



Title	A Study on Causal Discovery Considering Confounders
Author(s)	Song, Jing
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第13511号
Issue Date	2019-03-25
DOI	https://doi.org/10.14943/doctoral.k13511
Doc URL	https://hdl.handle.net/2115/74109
Type	doctoral thesis
File Information	Jing_Song.pdf



A Study on Causal Discovery Considering Confounders

Jing Song

A DISSERTATION

Submitted to the Division of Computer Science and Information
Technology

in partial fulfillment of the requirements for the

DOCTOR OF PHILOSOPHY

at the

GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY

HOKKAIDO UNIVERSITY



January 2019

A Study on Causal Discovery Considering Confounders

Jing Song

Abstract

There are a lot of observational data in the real world in which many variables are correlated with each other. Correlation is not equal to causality. The best way to demonstrate a causal relationship between variables is to conduct a controlled randomized experiment. However, real-world experiments are often expensive, unethical, or even impossible. Many researchers working in various fields are thus using statistical methods to analyze causal relationships between variables. Many studies have been conducted to infer causality from raw observational data, but most of them have been based on the assumption that all the variables (including confounders) affecting the causal relationships have been known. Today, however, emphasis has been placed on open data. In the open data environment, it is difficult to consider all related data beforehand, and an exploratory analysis is required to acquire data that can be confounding. Therefore, in this study, first, we analyzed how the existing methods which determines the causal direction between variables are influenced by unknown confounders. Through assessing the existing methods, we found that the existing methods are susceptible to confounding in different degrees. We thus investigated how to decide whether a third variable is confounding for two observed variables. Finally, we studied on a framework to perform causal analysis while considering the possible confounders. We have three purposes for the study. Firstly, investigating a general assessment method for causal discovery methods, especially investigating their performance when the data is confounded. Secondly, investigating how to determine a possible common cause variable. Thirdly, investigating how to do causal analysis of open data while considering the possible confounders.

Acknowledgments

I offer my sincerest gratitude to my respected academic supervisor, Assoc. Prof. Satoshi Oyama. Thanks for all his patient guidance, inspiring suggestions. I will not forget the encouraging words from Assoc. Prof. Satoshi Oyama which helped me to face the challenges in research. I feel very lucky to be one of his students in the past five years.

I also want to express my gratefulness to Prof. Masahiro Kurihara. I have learned many things through listening to comments of presentations in seminars. Thanks for all your helpful advice and your support in my research.

I feel very lucky to be a member of Laboratory of Intelligent Software. Thanks for the warm help from Mr. Duan, Mrs. Wang and Mrs. Hu when I first came to Japan. Thanks for the helpful advice from Mr. Abe, Mr. Furudo, Mr. Atarashi and Mr. Takase in daily research life. Thanks for my friends: Mr. Zhang and Mr. Ji and Mrs. Zhao. Thanks for making my life in Hokkaido so happy and colorful. Thanks for all my colleagues and friends who help me adapt to life here.

Finally, I want to express my gratitude for the love and care from my families. Thanks for all your encouragement and trust.

Contents

Abstract	i
Acknowledgments	ii
List of Figures	vii
List of Tables	viii
List of Algorithms	ix
1 Introduction	1
2 Background	5
2.1 Definition of Causality	5
2.2 Causal Bayesian Network	6
2.3 Cause Effect Pairs	7
2.4 Human Computation	8
2.5 Related Work	9
2.5.1 Causal Analysis	9
2.5.2 Human Computation	12
3 Evaluation of Existing Methods	14
3.1 Introduction	14
3.2 Models	15
3.2.1 ANM	15

3.2.2	PNL Model	16
3.2.3	IGCI	17
3.2.4	Implementation	19
3.3	Dataset	20
3.4	Results	21
3.4.1	Accuracy For Different Decision Rates	22
3.4.2	Area Under ROC Curve (AUC)	23
3.4.3	Algorithm Efficiency	26
3.4.4	Typical Examples of Model Failure	28
3.4.5	Response to Spurious Correlation Caused by “Confounding”	29
3.5	Summary	37
4	Common Cause Identification	38
4.1	Introduction	38
4.2	Intrinsic Dimension Estimation	40
4.2.1	Observable Common Cause Identification	41
4.3	Simulation	46
4.3.1	Simulation of Common Cause Case	47
4.3.2	Simulation of Selection Bias Case	52
4.3.3	Further Study	54
4.4	Real World Data	55
4.4.1	Altitude, Temperature, Sunshine	55
4.4.2	Cement, Blast Furnace Slag, Compressive Strength	56
4.5	Summary	56
5	Exploratory Causal Analysis of Open Data	58
5.1	Introduction	58
5.2	Preliminary Experiments	59
5.3	Proposed Framework	61
5.3.1	Explanation Generation	61
5.3.2	Keywords Extraction	62

5.3.3	Causal Analysis	64
5.4	Experiments	68
5.4.1	GDP and Carbon Dioxide Emmisions	68
5.4.2	College Entrance Rate and Current Liability	77
5.4.3	Summary of the Experiments	83
5.5	Summary	84
6	Conclusion and Future Work	87
A	Pairs of Variables Used in Experiment	90

List of Figures

1-1	Possible Relationships between X and Y.	4
2-1	Directed Acyclic Graph	7
2-2	A Multiluted DAG to Express the Intervention	8
3-1	Scatter Plot of Data in the Experiments	22
3-2	Accuracy of Three Models for Different Decision Rates. Decision rate changed when threshold was changed: the larger the threshold, the smaller the decision rate. In an ideal case, the accuracy of each model should improve with a decrease in the decision rate.	24
3-3	ROC of Three Models. Four graphs are shown because IGCI has two explicit forms.	27
3-4	Example of Discretized Data in CEP Dataset. Variable x_1 is between 0 (very female) and 14 (very male), reflecting features of artificial face. Variable x_2 is 0 (female) or 1 (male), reflecting gender decision. . . .	29
3-5	Example Scatter Plot of Noisy Data in CEP. Variable x_1 is female life expectancy at birth for 2000-2005. Variable x_2 is latitude of birth country's capital (data for China, Russia, and Canada were removed).	30
3-6	Scatter Plots of Generated Data.	32
3-7	Estimated Results of IGCI for Generated Data. Inference was “y causes x” when estimated result was larger than 0.	33
3-8	Test Statistics for PNL Model for Generated Data.	34
3-9	Results for “Common Cause” Case with Real World Data. Variable x_1 : “length”; variable x_2 : “diameter”.	35

3-10	Results for “Selection Bias” Case with Real World Data. Variable x_1 : “altitude”; variable x_2 : “longitude”	36
4-1	Simulation: Common Cause	45
4-2	Simulation: Selection Bias	45
4-3	Estimated Correlation Dimension for Linear Function	48
4-4	Estimated Intrinsic Dimension for Non-Linear Function	48
4-5	Scatter Plot of Artificial Data with Poisson Noise, Multiplicative Noise, and Gaussian Noise	50
4-6	Estimated Correlation Dimension for Added Gaussian Noise with Dif- ferent Variances	52
4-7	Count of Z in Sub-intervals of Interval (0, 2.0)	53
4-8	Scatter Plot of Selection Bias Data for Z in (0.9, 1.0)	53
4-9	Probability Density Function for Three Types of Distributions	54
5-1	Framework	61
5-2	Task Example	62
5-3	Collected Explanations by Crowdsourcing	63
5-4	Workflows.	65
5-5	Markov Equivalence Class.	67
5-6	Extracted Words Using LDA (Experiment i)	71
5-7	Recall and Precision of Causal Relationship Extraction Method and LDA (Experiment i)	72
5-8	Top Ten Words (Experiment ii)	80
5-9	Recall and Precision of Causal Relationship Extraction Method and LDA (Experiment ii)	81

List of Tables

3.1	Dataset Used in CEP.	21
3.2	AUC of Three Models.	26
3.3	Time to Make A Decision	26
3.4	P-value for ANM for Two Directions	34
4.1	Ideal Estimated Intrinsic Dimension for Two Confounding Cases. . .	44
4.2	Estimated Dimension for Different Types of Distributions.	54
5.1	Extracted Words Using Causal Relationship Extraction Method (Ex- periment i)	70
5.2	Extracted Words Using Causal Relationship Extraction Method (Ex- periment ii)	79
5.3	Summary of the Experiments	85
6.1	Pros. and Cons.	88
A.1	Pairs of Causal Variables	91

List of Algorithms

1	Additive Noise Model	16
2	Post Nonlinear Model	17
3	Entropy-based IGCI	19
4	Slope-based IGCI	19
5	Intrinsic Dimension Estimation	41

Chapter 1

Introduction

People are generally more concerned with causal relationships between variables than with statistical relationships between variables, and the concept of causality has been widely discussed [1, 2]. The best way to demonstrate a causal relationship between variables is to conduct a controlled randomized experiment. However, real-world experiments are often expensive, unethical, or even impossible. Many researchers working in various fields (economics, sociology, machine learning, etc.) are thus using statistical methods to analyze causal relationships between variables [3, 4, 5, 6, 7, 8, 9, 10, 11, 12].

Directed acyclic graphs (DAGs) have been used to formalize the concept of causality [13]. Although a conditional independence test cannot tell the full story of a causal relationship, it can be used to exclude irrelevant relationships between variables [14, 13]. However, a conditional independence test is impossible when there are only two variables. Several models have been proposed to solve this problem [15, 16, 17, 18, 19]. For two variables X and Y , there are at least six possible relationships between them (Figure 1-1). The top two diagrams show the independent case and the feedback case, respectively. The middle two show the two possible causal relationships between X and Y : “ X causes Y ” and “ Y causes X .” The bottom two show the “common cause” case and the “selection bias” case. Unobserved variables Z are “confounders”¹ for causal discovery between X and Y . The existence of “con-

¹For the definition of “confounding,” please refer to [20, 21].

founders”² creates a spurious correlation between X and Y . Distinguishing spurious correlation due to unobserved confounders from actual causality remains a challenging task in the field of causal discovery. Many models are based on the assumption that no unobserved confounders exist³.

In the work reported here, we first experimentally compared the performance of three state-of-the-art causal discovery model, the additive-noise model (ANM) [25], the post nonlinear (PNL) model [26], and the information geometric causal inference (IGCI) model [27]. We used three metrics: accuracy, area under the ROC curve (AUC), and time to make a decision. We compared their responses to confounding and found that these models suffered from confounding in different degrees. From the above experimental results, we think that finding and testing possible observable confounders is needed to correctly analyze causal relationship between two variables. Conditional independence is a general method to test whether a third variable Z is a possible common cause for two variables X and Y . However, many methods for conditional independence test have certain assumptions of data distributions, function forms or additive noise. When these assumptions are not satisfied, the conditional independence test may fail. Besides, causal relationship may be mixed with common causes in the real world problem. In this case, conditional independence test will fail to tell a possible common cause. To solve these problems, we proposed using intrinsic dimension estimation as an alternative to detect a possible common cause for two variables. The proposed method is a necessary condition and can be used to filter unrelated words in the candidates for confounders. It can be combined with other causal discovery methods to realize causal relationship analysis while considering specific confounders. The proposed method generally worked well for different types of added noise. In particular, it worked better than a kernel-based conditional independence test for Poisson noise. Besides, the proposed method can infer a possible common cause when it is mixed with causal relationship. Finally, we de-

²The number of “confounders” is not limited to one.

³There have been some efforts to deal with confounders. For example, Shimizu et al. extended the linear non-Gaussian acyclic model [22] to detect causal direction when there are “common causes” [23, 24].

veloped a framework for exploratory causal analysis of open data which combines human computation with machine learning methods. The proposed framework takes into account possible confounders when doing causal analysis and tests them using state-of-the-art machine learning methods. The proposed framework contains three main parts: collecting explanations of correlation in open data using a crowdsourcing marketplace, extracting keywords using natural language processing methods, and verifying the collected explanations using machine learning methods. The proposed framework explanatorily tested the possible confounders which were extracted from the collected texts and filtered the useful ones using machine learning methods. The related variables were then further analyzed using causal discovery methods. The testing step was included to filter the unrelated words. The causal discovery method was included to learn the causal directions between variables.

In Chapter 2, we discussed the background of our research, including the definition of causality, the causal Bayesian network, the task of cause effect pairs and human computation are discussed. Some related work about causal analysis and human computation is introduced as well.

In Chapter 3, In Chapter 3, we gave an assessment of three existing causal discovery models: the additive-noise model (ANM), the post nonlinear (PNL) model, and the information geometric causal inference (IGCI) model using the metrics of accuracy for different decision rates, area under ROC curve (AUC) and algorithm efficiency. Besides, we tested how the three models responded to spurious correlation caused by confounding using simulated and real world data. The experimental results showed that the existing methods are susceptible to confounding in different degrees.

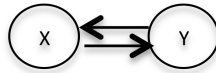
In Chapter 4, we proposed using intrinsic dimension estimation as a necessary condition to determine a possible common cause for two variables. We discussed its performance in several simulated experiments and experiments with real world data.

In Chapter 5, we proposed a framework for crowd-based causal analysis of open data. The proposed framework combined human computation with machine learning methods to solve the task of confounding in causal analysis.

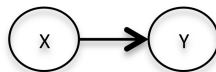
We concluded about our research and discussed our future work in Chapter 6.



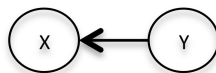
(a) Independent



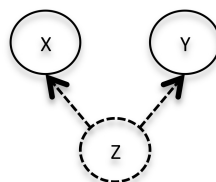
(b) Feedback



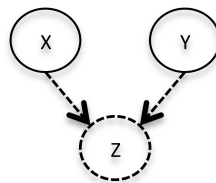
(c) X causes Y



(d) Y causes X



(e) "Common Cause"



(f) "Selection Bias"

Figure 1-1: Possible Relationships between X and Y.

Chapter 2

Background

2.1 Definition of Causality

Philosophers have defined two types of causality: general causality and actual causality.[2] The general causality, only pays attention to the general meaning of variables, for example, the relationship between rain and wet road. Compared with general causality, the actual causality focuses on specifically occurred things, for instance, it rains the whole day today that causes the road becomes wet. The definition of actual causality has been updated in [2]. To model causal relationships using statistical methods, Pearl gives two following definitions in [13].

Causal Structure: A causal structure of a set of variables V is a directed acyclic graph (DAG) in which each node corresponds to a distinct element of V , and each link represents direct functional relationship among the corresponding variables.

Causal Model: A causal model is a pair $M = \langle D, \theta_D \rangle$ consisting of a causal structure D and a set of parameters θ_D compatible with D . The parameters θ_D assign a function $x_i = f_i(pa_i, u_i)$ to each $x_i \in V$ and a probability measure $P(u_i)$ to each u_i , where PA_i are the parents of X_i in D and where each U_i is a random disturbance distributed according to $P(u_i)$, independently of all other u .

A directed acyclic graph (DAG) is usually used to represent a causal structure between variables (Definition 1). We use an example in [13] to illustrate how the causal structure works (Figure 2-1). The example shows us the causal relationship between five variables: season, rain, sprinkler, wet and slippery. The whole joint distribution can be expressed using the following equation (Equation 2.1):

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_4) \quad (2.1)$$

The causal relationship is different from the statistical relationship. We use an example in [13] to show how intervention works in a DAG (Figure 2-2). We take an intervention to turn the sprinkler on. For the causal relationships, there should be no more relationship between the variable x_3 and its parents. Thus the edge between x_1 and x_3 should be removed and a specified value needs to be given to the variable x_3 . On the contrast, for the statistical relationships, if we get the observation that “the sprinkler is on”, there are more chances that the season is dry and there is no rain right now, which means that the joint distribution of the other four variables have also been changed.

2.2 Causal Bayesian Network

Based on the discussion above, Pearl gives the following definition of causal bayesian network in [13].

Causal Bayesian Network Let $P(v)$ be a probability distribution on a set V of variables, and $P_x(v)$ denote the distribution resulting from the intervention $do(X = x)$ which sets a subset X of variables to constants x . Denote by P_* the set of all interventional distributions $P_x(v)$, $X \subseteq V$, including $P(v)$, which represents no intervention (i.e., $X = \emptyset$). A DAG G is said to be a causal Bayesian network compatible with P_* if and only if the following three conditions hold for every $P_x \in P_*$

1. $P_x(v)$ is Markov relative to G

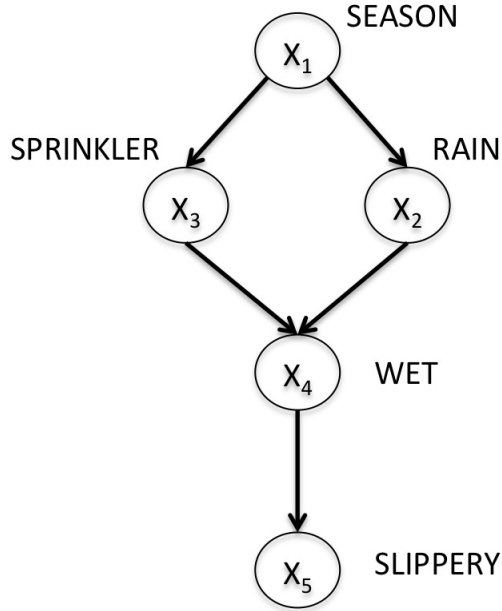


Figure 2-1: A DAG to Express Causal Relationships between Variables.

2. $P_x(V_i) = 1$ for all $V_i \in X$ whenever v_i is consistent with $X = x$
3. $P_x(v_i|pa_i) = P(v_i|pa_i)$ for all $V_i \notin X$ whenever pa_i is consistent with $X = x$,
i.e., each $P(v_i|pa_i)$ remains invariant to interventions not involving V_i .

The definition above gives constraints on the normal Bayesian network to deliver causal information between variables. It has two properties: firstly, the effects of intervention pa_i on its child node v_i can also be expressed using the conditional distribution $P(v_i|pa_i)$; secondly, once the direct causes of v_i has been controlled, no other interventions can affect the distribution of v_i . The causal bayesian network is the basic model in causality.

2.3 Cause Effect Pairs

The structure of causal bayesian network can be obtained using algorithms based on conditional independence tests. A well-known method can be the PC algorithm

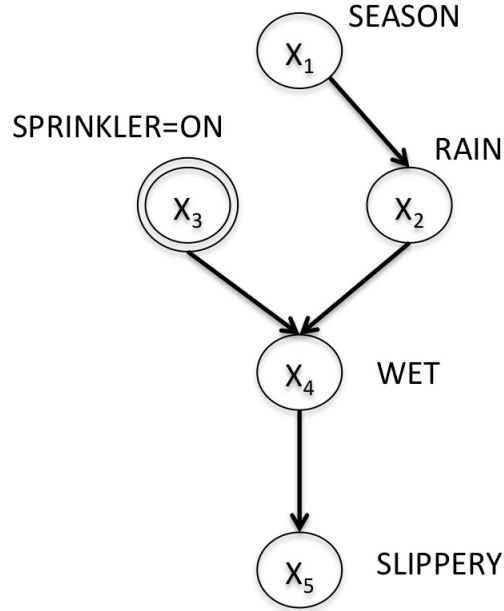


Figure 2-2: A Multilated Graph to Express the Intervention “Turning Sprinkler on”.

[14, 28]. However, these algorithms can only give Markov equivalent classes as the result. For the Markov equivalent class, they cannot decide the right causal structure. A simplest type of Markov equivalent class are pair of variables. The possible patterns for two variables are shown in Figure 1-1. The middle two diagrams show the possible causal relationships between X and Y . The remaining task is to decide the direction of the arrow. The bottom two diagrams represent “common cause” and “selection bias”, which are “confounders” for causal analysis. How to distinguish “confounders” from actual causality is a remaining task in this field.

2.4 Human Computation

Human intelligence or cognition aids in solving tasks which are difficult for machine learning methods. Many research has been done to resolve difficult machine learning takes using crowdsourcing. To solve the problem of confounding in causal analysis, human computation aids as well. Experts in specific domains can do causal analysis

while considering the effect of possible confounders with their domain knowledge. For instance, Huang et al. analyzed the causal relationship between energy consumption and GDP growth by dividing the data into four income categories [29]. Their experimental results showed that causal relationships differed among the four categories. The above research showed that the causal relationship may change after considering a certain confounder. Based on the above experimental results, we think analyzing causal relationships while considering the effect of possible confounders is of importance. Human expertise aids in the real world causal analysis task. Crowdsourcing is one way to collect them and solve some problems difficult for machine learning methods. Some related work about crowdsourcing and its application will be given in Section 2.5.

2.5 Related Work

2.5.1 Causal Analysis

Causal discovery [30, 31] has attracted the interest of researchers in various fields. The concept of causality was discussed and refined [32, 33]. Many causal discovery methods have been proposed until now. Some causal discovery models are based on cyclic models [34], which take into account feedback between variables in a causal relationship. Most causal discovery models use a directed acyclic graph (DAG) to express the causal relationships between variables [35, 30, 31]. The PC (Peter and Clark) algorithm [35] is a constraint-based algorithm to determine a set of DAGs in the same Markov equivalence class in which conditional independence test was used as a subroutine. Several methods have been proposed to test the conditional independence between variables and then apply them into causal discovery. [36] proposed using weakly additive noise models that use local permutation to measure conditional dependence. [37] proposed a kernel-based conditional independence (KCI) test in which there is no assumption about the specific functional form between variables and there is no constraint on the data distribution. The above methods either had

additive noise assumption or used the uncorrelatedness of residual function to do the conditional independence test. When these assumptions were not satisfied, the conditional independence test may fail. The PC algorithm did not take into account the effect of confounding which is sound and complete in the assumption of causal sufficiency and faithfulness. The latterly proposed FCI (Fast Causal Inference) algorithm and RFCI algorithms extends the PC algorithm and handles the common cause and selection bias.

Temporal information is useful for causal discovery modeling [38, 39]. Granger [40] proposed detecting the causal direction of time series data on the basis of the temporal ordering of the variables and used linear systems to make it more operational. He formulated the definition of causality in terms of conditional independence relations [41]. Chen et al. extended the linear stochastic systems he proposed [40] to work on nonlinear systems [42]. Shajarisales et al. proposed using the spectral independence criterion (SIC) for causal inference from time series data and a mechanism different from that used in Granger causality and compared the two methods [43]. For Granger causality [40] and extended Granger causality [42], temporal information is needed. When discovering causal relationship from time series data, the data resolution might be different from the true causal frequency. Gong et al. [44] discussed this issue and showed that using non-Gaussian of data can help identify the underlying model under some conditions.

Shimizu et al. [22] proposed using a linear non-Gaussian acyclic model (LiNGAM for short) to detect the causal direction of variables without the need for temporal information. LiNGAM works when the causal relationship between variables is linear, the distributions of disturbance variables are non-Gaussian, and the network structure can be expressed using a DAG. Several extensions of LiNGAM have been proposed [45, 23, 46, 24].

