009. 3, 13, 14, 38

[47] Edward H Simpson. **The interpretation of interaction in contingency tables**. *Journal of the Royal Statistical Society: Series B (Methodological)*, **13**(2):238–241, 1951. 4

[48] Marloes H Maathuis, Diego Colombo, et al. **A generalized backdoor criterion**. *The Annals of Statistics*, **43**(3):1060–1088, 2015. 4, 15, 17, 39, 79

[49] Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. **Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs**. *The Journal of Machine Learning Research*, **18**(1):8132–8193, 2017. 4, 5, 15, 18, 22, 25, 33, 34, 39, 40, 65, 105, 110, 127

[50] Benito van der Zander, Maciej Liśkiewicz, and Johannes Textor. **Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework**. *Artificial Intelligence*, **270**:1–40, 2019. 4, 15

[51] Donald B Rubin. **Estimating causal effects of treatments in randomized and nonrandomized studies.** *Journal of educational Psychology*, **66**(5):688, 1974. 4, 14, 41, 90, 91, 113

[52] Donald B Rubin. **The use of matched sampling and regression adjustment to remove bias in observational studies**. *Biometrics*, pages 185–203, 1973. 4

[53] Brian C Sauer, M Alan Brookhart, Jason Roy, and Tyler VanderWeele. **A review of covariate selection for non-experimental comparative effectiveness research**. *Pharmacoepidemiology and drug safety*, **22**(11):1139–1145, 2013. 4

[54] Tyler J VanderWeele and Ilya Shpitser. **A new criterion for confounder selection**. *Biometrics*, **67**(4):1406–1413, 2011. 4, 6, 13, 19, 38, 66, 82, 90, 91

[55] Ilya Shpitser, Tyler VanderWeele, and James M Robins. **On the validity of covariate adjustment for estimating causal effects**. *arXiv preprint arXiv:1203.3515*, 2012. 4, 15, 17, 18, 39

[56] Alberto Abadie and Guido W Imbens. **Large sample properties of matching estimators for average treatment effects**. *econometrica*, **74**(1):235–267, 2006. 5, 46, 51, 80

[57] Xavier De Luna, Ingeborg Waernbaum, and Thomas S Richardson. **Covariate selection for the nonparametric estimation of an average treatment effect**. *Biometrika*, **98**(4):861–875, 2011. 5, 15, 20, 25, 46, 66, 80, 82, 87, 90, 91

[58] Elizabeth A Stuart. **Matching methods for causal inference: A review and a look forward**. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **25**(1):1, 2010. 5, 14, 42, 61

[59] Yanyuan Ma and Liping Zhu. **Efficient estimation in sufficient dimension reduction**. *Annals of statistics*, **41**(1):250, 2013. 5

[60] Shujie Ma, Liping Zhu, Zhiwei Zhang, Chih-Ling Tsai, and Raymond J Carroll. **A robust and efficient approach to causal inference based on sparse sufficient dimension reduction**. *Annals of statistics*, **47**(3):1505, 2019. 5, 23, 24, 25, 46

[61] Emma Persson, Jenny Häggström, Ingeborg Waernbaum, and Xavier de Luna. **Data-driven algorithms for dimension reduction in causal inference**. *Computational Statistics & Data Analysis*, **105**:280–292, 2017. 5

[62] Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al. **Estimating high-dimensional intervention effects from observational data**. *The Annals of Statistics*, **37**(6A):3133–3164, 2009. 6, 7, 19, 25, 112

[63] Daniel Malinsky and Peter Spirtes. **Estimating bounds on causal effects in high-dimensional and possibly confounded systems**. *International Journal of Approximate Reasoning*, **88**:371–384, 2017. 6, 21, 25, 65, 73, 90, 91, 92, 109, 110, 112

[64] Tyler J VanderWeele. **Principles of confounder selection**. *European journal of epidemiology*, **34**(3):211–219, 2019. 6, 12, 13

[65] Jiji Zhang. **On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias**. *Artificial Intelligence*, **172**(16-17):1873–1896, 2008. 7, 15

[66] Doris Entner, Patrik Hoyer, and Peter Spirtes. **Data-driven covariate selection for nonparametric estimation of causal effects**. In *Artificial Intelligence and Statistics*, pages 256–264, 2013. 7, 13, 22, 25, 80, 81, 82, 90, 91, 92

[67] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. **Do-calculus when the True Graph Is Unknown.** In *UAI*, pages 395–404. Citeseer, 2015. 7, 21, 25

[68] David Maxwell Chickering. **Learning Bayesian networks is NP-complete**. In *Learning from data*, pages 121–130. Springer, 1996. 7, 27, 65

