

Fall 12-2019

## NONLINEAR FUNCTIONAL MODELS ON CAUSAL INFERENCE

RONG JIAO

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/uthsph\\_dissertsopen](https://digitalcommons.library.tmc.edu/uthsph_dissertsopen)



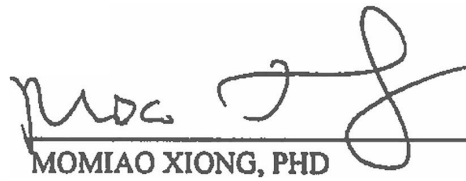
Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

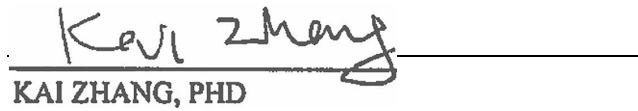
NONLINEAR FUNCTIONAL MODELS ON CAUSAL INFERENCE

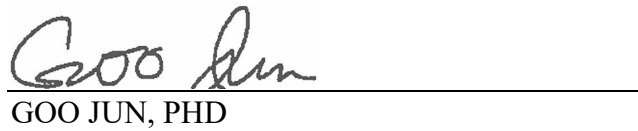
by

RONG JIAO, B.S, M.A

APPROVED:

  
MOMIAO XIONG, PHD

  
KAI ZHANG, PHD

  
GOO JUN, PHD

  
DEAN, THE UNIVERSITY OF TEXAS  
SCHOOL OF PUBLIC HEALTH

Copyright

by

RONG JIAO, M.A, Ph.D.

2019

## DEDICATION

To my family members:

Dianhua Jiao and Yujuan Ma

NONLINEAR FUNCTIONAL MODELS ON CAUSAL INFERENCE

by

RONG JIAO

B.S, Fudan University, 2011

M.A, Columbia University, 2012

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS

SCHOOL OF PUBLIC HEALTH

Houston, Texas

December, 2019

## PREFACE

The PhD training helps me to become a qualified statistical scientist, to investigate questions independently and to explore the new world. I am interested in statistical methods applied to medical research, which helps improve human healthcare. Causal inference, especially, has shown a great role in scientific research, e.g., network analysis for genetic imaging analysis and counterfactual modeling for go/no-go decision in clinical study. I am glad to have exciting results during the graduate research, and had several publications that show my work as milestones.

## ACKNOWLEDGMENTS

I here express gratitude to my dissertation advisor Dr. Momiao Xiong. He mentored me in methodological research and trained me to be a qualified biostatistician. Also, he led me to the field of causal inference, which makes my most contribution to this emerging research topic. I also want to appreciate the great help from my committee members: Dr. Kai Zhang, Dr. Goo Jun and Dr. Wenyaw Chan. As talented scholars in epidemiology, bioinformatics and biostatistics, they have offered valuable advice for my research. I also need to thank my lab fellows, Nan Lin, Zixin Hu, Wenjia Peng and Yuanyuan Liu. They offered helpful instructions in data analysis and great comments on innovative research, which helps me solve scientific questions and improve the novel methodologies.

# NONLINEAR FUNCTIONAL MODELS ON CAUSAL INFERENCE

Rong Jiao, M.A, PhD

The University of Texas

School of Public Health, 2019

Dissertation Chair: Momiao Xiong, PhD

## Abstract

Statistical analysis has experienced significant progress on association study, but it remains elusive to understand the etiology and mechanism of complex phenotypes. As a major analytical platform, association analysis may hamper the theoretic development of biomedical science and its application. Thus, many researchers suggest making the transition from association to causation.

The mainstream of research in genetics and epigenetics data analysis focuses on statistical association or exploring statistical dependence between variables. Despite significant progress in dissecting the genetic architecture of complex diseases by genome-wide association studies (GWAS), the signals identified by association analysis can only explain a small proportion of the heritability of complex diseases. A large fraction of risk genetic variants is still hidden. Finding causal SNPs only by searching the set of associated SNPs



may miss many causal variants. Using association analysis as a major analytical platform for the complex data analysis is a key issue that hampers the theoretic development of genomic science and its application in practice. Causality shapes how we view and understand mechanism of complex diseases.