LiNGAM is based on the assumption of linear relationships between variables. Hoyer et al. [25] proposed using an additive noise model (ANM) to deal with nonlinear relationships. If the regression function is linear, ANM works in the same way as LiNGAM. Zhang et al. [47, 48, 26, 49] proposed using a PNL model that takes

into account the nonlinear effect of causes, inner additive noise, and external sensor distortion. The ANM and PNL model are briefly introduced in the following section.

While the above models are based on structural equation modeling (SEM), which requires structural constraints on the data generating process, another research direction is based on the assumption that independent mechanisms in nature generate causes and effects. The idea is that the shortest description of joint distribution $p(\text{cause}, \text{effect})$ can be expressed by $p(\text{cause})p(\text{effect}|\text{cause})$. Compared with factorization into $p(\text{cause})p(\text{effect}|\text{cause})$, $p(\text{effect})p(\text{cause}|\text{effect})$ has lower total complexity. Although comparing total complexity is an intuitive idea, Kolmogorov complexity and algorithmic information could be used to measure it [4].

Janzing et al. [27, 50] proposed IGCI to infer asymmetry between cause and effect through the complexity loss of distributions. The IGCI model is briefly introduced in the following section. Zhang et al. [51] proposed using a bootstrap-based approach to detect causal direction. It is based on the assumption that the parameters of the causes involved in the causality data generation process are exogenous to those of the cause to the effect. Stegle et al. [52] proposed using a probabilistic latent variable model (GPI) to distinguish between cause and effect using standard Bayesian model selection.

The above methods follow the causal sufficiency [53, 54] which assumes no unobserved common cause exists. Some methods have been proposed to deal with the unobserved common cause when telling cause from effect. Janzing et al. extended the original additive noise model [55] to identify confounding cases [56]. The proposed model is identifiable under suitable conditions. Shimizu et al. proposed a variant of the linear non-Gaussian acyclic model [57] to detect causal direction between two variables when there are latent confounding variables [58]. The above machine learning methods handle unobserved common causes and did not make it clear that which variable is confounding.

In addition to the above studies on the causal relationship between two variables, there have been several reviews. Spirtes et al. discussed the main concepts of causal discovery and introduced several models based on SEM [59]. Eberhardt [60] discussed

the fundamental assumptions of causal discovery and gave an overview of related causal discovery methods.

2.5.2 Human Computation

Human intelligence or cognition aids in solving tasks which are difficult for machine learning methods. Many research has been done to resolve difficult machine learning tasks using crowdsourcing. Duan et al. [61] collected and aggregated emotion annotations for narrative texts. Human emotion is complex and the estimation of it is difficult if only machine learning methods are used while human intelligence aids in solving the task. Wang et al. [62] used crowdsourced labels to resolve entities. They evaluated the label quality of actual crowdsourced data for entity resolution and compared the prediction accuracy of two machine learning methods. Demartini et al. [63] combined query log mining, natural language processing, and crowdsourcing for query understanding. They used crowdsourcing to understand the query in a human-machine approach. Crowdsourcing has been used in dealing with open data task as well. Oyama et al. [64] proposed using crowdsourcing to improving open data from a one-star rating to a three-star rating in Berners-Lee’s rating scheme. The improved data were machine-readable. To reduce the errors of crowdsourcing workers, a quality control mechanism was proposed. Egami et al. [65] proposed a semi-automatic construction of linked data with the causality of urban problems. They used natural language processing (NLP) to extract causality words and used crowdsourcing to filter the causality words. Finally, they designed the schema containing causality based on the extracted words [66]. Human cognition or intelligence aid in resolving complex machine learning tasks and crowdsourcing is one way to collect it.

Crowdsourcing has been applied into social data analysis. It can be used to collect explanations for patterns or trends in data. The collected explanations may be unclear and of low quality since these workers have different backgrounds or skills. Willett et al. discussed seven strategies to improve the quality of generated explanations from workers [67]. They proposed a workflow to do social data analysis using crowdsourcing. Crowdsourcing has been used for generating predictors [68, 69] as

well. The experimental results in [68, 69] showed that the predictive features of homeowner electricity usage or body mass index can be successfully collected using crowdsourcing without domain experts. Other frameworks like the crowdsourcing data analytics system (CDAS) [70] have been proposed to solve complex problems by dividing tasks into computer-oriented tasks and human-oriented tasks. As a development of applying crowdsourcing in general machine learning tasks, the idea of human-in-the-loop [71, 72] has attracted the attention of researchers as well. An interactive matching learning for health informatics was proposed by Holzinger in [72], in which they discussed application of interactive machine learning and gave several examples. Human-in-the-loop usually helped the machine learning methods to resolve complex tasks.

Chapter 3

Evaluation of Existing Methods

Many researchers have interests in the topic of causal discovery from raw data. Causal discovery also aids in many machine learning problems[9]. As far as what we have known, the existing methods for causal discovery in bivariate case have two large directions: structural equation modeling (SEM), probability and causal inference. We will give an introduction in the following parts.

3.1 Introduction

Causal relationships differ from statistical relationships, and distinguishing cause from effect is a fundamental scientific problem that has attracted the interest of many researchers. Among causal discovery problems, discovering bivariate causal relationships is a special case. Causal relationships between two variables (“x causes y” or “y causes x”) belong to the same Markov equivalence class, and the well-known independence tests and conditional independence tests cannot distinguish directed acyclic graphs in the same Markov equivalence class. We empirically evaluated the performance of three state-of-the-art models for causal discovery in the bivariate case using both simulation and real world data: the additive-noise model (ANM), the post nonlinear (PNL) model, and the information geometric causal inference (IGCI) model. The performance metrics were accuracy, area under the ROC curve, and time to make a decision. The IGCI model was the fastest in terms of algorithm efficiency

even when the dataset was large while the PNL model took the most time to make a decision. In terms of decision accuracy, the IGCI model was susceptible to noise and thus performed well only under low-noise conditions. The PNL model was the most robust to noise. Simulation experiments showed that the IGCI model was the most susceptible to “confounding” while the ANM and PNL models were able to avoid the effects of confounding to some degree.

In the work reported here, we experimentally compared the performance of three state-of-the-art causal discovery model, the additive-noise model (ANM) [25], the post nonlinear (PNL) model [26], and the information geometric causal inference (IGCI) model [27]. We used three metrics: accuracy, area under the ROC curve (AUC), and time to make a decision.

In Section 3.2, we briefly describe the three models and the implementations of them. In Section 3.3, we describe the dataset we used. In Section 3.4, we present the results and give a detailed analysis of the performances of the three models. We conclude in Section 3.5 by summarizing the strengths and weaknesses of the three models and mentioning future tasks.

3.2 Models

We used the ANM, the PNL model, and the IGCI model in the comparison experiments. The ANM and PNL model define how causality data is generated in nature through SEM. The assumption of additive noise is enlightening. The IGCI model finds the asymmetry between cause and effect through the complexity loss of distributions. The assumption of IGCI is intuitive and how well it works needs to be further researched.

3.2.1 ANM

The additive noise model of Hoyer et al. [25] is based on two assumptions: 1) the observed effects (Y) can be expressed using functional models of the cause (X) and additive noise (N) (Equation 3.1); 2) the cause and additive noise are independent. If

$f()$ is a linear function and the noise has a non-Gaussian distribution, the ANM works in the same way as LiNGAM [22]. The model is learned by performing regression in both directions and testing the independence between the assumed cause and noise (residuals) for each direction. The decision rule is to choose the direction with the less dependence as the true causal direction. The algorithm of ANM was shown in Algorithm 1 in which only one direction fits the model is assumed. The ANM cannot handle the linear Gaussian case since the data can fit the model in both directions, so the asymmetry between cause and effect disappears. Gretton et al. improved the algorithm and extended the ANM to work even in the linear Gaussian case [73]. The improved model also works more efficiently in the multivariate case.

$$Y = f(X) + N \quad (3.1)$$

Algorithm 1 Additive Noise Model

Input: Observational data x and y

Output: The p-values for directions: $x \rightarrow y$ or $y \rightarrow x$.

```

if  $x$  and  $y$  are statistically dependent then
    if  $y = f(x) + n$  fits the data then
        Calculte the corresponding residuals  $n = y - f(x)$ .
        Test the independence between  $n$  and  $x$ .
    else if  $x = f(y) + n$  fits the data then
        Calculte the corresponding residuals  $n = x - f(y)$ .
        Test the independence between  $n$  and  $y$ .
    else
        The model cannot be fitted.
    end if
end if

```

3.2.2 PNL Model

In the post-nonlinear model of Zhang et al. [48, 26], effects are nonlinear transformations of causes with some inner additive noise, followed by external nonlinear distortion (Equation 3.2). From Equation 3.2, we obtain $N = f^{-1}(Y) - g(X)$, where X and Y are the two observed variables representing cause and effect, respectively. To identify the cause and effect, a particular type of constrained nonlinear ICA [74, 48]

is performed to extract two components that are as independent as possible. The two extracted components are the assumed cause and corresponding additive noise, respectively. The algorithm of PNL was shown in Algorithm 2. The estimation process shown in Algorithm 2 assumes that only one direction fits the model. In the real world problem, there are some possibility that both directions fit the model. The identification method of the model is described elsewhere ([48], Section 4). The identifiability of the causal direction inferred by the PNL model has been proven [26]. The PNL model can identify the causal direction of data generated in accordance with the model except for the five situations described in Table 1 in [26].

$$Y = f(g(X) + N) \quad (3.2)$$

Algorithm 2 Post Nonlinear Model

Input: Observational data x_1 and x_2 .

Output: Test statistics for directions: $x_1 \rightarrow x_2$ or $x_2 \rightarrow x_1$.

```

if  $x$  and  $y$  are statistically dependent then
  if  $y_1 = x_1, y_2 = g_2(x_2) - g_1(x_1)$  fits the model. then
    Test the independence between  $y_1$  and  $y_2$ .
  else if Fit the model by  $y_1 = x_2, y_2 = g_2(x_1) - g_1(x_2)$  fits the model. then
    Test the independence between  $y_1$  and  $y_2$ .
  else
    The model cannot be fitted.
  end if
end if

```

3.2.3 IGCI

The IGCI model [50, 27] is based on the hypothesis that if “X causes Y,” the marginal distribution $p(x)$ and the conditional distribution $p(y|x)$ are independent in a particular way. The IGCI model gives an information-theoretic view of additive noise and defines independence by using orthogonality. With ANM [25], if there is no additive noise, inference is impossible while it is possible with the IGCI model.

The IGCI model determines the causal direction on the basis of complexity loss. According to IGCI, the choice of $p(x)$ is independent of the choice of f for the re-

relationship $y = f(x) + n$. Let ν_x and ν_y be the reference distributions⁴ for X and Y .

$$D(P_x || \nu_x) := \int \log \frac{P(x)}{\nu(x)} P(x) dx$$

is the KL-distance between P_x and ν_x . $D(P_x || \nu_x)$ works as a feature of the complexity of the distribution. The complexity loss from X to Y is given by

$$V_{X \rightarrow Y} := D(P_x || \nu_x) - D(P_y || \nu_y)$$

.

The decision rule of the IGCi model is that, if $V_{X \rightarrow Y} < 0$, infer “ X causes Y ,” and if $V_{X \rightarrow Y} > 0$, infer “ Y causes X .” This rule is rather theoretical. An applicable and explicit form for the reference measure is entropy-based IGCi or slope-based IGCi.

1. Entropy-based IGCi:

$$\hat{S}(P_X) := \psi(m) - \psi(1) + \frac{1}{m-1} \sum_{i=1}^{m-1} \log |x_{i+1} - x_i| \quad (3.3)$$

where $\psi()$ is the digamma function⁵, and m is the number of data points.

$$\hat{V}_{X \rightarrow Y} := \hat{S}(P_Y) - \hat{S}(P_X) = -\hat{V}_{Y \rightarrow X} \quad (3.4)$$

2. Slope-based IGCi:

$$\hat{V}_{X \rightarrow Y} := \frac{1}{m-1} \sum_{i=1}^{m-1} \log \left| \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \right| \quad (3.5)$$

These explicit forms are simpler, and we can see that the two calculation methods coincide. The estimation process of IGCi was described in Algorithms 3 and 4. The calculation does not take much time even when dealing with a large amount of

⁴Reference distributions are used to measure the complexity of P_x and P_y . In [27], non-informative distributions like uniform and Gaussian ones are recommended.

⁵https://en.wikipedia.org/wiki/Digamma_function

data. However, the IGC model assumes that the causal process is noiseless and may perform poorly under high-noise conditions. We discuss the performance of the three models in Section 3.4.

Algorithm 3 Entropy-based IGC

Input: Observational data X and Y .

Output: The estimated result “ X causes Y ” or “ Y causes X ”.

Calculate the entropy estimator $\hat{S}(P_X)$ and $\hat{S}(P_Y)$.

Calculate the estimator for $V_{X \rightarrow Y}$ by $\hat{V}_{X \rightarrow Y}$.

if $\hat{V}_{X \rightarrow Y} < 0$ **then**

Infer that X causes Y .

else

Infer that Y causes X .

end if

Algorithm 4 Slope-based IGC

Input: Observational data X and Y .

Output: The estimated result “ X causes Y ” or “ Y causes X ”.

Calculate the estimator $V_{X \rightarrow Y}$ by $\hat{V}_{X \rightarrow Y} := \frac{1}{m-1} \sum_{i=1}^{m-1} \log \left| \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \right|$.

if $\hat{V}_{X \rightarrow Y} < 0$ **then**

Infer that X causes Y .

else

Infer that Y causes X .

end if

3.2.4 Implementation

We implemented the three models following the original work [25, 48, 27]. A brief introduction is given below.

ANM Using the reported experimental settings [25], we performed Gaussian processes for machine learning regression [75, 76, 77]. We then used the Hilbert-Schmidt Independence Criterion (HSIC) [78] to test the independence between the assumed cause and residuals. The dataset used had been labeled with the true causal direction for each pair with no cases of independence or feedback. Using the decision rule of ANM, we determined that the direction with the greater independence was the true causal direction.

PNL Model We used a particular type of constrained nonlinear ICA to extract the two components that would be the cause and noise if the model had been learned in the correct direction. The nonlinearities of $g()$ and $f^{-1}()$ in Equation 3.2 were modeled using multi-layer perceptrons. By minimizing the mutual information between the two output components, we made the output as independent as possible. After extracting two independent components, we tested their independence by using the HSIC [78, 79]. Finally, in the same way as for the ANM, we determined that the direction with the greater independence was the correct one.

IGCI (entropy, uniform) Compared with the first two models, the implementation of the IGCI (entropy, uniform) model was simpler. We used reported equations (3.3, 3.4) to calculate $\hat{V}_{X \rightarrow Y}$ and $\hat{V}_{Y \rightarrow X}$ and determined that the direction in which entropy decreased was the correct direction. If $\hat{V}_{X \rightarrow Y} < 0$, the inferred causal direction was “X causes Y”; otherwise, it was “Y causes X.” For the IGCI model, the data should be normalized before calculating $\hat{V}_{X \rightarrow Y}$ and $\hat{V}_{Y \rightarrow X}$. In accordance with the reported experimental results, we used the uniform distribution as the reference distribution because of its good performance. For the repetitive data in the dataset, we set $\log 0 = 0$.

IGCI (slope, uniform) The implementation of the IGCI (slope, uniform) model was similar to that of the IGCI (entropy, uniform) one. We used Equation 3.5 to calculate $\hat{V}_{X \rightarrow Y}$ and $\hat{V}_{Y \rightarrow X}$ and determined that the direction with a negative value was the correct one. For the same reason as above, we normalized the data to $[0,1]$ before calculating $\hat{V}_{X \rightarrow Y}$ and $\hat{V}_{Y \rightarrow X}$. To make Equation 3.5 meaningful, we filtered out the repetitive data.

3.3 Dataset

We used the CauseEffectPairs (CEP) [80] dataset, which contains 97 pairs of real world causal variables with the cause and effect labeled for each pair. The data set is publicly available online [80]. Some of the data were collected from the UCI Machine

Table 3.1: Dataset Used in CEP.

DWD climate data	Concrete compressive strength
Abalone data	Liver disorder dataset
Census-Income data	Pima Indians Diabetes Database
Auto Mpg data	UNdata
Cardiac arrhythmia database	Time series data

Learning Repository [81]. The data comes from various fields, including geography, biology, physics, and economics. The dataset also contains time series data. Some of the datasets contained in CEP are listed in Table 3.1. Most of the data are noisy. An appendix in [82] contains a detailed description of each pair of variables.

We used 91 of the pairs in our experiments since some of the data (e.g., pair0052) contains multi-dimensional variables⁶. The 91 pairs are listed in Table A.1 in Appendix A. Some contain the same variables collected in different countries or at different times⁷. The scatter plots of data are shown in Figure 3-1. The variables range in size from 126 to 16,382⁸. The variety of data types in the CEP dataset makes causal analysis using real world data challenging.

3.4 Results

Here, we first compare model accuracy for different decision rates⁹. We changed the threshold and calculated the corresponding decision rate and accuracy for each model. The accuracy of the models for different decision rates has been compared in the original study [50]. Compared with [50], we used more real world data in our experiments. Besides, how accuracy changed under different decision rates was showed. Our results are consistent with those shown of Mooij et al. [83]. The performance of the models for different decision rates is discussed in Subsection 3.4.1.

⁶The three models we evaluated cannot deal with multi-dimensional data.

⁷Country and time information is not included in the table.

⁸To avoid overfitting, we limited the size to 500 or less.

⁹Since all three models have two outputs, e.g., $\hat{V}_{X \rightarrow Y}$, $\hat{V}_{Y \rightarrow X}$ corresponding to the two possible causal directions, we set thresholds based on the absolute difference between them $|\hat{V}_{X \rightarrow Y} - \hat{V}_{Y \rightarrow X}|$ for use in deciding each decision rate and model accuracy.

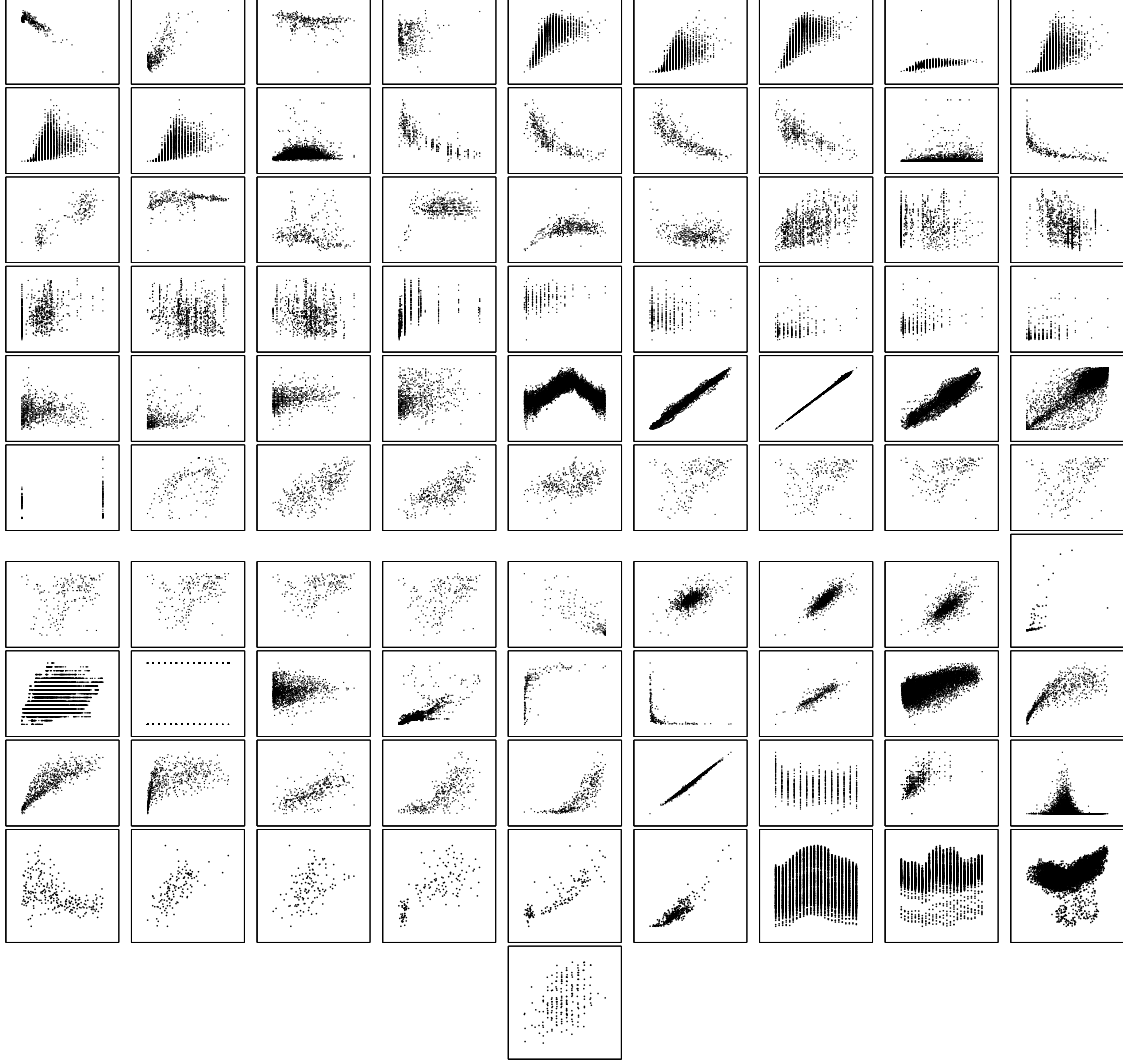


Figure 3-1: Scatter Plot of Data in the Experiments

Since causal discovery models in the bivariate case make a decision between two choices, we can regard these models as binary classifiers and evaluate them using AUC. We give the results for the entire (undivided) dataset in the following. Besides, we compare model efficiency by using the average time needed to make a decision. This is described in Subsection 3.4.3.

3.4.1 Accuracy For Different Decision Rates

We calculated the accuracy of each model for different decision rates using Equations 3.6 and 3.7. The results are plotted in (Figure 3-2). The decision rate changed when

the threshold was changed. The larger the threshold, the more stringent the decision rule. In an ideal situation, accuracy decreases as the decision rate increases, with the starting point at 1.0. However, the results with real world data were not perfect because the data were noisy.

As shown in Figure 3-2, the accuracy started from 1.0 for the ANM and IGCI and from 0.0 for the PNL model. This means that the PNL model made an incorrect decision when it had the highest confidence. Although the accuracies of the IGCI models started from 1.0, they dropped sharply when the decision rate was between 0 and 0.2. The reasons for this are discussed in detail in Subsection 3.4.4. After reaching a minimum, the accuracies increased almost continuously and stabilized. The accuracy of the ANM was more stable than those of the other models. When all decisions had been made, the model accuracies were ranked $IGCI > ANM > PNL$.

$$\text{Decision Rate} := \frac{N_{\text{Decision}}}{N_{\text{Data}}} \quad (3.6)$$

$$\text{Accuracy} := \frac{N_{\text{True Decision}}}{N_{\text{Decision}}} \quad (3.7)$$

3.4.2 Area Under ROC Curve (AUC)

Besides calculating the accuracy of the three models for different decision rates, we used the AUC to evaluate their performance. Some of the experimental results were presented in our previous work [84], for which the dataset was divided into two groups: inferred as “X causes Y” and inferred as “Y causes X.” Here we present updated experimental results for the entire dataset.