[69] R Dennis Cook. **Graphics for regressions with a binary response**. *Journal of the American Statistical Association*, **91**(435):983–992, 1996. 8, 23, 48

[70] R Dennis Cook. *Regression graphics: Ideas for studying regressions through graphics*, **482**. John Wiley & Sons, 2009. 8, 23, 48

[71] Nachman Aronszajn. **Theory of reproducing kernels**. *Transactions of the American mathematical society*, **68**(3):337–404, 1950. 9, 49

[72] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. **Kernel methods in machine learning**. *The annals of statistics*, pages 1171–1220, 2008. 9

[73] Debo Cheng, Jiuyong Li, Lin Liu, Thuc Duy Le, Jixue Liu, and Kui Yu. **Sufficient Dimension Reduction for Average Causal Effect Estimation**. *arXiv preprint arXiv:2009.06444*, 2020. 9, 46

[74] Debo Cheng, Jiuyong Li, Lin Liu, Jixue Liu, Kui Yu, and Thuc Duy Le. **Causal query in observational data with hidden variables**. In *24th European Conference on Artificial Intelligence*, 2020. 9, 65, 100

[75] Debo Cheng, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. **Local search for efficient causal effect estimation**. *arXiv preprint arXiv:2007.10014*, 2020. 10, 81

[76] Debo Cheng, Jiuyong Li, Lin Liu, Kui Yu, Thuc Duy Lee, and Jixue Liu. **Towards unique and unbiased causal effect estimation from data with hidden variables**. *arXiv preprint arXiv:2002.10091*, 2020. 10, 105

[77] John Concato, Nirav Shah, and Ralph I Horwitz. **Randomized, controlled trials, observational studies, and the hierarchy of research designs**. *New England Journal of Medicine*, **342**(25):1887–1892, 2000. 12

[78] Sander Greenland. **Quantifying biases in causal models: classical confounding vs collider-stratification bias**. *Epidemiology*, **14**(3):300–306, 2003. 12, 93, 105

[79] James M Robins. **Causal inference from complex longitudinal data**. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997. 12

[80] Tyler J VanderWeele and Ilya Shpitser. **On the definition of a confounder**. *Annals of statistics*, **41**(1):196, 2013. 12

[81] Fredrik Johansson, Uri Shalit, and David Sontag. **Learning representations for counterfactual inference**. In *International conference on machine learning*, pages 3020–3029, 2016. 13, 14

[82] Donald B Rubin. **Matching to remove bias in observational studies**. *Biometrics*, pages 159–183, 1973. 13, 43, 45, 46, 51, 53

[83] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 13

[84] Sebastian Schneeweiss, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun, and M Alan Brookhart. **High-dimensional propensity score adjustment in studies of treatment effects using health care claims data**. *Epidemiology (Cambridge, Mass.)*, **20**(4):512, 2009. 14, 18

[85] Donald B Rubin and Neal Thomas. **Combining propensity score matching with additional adjustments for prognostic covariates**. *Journal of the American Statistical Association*, **95**(450):573–585, 2000. 14, 44

[86] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015. 14, 41

[87] Paul R Rosenbaum. **A characterization of optimal designs for observational studies**. *Journal of the Royal Statistical Society: Series B (Methodological)*, **53**(3):597–610, 1991. 14

[88] Stefano M Iacus, Gary King, and Giuseppe Porro. **Causal inference without balance checking: Coarsened exact matching**. *Political analysis*, pages 1–24, 2012. 14

[89] Wei Luo, Yeying Zhu, and Debashis Ghosh. **On estimating regression-based causal effects using sufficient dimension reduction**. *Biometrika*, **104**(1):51–65, 2017. 14, 46

[90] Jianxuan Liu, Yanyuan Ma, and Lan Wang. **An alternative robust estimator of average treatment effect in causal inference**. *Biometrics*, **74**(3):910–923, 2018. 14, 23, 24, 25, 46

[91] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. **Estimation of regression coefficients when some regressors are not always observed**. *Journal of the American statistical Association*, **89**(427):846–866, 1994. 14

[92] Mark J van der Laan and Susan Gruber. **Collaborative double robust targeted maximum likelihood estimation**. *The international journal of biostatistics*, **6**(1), 2010. 14

[93] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. **Representation Learning for Treatment Effect Estimation from Observational Data**. In *Advances in Neural Information Processing Systems*, pages 2638–2648, 2018. 14, 128

[94] Judea Pearl. **Causal diagrams for empirical research**. *Biometrika*, **82**(4):669–688, 1995. 14, 16, 97