To explore bivariate causal discovery, I will introduce independence of cause and mechanism (ICM) as a basic principle, using additive noise model (ANM) as a major tool for bivariate causal discovery. Large-scale simulations will be performed to evaluate the feasibility of the ANM for bivariate causal discovery. Second, I will introduce machine-learning methods on confounder detection, to further analyze the case of no causation but having association. Last, I will expand causal analysis from bivariate discovery to network analysis, considering the causal relation between multiple variables. Entropy methods will be introduced to deal with the case of multiple factors and one cause, and structural equation models with nonlinear function scores will be applied to network analysis.

# CONTENTS

List of Tables .....	i
List of Figures .....	iv
1. Background .....	1
1.1 Literature Review .....	1
1.1.1 Bivariate Causal Inference .....	3
1.1.2 Hilbert-Schmidt Independence Criterion .....	10
1.1.3 Confounder Identification .....	12
1.1.4 Entropy Methods .....	12
1.1.5 Structural Equation Model .....	13
1.2 Public Health Significance .....	14
1.3 Specific Aims .....	16
2. Methods .....	18
2.1 Overall Study Design .....	18
2.1.1 Simulation Studies .....	18
2.1.2 Real Data Application .....	18
2.2 Methods for Aim 1(a): To Develop Bivariate Causal Inference Test for Continuous Variables .....	19
2.2.1 Statistical Modeling .....	19
2.2.2 Simulation Settings .....	21
2.3 Methods for Aim 1(b): To Develop Bivariate Causal Inference Test for Discrete Variables .....	24
2.3.1 Statistical Modeling .....	24
2.3.2 Simulation Settings .....	26
2.4 Methods for Aim 2: To Develop Nonlinear Functional Models for Causal Inference on Genomic Variables with Measured Confounders .....	32

2.4.1 Statistical Modeling .....	32
2.4.2 Simulation Settings .....	33
2.5 Methods for Aim 3(a): To Develop the Causal Inference Test for Multiple Causes and One Effect .....	34
2.5.1 Statistical Modeling .....	34
2.5.2 Simulation Settings .....	37
2.6 Methods for Aim 3(b): To Develop the Causal Network Model for High-dimensional Data .....	39
2.6.1 Statistical Modeling .....	39
2.6.2 Simulation Settings .....	42
2.7 Methods for Real Data Application on Proposed Aims .....	43
2.8 Declaration on Human Subjects .....	44
3. Results .....	45
3.1 Bivariate Causal Discovery for Continuous Variables in Genetic and Imaging Data Analysis .....	45
3.1.1 Introduction .....	45
3.1.2 Data and Notation .....	48
3.1.3 Linear Correlation and Causation .....	55
3.1.4 Application to KEGG Pathway .....	65
3.1.5 Application to Imaging Analysis .....	68
3.2 Bivariate Causal Discovery for Discrete Variables .....	77
3.2.1 Simulation Results .....	77
3.2.2 Application to Genome-wide Causal Study of Schizophrenia .....	82
3.2.3 Application to Disease Prediction .....	84
3.2.4 Application to Linkage Disequilibrium .....	85
3.3 Confounder Detection .....	91
3.3.1 Data and Notation .....	91
3.3.2 Simulation Results .....	93

3.3.3 Application to Gene Expression Data .....	97
3.4 Nonlinear Causal Network .....	98
3.4.1 Data and Notation .....	98
3.4.2 Simulation Results .....	100
3.4.3 Real Data Analysis .....	101
4. Discussion .....	103
5. Conclusion .....	108
6. References .....	112