The following steps were taken to get the experimental results:

1. Set X as the cause and Y as the effect in the input data.
2. Set $V_{X \rightarrow Y}$ and $V_{Y \rightarrow X}$ to be the outputs.
3. Calculate the absolute value of the difference between $V_{X \rightarrow Y}$ and $V_{Y \rightarrow X}$ (Equation 3.8) and map V_{diff} to $[0,1]$.

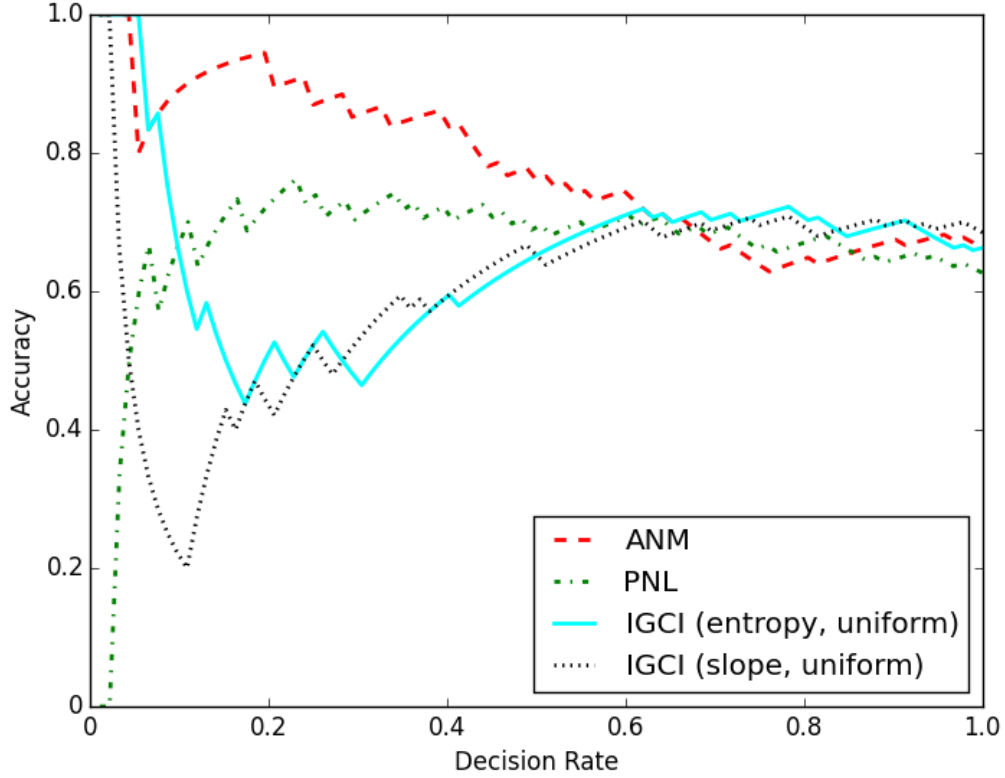


Figure 3-2: Accuracy of Three Models for Different Decision Rates. Decision rate changed when threshold was changed: the larger the threshold, the smaller the decision rate. In an ideal case, the accuracy of each model should improve with a decrease in the decision rate.

4. Assign a positive label to the pairs inferred as “X causes Y” and a negative one to the pairs inferred as “Y causes X.”
5. Use V_{diff} and the labels assigned in step 4 to calculate the true positive rate (TPR) and false positive rate (FPR) for different thresholds.
6. Plot the ROC curve and calculate the corresponding AUC value.

$$V_{\text{diff}} = |V_{X \rightarrow Y} - V_{Y \rightarrow X}| \quad (3.8)$$

In step 1), instead of dividing the data into two groups as done previously [84], we set the input matrix so that the first column was the cause and the second column

was the effect. Then, if the inferred causal direction for a pair was “X causes Y,” a positive label was assigned to that pair; otherwise, a negative label was assigned.

In step 3), we used the absolute value of the difference between $V_{X \rightarrow Y}$ and $V_{Y \rightarrow X}$ as the “confidence” of the model when making a decision. The larger the V_{diff} , the greater the confidence. We did not use division because, if one of $V_{X \rightarrow Y}$ and $V_{Y \rightarrow X}$ was very small, the division result would be very large. We mapped V_{diff} to $[0,1]$ to make the calculation more convenient. In this way, V_{diff} could be used in the same way as the output of a binary classifier. For causal discovery, the larger the V_{diff} , the greater the confidence in the decision. At the same time, more punishment should be given when the decision is incorrect.

In step 4), we labeled the data in accordance with the inferred causal direction. Since the correct label for all the pairs was “X causes Y,” if the inferred result for a pair was “Y causes X,” it was assigned a negative label.

In step 5), we used the normalized V_{diff} and the label assigned in step 4) to calculate TPR and FPR for different thresholds. We plotted TPR and FPR to get the ROC curve and calculated the corresponding AUC value.

The ROC results are plotted in Figure 3-3. The corresponding AUC values are shown in Table 3.2. Different from the results shown in Figure 3-2, both IGC models performed poorly in terms of AUC. The AUC values for IGC were smaller than 0.5, which means their performances were even worse than that of a random classifier. However, as described in Subsection 3.4.1, when we used different decision rates, the IGC models had better performance.

We checked the decisions made by the IGC models and found that they made several incorrect decisions when the threshold was large. Such decisions with a large threshold are punished severely when using the AUC metric. As shown in Figure 3-2, although the accuracies of the IGC models started from 1.0, they dropped sharply when the decision rate was between 0 and 0.2. An incorrect decision with a low decision rate was not punished much when evaluating accuracy for different decision rates. However, for the AUC, an incorrect decision when the threshold was large was punished more than when the threshold was small. For these reasons, the starting

Table 3.2: AUC of Three Models.

Model	AUC
ANM	0.6263
PNL	0.6120
IGCI (entropy)	0.4935
IGCI (slope)	0.4271

Table 3.3: Time to Make A Decision

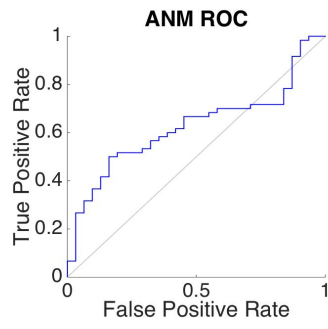
Model	Time (s)	
	Mean	Variance
ANM	10.7	7.4
PNL	60.0	20.3
IGCI (entropy)	0.0014	0.0019
IGCI (slope)	0.0014	0.0017

point of the ROC curve for the IGCI models in Figure 3-3 was shifted to the right, making the AUC less than 0.5. In Subsection 3.4.4, we will discuss why the IGCI models failed.

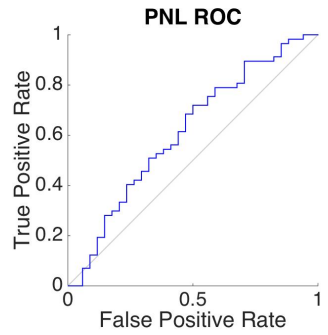
3.4.3 Algorithm Efficiency

Besides comparing the accuracy and ROC of the three models, we also compared the average time for the algorithm to make a decision¹⁰. We performed the experiment on the MATLAB platform with an Intel Core i7-4770 3.40 GHz CPU and 8 GB memory. From Table 3.3, we can see that the IGCI models were the most efficient one while the PNL model was the least efficient. The ANM was in the middle. The longer time to make a decision for the PNL model was due to the estimation procedure of f^{-1} and g in Equation 3.2.

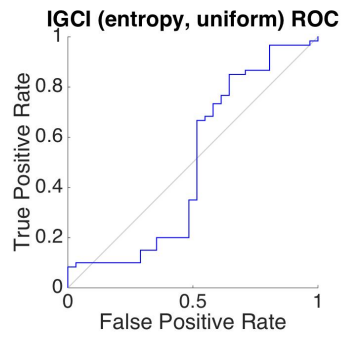
¹⁰Compared to our previous report [84], we reduced the program output so that the PNL model works faster. We have updated the results for time to make a decision for the PNL model accordingly.



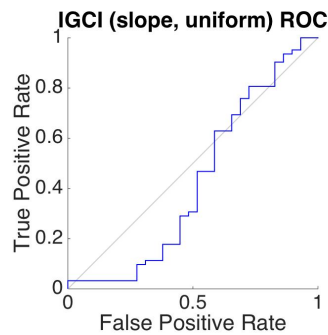
(a)



(b)



(c)



(d)

Figure 3-3: ROC of Three Models. Four graphs are shown because IGCI has two explicit forms.

3.4.4 Typical Examples of Model Failure

Discretization

In Subsection 3.4.1, we explained that the PNL model gives an incorrect decision when the threshold is set the highest, i.e., the accuracy for different decision rates starts from 0.0. We investigated the reason for the PNL model making an incorrect decision when the threshold was the highest. We found that it happens due to the discretization of data. A scatter plot for a pair of variables (pair0070) is shown in Figure 3-4. The data has two variables. Variable x_1 is a value between 0 and 14 reflecting the features of an artificial face. It is used to decide whether the face is that of a man or a woman. A value of 0 means very female, and a value of 14 means very male. Variable x_2 is a value of 0 (female) or 1 (male) reflecting the gender decision. Since variable x_2 has only two values, no matter what nonlinear transformation is made to x_2 , the transformation result is two discretized values. According to the mechanism of the PNL model, x_2 with two discretized values is inferred to be the cause since the independency is larger if x_2 is the cause. In fact, for this pair of variables, all three models made incorrect decisions in our experiments. For the ANM, the discretization of data makes regression analysis difficult. And the poor regression results negatively affect testing of the independence between the assumed cause and residuals using HSIC. For IGCI, to make Equations 3.3 and 3.5 meaningful, the repetitive data has to be filtered out. This means that only a few data points are actually used in the final IGCI calculation.

Noisiness

In Subsection 3.4.1, we showed that IGCI had good performance in general. However, its accuracy dropped sharply when the decision rate was between 0 and 0.2. The reason for this is that it made incorrect decisions with high confidence when dealing with pair0056-pair00

63. These eight pairs contain much noise, which degraded model performance. Moreover, there were outliers for the eight pairs, which greatly affected the decision result.

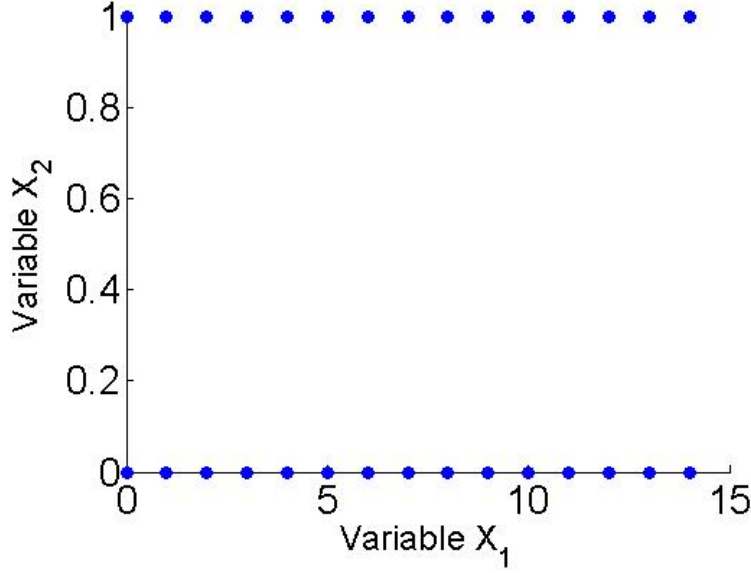


Figure 3-4: Example of Discretized Data in CEP Dataset. Variable x_1 is between 0 (very female) and 14 (very male), reflecting features of artificial face. Variable x_2 is 0 (female) or 1 (male), reflecting gender decision.

A scatter plot for one example pair is shown in Figure 3-5 (pair0056). It shows that the two variables have relatively small correlation and that there are outliers in the data. The calculation method used in IGCi is such that these kinds of outliers affect the inference result more than the other data points. The incorrect decisions that IGCi made about pair0056-pair0063 account for the small AUC value for the IGCi models given in Subsection 3.4.2. The ANM also made an incorrect inference about these pairs. This is because noise and outliers make overfitting more likely to occur for these variables. For these noisy pairs, the PNL model had the best performance.

3.4.5 Response to Spurious Correlation Caused by “Confounding”

A causal relationship differs from a statistical one, and a statistical relationship is usually not enough to explain a causal one. Even if we observe that two variables are highly correlated, we cannot say that they have a causal relationship. As shown in Table A.1 in Appendix A, the causal direction of the variables in CEP is obvious by

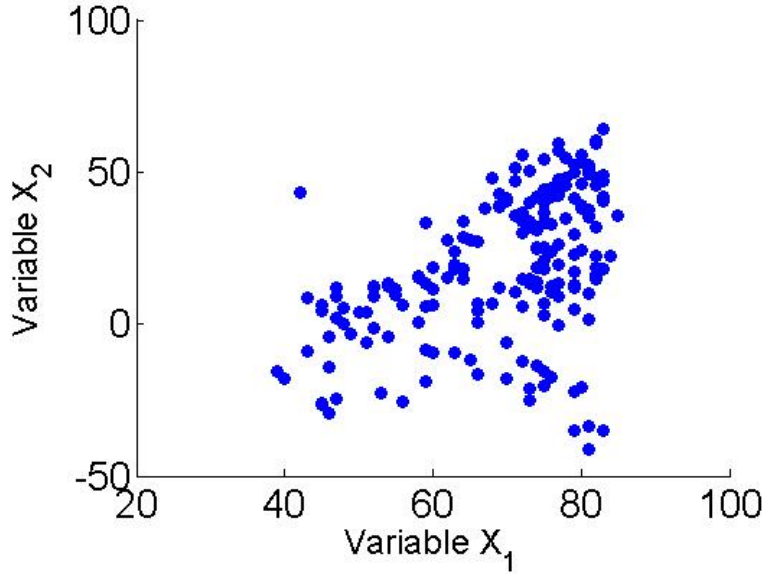


Figure 3-5: Example Scatter Plot of Noisy Data in CEP. Variable x_1 is female life expectancy at birth for 2000-2005. Variable x_2 is latitude of birth country’s capital (data for China, Russia, and Canada were removed).

common sense. The dataset has been collected for evaluating causal discovery models, and the causal direction of most pairs is obvious from knowledge. However, the relationships of variables that are of general interest in the real world are usually more controversial, e.g., smoking and lung cancer, for which the existence of confounding is usually a bone of contention. A good causal discovery model for two variables should have the ability to avoid the effect of “confounding” to some degree. To test how the ANM, the PNL model, and IGCI perform when dealing with spurious correlation caused by confounding, we first simulated the “common cause” case shown in Figure 1-1. We controlled the data generating process to simulate different degrees of confounding. In addition to simulation, we used real world data from the CEP to evaluate model performance when dealing with real world data.

Simulation

We conducted simulation experiments of the “common cause” confounding case shown in Figure 1-1. We generated data using two equations: $x = a * z^3 + b * n_1$ and $y = a * z + b * n_2$, where $z, n_1, n_2 \in U(0, 1)$. We used the quotient a/b to control

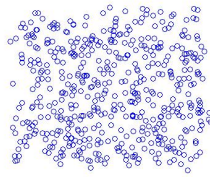
the degree of confounding. There was no direct causal relationship between variables x and y except for the various degrees of confounding. Scatter plots of the data generated using five different quotients are shown in Figure 3-6.

We used the generated data to test the performance of the three models. We performed ten random experiments for each quotient and calculated the average of the inferred results. The experimental results are shown in Figure 3-7, Figure 3-8, and Table 3.4. For IGCI, when the degree of confounding was low, the mean of the estimated results¹¹ was almost zero. Each estimated result got a positive or negative value randomly. As the degree of confounding increased, the estimated results tended to approve “ y causes x .” For the PNL model, the independence assumptions for both directions were accepted when $a/b = 0.1, 1, 10$ ($\alpha = 0.01$) and the means of the test statistics (Equation 4 in [79]) were almost equal. When $a/b = 100$, the PNL model rejected the independence assumption for direction “ x causes y ” and accepted that “ y causes x .” When a/b was even larger, the independence assumptions were rejected for both directions, especially that of “ x causes y .” For the ANM, when the degree of confounding was low, the independence assumptions were accepted for both directions. When the degree was high, the independence assumptions were rejected for both directions, especially “ x causes y .” From these results, we can see that the IGCI is the most susceptible of the three models tested to confounding while the PNL model and the ANM can avoid the effect of confounding to a certain extent.

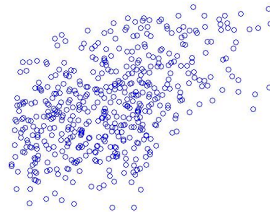
Real World Data

In addition to the simulation experiments described above, we conducted experiments using real world data. The generation of real world data can be very complex, which increases the difficulty of the causal discovery task. Here we describe the “common cause” and “selection bias” cases (Figure 1-1).

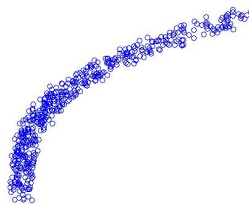
¹¹We used the difference between $\hat{V}_{X \rightarrow Y}$ and $\hat{V}_{Y \rightarrow X}$ (Equations 3.4, 3.5) as the estimated result. For IGCI, $\hat{V}_{X \rightarrow Y}$ is the opposite of $\hat{V}_{Y \rightarrow X}$ if there is no repetitive data. Thus, we can infer that the correct causal direction is $X \rightarrow Y$ if the estimated result is negative and that $Y \rightarrow X$ is correct if the estimated result is positive.



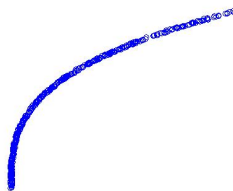
(a) a/b : 0.1



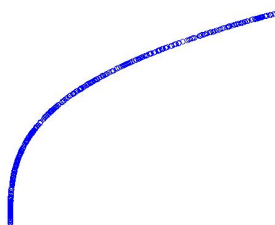
(b) a/b : 1



(c) a/b : 10

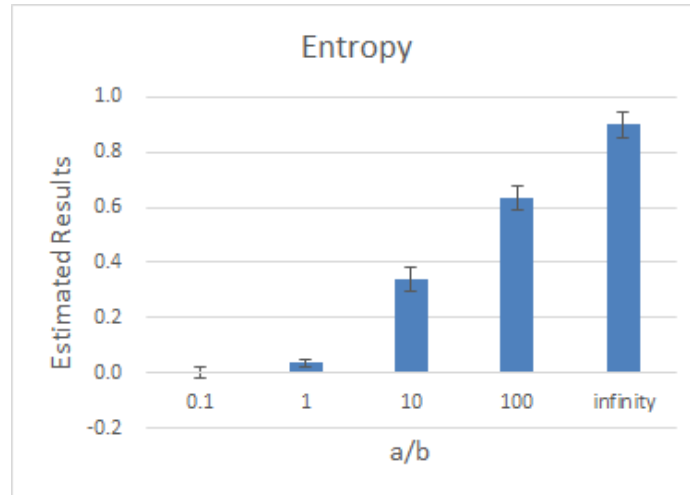


(d) a/b : 100

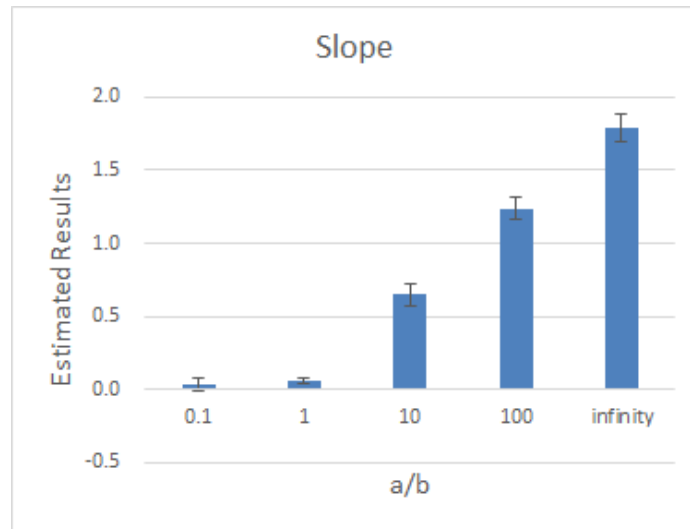


(e) a/b : 1000

Figure 3-6: Scatter Plots of Generated Data.



(a)



(b)

Figure 3-7: Estimated Results of IGCI for Generated Data. Inference was “y causes x” when estimated result was larger than 0.

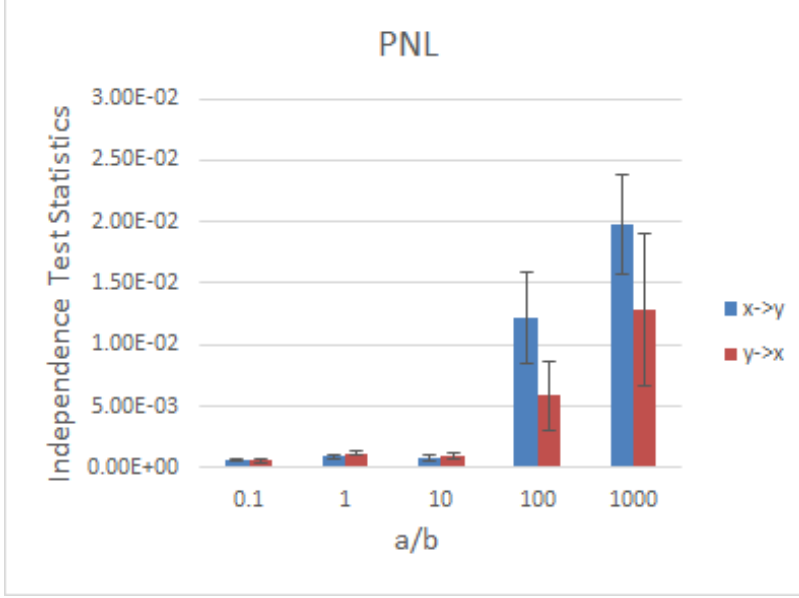


Figure 3-8: Test Statistics for PNL Model for Generated Data.

Table 3.4: P-value for ANM for Two Directions

a/b	$x \rightarrow y$ assumed	$y \rightarrow x$ assumed
0.1	0.7044	0.7109
1	0.2837	0.1038
10	1.03E-49	1.47E-16
100	4.32E-103	7.24E-18
1000	4.46E-179	7.70E-21

Common Cause Case Although data for “common cause” are not included in the CEP dataset, some CEP pairs contain the same cause, as shown in Appendix A. We combined data containing the same cause to obtain pairs of variables, such as “length and diameter”¹². A scatter plot of the results is shown in Figure 3-9 for the pair “length, diameter.” For the ANM, the p-value for the forward direction, “length causes diameter,” was 8.28×10^{-5} while that for the backward direction was 1.05×10^{-3} . For the PNL model, the independence test statistic was 1.11×10^{-3} for the forward direction and 9.50×10^{-4} for the backward direction. For IGCI, the

¹²The pair “length, diameter” was created from “cause: rings (abalone), effect: length” and “cause: rings (abalone), effect: diameter” using data for abalone [81].