[95] Felix Elwert. **Graphical causal models**. In *Handbook of causal analysis for social research*, pages 245–273. Springer, 2013. 14

[96] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015. 15

[97] Daniel Malinsky, Ilya Shpitser, and Thomas Richardson. **A potential outcomes calculus for identifying conditional path-specific effects**. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088. PMLR, 2019. 15

[98] Thomas Richardson, Peter Spirtes, et al. **Ancestral graph Markov models**. *The Annals of Statistics*, **30**(4):962–1030, 2002. 17, 33, 35, 66

[99] Peter Spirtes. **Causal inference via ancestral graph models**. *Highly structured stochastic systems*, **27**:83, 2003. 17, 33

[100] Benito van der Zander, Maciej Liskiewicz, and Johannes Textor. **Constructing Separators and Adjustment Sets in Ancestral Graphs.** In *CI@ UAI*, pages 11–24, 2014. 18, 33, 39, 40, 82

[101] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. **Variable selection for propensity score models**. *American journal of epidemiology*, **163**(12):1149–1156, 2006. 19

[102] M Maria Glymour, Jennifer Weuve, and Jarvis T Chen. **Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: measurement, selection, and bias**. *Neuropsychology review*, **18**(3):194–213, 2008. 19

[103] Preetam Nandy, Marloes H Maathuis, Thomas S Richardson, et al. **Estimating the effect of joint interventions from observational data in sparse high-dimensional settings**. *The Annals of Statistics*, **45**(2):647–674, 2017. 20, 25

[104] Zhuangyan Fang and Yangbo He. **IDA with Background Knowledge**. In *Conference on Uncertainty in Artificial Intelligence*, pages 270–279. PMLR, 2020. 20, 25

[105] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. **The max-min hill-climbing Bayesian network structure learning algorithm**. *Machine learning*, **65**(1):31–78, 2006. 20, 25, 27, 28

[106] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. **Constraint-based Causal Discovery: Conflict Resolution with Answer Set Programming.** In *UAI*, pages 340–349, 2014. 21

[107] Ilya Shpitser and Judea Pearl. **Identification of joint interventional distributions in recursive semi-Markovian causal models**. In *Proceedings of the National Conference on Artificial Intelligence*, **21**, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006. 21

[108] Craig A Tovey. **A simplified NP-complete satisfiability problem**. *Discrete applied mathematics*, **8**(1):85–89, 1984. 21

[109] Wei Luo and Yeying Zhu. **Matching using sufficient dimension reduction for causal inference**. *Journal of Business & Economic Statistics*, pages 1–13, 2019. 23, 25, 53

[110] Trinetri Ghosh, Yanyuan Ma, and Xavier de Luna. **Sufficient dimension reduction for feasible and robust estimation of average causal effect**. *Statistica Sinica*, 2020. 23, 24, 25, 46, 53

[111] Kenji Fukumizu, Francis R Bach, Michael I Jordan, et al. **Kernel dimension reduction in regression**. *The Annals of Statistics*, **37**(4):1871–1905, 2009. 23, 49

[112] Kenji Fukumizu and Chenlei Leng. **Gradient-based kernel dimension reduction for regression**. *Journal of the American Statistical Association*, **109**(505):359–370, 2014. 23, 24

[113] R Dennis Cook and Sanford Weisberg. **Sliced inverse regression for dimension reduction: Comment**. *Journal of the American Statistical Association*, **86**(414):328–332, 1991. 23

[114] Bing Li, Yuexiao Dong, et al. **Dimension reduction for nonelliptically distributed predictors**. *The Annals of Statistics*, **37**(3):1272–1298, 2009. 23

[115] David Heckerman, Dan Geiger, and David M Chickering. **Learning Bayesian networks: The combination of knowledge and statistical data**. *Machine learning*, **20**(3):197–243, 1995. 25

[116] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. **DAGs with NO TEARS: Continuous optimization for structure learning**. *Advances in Neural Information Processing Systems*, **31**:9472–9483, 2018. 25, 27

[117] Maxime Gasse, Alex Aussem, and Haytham Elghazel. **A hybrid algorithm for Bayesian network structure learning with application to multi-label learning**. *Expert Systems with Applications*, **41**(15):6755–6772, 2014. 25

[118] A Philip Dawid. **Conditional independence in statistical theory**. *Journal of the Royal Statistical Society: Series B (Methodological)*, **41**(1):1–15, 1979. 25

[119] Judea Pearl and Thomas S Verma. **A theory of inferred causation**. In *Studies in Logic and the Foundations of Mathematics*, **134**, pages 789–811. Elsevier, 1995. 26