## LIST OF TABLES

Table 2.1 Type I error rates of the ANMs between two continuous variables, assuming independence.....	22
Table 2.2 Type I error rates of the ANMs between two continuous variables in the presence of association.....	22
Table 2.3 Power of the ANMs between two continuous variables .....	23
Table 2.4 Rules to determine causal direction .....	25
Table 2.5 Type I error rates of the ANMs between two discrete variables, assuming no association and no causation .....	27
Table 2.6 Probability mass functions for X and Y .....	28
Table 2.7 Power of the ANMs between two discrete variables, assuming no association but having causation .....	28
Table 2.8 Type I error rates of the ANMs between two discrete variables in the presence of association .....	29
Table 2.9 Power of the ANMs between two discrete variables in the presence of association .....	30
Table 2.10 Type I error rates of the ANMs between two discrete variables in the presence of linkage disequilibrium .....	32
Table 2.11 Type I error rates of causal inference test on multiple causes and one effect .....	38
Table 2.12 Powers of causal inference test on multiple causes and one effect .....	39

Table 3.1 Type I error rates of the ANMs for testing causation, assuming no association .....	63
Table 3.2 Type I error rates of the ANMs for testing causation in the presence of association .....	64
Table 3.3 Power of the ANMs for detecting causation between two variables .....	65
Table 3.4 Accuracy of the ANMs and other six methods for inferring Wnt pathway .....	66
Table 3.5 P-values for assessing association and causal relationships between the cholesterol and brain region .....	71
Table 3.6 P-values for assessing association and causal relationships between the working memory and brain region .....	75
Table 3.7 Average type 1 error rates of the statistics for testing causal relationships between SNP and disease in the presence of association .....	79
Table 3.8 P-values of top 15 SNPs that had significant causal relationships with schizophrenia .....	83
Table 3.9 Ten-fold cross-validated accuracy and AUC for SCZ risk prediction of using top 15 causal SNPs and association SNPs .....	85
Table 3.10 Power to detect association between SNP1 and Disease .....	89
Table 3.11 Power to detect causation between SNP1 and disease .....	90
Table 3.12 Type I error rates of causal test between SNP2 and disease .....	90
Table 3.13 Power of test for association between SNP2 and disease .....	90
Table 3.14 P-values for causation and association tests of 20 neighboring SNPs of causal SNP rs6578689 .....	91

Table 3.15 Variables with possible hidden confounder in RIG-I-like receptor signaling pathway .....	97
Table 3.16 Power and FDR of four methods in causal network analysis .....	101
Table 3.17 Power of detection of four causal methods on two datasets .....	102

## LIST OF FIGURES

Figure 2.1 Two independent continuous variables.....	21
Figure 2.2 Two associated continuous variables without causation .....	22
Figure 2.3 Two independent continuous variables .....	23
Figure 2.4 Two independent discrete variables .....	26
Figure 2.5 Two causal discrete variables that do not show association .....	27
Figure 2.6 Two associated discrete variables without causation .....	29
Figure 2.7 Two causal discrete variables .....	29
Figure 2.8 One casual SNP and its neighboring associated SNP .....	31
Figure 2.9 No confounder exists .....	33
Figure 2.10 Confounder exists .....	34
Figure 2.11 Multiple covariates are independent of the response .....	37
Figure 2.12 Multiple covariates causes the response .....	38
Figure 2.13 Causal network .....	42
Figure 3.1 An example of joint distribution $p(x, y)$ generated by $Y := f(X) + E_Y$ , where $f(X) = X^3$ and $E_Y$ is uniformly distributed in $[-1, 1]$ . The interval of the red line represents the bandwidth of the conditional distribution $p_{Y X}$ .....	49
Figure 3.2 An example of joint distribution $p(x, y)$ generated by $Y := f(X) + E_Y$ , where $f(X) = X^3$ and $E_Y$ is uniformly distributed in $[-1, 1]$ . The interval of the red line the represents bandwidth of the conditional distribution $p_{X Y}$ .....	50