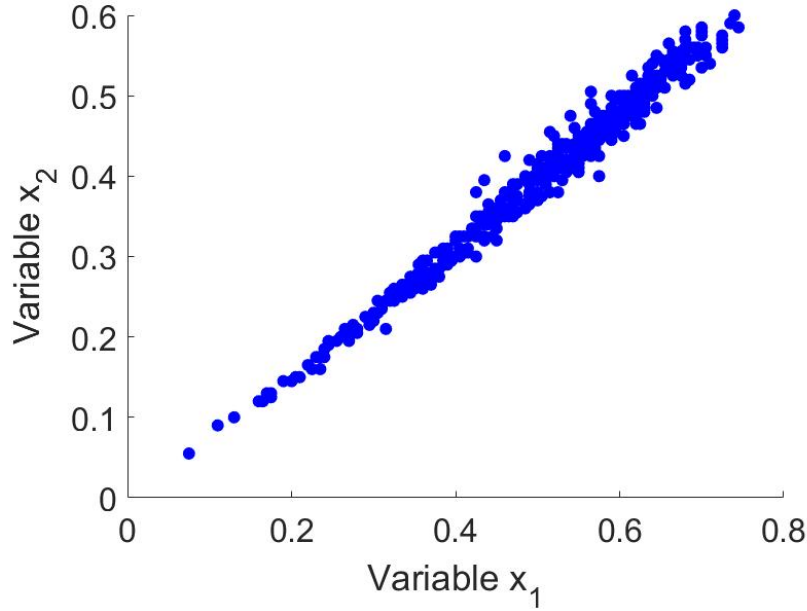


Figure 3-9: Results for “Common Cause” Case with Real World Data. Variable x_1 : “length”; variable x_2 : “diameter”.

result estimated by calculating the entropy was 0.2197 while the other was 0.056. For this pair, although the tendency was not strong, the three models tended to approve “diameter is the cause of length.”

Selection Bias Case Although there is no causal relationship between X and Y in the selection bias case, independence does not hold between X and Y when conditioned on variable Z. This is the well-known Berkson paradox¹³. We used the variables “altitude and longitude” (Figure 3-10) contained in CEP to test how the three models perform when dealing with the “selection bias” case. The variables were obtained by combining “cause: altitude, effect: temperature” and “cause: longitude, effect: temperature.” The data came from 349 weather stations in Germany [80]. A scatter plot of the results is shown in Figure 3-10. For the ANM, the p-value for the forward direction, “altitude causes longitude,” was 4.21×10^{-2} while that for the backward direction was 8.73×10^{-2} . The independence assumptions were accepted for both directions although “longitude causes altitude” was favored. For the PNL model, the

¹³https://en.wikipedia.org/wiki/Berkson's_paradox.

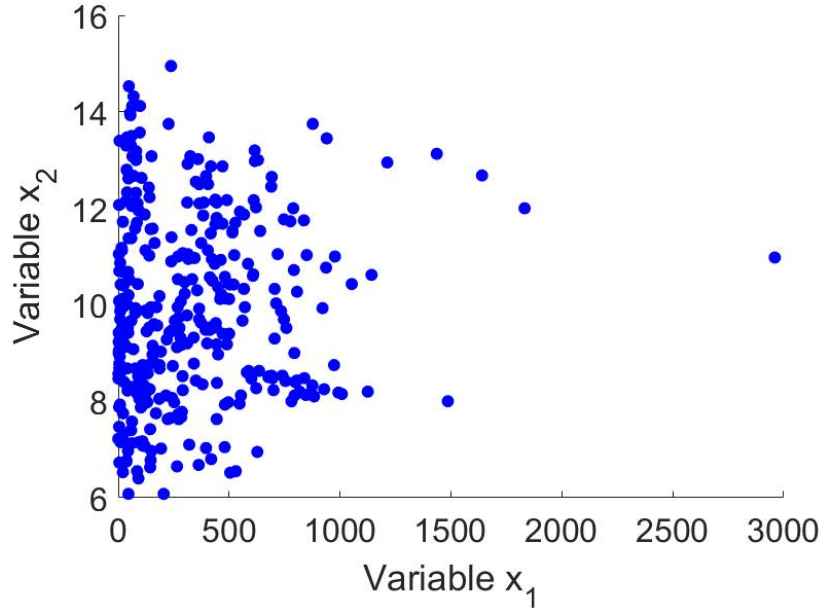


Figure 3-10: Results for “Selection Bias” Case with Real World Data. Variable x_1 : “altitude”; variable x_2 : “longitude”.

test statistic was 2.46×10^{-3} for the forward direction and 3.30×10^{-3} for the backward direction. The independence assumptions were accepted for both directions, and the independence test results were similar. For IGCI, the result estimated by calculating the entropy was 1.2742 and that estimated by calculating the slope was 1.9032. Both results were positive, and the estimated causal direction was “longitude causes altitude.” Although there should be no causal relationship between “altitude” and “longitude,” it was hard for the three models to determine that from the limited amount of observational data available.

For most cases of causal discovery in the real world, only limited observational data can be obtained, and in some cases the data are incomplete. Moreover, for the case of two variables, the causal sufficiency assumption [85] is easily violated if there is an unobserved common cause. The limited amount of data and unobserved confounders make causal discovery in the bivariate case challenging.

3.5 Summary

We compared three state-of-the-art models (ANM, PNL model, IGCI) for causal discovery in the binary case using simulation and real world data. Testing using different decision rates showed that the three models had similar accuracies when all the decisions were made. To check whether the decisions made were reasonable, we used a binary classifier metric: the area under the ROC curve (AUC). The IGCI model had a small AUC value because it made several incorrect decisions when the threshold was high. Compared with those of the other models, the accuracy of the ANM was relatively stable. A comparison of the time to make a decision showed that IGCI was the fastest even when the dataset was large. The PNL model took the most time to make a decision.

Of the three models, the IGCI one had the best performance when there was little noise and the data were not discretized much. Improving the performance of the IGCI model when there is much noise and how to deal with discretized data are future tasks. Although the performance of the ANM was relatively stable, overfitting should be avoided for ANM because it will negatively affect the subsequent independence test. Of the three models, the PNL model is the most generalized one as it takes into account the nonlinear effect of causes, inner additive noise, and external sensor distortion. However, estimation procedure of $g()$ and $f^{-1}()$ is a lengthy procedure. Finally, testing the responses of the models to “confounding” showed that the ANM and the PNL model can avoid the effect of “confounding” to some degree while IGCI is the most susceptible to confounding.

Chapter 4

Common Cause Identification

4.1 Introduction

Confounding may exist in causal analysis of real world data. Stratification or matching can be used to deal with confounding in causal analysis [86, 87]. Stratification was used to fix the value of confounders in certain subgroups through which the effect of confounders can be reduced or vanishes [88]. Matching can be viewed as a special case of stratification and is frequently used in case-control studies [89]. Stratification or matching variable should be carefully chosen since bad choice of it would induce extra errors [89]. Some causal relationships have been analyzed using stratification. [29] analyzed the causal relationship between energy consumption and GDP growth by dividing the data into four income categories. Their experimental results showed that causal relationships differed among the four categories. Coondoo and Dinda analyzed the causal relationship between income and emissions by dividing countries into specific groups [90]. The experimental results showed that the causal relationships varied among the groups as well. The above experiments showed that the effects of confounders should be considered when doing causal analysis since the analysis result may change after considering certain confounder. A common cause or selection bias (Figure 1-1) can induce spurious correlation between two variables. If the data was unbiased, the possible confounding comes from unconsidered common cause. [91] discussed the responses of several causal discovery models to confounding which

showed that some models can avoid the effect of confounding to a certain degree and some cannot. To correctly analyze causal relationship between two variables, finding and testing possible observable confounders is needed. Conditional independence is a necessary condition for a common cause. To test whether a third variable Z is a possible confounder for two variables X and Y , a conditional independence test can be conducted. [92] proposed a framework for crowd-based causal analysis of open data in which conditional independence test was used as an initial step to filter unrelated words. The remained related words were then used to further learn causal relationship between variables. To filter unrelated words, methods like conditional independence test are needed to identify possible common causes. However, many methods for conditional independence test have certain assumptions of data distributions, function forms or additive noise. When these assumptions are not satisfied, the conditional independence test may fail. Besides, causal relationship may be mixed with common causes in the real world problem. In this case, conditional independence test will fail to tell a possible common cause. Here we propose using intrinsic dimension estimation as an alternative to detect a possible common cause for two variables. The proposed method is a necessary condition and can be used to filter unrelated words in the candidates for confounders. It can be combined with other causal discovery methods to realize causal relationship analysis while considering specific confounders. The proposed method generally worked well for different types of added noise. In particular, it worked better than a kernel-based conditional independence test for Poisson noise. Besides, the proposed method can infer a possible common cause when it is mixed with causal relationship.

In Section 4.2, we introduce intrinsic dimension estimation and its application to causal discovery. In Section 4.3, we first present the simulation results for different types of function forms. We then present the noise test results. Besides, we present the results of how the type of distribution affects the estimated dimension. Finally, we present the simulation results of mixed pattern. In Section 4.4, we present test results for real world data. In Section 4.5, we summarize the key points.

4.2 Intrinsic Dimension Estimation

We propose using intrinsic dimension estimation to test possible common causes. The proposed method was inspired by the IGC causal discovery method [93] in which the asymmetries between cause and effect are identified by analyzing the complexity loss of their distributions. We studied the manifold of data in ordinary three-dimensional Euclidean space for the common cause case, as shown in Figure 1-1. After introducing the intrinsic dimension estimation method we used, we describe its application to causal discovery.

The intrinsic dimension can be interpreted to mean the smallest number of variables needed to express the entire dataset; it is usually less than the total number of observed variables [94] and can be used to characterize fractals. Estimators of the intrinsic dimension are related to fractal geometry, e.g., the correlation dimension. We use the correlation dimension estimator introduced by Grassberger and Procaccia [95]. There are several methods to estimate the intrinsic dimension. We chose correlation dimension estimation because it is easy to be calculated and the estimated result can be proven theoretically. The correlation dimension is based on the assumption that the number of data points in a hypersphere with radius ε is proportional to the dimension ε^d [94]. The number of data points lying in a hypersphere within radius ε is given by

$$C(\varepsilon) = \lim_{N \rightarrow \infty} \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N H(\varepsilon - \|x_i - x_j\|), \quad (4.1)$$

where $H(x)$ is the Heaviside step function. Dimension d is estimated as ε approaches zero:

$$d = \lim_{\varepsilon \rightarrow 0} \frac{\log C(\varepsilon)}{\log \varepsilon}. \quad (4.2)$$

Since calculating the limit is impossible in practice, an approximate value is estimated in accordance with the L'Hospital rule, where ε_1 and ε_2 are set between the minimal and maximal pairwise distances measured in the data set [94]. An example

of pseudocode was shown in Algorithm 5.

$$d \approx \frac{\log C(\varepsilon_2) - \log C(\varepsilon_1)}{\log \varepsilon_2 - \log \varepsilon_1} \quad (4.3)$$

Algorithm 5 Intrinsic Dimension Estimation

Input: Normalized input data D without duplicated data points

Output: Estimated intrinsic dimension d

```

[m, n] ← size(D)
distance ← list [ ] [ ]
while i < m do
  if i ≠ m then
    j = i + 1
    l ← zeros(m)
    while j ≤ m do
      Δx = ||xi - xj||
      l[j] ← Δx
      j = j + 1
    end while
  end if
  distance.append(l)
  i = i + 1
end while
v ← no zero values of distance
ε1 ← median(v), ε2 ← max(v)
s1 ← length(v < ε1), s2 ← length(v < ε2)
Cr1 ← (2/(n * (n - 1))) * s1
Cr2 ← (2/(n * (n - 1))) * s2
d ← (log(Cr2) - log(Cr1))/(log(ε2) - log(ε1))

```

4.2.1 Observable Common Cause Identification

The estimation results for the common cause case differ from that for the selection bias case (Table 4.1). For the common cause case, the data are distributed in a one-dimensional manifold (a line), and the estimated intrinsic dimension is 1 ideally. For the selection bias case, the data are distributed in a two-dimensional manifold (a surface), and the estimated intrinsic dimension is 2 ideally.

For the common cause case, suppose that $X_i = \alpha m_i, Y_i = \beta m_i, Z_i = m_i$. The distance between two points x_i and x_j with coordinates (X_i, Y_i, Z_i) and (X_j, Y_j, Z_j)

respectively can be calculated using

$$\begin{aligned}
\|x_i - x_j\| &= \sqrt{\alpha^2 (m_i - m_j)^2 + \beta^2 (m_i - m_j)^2 + (m_i - m_j)^2} \\
&= \sqrt{\alpha^2 + \beta^2 + 1} |m_i - m_j| \\
&= \sqrt{\alpha^2 + \beta^2 + 1} \Delta m \propto \Delta m.
\end{aligned} \tag{4.4}$$

Suppose that the data points exist in a line with unit length. The distance between two adjacency disappoints would be $\frac{1}{N}$. Thus, the convergence rate of ε is proportional to that of $\frac{1}{N}$. For the common cause case, a suitable value proportional to $\frac{1}{N}$ can be found for ε which makes that $H(\varepsilon - \|x_i - x_j\|)$ equals one only when x_j is the next nearest point ($\Delta m \rightarrow 0$) except the last point. This means that $\sum_{j=i+1}^N H(\varepsilon - \|x_i - x_j\|) = 1 + \underbrace{0 + \dots + 0}_{N-i-1}$. We thus get

$$\begin{aligned}
&\log C \left(\varepsilon \rightarrow 0, \varepsilon \propto \frac{1}{N} \right) = \\
&\lim_{N \rightarrow \infty} \log \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \left(1 + \underbrace{0 + \dots + 0}_{N-i-1} \right).
\end{aligned} \tag{4.5}$$

The estimated dimension d can then be derived:

$$\begin{aligned}
d &= \lim_{\varepsilon \rightarrow 0, \varepsilon \propto 1/N} \log C(\varepsilon) / \log \varepsilon \\
&= \lim_{\varepsilon \rightarrow 0, \varepsilon \propto 1/N} \left[\lim_{N \rightarrow \infty} \log (N-1) / N(N-1) / \log \varepsilon \right] \\
&= \lim_{\varepsilon \rightarrow 0, \varepsilon \propto 1/N} \left[\lim_{N \rightarrow \infty} \log (1/N) / \log \varepsilon \right]
\end{aligned} \tag{4.6}$$

which is approximately equal to 1.

For the selection bias case, suppose $X_i = m_i, Y_i = n_i, Z_i = \alpha m_i + \beta n_i$. The distance between x_i and x_j is given by

$$\begin{aligned}
\|x_i - x_j\| &= \sqrt{(\alpha^2 + 1) \Delta m^2 + (\beta^2 + 1) \Delta n^2 + 2\alpha\beta \Delta m \Delta n} \\
&\propto \sqrt{\Delta m^2 + \Delta n^2 + \Delta m \Delta n},
\end{aligned} \tag{4.7}$$

where $\Delta m = m_i - m_j$, $\Delta n = n_i - n_j$. In this case, the value of $\|x_i - x_j\|$ is affected by the values of Δm and Δn . The x_j could be the next nearest point in two directions: $\Delta m \rightarrow 0, \Delta n = 0$ or $\Delta n \rightarrow 0, \Delta m = 0$. The data points distributed in a surface and the convergence rate of ε is proportional to that of $\frac{1}{\sqrt{N}}$. In this case, suitable values $\varepsilon_1, \varepsilon_2$ proportional to $\frac{1}{\sqrt{N}}$ can be found which makes that $H(\varepsilon - \|x_i - x_j\|)$ equals one when x_j is the next nearest point in the two directions. If $\varepsilon_1 = \varepsilon_2$, we can thus get that $\sum_{j=i+1}^N H(\varepsilon - \|x_i - x_j\|) = 1 + 1 + \underbrace{0 + \dots + 0}_{N-i-2}$ when $i \neq N-1$. When $i = N-1$, $\sum_{j=i+1}^N H(\varepsilon - \|x_i - x_j\|) = 1$. In this case,

$$\begin{aligned} & \log C \left(\varepsilon \rightarrow 0, \varepsilon \propto \frac{1}{\sqrt{N}} \right) = \\ & \lim_{N \rightarrow \infty} \log \frac{1}{N(N-1)} \left[\sum_{i=1}^{N-2} \left(1 + 1 + \underbrace{0 + \dots + 0}_{N-i-2} \right) + 1 \right]. \end{aligned} \quad (4.8)$$

The estimated dimension d is given by

$$\begin{aligned} d &= \lim_{\varepsilon \rightarrow 0, \varepsilon \propto 1/\sqrt{N}} \log C(\varepsilon) / \log \varepsilon \\ &= \lim_{\varepsilon \rightarrow 0, \varepsilon \propto 1/\sqrt{N}} \left[\lim_{N \rightarrow \infty} \log [2(N-2) + 1] / N(N-1) / \log \varepsilon \right] \\ &= \lim_{\varepsilon \rightarrow 0, \varepsilon \propto 1/\sqrt{N}} \left[\lim_{N \rightarrow \infty} \log [2/N - 1/N(N-1)] / \log \varepsilon \right] \\ &\propto \lim_{N \rightarrow \infty} \log [2/N - 1/N(N-1)] / \log N^{-0.5} \\ &\propto \lim_{N \rightarrow \infty} \log N^{-1} / \log N^{-0.5}. \end{aligned} \quad (4.9)$$

If $\varepsilon_1 \neq \varepsilon_2$, the next nearest point would lie in the direction with the smaller ε . We can get that

$$\begin{aligned} & \log C \left(\varepsilon \rightarrow 0, \varepsilon \propto \frac{1}{\sqrt{N}} \right) = \\ & \lim_{N \rightarrow \infty} \log \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \left(1 + \underbrace{0 + \dots + 0}_{N-i-1} \right). \end{aligned} \quad (4.10)$$

Table 4.1: Ideal Estimated Intrinsic Dimension for Two Confounding Cases.

Variable 1	Variable 2	Variable 3	Estimated Dimension
$x=f(z)$	$y=g(z)$	z	around 1
x	y	$z=h(x,y)$	around 2

The estimated dimension d is given by

$$\begin{aligned}
d &= \lim_{\varepsilon \rightarrow 0, \varepsilon \propto 1/\sqrt{N}} \log C(\varepsilon) / \log \varepsilon \\
&= \lim_{\varepsilon \rightarrow 0, \varepsilon \propto 1/\sqrt{N}} \left[\lim_{N \rightarrow \infty} \log (N-1) / N(N-1) / \log \varepsilon \right] \\
&= \lim_{\varepsilon \rightarrow 0, \varepsilon \propto 1/\sqrt{N}} \left[\lim_{N \rightarrow \infty} \log (1/N) / \log \varepsilon \right] \\
&\propto \lim_{N \rightarrow \infty} \log N^{-1} / \log N^{-0.5} .
\end{aligned} \tag{4.11}$$

The estimated dimension was not affected by whether ε_1 equals ε_2 or not. Eqs. 4.9 and 4.11 are both approximately equal to 2. Scatter plots of the simulation data¹ are shown in Figures 4-1 and 4-2. The manifold for the common cause case is a line with an intrinsic dimension around 1 while that for the selection bias case is a surface with an intrinsic dimension around 2. These characteristics can be used to detect possible confounders for variables x and y .

Suppose that variables x and y are dependent. To test whether a variable z is a possible common cause, z can be added into the data. If the estimated correlation dimension is around 1, there is high possibility that z is a confounding common cause variable for x and y . Here it is hard to assert that z is a common cause variable for x and y because the Markov equivalence class exists.

Like a conditional independence test, correlation dimension estimation cannot distinguish the following three patterns belonging to the same Markov equivalence class.

1. A chain: $X \rightarrow Z \rightarrow Y$
2. Another chain: $X \leftarrow Z \leftarrow Y$

¹Simulation data for the common cause case were generated using $x = z^2, y = z^3$; those for the selection bias case were generated using $z = x + y$.

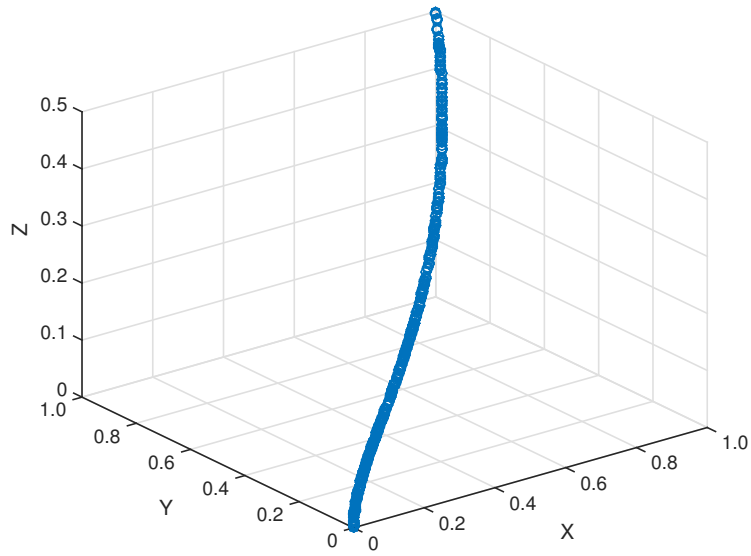


Figure 4-1: Simulation: Common Cause

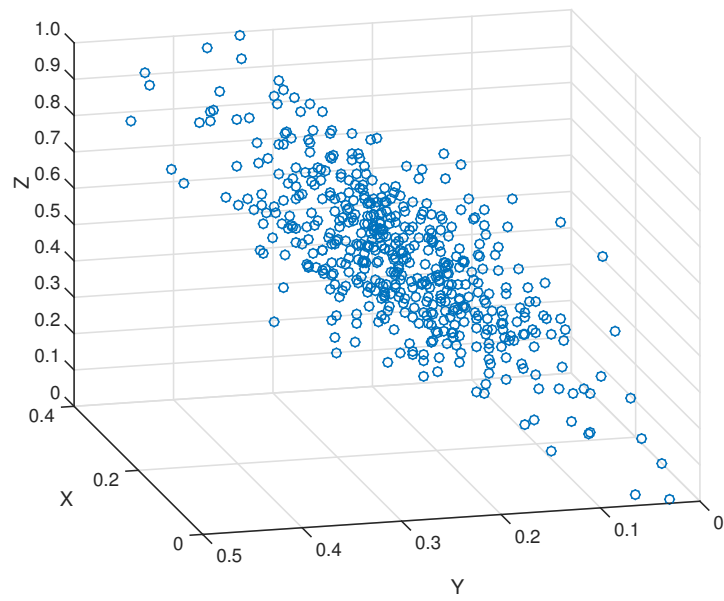


Figure 4-2: Simulation: Selection Bias

3. A fork: $X \leftarrow Z \rightarrow Y$

For the first two patterns, X and Y have an indirect causal relationship via variable Z . For the last pattern, Z works as a common cause for X and Y . For all three patterns, if variable Z is fixed, the path between X and Y can be blocked [30].