[120] Gregory F Cooper and Edward Herskovits. **A Bayesian method for the induction of probabilistic networks from data**. *Machine learning*, **9**(4):309–347, 1992. 27

[121] Gideon Schwarz et al. **Estimating the dimension of a model**. *Annals of statistics*, **6**(2):461–464, 1978. 27

[122] Wray Buntine. **Theory refinement on Bayesian networks**. In *Uncertainty Proceedings 1991*, pages 52–60. Elsevier, 1991. 27

[123] David Maxwell Chickering. **Optimal structure identification with greedy search**. *Journal of machine learning research*, **3**(Nov):507–554, 2002. 27

[124] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. **A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images**. *International journal of data science and analytics*, **3**(2):121–129, 2017. 27

[125] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. **Learning functional causal models with generative neural networks**. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer, 2018. 27

[126] Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, and Yixin Chen. **D-vae: A variational autoencoder for directed acyclic graphs**. In *Advances in Neural Information Processing Systems*, pages 1588–1600, 2019. 27

[127] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. **DAG-GNN: DAG Structure Learning with Graph Neural Networks**. In *International Conference on Machine Learning*, pages 7154–7163, 2019. 27

[128] Yoshua Bengio, Aaron Courville, and Pascal Vincent. **Representation learning: A review and new perspectives**. *IEEE transactions on pattern analysis and machine intelligence*, **35**(8):1798–1828, 2013. 27

[129] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, **1**. MIT press Cambridge, 2016. 27

[130] Arthur Gretton, Peter Spirtes, and Robert Tillman. **Nonlinear directed acyclic structure learning with weakly additive noise models**. *Advances in neural information processing systems*, **22**:1847–1855, 2009. 27, 127

[131] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. **Nonlinear causal discovery with additive noise models**. *Advances in neural information processing systems*, **21**:689–696, 2008. 27, 127

[132] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. **Causal discovery with continuous additive noise models**. *The Journal of Machine Learning Research*, **15**(1):2009–2053, 2014. 27, 127

[133] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. The MIT Press, 2017. 27, 127

[134] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. **D'ya like DAGs? A Survey on Structure Learning and Causal Discovery**. *arXiv preprint arXiv:2103.02582*, 2021. 27

[135] Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. **Algorithms for Large Scale Markov Blanket Discovery.** In *FLAIRS conference*, **2**, pages 376–380, 2003. 28

[136] Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. **Time and sample efficient discovery of Markov blankets and direct causal relations**. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678. ACM, 2003. 29

[137] Richard E Neapolitan et al. *Learning bayesian networks*, **38**. Pearson Prentice Hall Upper Saddle River, NJ, 2004. 32

[138] Emilija Perković, Markus Kalisch, and Maloes H Maathuis. **Interpreting and using CPDAGs with background knowledge**. *arXiv preprint arXiv:1707.02171*, 2017. 32

[139] Elias Bareinboim, Jin Tian, and Judea Pearl. **Recovering from selection bias in causal and statistical inference**. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014. 33

[140] R Ayesha Ali, Thomas S Richardson, Peter Spirtes, et al. **Markov equivalence for ancestral graphs**. *The Annals of Statistics*, **37**(5B):2808–2837, 2009. 36

[141] James Robins. **A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect**. *Mathematical modelling*, **7**(9-12):1393–1512, 1986. 41

[142] Donald B Rubin. **Randomization analysis of experimental data: The Fisher randomization test comment**. *Journal of the American statistical association*, **75**(371):591–593, 1980. 41

[143] Brian K Lee, Justin Lessler, and Elizabeth A Stuart. **Improving propensity score weighting using machine learning**. *Statistics in medicine*, **29**(3):337–346, 2010. 43

[144] Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. **Propensity score estimation with boosted regression for evaluating causal effects in observational studies.** *Psychological methods*, **9**(4):403, 2004. 43

[145] Paul R Rosenbaum. **Overt bias in observational studies**. In *Observational studies*, pages 71–104. Springer, 2002. 44

[146] Jasjeet S Sekhon. **Multivariate and propensity score matching software with automated balance optimization: the matching package for R**. 2011. 44

[147] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. **Causal effect inference with deep latent-variable models**. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017. 53, 56, 97, 117, 128

[148] Alfred F Connors Jr, Neal V Dawson, Charles Thomas, Frank E Harrell Jr, Norman Desbiens, William J Fulkerson, Peter Kussin, Paul Bellamy, Lee Goldman, and William A Knaus. **Outcomes following acute exacerbation of severe chronic obstructive lung disease. The SUPPORT investigators (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments)**. *American journal of respiratory and critical care medicine*, **154**(4):959–967, 1996. 53