Figure 3.3 The data generated by $Y = 5X^2 + \varepsilon$ , where $X$ follows a uniform distribution between .....	57
Figure 3.4 False decision rates as a function of the parameter $b$ for the model 1 .....	60
Figure 3.5 False decision rates as a function of the parameter $b$ for the model 2 .....	60
Figure 3.6 The false decision rates of the ANMs for detecting the true causal direction $X \rightarrow Y$ for the model 3 .....	61
Figure 3.7 The false decision rates of the ANMs for detecting the true causal direction $X \rightarrow Y$ for the model 4 .....	61
Figure 3.8 The false decision rates of the ANMs for detecting the true causal direction $X \rightarrow Y$ for the model 5 .....	61
Figure 3.9 The false decision rates of the ANMs for detecting the true causal direction $X \rightarrow Y$ for the model 6 .....	61
Figure 3.10 The ANM-inferred network structure of the Wnt pathway. The green lines represented the inferred paths consistent to the KEGG while the gray ones represented the inferred edges absent in the KEGG .....	67
Figure 3.11 (A) A slice of the FA map from a single individual's DTI data .....	69
Figure 3.11 (B) FA map reconstruction with the first two 3D-FPC scores .....	69
Figure 3.12 Imputed FA map in Figure 11A using 3D-FPC scores and matrix completion ...	70
Figure 3.13 (A) AD and normal individuals' CHL curves .....	72
Figure 3.13 (B) Images of temporal L hippocampus region .....	72
Figure 3.14 (A) AD and normal individuals' CHL curves .....	73

Figure 3.14 (B) AD and normal individuals' working memory .....	73
Figure 3.14 (C) Images of temporal R hippocampus region .....	73
Figure 3.15 (A) AD and normal individuals' CHL curves .....	74
Figure 3.15 (B) Images of Occipital Lobe Region .....	74
Figure 3.16 The power curves of the causation test as a function of the parameter $p$ with significance levels $\alpha = 0.05$ .....	79
Figure 3.17 The power curves of the causation test as a function of the parameter $p$ with significance levels $\alpha = 0.01$ .....	80
Figure 3.18 Scatterplot of model $X = \log(T + 1) + N_X$ and $Y = (T - 1)^2 - 4 + N_Y$ .....	93
Figure 3.19 Scatterplot of $\hat{T}$ (estimated $T$ ) vs. $T$ .....	94
Figure 3.20 Scatterplot of model $X = e^T + N_Y$ and $Y = (T - 1)^3 - 4 + N_X$ .....	95
Figure 3.21 (A) Scatterplot of model $X = T^2 - T + N_X$ and $Y = \sin T + N_Y$ .....	96
Figure 3.21 (B) Scatterplot of $\hat{T}$ (estimated $T$ ) vs. $T$ .....	96
Figure 3.22 Part of RIG-I-like receptor signaling pathway .....	97
Figure 3.23 Causal Network .....	100

# **1 Background**

## **1.1 Literature Review**

Despite significant progress in dissecting the genetic architecture of complex diseases by association analysis, understanding the etiology and mechanism of complex diseases remains elusive. Using association analysis and machine learning systems that operate, almost exclusively, in a statistical, or model-free modes as a major analytic platform for genetic studies of complex diseases is a key issue that hampers the discovery of mechanisms underlying complex traits (Pearl 2018).

As an alternative to association analysis, causal inference may provide tools to unravel principles underlying complex traits. Power of causal inference is its ability to predict the effects of actions on the system (Mooij et al. 2016). Typical methods to unravel cause-and-effect relationships are interventions and controlled experiments. Unfortunately, the experiments in human genetics are unethical and technically impossible. Next generation genomic, epi-genomic, sensing and image technologies produce ever deeper multiple omic, physiological, imaging, environmental and phenotypic data with millions of features. These data are almost all “observational”, which have not been randomized or otherwise experimentally controlled (Glymour 2015). In the past decades, a variety of statistical methods and computational algorithms for causal inference that attempt to abstract causal knowledge from purely observational data, referred to as causal discovery, have been