An estimated dimension of 1 works as a necessary condition for variable z to be a common cause for x and y . The necessity can be used to determine whether a third variable is a possible common cause for two observed variables. If the estimated dimension is around 1, the third variable is a possible common cause for the two observed variables. As far as we have known, no machine learning methods are sufficient to decide a common cause even the generally used (conditional) independence test. The necessity can help filter the unrelated variables. To distinguish the exact DAG, additional assumptions are needed. For example, an additive noise model [55] needs to be assumed. Nevertheless, the proposed method is effective and efficient as an initial step in detecting possible confounding variables and is easy to implement.

4.3 Simulation

We simulated application of the proposed method for linear and non-linear functions and their combination and compared the results with those for the KCI test proposed by [37]. Our proposed method worked better in the Poisson noise case and slightly better in the multiplicative noise case. Simulation of the selection bias case (Figure 1-1) showed that the proposed method was not affected by Berkson’s paradox. To test how the type of distribution affects the estimated correlation dimension, we generated simulation data using three types of distributions: logistic, Gaussian, and uniform. Finally, we simulated the mixed pattern with causal relationship and showed that the proposed method can tell a possible common cause even when it is mixed with causal relationship.

4.3.1 Simulation of Common Cause Case

We used linear and non-linear functions for the common cause simulation. The artificial data were generated using two equations,

$$\begin{aligned} X &= \alpha * Z + \beta * Z^2, \\ Y &= 2 * \alpha * Z + \beta * Z^3, \end{aligned} \tag{4.12}$$

where α and β are used to control the nonlinearity of the function. We set parameter β to 0 and varied parameter α from 0.1 to 1 for the linear function. For the nonlinear function, we set α to 0 and varied the value of β instead. The experimental results are presented below.

Different Types of Function Forms

Linear Function In the linear function simulation, we set β to 0 and varied the value of α . For each α , we repeated the randomized experiment 20 times and took the average of the estimated correlation dimension. As shown by the plot in Figure 4-3, the estimated dimension was negligibly affected by the value of alpha. In all cases, the estimated dimension was very close to 1. We did the same simulation with the method proposed by [37] in which α was varied and for each α 20 randomized experiments were run. The p-value was not affected by the value of α either. However, in 200 randomized experiments, the independence of X and Y given Z was accepted only eight times ($\alpha = 0.05$).

Non-linear Function In the non-linear function simulation, we set α to 0 and varied the value of β to simulate a non-linear function in the no additive noise case. As shown in Figure 4-4, the estimated intrinsic dimension was very close to 1. The same simulation was run for the KCI test [37] with varied β values. The p-value was not affected by the value of β either. In 200 randomized experiments, the independence of X and Y given Z was accepted ten times.

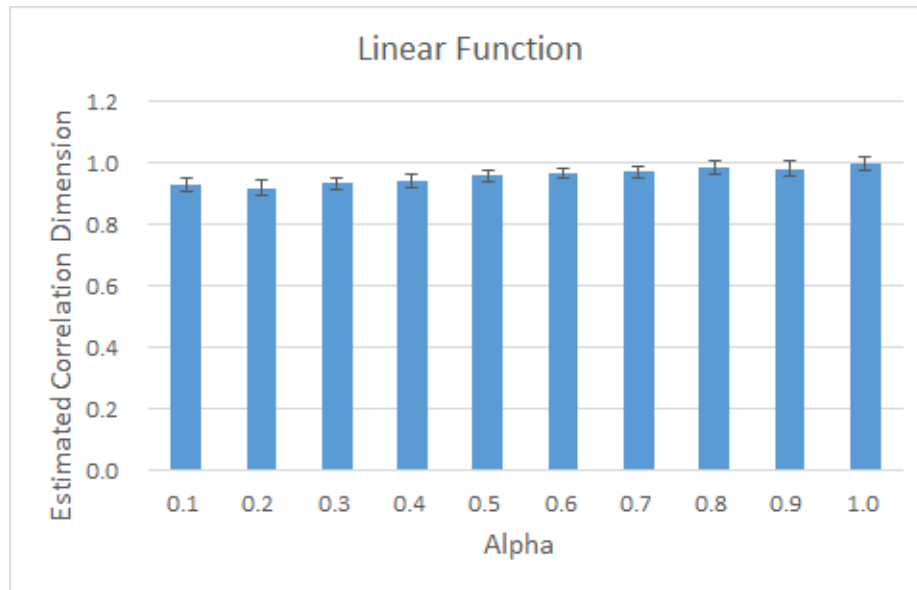


Figure 4-3: Estimated Correlation Dimension for Linear Function

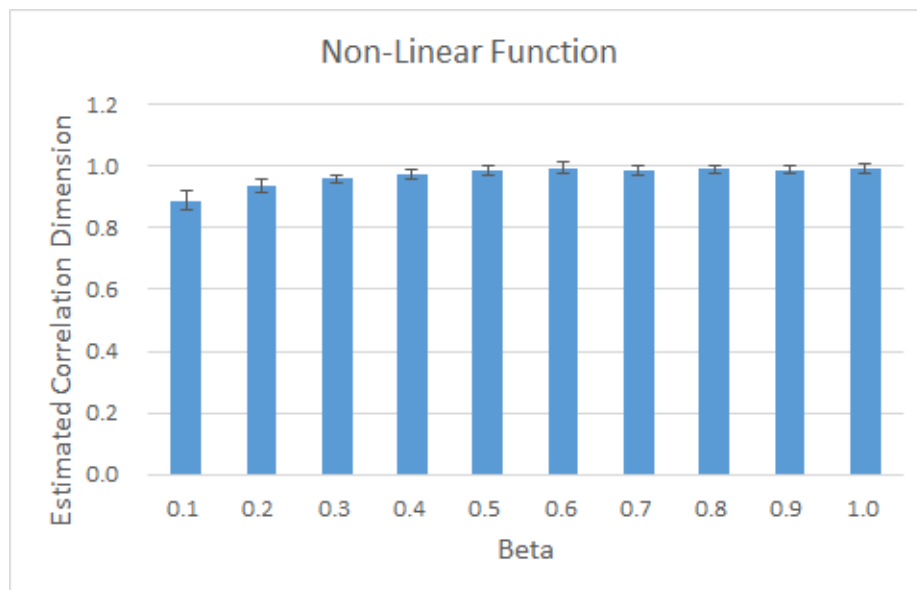


Figure 4-4: Estimated Intrinsic Dimension for Non-Linear Function

Combination of Linear and Non-Linear Functions In the combination simulation, both α and β were set to 1, and the randomized experiment was repeated 20 times. The average estimated intrinsic dimension was 1.0082 with a standard deviation of 0.0221, indicating that the estimated dimension is negligibly affected by the type of function between variables. The same simulation was run for the KCI test [37]. In all 20 randomized experiments, the independence of X and Y given Z was not accepted. KCI test assumed uncorrelatedness of residual function given the common cause. Less of residuals made KCI test fail. Compared with KCI test, our proposed method worked better in all the above cases.

Noise Test

We tested the effect of three different types of noise: Poisson noise, multiplicative noise, and Gaussian noise (Figure 4-5).

Poisson noise Poisson noise is a general kind of image sensor noise that is not stationary and neither additive nor multiplicative. We simulated data using Eq. 4.12, where α and β were set to 1, and Z was generated in a standard normal distribution. We added noise into the generated data through the Poisson process. If the no-noise data was 1, the noised data was generated from a Poisson distribution with expectation 1. Randomized experiments were run 100 times, and the average estimated intrinsic dimension was 0.9053 with a standard variance of 0.0239. The proposed method worked well in the Poisson noise case. For comparison, we ran the KCI test with the same experiment settings. In the 100 randomized experiments, the conditional independence between X and Y given Z was accepted 47 times ($\alpha = 0.05$). That is, about half of them were rejected. These experimental results showed that the proposed method worked better than the KCI test when there was Poisson noise.

Multiplicative noise We added multiplicative noise to the data generation process using the equation $J = I + n * I$, where n is random noise with mean 0. We added noise with different variances from 0.01 to 0.05 into the generated artificial

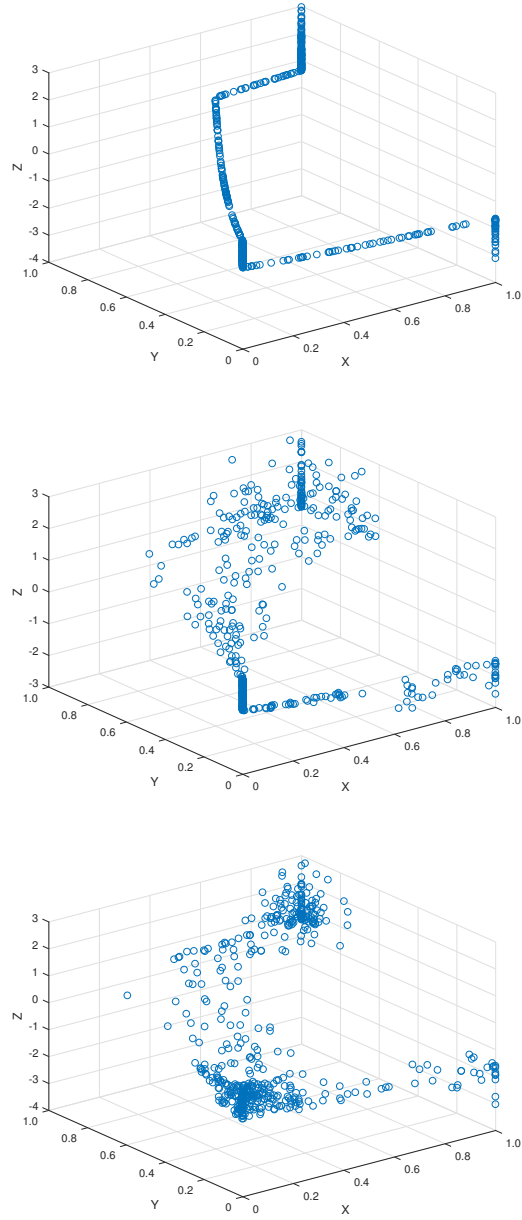


Figure 4-5: Scatter Plot of Artificial Data with Poisson Noise, Multiplicative Noise, and Gaussian Noise

data. The experimental results showed that the estimated correlation dimension was negligibly affected by the variance of the noise. We conducted randomized experiments 100 times; the average estimated correlation dimension was 0.8972 with a standard variance of 0.0445. We did the same simulation for the KCI test. In the 100 randomized tests, the conditional independence of X and Y given Z was accepted 95 times ($\alpha = 0.05$).

Gaussian noise We added Gaussian noise with different variances to the generated artificial data. The estimated correlation dimension with Gaussian noise for variances from 0.01 to 0.05 is shown in Figure 4-6. We ran randomized experiments 20 times for each variance. The estimated correlation dimension was a little larger than 1; the estimated intrinsic dimension tended to initially increase and then decrease as the degree of noise increased. We conducted additional randomized experiments with variances from 0.0001 to 10. When the variance was 0.0001, the average estimated correlation dimension was 1.1361 with a standard variance of 0.0608. When the variance was 10, the average of estimated correlation dimension was 1.0208 with a standard variance of 0.0173. The estimated intrinsic dimension was close to one when the added Gaussian noise was very little or very large. In all the randomized experiments, the estimated intrinsic dimension was less than 1.7. We did the same simulation for the KCI test. In the 100 randomized tests, the conditional independence of X and Y given Z was accepted 94 times.

These results show that the proposed method generally worked well for different types of added noise. The estimated value of the correlation dimension is affected by the type of noise and number of samples. The estimated correlation dimension after adding Gaussian noise was below 1.7 (Figure 4-6). Thus, we think 1.7 would be a suitable upper limit value for the proposed method. The estimated correlation dimension was above 0.8 in our simulation (Figs. 4-3 and 4-4). Since the real world data has limited data points and the observed data is discrete, the estimated dimension tends to be lower. Therefore, we reduced the lower limit by 0.1 to 0.7. According to the above experimental results and considerations, we set the confidence interval

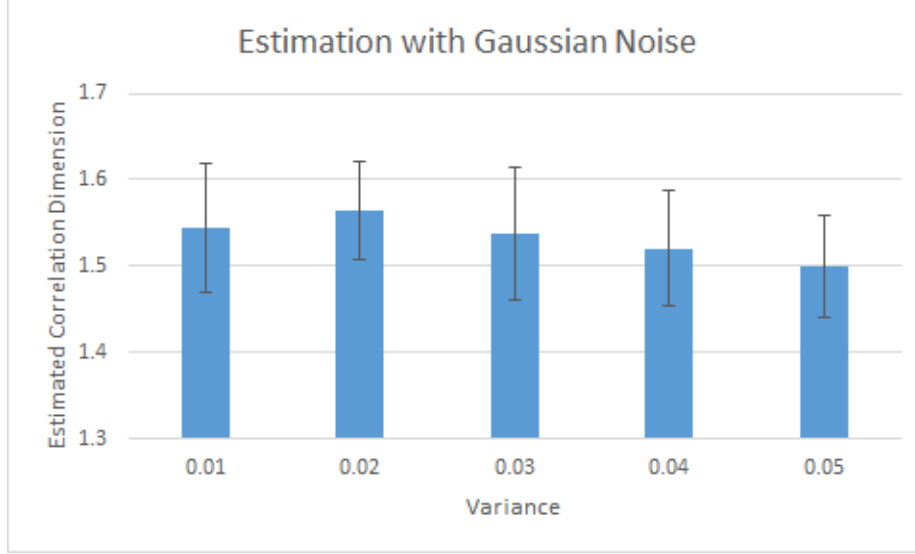


Figure 4-6: Estimated Correlation Dimension for Added Gaussian Noise with Different Variances

to be $[0.7, 1.7]$ when applying the correlation dimension to causal discovery.

4.3.2 Simulation of Selection Bias Case

We simulated the selection bias case (Figure 1-1) using $Z = X + Y^3$. We first randomly generated 2000 pairs of X-Y data values from a uniform distribution $(0, 1)$ and calculated the corresponding Z values. Next, we counted the instances of Z in each sub-interval. Finally, we selected the X and Y with Z in the interval with the most instances. Figure 4-7 shows the count of Z in the sub-intervals of interval $(0, 2.0)$. Figure 4-8 shows a scatter plot of 205 instances of the selection bias data with Z in the interval $(0.9, 1.0)$. We used the KCI test [37] to test the independence of X and Y. The p-value was 0, which means that X and Y were not independent although they were randomly generated and thus should have been independent. However, selection bias can make two randomly generated variables dependent (Berkson's paradox). Intrinsic dimension estimation using the selected biased data shown in Figure 4-8 gave an estimated value of 1.8322. The average estimated dimension for 20 randomized experiments was 1.8273 with a variance of 0.0568, demonstrating that the proposed method can give a correct estimation even when the data is biased.

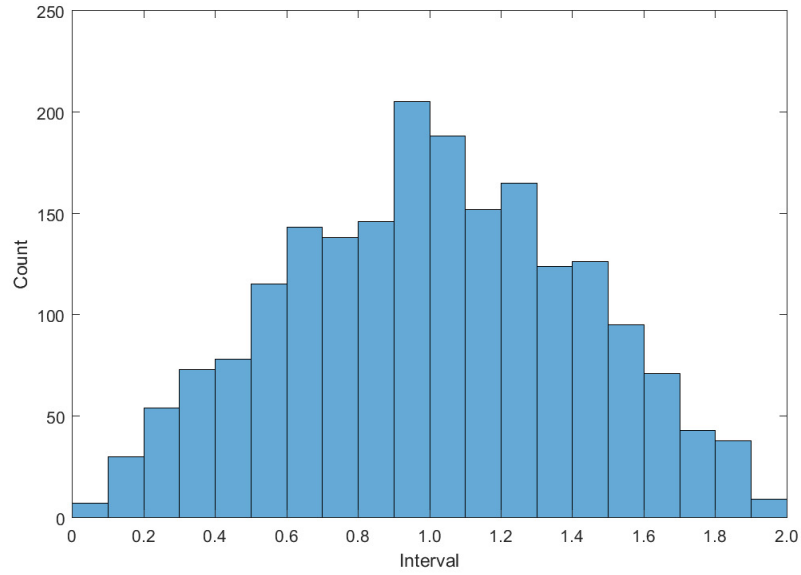


Figure 4-7: Count of Z in Sub-intervals of Interval (0, 2.0)

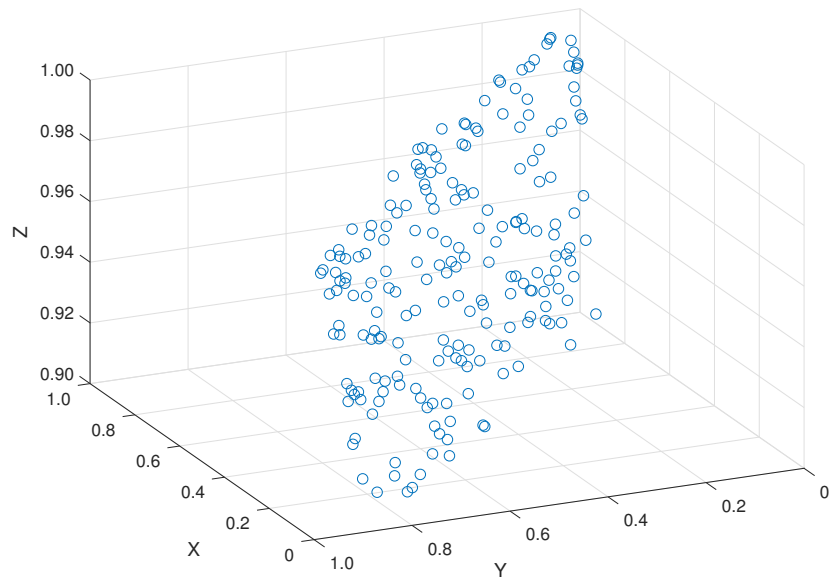


Figure 4-8: Scatter Plot of Selection Bias Data for Z in (0.9, 1.0)

Table 4.2: Estimated Dimension for Different Types of Distributions.

Distribution	Kurtosis (mean: 0, variance: 1)	Estimated Dimension
Logistic	1.2	0.9204
Gaussian	0	0.9207
Uniform	-1.2	1.0009

4.3.3 Further Study

Different Types of Distributions

To test how the type of distribution affects the estimated intrinsic dimension, we used three different types of distributions: logistic, Gaussian, and uniform (Figure 4-9) with kurtosis of 1.2, 0, and -1.2 for normalized data with mean 0 and variance 1². The estimated dimensions are shown in Table 4.2. For all three types, the estimated dimension was close to 1, indicating that the estimated dimension is slightly affected by the data distribution.

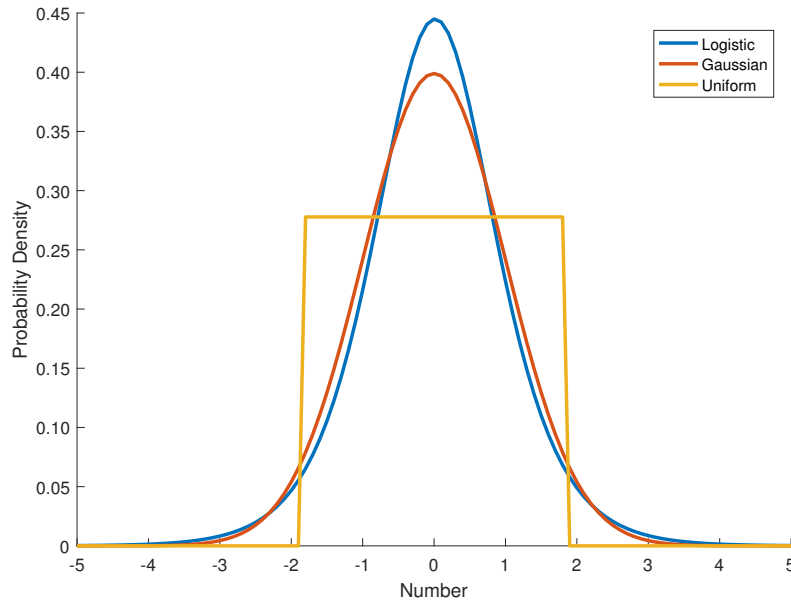


Figure 4-9: Probability Density Function for Three Types of Distributions

²<https://en.wikipedia.org/wiki/Kurtosis>

Mixed Pattern of Common Cause and Causal Relationship

Common cause and causal relationship may be mixed with each other in the real world. If causal relationship exists between X and Y , the conditional independence test cannot tell a possible common cause of X and Y any more. However, even when real causal relationship exists, the effect of confounding should be considered as well. As shown by the existing researches, causal relationship may change after considering the effect of specific confounder [29, 90]. Our proposed method can still decide possible common cause when it is mixed with causal relationship. We ran simulation to show the experimental results. The artificial data was generated by $X = Z + Z^2, Y = 2 * Z + Z^3 + X$ in which Z was a common cause for X and Y , X causes Y . The average of estimated intrinsic dimension was 0.9410 with standard variance 0.0299. The simulation results showed that the proposed method can give correct inference when causal relationship and common cause are mixed with each other.

4.4 Real World Data

Finally, we used real world data obtained from the CauseEffectPairs (CEP) dataset [80] to test our proposed method.

4.4.1 Altitude, Temperature, Sunshine

The pairs “altitude→temperature” and “altitude→sunshine” are two cause effect pairs in the CEP dataset. They were taken from the UCI Machine Learning Repository [81]. Temperature and sunshine are variables confounded by a common cause, altitude. The p-value from a KCI test for temperature and sunshine was 0.002. Given the altitude, the p-value for temperature and sunshine was 3.2779e-04. The independence of temperature and sunshine given the altitude was thus rejected ($\alpha = 0.05$). For the three variables, although altitude is a common cause for the other two variables, sunshine causes temperature as well in human cognition. Conditional inde-

pendence test cannot tell a possible common cause anymore when the two are mixed with each other. Although altitude is a common cause for temperature and sunshine, the KCI test failed to detect it. With our proposed method, the estimated intrinsic dimension was 0.7132, which lies in the interval $[0.7, 1.7]$, showing that the variable “altitude” can be a common cause for “temperature” and “sunshine.”

4.4.2 Cement, Blast Furnace Slag, Compressive Strength

The pairs “cement→compressive strength” and “blast furnace slag→compressive strength” are two more cause effect pairs in the CEP dataset. We combined the data to get a real world example of “selection bias” (Figure 1-1). The p-value from a KCI test for cement and blast furnace slag was 0, which shows that the data was biased. The estimated intrinsic dimension was 2.0156. Application of dimension estimation to finding possible confounding variables did not suffer from the effects of Berkson’s paradox and was useful in common cause identification.