[149] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. **Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference**. *Political analysis*, **15**(3):199–236, 2007. 53, 113

[150] Kosuke Imai and Marc Ratkovic. **Covariate balancing propensity score**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(1):243–263, 2014. 53, 56, 57, 66, 72, 73, 113

[151] Stefan Wager and Susan Athey. **Estimation and inference of heterogeneous treatment effects using random forests**. *Journal of the American Statistical Association*, **113**(523):1228–1242, 2018. 53, 74, 82, 113

[152] Lee D S Almond D, Chay K Y. **The costs of low birth weight**. *The Quarterly Journal of Economics*, **120**(3):1031–1083, 2005. 55, 58, 59, 72, 117, 118

[153] Matias D Cattaneo. **Efficient semiparametric estimation of multi-valued treatment effects under ignorability**. *Journal of Econometrics*, **155**(2):138–154, 2010. 58, 118

[154] Alfred F Connors, Theodore Speroff, Neal V Dawson, Charles Thomas, Frank E Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W Wu, Robert M Califf, et al. **The effectiveness of right heart catheterization in the initial care of critically III patients**. *Jama*, **276**(11):889–897, 1996. 59, 118

[155] Wen Wei Loh and Stijn Vansteelandt. **Confounder selection strategies targeting stable treatment effect estimators**. *arXiv preprint arXiv:2001.08971*, 2020. 60

[156] Jasjeet S Sekhon. **Multivariate and propensity score matching software with automated balance optimization: the matching package for R**. *Journal of Statistical Software, Forthcoming*, 2008. 69

[157] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, Peter Bühlmann, et al. **Causal inference using graphical models with the R package pcalg**. *Journal of Statistical Software*, **47**(11):1–26, 2012. 71, 100, 112, 113

[158] Robert Tibshirani. **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 74

[159] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. **BART: Bayesian additive regression trees**. *The Annals of Applied Statistics*, **4**(1):266–298, 2010. 74

[160] Mark J Van Der Laan and Daniel Rubin. **Targeted maximum likelihood learning**. *The International Journal of Biostatistics*, **2**(1), 2006. 74

[161] Marco Scutari. **Learning Bayesian networks with the bnlearn R package**. *arXiv preprint arXiv:0908.3817*, 2009. 77, 96, 114

[162] Johannes Textor and Maciej Liśkiewicz. **Adjustment criteria in causal diagrams: an algorithmic perspective**. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 681–688. AUAI Press, 2011. 87

[163] Rakesh Agrawal, Ramakrishnan Srikant, et al. **Fast algorithms for mining association rules**. In *Proc. 20th int. conf. very large data bases, VLDB*, **1215**, pages 487–499. Citeseer, 1994. 87, 89, 90, 110, 112

[164] Arvid Sjölander. **Regression standardization with the R package stdReg**. *European journal of epidemiology*, **31**(6):563–574, 2016. 92

[165] Joseph Berkson. **Limitations of the application of fourfold table analysis to hospital data**. *Biometrics Bulletin*, **2**(3):47–53, 1946. 93

[166] David M Chickering, Dan Geiger, David Heckerman, et al. **Learning Bayesian networks is NP-hard**. Technical report, Citeseer, 1994. 100

[167] Keyvan Sadeghi and Giovanni M Marchetti. **Graphical Markov models with mixed graphs in R**. *The R Journal*, **4**(2):65–73, 2012. 100

[168] Edwin P Martens, Wiebe R Pestman, Anthonius de Boer, Svetlana V Belitser, and Olaf H Klungel. **Instrumental variables: application and limitations**. *Epidemiology*, pages 260–267, 2006. 106

[169] Miguel A Hernán and James M Robins. **Instruments for causal inference: an epidemiologist's dream?** *Epidemiology*, pages 360–372, 2006. 106

[170] Johannes Textor, Juliane Hardt, and Sven Knüppel. **DAGitty: a graphical tool for analyzing causal diagrams**. *Epidemiology*, **22**(5):745, 2011. 113

[171] Razieh Nabi, Todd McNutt, and Ilya Shpitser. **Semiparametric Causal Sufficient Dimension Reduction Of High Dimensional Treatments**. *arXiv preprint arXiv:1710.06727*, 2017. 127

[172] Juan D Correa and Elias Bareinboim. **Causal effect identification by adjustment under confounding and selection biases**. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 127

[173] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. **GANITE: Estimation of individualized treatment effects using generative adversarial nets**. In *International Conference on Learning Representations*, 2018. 128

[174] Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. **Learning overlapping representations for the estimation of individualized treatment effects**. *arXiv preprint arXiv:2001.04754*, 2020. 128