developed (Zhang et al. 2018). Causal inference is one of the most useful tools developed in the past century. The classical causal inference theory explores conditional independence relationships in the data to discover causal structures. The PC algorithms and the fast causal inference (FCI) algorithms developed at Carnegie Mellon University by Peter Spirtes and Clark Glymour are often used for cause discovery (Le et al. 2016). Despite its fundamental role in science, engineering and biomedicine, the conditional independence-based classical causal inference methods can only identify the graph up to its Markov equivalence class, which consists of all DAGs satisfying the same conditional independence distributions via the causal Markov conditions (Nowzohour and Bühlmann 2016). For example, consider three simple DAGs:  $x \rightarrow y \rightarrow z$ ,  $x \leftarrow y \leftarrow z$  and  $x \leftarrow y \rightarrow z$ . Three variables  $x, y$  and  $z$  in all three DAGs satisfy the same causal Markov condition:  $x$  and  $z$  are independent, given  $y$ . This indicates that these three DAGs form a Markov equivalence class. However, these three DAGs represent three different causal relationships among variables  $x, y$  and  $z$ , which prohibits unique causal identification. These non-unique causal solutions seriously limit their translational application.

In the past decade, causal inference theory is undergoing exciting and profound changes from discovering only up to the Markov equivalent class to identify unique causal structure (Peters et al. 2012; Peters and Bühlman, 2014). A class of powerful algorithms to find a unique causal solution are based on properly defined functional causal models (FCMs). They include the linear, non-Gaussian, acyclic model (LiNGAM) (Zhang et al. 2018;

Shimizu et al. 2006), the additive noise model (ANM) (Hoyer et al. 2009; Peters et al. 2014), and the post-nonlinear (PNL) causal model (Zhang and Hyvärinen 2009).

### **1.1.1 Bivariate Causal Inference**

In genomic and epi-genomic data analysis, we usually consider four types of associations: association of discrete variables (DNA variation) with continuous variables (phenotypes, gene expressions, methylations, imaging signals and physiological traits), association of continuous variables (expressions, methylations and imaging signals) with continuous variables (gene expressions, imaging signals, phenotypes and physiological traits), association of discrete variables (DNA variation) with binary trait (disease status) and association of continuous variables (gene expressions, methylations, phenotypes and imaging signals) with binary trait (disease status). All these four types of associations can be extended to four types of causations. This dissertation focuses on studying causal relationships between two continuous variables and two discrete variables respectively.

The many causal inference algorithms using observational data require that two variables being considered as cause-effect relationships are part of a larger set of observational variables (Mooij et al. 2016). Similar to genome-wide association studies where only two variables are considered, bivariate causal discovery is investigated to infer cause-effect relationships between two observed variables. To simplify the cause discovery studies, it is usually assumed that there is no selection bias, no feedback and no confounding. It also

assumes that nature consists of autonomous and independent causal generating process modules and attempts to replace causal faithfulness by the assumption of Independence of Cause and Mechanism (ICM) (Peters et al. 2017; Besserve et al. 2017; Schölkopf et al. 2012; Janzing et al. 2010; Lemeire et al. 2012).

The philosophical causal principle assumes that nature consists of independent, autonomous causal generating process modules (Peters et al. 2017; Shajarisales). In other words, causal generating processes of a system's variables are independent. If we consider two variables: cause  $X$  and effect  $Y$ , then the mechanism that generates cause  $X$  and the mechanism that generates effect  $Y$  from the cause  $X$  are independent. Or, the process that generates the effect  $Y$  from the cause  $X$  contains no information about the process that generates the cause  $X$ . In the probability setting, this indicates that the cause distribution  $P(X)$  and the conditional distribution  $P(Y|X)$  of  $Y$  given  $X$  are independent. Statistics provides definition of independence between two random variables, but provides no tool for defining independence between two distributions (Peters et al. 2017). Algorithmic information theory can offer notion and mathematical formulation of independence between two distributions or independence of mechanisms (Janzing et al. 2010; Parascandolo 2017).

Assume no confounding, no selection bias and no feedback. Consider the bivariate additive noise models  $X \rightarrow Y$  and  $Y \rightarrow X$ . If the density  $P_{X,Y}$  is induced by the ANM  $X \rightarrow Y$ , but not by the ANM  $Y \rightarrow X$ , then the ANM  $X \rightarrow Y$  is identifiable. Independence of cause

and mechanism states that the conditional distribution  $P_{Y|X}$  contains no information about the distribution of causal  $P_X$ .