4.5 Summary

We proposed using intrinsic dimension estimation to detect a possible common cause for two variables. The proposed method does not need certain assumptions of function forms, data distributions. Comparison experiment was conducted with a kernel-based independence test. The experimental results showed that the proposed method worked comparatively or better than the existing method. Besides, the proposed method can still tell a possible common cause when it is mixed with causal relationship while conditional independence test cannot. The proposed method does not suffer from the effects of Berkson’s paradox. The estimated intrinsic dimension was around 1 for the common cause case and around 2 for the selection bias case. In a scatter plot of data, the topology of the data generated for the common cause case would be a line while that for the selection bias case would be a surface. Simulation testing showed that the estimated correlation dimension is slightly affected by the type of distribution. Testing using variables from the CauseEffectPairs dataset showed that

the proposed method can estimate the common cause for real world data.

Chapter 5

Exploratory Causal Analysis of Open Data

5.1 Introduction

Open data are becoming increasingly available in various domains owing to many preceding studies to promote them [96, 97, 98]. Open data are important public goods and people can obtain useful insights to solve problems in the society by analyzing them [99]. In analyzing open data for policy making, it is important to distinguish causality from correlation because what we really want to know is the actual cause of the problem and the effect of the remedy for it. Causal analysis is a fundamental problem in data science and has attracted interests from researchers in various fields [29, 100]. Causal analysis of open data can give insights on some focused issues, e.g. carbon dioxide emissions. However, causal relationship is complex. The relationship between variables can be direct or indirect. Besides, spurious correlation may exist in real world. For open data, spurious correlation generally exists and it is hard to include all related variables into the analysis process at one time since the data may be provided by different organizations. Correct causal analysis result may be difficult to be obtained when only machine learning methods are included.

Domain knowledge can aid in reducing the effect of confounding in causal analysis. Huang et al. analyzed the causal relationship between energy consumption and GDP

growth by dividing the data into four income categories [29]. Their experimental results showed that causal relationships differed among the four categories. The above research showed that the causal relationship may change after considering a certain confounder. Based on the above experimental results, we think analyzing causal relationships while considering the effect of possible confounders is of importance. When finding those confounders, human expertise will be helpful.

In our work, we developed a framework for exploratory causal analysis of open data which combines human computation with machine learning methods. The proposed framework takes into account possible confounders when doing causal analysis and tests them using state-of-the-art machine learning methods. The proposed framework contains three main parts: collecting explanations of correlation in open data using a crowdsourcing marketplace, extracting keywords using natural language processing methods, and verifying the collected explanations using machine learning methods. The proposed framework explanatorily tested the possible confounders which were extracted from the collected texts and filtered the useful ones using machine learning methods. The related variables were then further analyzed using causal discovery methods. The testing step was included to filter the unrelated words. The causal discovery method was included to learn the causal directions between variables.

In Section 5.2, we introduce the preliminary experiments and discuss the performance of existing methods. In Section 5.3, we propose our framework and introduce the related matching learning methods used in the framework. In Section 5.4, we present the experimental results and discuss the performance of proposed framework. We conclude in Section 5.5.

5.2 Preliminary Experiments

Spurious Correlation

Correlation is easy to be observed in real world data. Some correlation between variables e.g. chocolate consumption and Nobel laureates are debatable. There are

generally four types of relationships between two variables (Figure 1-1) except independence and feedback. The bottom two in Figure 1-1 describes two confounding cases: “common cause” and “selection bias”. The selection bias case was confounded only when the value of variable Z was fixed which means the data are biased. Here we focused on the common cause case in which spurious correlation are induced by the unconsidered common cause Z . The following parts discussed the performance of existing methods when the data is confounded by a common cause.

Model with No Confounding Assumption

We generated the simulated data using $X = 2*Z, Y = Z^2$ with Gaussian noise added. X and Y are two observed variables while Z are the confounder. We conducted experiment with the PNL model which works under no confounding assumption and is stable to noise. We used the simulated data of X and Y which have spurious correlation. The PNL model inferred that Y caused X . Even when X and Y are confounded by Z , it is hard for existing methods to learn the correct causal structure while the confounding variable is unconsidered.

Model with Confounding Assumption

We ran another simulation using the method which takes into account of unobserved common causes. We took the BMLiNGAM method as an example. BMLiNGAM was an extension of LiNGAM which takes into account of the unobserved confounder in causal analysis. We conducted experiments using the BMLiNGAM method with the same experiments settings as above. BMLiNGAM inferred that Y caused X as well.

The above experimental results showed that it was hard for the existing methods to give correct inference when the correlation in data was induced by a third unconsidered variable. Finding possible common cause and completing the causal structure is needed for correct inference.

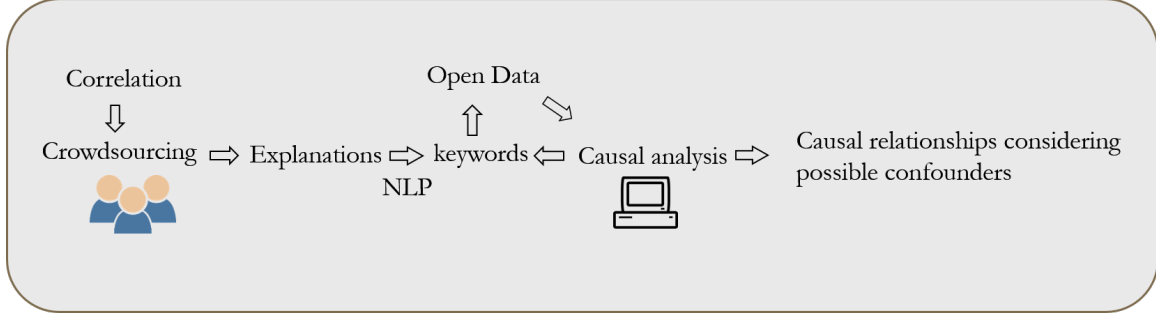


Figure 5-1: Framework

5.3 Proposed Framework

To address the above problem, we proposed an exploratory approach to do causal analysis using open data. The proposed framework mainly contains three parts: collecting explanations of correlation using crowdsourcing, extracting keywords using the NLP methods and doing causal analysis using state-of-the-art machine learning methods. Correlation are easily observed in open data. Variables with high correlation are focused on and the explanations about the correlation are collected using general crowdsourcing platform. General NLP methods are used to extract related keywords which may represent possible confounders. Finally, the collected explanations are analyzed using several machine learning methods with open data.

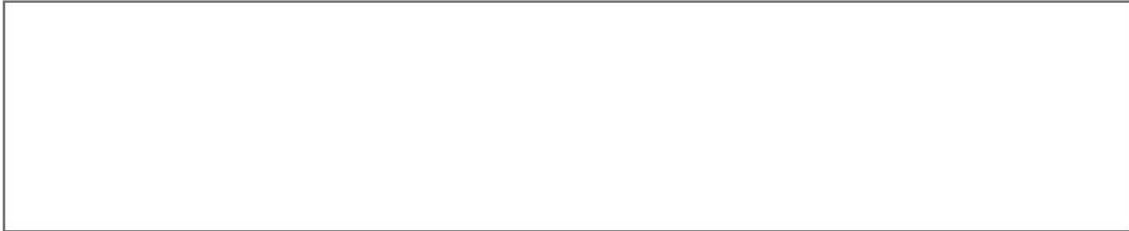
5.3.1 Explanation Generation

Correlation is easily observed and some of them are interesting but confusing. People may have different explanations about the same correlation. The explanations are usually based on their life experience and needs further test about their correctness. Compared with the explanation given by experts in different fields, the explanations collected using the crowdsourcing platform can be quickly generated in low cost. We collected explanations of high correlation in open data in a Japanese crowdsourcing marketplace. One of the task example about GDP and carbon dioxide emissions was shown in Fig. 5-2. We instructed the workers to write their ideas freely, with a word count of 50 to 300. The collected explanations were in Excel format and were

Explanation Collection

Countries with high GDP often have high carbon dioxide emissions.

Please think about the reasons for this and write them in the text area below (50-300 words)



Thank you.

Figure 5-2: Task Example

machine readable (Fig. 5-3). Many workers wrote more than one reason (as shown in the example below), indicating that crowdsourcing workers are willing to share their ideas about a topic.

“The fact that GDP is high means that people in that country are wealthy, so the car ownership rate is relatively high, which increases carbon dioxide emissions. In addition, the manufacturing industry is thriving, and a large amount of electricity is needed to operate a factory. Thermal power generation is the easiest way to generate electricity, so carbon dioxide emissions will increase. For families in countries with high GDP, there are few cases where family members get together in one room. Each person goes to their own room, turns on the electricity, and uses the air conditioner, TV, and PC in their room, so electricity consumption increases, which leads to an increase in carbon dioxide emissions.”

5.3.2 Keywords Extraction

We extracted keywords using causal relationship extraction methods [101] and latent Dirichlet allocation (LDA) model [102].

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	GDPが高い国は、経済活動が活発だから必然的にエネルギーの消費量も多く、	In countries with high GDP, economic activities are more active which consumes energy. As																	
2	経済活動が盛んだから自然と二酸化炭素も増える。森林が減り、産業が盛ん、	Carbon dioxide emissions increase automatically when economic activity is thriving. Carbon																	
3	GDPが高いということは、人や物の生産力も高いということ。そして生産力が高	A high GDP means that the productivity of people and goods is also high. And if the produc																	
4	GDPが高い国では、国内で生産したり消費したりする量が多いので、自然と	In countries with high GDP, there are many goods to be produced and consumed domestic																	
5	GDPが高いということは、比較的、国民が裕福で車の保有率が高いので、二	The fact that GDP is high means that people in that country are wealthy, so the car ownershi																	
6	GDP(国内総生産)の高い国は、生活の豊かさが一般的に見ても安定しており、	The life in countries with high GDP (Gross Domestic Product) are generally stable, and there																	
7	経済活動が高い国はエネルギー消費量が多いため、二酸化炭素の排出量が多	In countries where economic activity is high, energy consumption is also high, so I think that																	
8	GDPが高いということは社会が発展しているということである。発展とは便利	A high GDP means that society is developing. Development is a process that is convenient a																	
9	産業の生産量が高くなる為、それなりのエネルギー消費が発生し、エネルギ	As the production volume of the industry increases, appropriate energy consumption occurs																	
10	GDPが高い国はたくさんのもをを生産し、それにより利益を得ており、た	A country with a high GDP produces a lot of things, and therefore benefits from it, and it is r																	

Figure 5-3: Collected Explanations by Crowdsourcing

Causal Relationship Extraction Method

Causal relationships can be extracted from general natural language. We extracted the causal relationships contained in the phrases with the word “tame”. For the sentences with the word “tame”, the dependency structure between bunsetsu-phrases is constructed by Cabocha¹ [103]. The modifier bunsetsu-phrase of tame in subordinate clause and modified bunsetsu-phrase of tame in matrix clause are then identified. Sentences are classified as “adverbial verb phrase” if the modifier bunsetsu-phrase and the modified bunsetsu-phrase fulfill the four morphological conditions [103]. Finally, we extracted the nouns in adverbial verb phrase except the “GDP” and “carbon dioxide emissions”.

LDA

Since the collected explanations were sentences about the same topic, we could treat all the collected explanations in one task as a document and determine its topic. There are limited words for each collected explanation by different workers and it is preferable to learn the topic of all the collected explanations. We assume that the collected explanations were about the same topic and the topic number of the document was set to be 1. The correlation of different topics [104] are not involved in the task. We chose LDA model to extract the related keywords. Through extracted the related words, we learned the most probable words for the topic. The online LDA model proposed by Hoffman et al. [102] was used in the experiment. The topic and most probable words were obtained using the following steps.

Morphological Analysis

¹<http://taku910.github.io/cabocha/>

The explanations were in Japanese and we used the Janome morphological analysis engine in Python².

Document Cleaning

The assigned two variables with high correlation in each task were filtered out. Other meaningless stops words were filtered out either to enable creation of a word list containing only nouns, i.e., a bag-of-words.

Document-term Matrix Building

After getting the bag-of-words for a “document,” the term frequency matrix was calculated and used to determine the feature of the document.

LDA Model Application

The online LDA model was used to determine the topic of each document. The most probable words were obtained in this step.

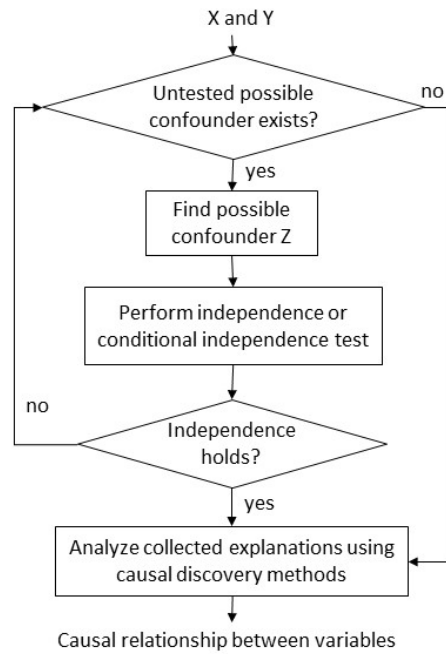
5.3.3 Causal Analysis

We used machine learning methods to analyze the explanations collected from the crowdsourcing platform. According to our empirical study [105], we proposed two workflows to analyze causal relationships (Fig. 5-4). The top workflow in Fig. 5-4 used independence test or conditional independence test to filter the unrelated words. Our previous study showed that it worked well in the multiplicative noise or Gaussian noise case. The bottom workflow used intrinsic dimension estimation to filter the unrelated words. Our previous study showed that it worked well in the no noise case or the Poisson noise case. Since we cannot assume how the open data was generated, we included the two workflows in the proposed framework.

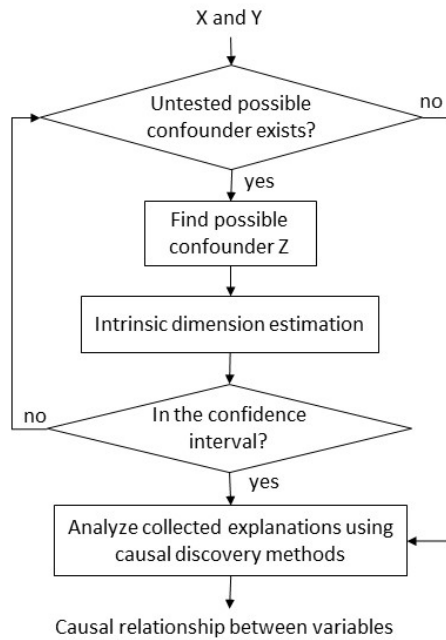
Decision of Possible Confounder

The difference of two workflows is how to filter unrelated words and decide possible confounders. Two different methods were used to decide possible confounders. The

²<https://pypi.python.org/pypi/Janome>



(a) Multiplicative Noise Case or Gaussian Noise Case



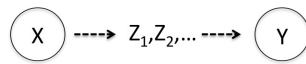
(b) No Noise Case or Poisson Noise Case

Figure 5-4: Workflows.

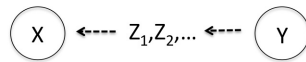
performance of them were discussed in the previous work [105].

Intrinsic Dimension Estimation We first used the intrinsic dimension estimation method discussed in Chapter 4 to decide whether a variable is a possible confounder. As discussed in Chapter 4, if a third variable is a common cause for two variables, the data of all the three variables are distributed in a one-dimensional manifold (a line). The estimated intrinsic dimension is around 1 ideally. If the third variable is not confounding, the estimated intrinsic dimension would be around 2. Empirical study showed that the estimated dimension lies in the confidence interval $[0.7, 1.7]$ after adding different types of noise.

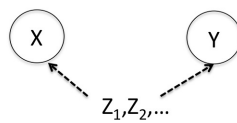
Independence or Conditional Independence Test Conditional independence test can be generally used to decide whether a third variable is a possible common cause. We used the KCI test proposed by Zhang et al. [37] in the framework. The KCI test used the uncorrelatedness of residual function to do the conditional independence test. Thus, additive errors are needed. If conditional independence does not hold, the third variable is not a possible common cause. However, if conditional independence holds, further analysis of three possibilities is needed. The third variable can be a common cause. It can also stand in the path between X and Y (Figure 5-5). The three patterns in Figure 5-5 belong to the same Markov equivalence class [106], and a conditional independence test cannot distinguish them. We first tested the conditional independence between the words with spurious correlation given the extracted word Z . If the conditional independence existed, we then used the causal direction inference method described in the following to infer the causal direction between extracted word with the other two variables. If the conditional independence was not accepted, we changed the order between variables and further tested the independence or conditional independence between them. The process ended until one order of variables was accepted. If there was one order accepted, we then used the following causal direction inference method to do further analysis. If no order was accepted, we decided that the extracted word was unrelated.



(a) “X Causes Y”



(b) “Y Causes X”



(c) “Common Cause”

Figure 5-5: Markov Equivalence Class.

Causal Direction Inference

Although the cause effect pairs belong to the same Markov equivalence class, some methods have been proposed to tell cause from effect by adding further assumptions. We used the IGCI method introduced before.

5.4 Experiments

We conducted experiments using open data and collected explanations about the correlation in data. The tasks were simple and they were finished in one day. Topic words were extracted and further causal analysis was done using the workflows proposed in Fig. 5-4. Since we cannot tell whether the real world data were generated, we analyzed the open data using both of the two workflows proposed in Fig. 5-4. In general, the two workflows will give the same analysis result. When there were conflicts in deciding whether to do further analysis using causal discovery methods, we chose to do it at least one workflow agreed. We did comparison experiments with the existing state-of-the-art methods under no confounding assumption and confounding assumption. The experimental results showed that finding possible confounding variables can help better understand the causal relationships between variables.

5.4.1 GDP and Carbon Dioxide Emmisions

We conducted experiments using the open data of GDP and carbon dioxide emissions. The relationship between GDP (Gross Domestic Product) and carbon dioxide emissions has been studied in [29, 90]. In [29, 90], countries have been allocated in specific groups according to GNI (Gross National Income) or GDP. This means that considering the effect of specific economic indicators are important when analyzing the causal relationship between GDP and carbon dioxide emissions. We first used existing causal discovery methods which work in the no confounding assumption and confounding assumption to analyze the causal relationship between the two. The PNL model can tell cause from effect in the no confounding assumption. We implemented

it and used it to infer the causal direction between carbon dioxide emissions and GDP. The result was that PNL inferred that carbon dioxide caused GDP which did not follow the existing research result. BMLiNGAM was an extension of LiNGAM which takes the unobserved confounders in causal analysis into account. BMLiNGAM inferred that carbon dioxide emissions caused GDP as well. The assumption of unobserved confounders did not change the inferred result.

Word Extraction

Causal Relationship Extraction Method We conducted experiments by collecting explanations of the correlation between GDP and carbon dioxide emissions. The task was shown in Fig. 5-2. We collected explanations from 50 workers. The number of collected characters totally was 4713. For this task, the average number of characters each worker answered was 94.26. Most workers answered several sentences. We extracted the causal relationships of the collected texts using the keywords “tame” [101]. Nineteen sentences with the word “tame” were extracted. We selected the adverbial verb phrases which fulfill one of the four morphological conditions as discussed in [101]. In the nineteen sentences, eleven adverbial verb phrases fulfilled one of the four morphological conditions. The following causal relationships were extracted:

- To operate the factory, carbon dioxide emissions increase.
- To maintain the level of living and production, an energy source is necessary.
- Since the energy consumption is high, carbon dioxide emissions increase.
- To produce goods, carbon dioxide emissions increase.
- To open a civilization, cut the trees.
- To transport, carbon dioxide emissions increase.
- Since gasoline is burned, carbon dioxide emissions increase.
- To supply electricity, carbon dioxide emissions increase.

Table 5.1: Extracted Words Using Causal Relationship Extraction Method (Experiment i)

factory
level
living
production
energy
source
consumption
goods
civilization
trees
gasoline
electricity

- To generate energy, carbon dioxide emissions increase.
- Since consumptions increase, carbon dioxide emissions increase.
- Since energy is needed, carbon dioxide emissions increase.

About 80 percentage of the extracted causal relationships using the word “tame” were about carbon dioxide emissions. Since only phrases with “tame” were extracted, some background information was ignored which may be possible confounders. We extracted the nouns contained in the adverbial verb phrases except “GDP” and “carbon dioxide emissions”. The extracted words were shown in Table 5.1.

LDA We used LDA model to extract keywords related to the topic of GDP and carbon dioxide emissions. The ten words with the highest probability were shown in Figure 5-6. We searched for data about the ten related words and got data about the following variables in the World Bank Open Data⁴. In the extracted 10 words, five of them were found about the related data.

We filtered the unrelated words using machine learning methods. The detailed results will be shown in the following. We compared the performance of causal relationship extraction method and LDA using the metrics of recall and precision. The

⁴<https://data.worldbank.org/>

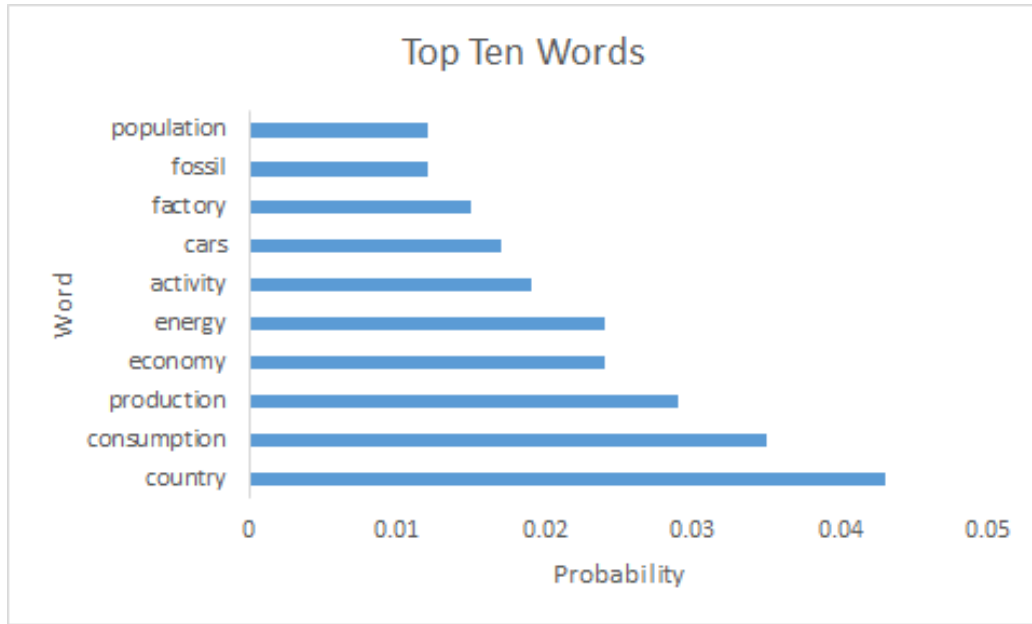


Figure 5-6: Extracted Words Using LDA (Experiment i)

experimental results were shown in Fig. 5-7. Here, the causal relationships learned by machine learning methods were assumed to be the true ones. For this pair, causal relationship extraction method had better recall. The precision of the two are the same with each other which was about 40 percentage. Totally, causal relationship extraction method had better performance for exploratory causal analysis of GDP and carbon dioxide emissions.

Cause Analysis Result

The following words has been extracted using the above two methods. Some of them are decided to be useful since extensive causal structure has been learned after adding the mentioned variable. Some of them cannot be found open data and some others are decided to be unrelated. In the verifying process, the existing methods may give reverse decisions. The extracted words and detailed causal analysis results are discussed in the following.

Access to electricity We searched for data according to the extracted word “electricity”. We got the data “access to electricity (% of population)” from the world

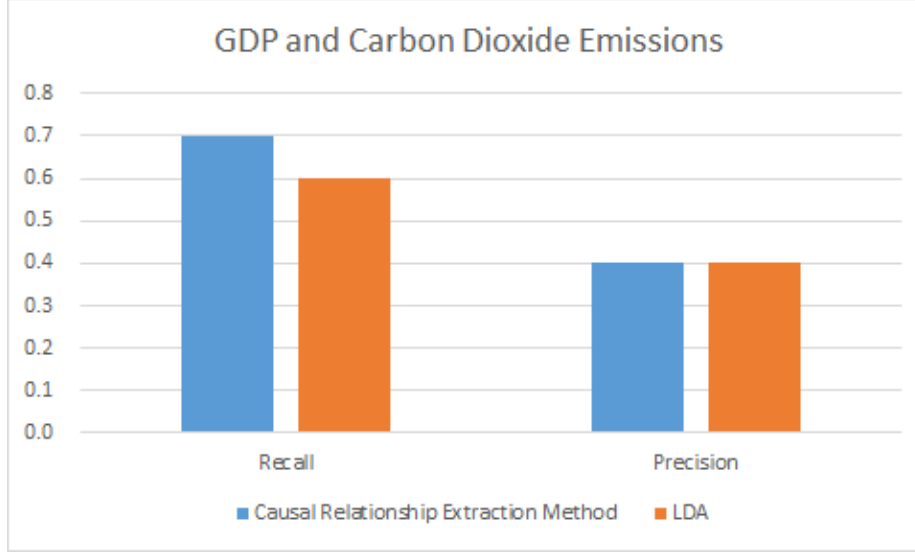


Figure 5-7: Recall and Precision of Causal Relationship Extraction Method and LDA (Experiment i)

bank dataset. We first tested the conditional independence between carbon dioxide emissions and GDP given access to electricity. The null hypothesis of conditional independence was accepted with p-value ⁵ 0.04. Besides, the intrinsic dimension estimation result of the three variables was 0.8 which lies in the interval [0.7, 1.7]. Both of the methods assume that “access to electricity” may be a possible common cause of GDP and carbon dioxide emissions. We then inferred the causal direction between carbon dioxide emissions and access to electricity, GDP and access to electricity. The IGCI method inferred that “access to electricity” caused GDP and carbon dioxide emissions. Thus the following causal relationships were obtained through the proposed framework with “access to electricity” as a common cause:

$$\begin{array}{c} \text{GDP} \leftarrow \text{access to electricity} \\ \rightarrow \text{carbon dioxide emissions} \end{array} .$$

Exports of goods and services We searched for the data according to the keyword “goods”. We got the data of “exports of goods and services (% of GDP)”.

⁵The p-value showed the likelihood of independence. We set the significance level to be 0.02 in the experiment. When the p-value was over 0.02, we concluded that the variables were independent.

Testing the conditional independence showed that the conditional independence did not exist between GDP and carbon dioxide emissions given “exports of goods and services”. We then analyzed the data using the intrinsic dimension estimation. The estimated intrinsic dimension was 0.9 which lies in the confidence interval. Although conditional independence rejected the assumption, the intrinsic dimension result showed that “exports of goods and services” could still be a possible common cause. We then inferred the causal direction between carbon dioxide emissions and exports of goods and services, GDP and exports of goods and services. The result was that the IGCI method inferred that “exports of goods and services” caused the carbon dioxide emissions and GDP. The following causal relationships were obtained through the proposed framework:

$$\begin{aligned} \text{GDP} &\leftarrow \text{exports of goods and services} \\ &\rightarrow \text{carbon dioxide emissions} \quad . \end{aligned}$$

Imports of goods and services The data was searched for according to the extracted word “goods”. We first tested the conditional independence between GDP and carbon dioxide emissions given “imports of goods and services”. The null hypothesis of conditional independence was rejected with p-value 1.97×10^{-5} . When analyzing the data using intrinsic dimension estimation, the estimated intrinsic dimension was 0.9 which lies in the confidence interval. The intrinsic dimension estimation results showed that “imports of goods and services” could be a possible common cause. We inferred the causal direction between carbon dioxide emissions and imports of goods and services. The IGCI method inferred that “imports of goods and services” caused GDP and carbon dioxide emissions. The following causal relationships were learned through the proposed framework:

$$\begin{aligned} \text{GDP} &\leftarrow \text{imports of goods and services} \\ &\rightarrow \text{carbon dioxide emissions} \quad . \end{aligned}$$

Price level ratio of PPP conversion factor (GDP) to market exchange rate

We searched for data according to the extracted word “level”. We obtained the data of “price level ratio of PPP conversion factor (GDP) to market exchange rate” from the world bank. We tested the conditional independence between GDP and carbon dioxide emissions given “price level ratio of PPP conversion factor (GDP) to market exchange rate”. The null hypothesis of conditional independence was accepted with the p-value 0.0289. Conditional independence showed that “price level ratio of PPP conversion factor (GDP) to market exchange rate” could be a possible common cause of GDP and carbon dioxide emissions. Besides, we estimated the intrinsic dimension of the three variables. The estimated result was 1.1 which lies in the confidence interval. The intrinsic estimation result also showed that “price level ratio of PPP conversion factor (GDP) to market exchange rate” could be a possible common cause of GDP and carbon dioxide emissions. Finally, we used the IGCI method to infer the causal direction between “price level ratio of PPP conversion factor (GDP) to market exchange rate” and GDP, “price level ratio of PPP conversion factor (GDP) to market exchange rate” and carbon dioxide emissions. The IGCI method inferred that “price level ratio of PPP conversion factor (GDP) to market exchange rate” causes GDP and carbon dioxide emissions. The following causal relationships were obtained by the proposed framework:

$$\begin{array}{c} \text{GDP} \leftarrow \text{price level ratio of PPP} \\ \text{conversion factor (GDP) to market exchange rate} \\ \rightarrow \text{carbon dioxide emissions} \end{array} .$$

GNI The GNI was searched as an indicator of economy which was often mentioned in the collected explanations. We first analyzed the data under the additive noise assumption. We tested the independence between the three variables. The p-value for the test of conditional independence between GDP and carbon dioxide emissions conditioned on GNI was 0.0178, meaning that they were not independent given GNI

($\alpha = 0.02$). We then calculated the p-value for the test of conditional independence between GNI and carbon dioxide emissions conditioned on GDP. It was 0.0266, indicating conditional independence. For further study, we calculated the p-value for the test of conditional independence between GDP and GNI conditioned on carbon dioxide emissions. It was 4.38×10^{-5} , indicating that they were not independent. When analyzing the data under the no additive noise assumption, the estimated intrinsic dimension was 0.5. From these results, we can assume that GDP may be a common cause of GNI and carbon dioxide emissions. To further study the relationship between them, we calculated the causal direction between GDP and GNI and between GDP and carbon dioxide emissions. To further study the relationship between them, we calculated the causal direction between GDP and GNI and between GDP and carbon dioxide emissions. The IGCI method revealed that GDP growth causes carbon dioxide emissions and that GNI causes GDP growth. That is, the crowd identified a variable that affects the causal relationship analysis between GDP and carbon dioxide emissions, as was done in a previous study [29]:

$$\text{GNI} \rightarrow \text{GDP} \rightarrow \text{carbon dioxide emissions} \quad .$$

Fossil fuel energy consumption The fossil energy consumption was searched according to the words: “consumption”, “energy” and “fossil”. We first analyzed the data under the additive noise assumption. The p-value for the test of conditional independence between GDP and carbon dioxide emissions given energy consumption was 5.90×10^{-4} , so the null hypothesis of independence was rejected. The p-value for the test of conditional independence between carbon dioxide emissions and fossil fuel energy consumption given GDP was 0.0763, indicating that they were conditionally independent. However, the p-value for the test of independence between carbon dioxide emissions and GDP was 1.00×10^{-3} which showed that they were dependent on each other. When analyzing the data under the no additive noise assumption, the estimated intrinsic dimension was 0.7 which lies in the confidence interval. These re-

sults indicate that GDP can be a common cause or work as a middle variable between carbon dioxide emissions and fossil fuel energy consumption. Further analysis using the IGCI method showed that the causal direction between carbon dioxide emissions and GDP was backward and that the one between fossil fuel energy consumption and GDP was forward:

$$\begin{array}{c} \text{fossil fuel energy consumption} \rightarrow \text{GDP} \\ \rightarrow \text{carbon dioxide emissions} \end{array} .$$

Electric power consumption Electric power consumption was searched according to the extracted word “consumption”. We first analyzed the data under the additive noise assumption. The p-value for a test of conditional independence between GDP and carbon dioxide emissions given electric power consumption was 5.08×10^{-4} . GDP and carbon dioxide emissions were thus not independent given electric power consumption. The p-value for a test of independence between electric power consumption and GDP was 0.0410, which means that they were independent ($\alpha = 0.02$). However, when conditioned on carbon dioxide emissions, they were not independent (p-value 0.0033). That is, carbon dioxide emissions can be a common effect between electric power consumption and GDP. When analyzing the data under the no additive noise assumption, the estimated intrinsic dimension was 0.7. Further analysis using IGCI was did to infer the causal relationship of the three variables. The IGCI method revealed that electric power consumption causes carbon dioxide emissions and that GDP causes carbon dioxide emissions. For these variables, the following relationship was learned:

$$\begin{array}{c} \text{GDP} \rightarrow \text{carbon dioxide emissions} \leftarrow \\ \text{electronic power consumption} \end{array} .$$

Others We searched for data according to the extracted word “population”. We got the data of population growth. The p-value for the conditional independence test of carbon dioxide emissions and population growth given GDP was 0.2770 while that for the independence test of carbon dioxide emissions and population growth was 0.0040. Under the no additive noise assumption, the estimated intrinsic dimension was 0.8. For further analysis, IGCi inferred that population growth causes GDP and carbon dioxide emissions caused GDP as well. For these variables, the conditional independence test and causal direction inference method gave different results. For the variable population total, the conditional independence test and causal direction inference method gave different results as well. Besides, we obtained the data of “energy intensity level of primary energy”, “PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total)” and “renewable electricity output (% of total electricity output)”. For these variables, after deleting the incomplete data, IGCi inferred that carbon dioxide emissions caused GDP which was inconsistent with the existing research result. These results indicate that each method has limitations and cannot give correct results for all real world data.

5.4.2 College Entrance Rate and Current Liability

We did experiments using the government data of Japan. We found that the college entrance rate is highly correlated with the current liability. We did independence test of the two and found that they are dependent on each other. For the two variables, areas with high current liability tend to have high college entrance rate. We first used existing causal discovery methods to analyze the relationship between them. We analyzed the causal relationship between college entrance rate and current liability using the PNL model under the no confounding assumption. The PNL model inferred that college entrance rate caused current liability. We further analyzed the causal relationship between the two variables under the confounding assumption. We used the BMLiNGAM to infer the causal direction between college entrance rate and current liability. The BMLiNGAM inferred that college entrance rate caused current liability as well. The inferred results of the two methods are consistent with each

other. However, the correlation may be induced by other unconsidered variables and direct causal relationships may not exist. We collected explanations and used our proposed framework to analyze the relationship between the two. The collected explanations were from 20 workers. The number of collected characters totally was 1818. For this task, the average number of characters each worker answered was 90.9.

Word Extraction

Causal Relationship Extraction Method The explanations of the correlation between college entrance rate and current liability were collected using crowdsourcing platform. We first extracted the causal relationships of the collected texts using the keywords “tame” [101]. Six phrases were extracted with the word “tame”. In the phrases with the word “tame”, about 67 percentage of them fulfilled one of the four morphological conditions as discussed in [101]. For this pairs, the following causal relationships were extracted:

- Cities are blessed with infrastructure.
- Since the housing environment is in place, housing loan is heavily borrowed.
- Since (poor people) do not have expensive shopping, the household is usually above the average in finance.
- Enter higher school for employment.

Compared with GDP and carbon dioxide emissions, the phrases with “tame” decreased. Workers can give more explanations if the topic is generally discussed. For some unfamiliar spurious correlation, workers may give guesses with fewer “tame” phrases. We extracted the nouns contained in the adverbial verb phrases. The extracted words were shown in Table 5.2.

LDA We obtained the ten most related words using the LDA model. The extracted words were shown in Figure 5-8. We searched for the data about the related words. In the extracted 10 words, 40 percentage of them have been found the related data.

Table 5.2: Extracted Words Using Causal Relationship Extraction Method (Experiment ii)

city
infrastructure
housing
loan
shopping
finance
average
household
employment
school

We filtered the unrelated words using machine learning methods. The detailed results will be shown in the following. We compared the performance of causal relationship extraction method and LDA using the metrics of recall and precision. The experimental results were shown in Figure 5-9. Causal relationship extraction method and LDA have the same precision and recall for this pair.

Causal Analysis Result

The following words has been extracted using the above two methods in which some are decided to be useful since extensive causal structure has been learned after adding the mentioned variable. The extracted words and detailed causal analysis results are discussed in the following.

Number of high schools The data was searched for according to the word “school”. The data was obtained from e-Stat which is a portal site for Japanese government statistics⁶. We tested the conditional independence between college entrance rate and current liability given number of high schools. The null hypothesis of conditional independence was accepted with the p-value 0.0363. The intrinsic estimation result showed that number of high schools was a possible confounder as well. We then inferred the causal direction between college entrance rate and number of high

⁶<http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001052235&cycode=0>

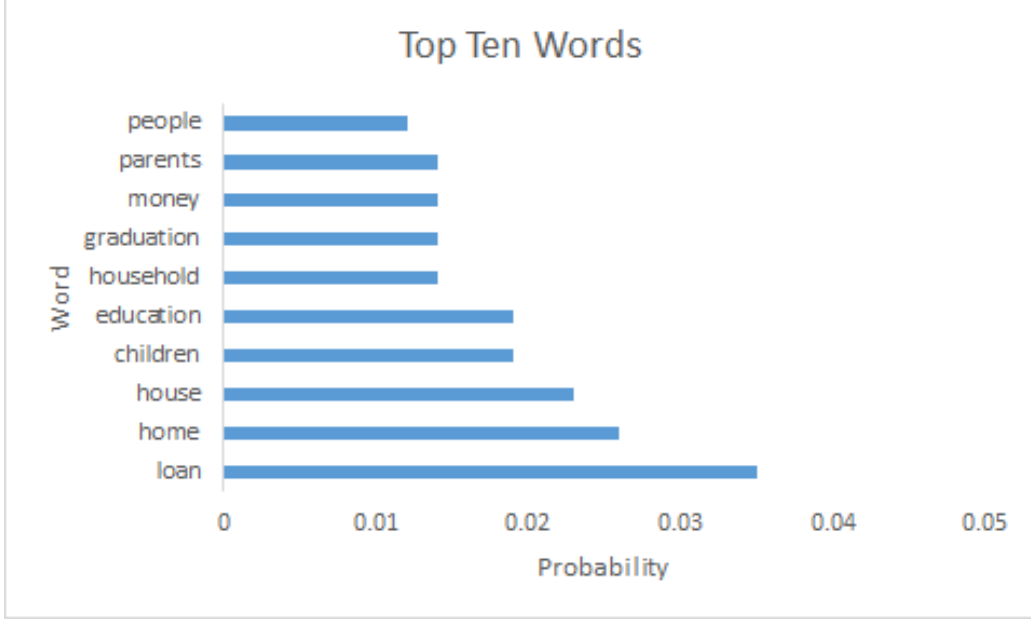


Figure 5-8: Top Ten Words (Experiment ii)

schools. The IGCi method inferred that college entrance rate caused the number of high schools. For the causal direction between current liabilities and number of high schools, the IGCi method inferred that the number of high schools caused current liabilities. The following causal relationships were obtained through the proposed framework:

$$\begin{aligned} &\text{college entrance rate} \rightarrow \text{number of high schools} \\ &\quad \rightarrow \text{current liability} \end{aligned}$$

Salary The data of “salary” was searched for according to the word “money” and we obtained the data from e-Stat. We first analyzed the data under the additive noise assumption. We tested the independence between college entrance rate and current liability given salary. The p-value was 0.0094 and the null hypothesis of conditional independence was rejected. The result showed that “salary” was not a common cause for the two and further analysis was needed. We then tested the independence between salary and current liability. The p-value was 0 and the two were dependent on each other. When further testing the conditional independence between salary

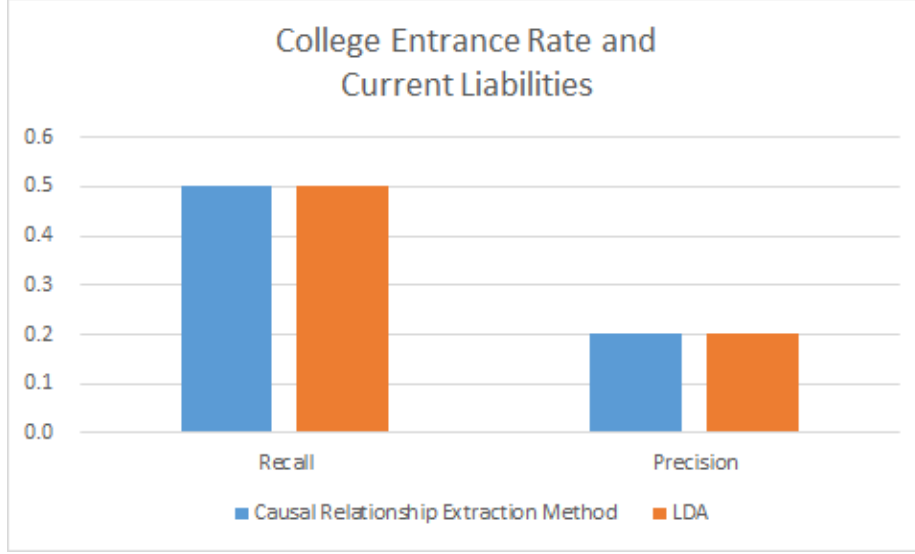


Figure 5-9: Recall and Precision of Causal Relationship Extraction Method and LDA (Experiment ii)

and current liability given college entrance rate, the null hypothesis was accepted and the two were conditionally independent. When analyzing the data under no additive noise assumption, the estimated intrinsic dimension was 1. The experimental result showed that entrance rate can be a common cause or lies in the middle between salary and current liability. For further analysis, we used causal direction inference method IGCI to infer the causal direction between salary and college entrance rate, current liability and college entrance rate. IGCI inferred that college entrance rate causes salary and current liability. The following relationship was learned through the proposed framework:

$$\text{salary} \leftarrow \text{college entrance rate} \rightarrow \text{current liability} \quad .$$

Debt ratio for housing and land (vs. current liability) This data was searched for according to the words: “housing” and “house”. We used the causal discovery workflow proposed in Figure 5-4 to analyze the relationship between the three. We first analyzed the data under the additive noise assumption. We tested the independence of college entrance rate and current liability given debt ratio for

housing and land. The p-value was 5.40×10^{-4} which means the null hypothesis was rejected. We further tested the conditional independence between current liability and debt ratio for housing and land given college entrance rate. The p-value was 0.6408. When we tested the independence between the current liability and debt ratio for housing, the p-value was 0.0410. When analyzing the data under no additive noise assumption, we got the estimated intrinsic dimension 1.6. For further study, we calculated the causal direction between current liability and college entrance rate, debt ratio for housing and land and college entrance rate. IGCI inferred that debt for housing and land causes college entrance rate and college entrance rate causes current liability. The learned causal relationship was as the following:

$$\begin{aligned} &\text{debt ratio for housing and land} \rightarrow \\ &\text{college entrance rate} \rightarrow \text{current liability} \quad . \end{aligned}$$

Education expenses ratio This data was searched for according to the word “education”. We first analyzed the data using independence test under the additive noise assumption. The result showed that education expenses ratio was independent from the college entrance rate and current liability. When estimating intrinsic dimension of the three variables, the result was 1.7. The decisions under additive noise assumption and no additive noise assumption were not consistent with each other. In our experiments, we chose to do further analysis if one method decided that the variable is a possible common cause. When further analyzing the causal direction between current liability and education expenses ratio, IGCI decided that education expenses ratio causes current liability. When further analyzing the causal direction between college entrance rate and education expenses ratio, IGCI decided that education expenses ratio caused college entrance rate.

$$\begin{aligned} &\text{current liability} \leftarrow \text{education expenses ratio} \\ &\quad \rightarrow \text{college entrance rate} \quad . \end{aligned}$$

Others We searched for data according to the keyword “house” and the data of “housing expenses ratio” was got. We tested the independence between housing expenses ratio and college entrance rate or current liability. The result showed that housing expenses ratio was independent from the two variables. When we estimate the intrinsic dimension of the data, the result was 2.1 which did not lie in the confidence interval $[0.7, 1.7]$. Although unrelated words may be included in the collected explanations, the verification parts using machine learning methods can detect them. Besides, we obtained the data of consumption expenditure (1month per household) according to the extracted word “household” and the data of employment rate according to the extracted word “employment”. The experimental results of conditional independence tests are reverse to that of the causal direction inference method. Each method has limitations and cannot give correct inference for all the real world data. For these data, the proposed framework cannot figure it out either.

5.4.3 Summary of the Experiments

We conducted experiments using the open data of world bank and Japanese government. The causal relationships between GDP and carbon dioxide emissions, college entrance rate and current liability were analyzed. The intrinsic dimension estimation results and obtained causal relationships were summarized in Table 5.3. In the eleven relationships we obtained, five of them contained new common causes. The other ones helped understand the relationship between GDP and carbon dioxide emissions, current liability and college entrance rate as well. We collected explanations from 50 workers for the task of GDP and carbon dioxide emissions with seven causal relationships extracted. The useful causal relationships (% of each worker) was 0.14. For the second task, we obtained four causal relationships from 20 workers. The useful causal relationships (% of each worker) was 0.2. More causal relationships were collected after increasing the number of workers but the percentage decreased a little. The appropriate number of workers may vary between different tasks. From our empirical study, the least number of workers would be 20 since insufficient texts would affect the performance of machine learning methods used in the proposed framework, e.g.

LDA model. There is no upper limits of crowdsourcing workers generally but many workers would give repeated answers when the number is large.