Mutual information of zero between the cause  $X$  and residual variable  $E_Y$  shows that  $X$  and  $E_Y$  are independent. Therefore, algorithmic independence between the distribution of cause  $X$  and conditional distribution  $P_{Y|X}$  of effect given the cause is equivalent to the independence of two random variables  $X$  and  $E_Y$  in the ANM. Peters et al. (2017) showed that a joint distribution  $P_{X,Y}$  does not admit an ANM in both directions at the same time under some quite generic conditions.

Empirically, if the ANM  $X \rightarrow Y$  fits the data, then we infer that  $X$  causes  $Y$ , or if the ANM  $Y \rightarrow X$  fits the data, then  $Y$  causes  $X$  will be concluded. Although this statement cannot be rigorously proved, in practice, this principle will provide the basis for bivariate cause discovery (Mooij et al. 2016). To implement this principal, we need to develop statistical methods for assessing whether the additive noise model fits the data or not.

Besides Additive Noise Model, distance correlation is also popular in bivariate causal inference. The basis principal for assessing causation  $X \rightarrow Y$  is that the distribution  $P(X)$  of causal  $X$  is independent of the causal mechanism or conditional distribution  $P(Y | X)$  of the effect  $Y$ , given causal  $X$ . The question is how to assess their independence. Recently, distance correlation is proposed to measure dependence between random vectors, which

allows for both linear and nonlinear dependence (Sze'kely et al. 2007; Sze'kely and Rizzo 2009). Distance correlation extends the traditional Pearson correlation in two remarkable directions:

- (1) Distance correlation extends the Pearson correlation defined between two random variables to the correlation between two sets of variables with arbitrary numbers;
- (2) Zero of distance correlation indicates independence of two random vectors.

Discretizing distributions  $P(X)$  and  $P(Y | X)$ , and viewing their discretized distributions as two vectors  $P(X)$  and  $P(Y | X)$ , the distance correlation between  $P(X)$  and  $P(Y | X)$  can be used to assess causation between  $X$  and  $Y$ .

Consider two vectors of random variables:  $p$  - dimensional vector  $X$  and  $q$  - dimensional vector  $Y$ . Let  $P(x)$  and  $P(y)$  be density functions of the vectors  $X$  and  $Y$ , respectively. Let  $P(x, y)$  be the joint density function of  $X$  and  $Y$ . There are two ways to define independence between two vectors of variables: (1) density definition and (2) characteristic function definition. In other words, if  $X$  and  $Y$  are independent then either

$$(1) P(x, y) = P(x)P(y) \text{ or}$$

$$(2) f_{X,Y}(t, s) = f_X(t)f_Y(s),$$

where  $f_{X,Y}(t, s) = E[e^{i(t^T x + s^T y)}]$ ,  $f_X(t) = E[e^{it^T x}]$  and  $f_Y(s) = E[e^{is^T y}]$  are the characteristic functions of  $(X, Y)$ ,  $X$  and  $Y$ , respectively. Therefore, we can use both distances



$\|P(x, y) - P(x)P(y)\|$  and  $\|f_{X,Y}(t, s) - f_X(t)f_Y(s)\|$  to measure dependence between two vectors  $X$  and  $Y$ . Distance correlation (Sze'kely et al. 2007) uses distance between characteristic functions to define the dependence measure.

Distance correlation can be used to test independence between causal and causal generating mechanisms (Liu and Chan 2016). Consider  $p$ -dimensional random vector  $X$  and  $q$ -dimensional random vector  $Y$ . Let  $P(X, Y)$  be their joint distribution. Let  $P(X)$  and  $P(Y|X)$  be the density function of  $X$  and conditional density function of  $Y$ , given  $X$ , respectively. Similarly, we can define  $P(Y)$  and  $P(X|Y)$ . Unlike association analysis where dependence is measured between two random vectors, in causal analysis, dependence is measured between two distributions.

Given that causation relationship exists between two variables  $X$  and  $Y$ , additive noise model and distance correlation can be used to determine the causal direction with high accuracy. Recently, a new method called stochastic complexity (Budhathoki and Vreeken 2017) has been proposed to identify causal direction with the lowest Kolmogorov complexity. Simulations in the paper show that it has higher accuracy than additive noise model and distance correlation. It is also applied to identifying cause and effect in data that was not collected through carefully controlled randomized trials.

A Turing machine is a hypothetical machine developed by Alan Turing in 1936. Turing machine is designed to simulate any computer algorithm, no matter how complicated it is (Ashrafian et al. 2015). Consider a universal Turing Machine  $T$ . For any binary string  $s$ , we define Kolmogorov complexity  $K_T(s)$  as the length of the shortest program that generates  $s$ , denoted as  $s^*$ , using universal prefix Turing machine  $T$  that outputs  $s$  and then stops (Peters et al. 2017; Kolmogorov 1965). Therefore, we have  $K_T(s) = |s^*|$ , where  $|\cdot|$  denotes the number of bits of a binary string. Intuitively, the Kolmogorov complexity measures the minimal amount of information required to generate  $s$  by any effective process. Similar to conditional probability, we can also define conditional Kolmogorov complexity. The conditional Kolmogorov complexity  $K(t|s)$  of string  $t$  given  $s$ , is defined as the length of the shortest program that can generate  $t$  from  $s$  and then stops. The Kolmogorov complexity  $K(t, s)$  of the concatenation of two strings,  $t$  and  $s$  is defined as the length of the shortest program that generate string  $t's$ , where  $t'$  is the prefix code of  $t$ .

Now we introduce “additivity of complexity” property. It can be shown that (Grunwald and Vitanyi 2004):

$$K(t, s) = K(t) + K(s|t^*),$$

where  $t^*$  denotes the first shortest prefix program that generates  $t$  and then stops and is in general uncomputable.

Algorithmic mutual information is defined as

$$I(s:t) = K(s) - K(s|t^*).$$

Substituting  $K(s|t^*)$  in equation (A1) into equation (A2), we obtain

$$I(s:t) \pm K(s) + K(t) - K(s,t),$$

where the symbol  $\pm$  implies that the equation can hold for up to constants. Equation (A3) states that this information is symmetrical:  $I(s:t) = I(t:s)$ . Therefore,  $I(s:t)$  is called algorithmic mutual information between  $s$  and  $t$ . The algorithmic mutual information quantifies the amount of information two strings or objects have in common, or the amount of bits saved when compressing  $s, t$  jointly rather than compressing  $s, t$  independently.

Similar to mutual information  $I(s:t)$  between two random variables where mutual information of zero implies independence of two variables, the algorithmic mutual information of zero  $I(s:t)$  indicates algorithmically independence of two distributions of random variables. We also can define algorithmic conditional mutual information as

$$I(s:t|z) \pm K(s|z) + K(t|z) - K(s,t|z).$$

In statistics, although dependence between two random variables can be measured, there are no measures to quantify dependence between two distributions. We use algorithmic mutual information to measure independence between two distributions that can be used to assess causal relationships between two variables. Consider two variables  $X$  and  $Y$  and assume  $X$

causes  $Y$  ( $X \rightarrow Y$ ). Let the marginal distribution of cause  $X$  and conditional distribution of effect  $Y$  given  $X$  be  $P_X$  and  $P_{Y|X}$ , respectively. The independence of cause and mechanism (ICM) states that the distributions  $P_X$  and  $P_{Y|X}$  are independent and hence  $P_X$  and  $P_{Y|X}$  are algorithmically independent, which implies that their algorithmic mutual information should be equal to zero (Peters et al. 2017):