We conducted experiments on some other topics such as “overseas travel rate” and “college entrance rate” as well. However, open data cannot be found for most of the extracted keywords such as “chance”, and “richness”. The following factors can limit the applicability of the proposed framework: noised answers from workers, limited open data, conflicts between machine learning methods. Although the noise of answers can be filtered by machine learning methods to some degree, improving the quality of collected answers will improve the performance of the proposed framework. In our experiments, only half of them can be found. Although not all the variables can be found about the open data, we can find more data if we use semantic search using RDF [107]. Finally, improving the accuracy of machine learning methods would improve the performance of the proposed framework as well.

5.5 Summary

We developed a framework for exploratory causal analysis of open data. The proposed framework mainly contains three parts: crowd-based explanation collections, keywords extraction with NLP methods and exploratory causal analysis using open data. The task of crowd-based explanation generation was simple and can be conducted in the general crowdsourcing marketplace. Causal relationship extraction methods and the LDA model were used to extract the most related words to the topic. New data was searched for according to the extracted words. Finally, machine learning methods were used to filter the unrelated words and learn the causal relationship between the variables. Compared with the existing methods, extended causal structure can be got by using the proposed framework. When analyzing causal relationships between variables, some conflicts may exist. Since no methods can give all correct inference, we conducted further analysis at least one method decided that the extracted word may be confounding. Although the proposed framework has some limitations, the experimental results showed its effectiveness when doing causal dis-

Table 5.3: Summary of the Experiments

Experiment	Variable	Intrinsic Dimension Result	Obtained Causal Relationships
i	Access to electricity	0.8	$GDP \leftarrow$ Access to electricity $\rightarrow$ Carbon dioxide emissions
i	Exports of goods and services	0.9	$GDP \leftarrow$ Exports of goods and services $\rightarrow$ Carbon dioxide emissions
i	Imports of goods and services	0.9	$GDP \leftarrow$ Imports of goods and services $\rightarrow$ Carbon dioxide emissions
i	Price level ratio of PPP conversion factor (GDP) to market exchange rate	1.1	$GDP \leftarrow$ Price level ratio of PPP conversion factor (GDP) to market exchange rate $\rightarrow$ Carbon dioxide emissions
i	GNI	0.5	$GNI \rightarrow GDP \rightarrow$ Carbon dioxide emissions
i	Fossil fuel energy consumption	0.7	Fossil fuel energy consumption $\rightarrow GDP \rightarrow$ Carbon dioxide emissions
i	Electric power consumption	0.7	$GDP \rightarrow$ Carbon dioxide emissions $\leftarrow$ Electronic power consumption
ii	Number of high schools	1.4	College entrance rate $\rightarrow$ Number of high schools $\rightarrow$ Current liability
ii	Salary	1.0	$Salary \leftarrow$ College entrance rate $\rightarrow$ Current liability
ii	Debt ratio for housing and land(vs. current liability)	1.6	Debt ratio for housing and land(vs. current liability) $\rightarrow$ College entrance rate $\rightarrow$ Current liability
ii	Education expenses ratio	1.7	Current liability $\leftarrow$ Education expenses ratio $\rightarrow$ College entrance rate

covery in open data.

Chapter 6

Conclusion and Future Work

We compared three state-of-the-art existing methods: ANM, PNL, IGCI for causal discovery in binary case with real world data. We first showed the performance of current methods under different decision rates. The experiment result is that the latest IGCI model gives the best performance. To check whether the decision has been made reasonably, we adopt the evaluating methods for a binary classifier: ROC and AUC. The result is that IGCI get small AUC value for the decision “X causes Y” because it makes several wrong decisions when the threshold is set large. Compared with other methods, the experiment results of ANM are relatively stable. Finally, we compared the time cost for ANM, PNL and IGCI when making a decision. The result is that IGCI is the fastest even when the dataset is large. PNL costs the most time to give a decision.

In the three models, IGCI can give good performance when the noise is little and the data is not much discretized. How to improve the performance of IGCI in much noise case and how to deal with discretized data is a remaining task. Compared with other methods, the performance of ANM is relatively stable. However, overfitting should be avoided for ANM. This is because overfitting will change the value of residuals and affect the result of independence test using HSIC. In the three models, PNL is the most generalized model which takes into account nonlinear effect of causes, additive inner noise and sensor distortion outside. However, modeling nonlinearities of f_1 and f_2^{-1} takes much time for PNL. We concluded the strong points and weaknesses

Table 6.1: Pros. and Cons.

Model	Strong Ponits	Weaknesses
ANM	stable performance in general	overfitting could occur
IGCI	good performance in low noise case, high efficiency	susceptible to noise, unsuitable to discrete data
PNL	relatively effective to noise	time cost is large

of the three models in Table 6.1.

Testing the responses of the models to “confounding” showed that the ANM and the PNL model can avoid the effect of “confounding” to some degree while IGCI is the most susceptible to confounding. Neither of these models can avoid the effect of confounding totally. Thus, we proposed using intrinsic dimension estimation to determine a possible common cause for two variables. Simulated application showed that the proposed method worked well for both linear and non-linear functions. Testing using different types of noise showed that it generally worked well for different types of added noise. In particular, it worked better than a kernel-based conditional independence test for Poisson noise. Testing of how the estimated intrinsic dimension is affected by different types of distributions showed that the estimated dimension is nearly not affected by the type of distribution. Simulation of mixed pattern showed that the proposed method can still tell a possible common cause when it is mixed with causal relationship. Finally, experiments using variables from the CauseEffectPairs dataset showed that the proposed method can give correct inferred results for real world data.

Finally, we proposed a framework for exploratory causal analysis of open data, in which possible confounding variables are collected and incrementally tested from a large volume of open data. The proposed framework takes into account the effect of possible confounding in causal analysis first by collecting explanations of the correlation between variables using a crowdsourcing platform. Keywords are then extracted using natural language processing methods. The related open data are searched for according to the extracted keywords. Finally, the collected explanations are tested

using several machine learning methods. We conducted experiments using World Bank data and Japanese open government data. The experimental results showed that our proposed framework can do causal analysis while considering the effect of possible confounders. We used two natural language processing methods in the keyword extraction process and the experimental results showed that specific causal relationship extraction method worked better than the LDA model. Several machine learning methods were included in the causal analysis workflow. (Conditional) independence test and intrinsic dimension estimation were used to filter the unrelated words and causal direction inference method was used to learn the exact causal patterns. Conflicts may exist between the inference results of these methods since the real world causal relationships are usually complex and no machine learning methods can give all correct inferences. Although conflicts exist, including of different types of machine learning methods are needed for correct inference.

We will further investigate how to detect and reduce the effect of confounding when multiple common causes exist. Currently, the proposed method can deal with a specific common cause. The estimation results may be affected when multiple common causes exist. How the inferred result is affected and how to avoid it are still needed to be researched. Besides, the proposed framework for crowd-based causal analysis of open data can be further developed with linked open data included. The automation can be improved after including the linked open data. To develop a human computation system, the conflicts between machine learning methods are still needed to be resolved. Besides, some machine learning methods are time consuming. Improving this weakness remains as our future work as well.

Appendix A

Pairs of Variables Used in Experiment

Table A.1: Pairs of Causal Variables

Cause	Effect
altitude	temperature
altitude	precipitation
longitude	temperature
altitude	sunshine
rings (abalone)	length
rings (abalone)	shell weight
rings (abalone)	diameter
rings (abalone)	height
rings (abalone)	whole weight
rings (abalone)	shucked weight
rings (abalone)	viscera weight
age	body mass index
age	2-hour serum insulin
age	diastolic blood pressure
age	plasma glucose concentration at 2 hours in an oral glucose tolerance test
age	height
age	weight
population with sustainable access to improved drinking water sources	infant mortality rate
age	wage per hour

Table A.1 (continued)

age	heart rate
stock return of Hang Seng Bank	stock return of HSBC Hldgs
stock return of Hutchison Holdings	stock return of Cheung Kong Holdings
stock return of Cheung Kong Holdings	stock return of Sun Hung Kai Properties
open http connections during that minute	bytes sent during that minute
parameter between 0 and 14	gender decision
superplasticizer	compressive strength
coarse aggregate	compressive strength
water	compressive strength
age (day)	compressive strength
alcohol consumption	mcv (mean corpuscular volume)
alcohol consumption	alkphos (alkaline phosphatas)
alcohol consumption	sgpt (alanine aminotransferase)
alcohol consumption	sgot (aspartate aminotransferase)
alcohol consumption	gammagt (gamma-glutamyl transpeptidase)
displacement	mpg
outdoor temperature	indoor temperature
temperature	ozone concentration
sunspot area	global mean temperature anomalies
latitude of birth country's capital	life expectancy at birth

Table A.1 (continued)

energy use	CO2 emissions
GNI (gross national income)	life expectancy
GNI (gross national income)	under-5 mortality rate
average annual rate of change in population	average annual rate of change in total dietary consumption for total population
solar radiation	daily average temperature
PPFD (photosynthetic photon flux density)	NEP (net ecosystem productivity)
PPFDdif (photosynthetic photon flux density, diffusive)	NEP (net ecosystem productivity)
PPFDdir (photosynthetic photon flux density, direct)	NEP (net ecosystem productivity)
horsepower	acceleration
age	dividends from stock
age of child in years	concentration of GAG (glycosaminoglycan)
duration of eruption in minutes	time to next eruption in minutes
latitude	temperature
longitude	precipitation
cement	compressive strength
blast furnace slag	compressive strength
temperature	CO2 flux at night
population	employment
size of room	rent
categorical (1: Sundays+Holidays or 2: working days)	number of cars

Table A.1 (continued)

mean temperature	total snow
age	relative spinal bone mineral density
mass loss April	mass loss October
day of the year	mean daily temperature in Furtwangen
weight	mpg
clay content in soil	soil moisture at 10 cm depth
hour of the day	temperature in degrees Celsius
clay content	organic C content in soil
hour of the day	load (total electricity consumption in a region of Turkey)
fine aggregate	compressive strength
initial speed	subsequent speed
average precipitation from 1948 to 2004 in mm/day	average runoff from 1948 to 2004 in mm/day
(mean daily air temperature) year 2000, day 50	(mean daily air temperature) year 2000, day 51
(mean daily pressure at surface) year 2000, day 50	(mean daily pressure at surface) year 2000, day 51
(mean daily sea level pressure) year 2000, day 50	(mean daily sea level pressure) year 2000, day 51
(mean daily relative humidity) year 2000, day 50	(mean daily relative humidity) year 2000, day 51
time to take weekly measurements	protein content of milk produced by each cow at time X

Bibliography

- [1] Clive WJ Granger. Some recent development in a concept of causality. *Journal of econometrics*, 39(1):199–211, 1988.
- [2] Joseph Y Halpern. A modification of the halpern-pearl definition of causality. *arXiv preprint arXiv:1505.00162*, 2015.
- [3] Xiaohai Sun, Dominik Janzing, Bernhard Schölkopf, and Kenji Fukumizu. A kernel-based causal learning algorithm. In *Proceedings of the 24th international conference on Machine learning*, pages 855–862. ACM, 2007.
- [4] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.
- [5] Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf. Causality: Objectives and assessment. In *Causality: Objectives and Assessment*, pages 1–42, 2010.
- [6] Kun Zhang, Bernhard Schölkopf, and Dominik Janzing. Invariant gaussian process latent variable models and application in causal discovery. *arXiv preprint arXiv:1203.3534*, 2012.
- [7] Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.
- [8] Zhitang Chen, Kun Zhang, and Laiwan Chan. Causal discovery with scale-mixture model for spatiotemporal variance dependencies. In *Advances in Neural Information Processing Systems*, pages 1727–1735, 2012.
- [9] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.
- [10] Zhitang Chen, Kun Zhang, Laiwan Chan, and Bernhard Schölkopf. Causal discovery via reproducing kernel hilbert space embeddings. *Neural computation*, 26(7):1484–1517, 2014.

- [11] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning, JMLR: W&CP, Lille, France*, 2015.
- [12] Biwei Huang, Kun Zhang, and Bernhard Schölkopf. Identification of time-dependent causal model: A gaussian process treatment. In *the 24th International Joint Conference on Artificial Intelligence., Machine Learning Track-Buenos, Argentina*, pages 3561–3568, 2015.
- [13] Judea Pearl et al. Models, reasoning and inference, 2000.
- [14] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.
- [15] Xiaohai Sun, Dominik Janzing, and Bernhard Schölkopf. Causal inference by choosing graphs with most plausible markov kernels. In *ISAIM*, 2006.
- [16] Xiaohai Sun, Dominik Janzing, and Bernhard Schölkopf. Distinguishing between cause and effect via kernel-based complexity measures for conditional distributions. In *ESANN*, pages 441–446, 2007.
- [17] Xiaohai Sun, Dominik Janzing, and Bernhard Schölkopf. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomputing*, 71(7):1248–1256, 2008.
- [18] Dominik Janzing, Patrik O Hoyer, and Bernhard Schölkopf. Telling cause from effect based on high-dimensional observations, 2009.
- [19] Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *AISTATS*, 2015.
- [20] Clarice R Weinberg. Toward a clearer definition of confounding. *American Journal of Epidemiology*, 137(1):1–8, 1993.
- [21] Penelope P Howards, Enrique F Schisterman, Charles Poole, nzy S Kaufman, and Clarice R Weinberg. “toward a clearer definition of confounding” revisited with directed acyclic graphs. *American journal of epidemiology*, 176(6):506–511, 2012.
- [22] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.
- [23] Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

- [24] Shohei Shimizu and Kenneth Bollen. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *Journal of Machine Learning Research*, 15(1):2629–2652, 2014.
- [25] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.
- [26] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.
- [27] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.
- [28] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.
- [29] Bwo-Nung Huang, Ming Jeng Hwang, and Chin Wei Yang. Causal relationship between energy consumption and GDP growth revisited: a dynamic panel data approach. *Ecological Economics*, 67(1):41–54, 2008.
- [30] Judea Pearl. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.
- [31] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [32] Clive WJ Granger. Some recent development in a concept of causality. *Journal of Econometrics*, 39(1-2):199–211, 1988.
- [33] Joseph Y Halpern. A modification of the Halpern-Pearl definition of causality. In *International Joint Conference on Artificial Intelligence*, pages 3022–3033, 2015.
- [34] Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(Nov):3387–3439, 2012.
- [35] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.
- [36] Robert E Tillman, Arthur Gretton, and Peter Spirtes. Nonlinear directed acyclic structure learning with weakly additive noise models. In *Advances in Neural Information Processing Systems*, pages 1847–1855, 2009.

- [37] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.
- [38] Jonas Peters, Dominik Janzing, Arthur Gretton, and Bernhard Schölkopf. Detecting the direction of causal time series. In *Proceedings of the 26th annual international conference on machine learning*, pages 801–808. ACM, 2009.
- [39] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162, 2013.
- [40] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.
- [41] Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.
- [42] Yonghong Chen, Govindan Rangarajan, Jianfeng Feng, and Mingzhou Ding. Analyzing multiple nonlinear time series with extended granger causality. *Physics Letters A*, 324(1):26–35, 2004.
- [43] Naji Shajarisales, Dominik Janzing, Bernhard Schoelkopf, and Michel Besserve. Telling cause from effect in deterministic linear dynamical systems. *arXiv preprint arXiv:1503.01299*, 2015.
- [44] Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *Proceedings of 32th International Conference on Machine Learning (ICML 2015)*, 2015.
- [45] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(May):1709–1731, 2010.
- [46] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248, 2011.
- [47] Kun Zhang and Lai-Wan Chan. Extensions of ica for causality discovery in the hong kong stock market. In *International Conference on Neural Information Processing*, pages 400–409. Springer, 2006.
- [48] Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using non-linear acyclic causal models. In *Journal of machine learning research, workshop and conference proceedings (NIPS 2008 causality workshop)*, volume 6, pages 157–164, 2008.

- [49] Kun Zhang, Zhikun Wang, and Bernhard Schölkopf. On estimation of functional causal models: Post-nonlinear causal model as an example. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 139–146. IEEE, 2013.
- [50] Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*, 2012.
- [51] Kun Zhang, Jiji Zhang, and Bernhard Schölkopf. Distinguishing cause from effect based on exogeneity. *arXiv preprint arXiv:1504.05651*, 2015.
- [52] Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pages 1687–1695, 2010.
- [53] Judea Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):1–62, 2010.
- [54] Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(May):1643–1662, 2010.
- [55] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696, 2009.
- [56] Dominik Janzing, Jonas Peters, Joris Mooij, and Bernhard Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 249–257. AUAI Press, 2009.
- [57] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.
- [58] Shohei Shimizu and Kenneth Bollen. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *Journal of Machine Learning Research*, 15(1):2629–2652, 2014.
- [59] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, page 3. Springer Berlin Heidelberg, 2016.
- [60] Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, pages 1–11, 2017.

- [61] Lei Duan, Satoshi Oyama, Haruhiko Sato, and Masahito Kurihara. Multi-emotion estimation in narratives from crowdsourced annotations. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 91–100. ACM, 2015.
- [62] Jingjing Wang, Satoshi Oyama, Masahito Kurihara, and Hisashi Kashima. Learning an accurate entity resolution model from crowdsourced labels. In *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, pages 103:1–103:8. ACM, 2014.
- [63] Gianluca Demartini, Beth Trushkowsky, Tim Kraska, and Michael J Franklin. CrowdQ: crowdsourced query understanding. In *Proceedings of the 6th Biennial Conference on Innovative Data Systems Research*, 2013.
- [64] Satoshi Oyama, Yukino Baba, Ikki Ohmukai, Hiroaki Dokoshi, and Hisashi Kashima. Crowdsourcing chart digitizer: task design and quality control for making legacy open data machine-readable. *International Journal of Data Science and Analytics*, 2(1-2):45–60, 2016.
- [65] Shusaku Egami, Takahiro Kawamura, Kouji Kozaki, and Akihiko Ohsuga. Construction of linked urban problem data with causal relations using crowdsourcing. In *Proceedings of 2017 IIAI International Congress on Advanced Applied Informatics*, pages 814–819. IEEE, 2017.
- [66] Shusaku Egami, Takahiro Kawamura, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. Schema design of illegally parked bicycles LOD. In *Proceedings of 2016 IIAI International Congress on Advanced Applied Informatics*, pages 692–697. IEEE, 2016.
- [67] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. Strategies for crowdsourcing social data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 227–236. ACM, 2012.
- [68] Josh C Bongard, Paul DH Hines, Dylan Conger, Peter Hurd, and Zhenyu Lu. Crowdsourcing predictors of behavioral outcomes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(1):176–185, 2013.
- [69] Kirsten E Bevelander, Kirsikka Kaipainen, Robert Swain, Simone Dohle, Josh C Bongard, Paul DH Hines, and Brian Wansink. Crowdsourcing novel childhood predictors of adult obesity. *PloS one*, 9(2):e87756, 2014.
- [70] Xuan Liu, Meiyu Lu, Beng Chin Ooi, Yanyan Shen, Sai Wu, and Meihui Zhang. CDAS: a crowdsourcing data analytics system. *Proceedings of the VLDB Endowment*, 5(10):1040–1051, 2012.
- [71] Andreas Holzinger. Human-computer interaction and knowledge discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In *International Conference on Availability, Reliability, and Security*, pages 319–328. Springer, 2013.

- [72] Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.
- [73] Robert E Tillman, Arthur Gretton, and Peter Spirtes. Nonlinear directed acyclic structure learning with weakly additive noise models. In *Advances in Neural Information Processing Systems*, pages 1847–1855, 2009.
- [74] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.
- [75] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.
- [76] Carl Edward Rasmussen and Hannes Nickisch. GPML code, 2010.
- [77] Carl Edward Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (gpml) toolbox. *Journal of Machine Learning Research*, 11(Nov):3011–3015, 2010.
- [78] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129, 2005.
- [79] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. A kernel statistical test of independence. In *NIPS*, volume 20, pages 585–592, 2008.
- [80] Joris M Mooij, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. CauseEffectPairs repository, 2014.
- [81] M. Lichman. UCI machine learning repository, 2013.
- [82] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773*, 2014.
- [83] Joris M Mooij, Dominik Janzing, and B Schölkopf. Distinguishing between cause and effect. In *NIPS Causality: Objectives and Assessment*, pages 147–156, 2010.
- [84] Song Jing, Oyama Satoshi, Sato Haruhiko, and Kurihara Masahito. Evaluation of causal discovery models in bivariate case using real world data. In *Proceedings of the International MultiConference of Engineers and Computer Scientists 2016*, pages 291–296, 2016.
- [85] Richard Scheines. An introduction to causal inference. *Causality in crisis*, pages 185–199, 1997.
- [86] Christian Scholz. Strata-a method for strategic analysis of complex systems. *European Journal of Operational Research*, 24(1):168–177, 1986.

- [87] 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):1–21, 2010.
- [88] Mohamad Amin Pourhoseingholi, Ahmad Reza Baghestani, and Mohsen Vahedi. How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology From Bed to Bench*, 5(2):79, 2012.
- [89] KJ Jager, C Zoccali, A Macleod, and FW Dekker. Confounding: what it is and how to deal with it. *Kidney International*, 73(3):256–260, 2008.
- [90] Dipankor Coondoo and Soumyananda Dinda. Causality between income and emission: a country group-specific econometric analysis. *Ecological Economics*, 40(3):351–367, 2002.
- [91] Jing Song, Satoshi Oyama, and Masahito Kurihara. Tell cause from effect: models and evaluation. *International Journal of Data Science and Analytics*, 4(2):99–112, 2017.
- [92] Jing Song, Satoshi Oyama, and Masahito Kurihara. A framework for crowd-based causal analysis of open data. In *Proceedings of the 2018 International Conference on Systems, Man, and Cybernetics*. IEEE, 2018.
- [93] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.
- [94] John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.
- [95] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *The Theory of Chaotic Attractors*, pages 170–189, 2004.
- [96] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer, 2007.
- [97] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009.
- [98] Carl Hartung, Adam Lerer, Yaw Anokwa, Clint Tseng, Waylon Brunette, and Gaetano Borriello. Open data kit: tools to build information services for developing regions. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, pages 18:1–18:12. ACM, 2010.

- [99] Christopher Bryan Davis. *Making sense of open data: from raw data to actionable insight*. PhD thesis, Delft University of Technology, 2012.
- [100] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning, 2012.
- [101] Takashi Inui, Kentaro Inui, and Yuji Matsumoto. Acquiring causal knowledge from text using the connective marker tame. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(4):435–474, 2005.
- [102] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 856–864, 2010.
- [103] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics, 2002.
- [104] David M Blei and John D Lafferty. Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, pages 147–154. MIT Press, 2005.
- [105] Jing Song, Satoshi Oyama, and Masahito Kurihara. Identification of possible common causes by intrinsic dimension estimation. In *Proceedings of the 6th International Conference on Big Data and Smart Computing*. IEEE, 2019.
- [106] Steen A Andersson, David Madigan, Michael D Perlman, et al. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.
- [107] Ramanathan Guha, Rob McCool, and Eric Miller. Semantic search. In *Proceedings of the 12th International Conference on World Wide Web*, pages 700–709. ACM, 2003.