$$I(P_X: P_{Y|X}) \pm 0,$$

or , equivalently,

$$K(P_{X,Y}) \pm K(P_X) + K(P_{Y|X}).$$

In other words, distributions  $P_X$  and  $P_{Y|X}$  have no common information. If  $X$  causes  $Y$ , then the conditional distribution  $P_{Y|X}$  of the effect  $Y$  given cause  $X$  contains no information about cause  $X$ . Thus, the algorithmic mutual information can be used to infer whether  $X \rightarrow Y$  or  $Y \rightarrow X$ . If  $I(P_X: P_{Y|X}) < I(P_Y: P_{X|Y})$  then  $X \rightarrow Y$ . Similarly, if  $I(P_X: P_{Y|X}) > I(P_Y: P_{X|Y})$  then  $Y \rightarrow X$ . Cause and effect cannot be identified from their joint distribution. Cause and effect are asymmetric. The joint distribution is symmetric. It can be factorized to  $P_{X,Y} = P_X P_{Y|X} = P_Y P_{X|Y}$ .

### 1.1.2 Hilbert-Schmidt Independence Criterion

Covariance can be used to measure association, but cannot be used to test independence between two variables. A covariance operator can measure the magnitude of dependence, and

is a useful tool for assessing dependence between variables. Specifically, we will use the

Hilbert-Schmidt norm of the cross-covariance operator or its approximation, the Hilbert-Schmidt independence criterion (HSIC) to measure the degree of dependence between the residuals and potential causal variable (Gretton et al. 2005; Mooij et al. 2016).

*Calculation of the HSIC consists of the following steps.*

Step 1: Use test data set to compute

$$y_i = \hat{f}(x_i) + E_Y(i), i = 1, \dots, m.$$

Step 2: Compute the residuals:

$$\varepsilon_i = E_Y(i) = y_i - \hat{f}(x_i), i = 1, \dots, m.$$

Step 3: Select two kernel functions  $k_E(\varepsilon_i, \varepsilon_j)$  and  $k_x(x_i, x_j)$ . In practice, we often use the

Gaussian kernel function. Compute the Kernel matrices:

$$K_{E_Y} = \begin{bmatrix} k_E(\varepsilon_1, \varepsilon_1) & \cdots & k_E(\varepsilon_1, \varepsilon_m) \\ \vdots & \ddots & \vdots \\ k_E(\varepsilon_m, \varepsilon_1) & \cdots & k_E(\varepsilon_m, \varepsilon_m) \end{bmatrix}, K_x = \begin{bmatrix} k_x(x_1, x_1) & \cdots & k_x(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k_x(x_m, x_1) & \cdots & k_x(x_m, x_m) \end{bmatrix}.$$

Step 4: compute the HSCI for measuring dependence between the residuals and potential causal variable.

$$HSIC^2(E_Y, X) = \frac{1}{m^2} \text{Tr}(K_{E_Y} H K_X H),$$

where  $H = I - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T$ ,  $\mathbf{1}_m = [1, 1, \dots, 1]^T$  and  $\text{Tr}$  denotes the trace of the matrix.

### 1.1.3 Confounder Identification

A hidden confounder usually results in the case of no causation but having association. One modern method to detect confounder using additive noise model (Janzing 2009) consists of two main steps: 1) Initial dimension reduction, and 2) minimization of dependence criterion.

This method is proposed for inferring the existence of a latent common cause (“confounder”) of two observed random variables, which assumes that the two effects of the confounder are functions of the confounder plus independent additive noise. That is,  $X = f_1(T) + N_1$  and  $Y = f_2(T) + N_2$ . Gaussian Process Regression (Rasmussen and Williams, 2006) is applied to fit the regression model, which maximizes marginal likelihood in parameter settings. To measure the dependence, Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005, 2008) is in use. With this method, scientists will be able to distinguish between i)  $X \rightarrow Y$  ii)  $Y \rightarrow X$  and iii)  $X \leftarrow T \rightarrow Y$ .

### 1.1.4 Entropy Methods

In statistics, although dependence between two random variables can be measured, there is no measure to quantify dependence between two distributions. Algorithmic mutual information can be used to measure independence between distributions. If  $X \rightarrow Y$ , then we have  $I(P(X): P(Y|X))=0$ .

When conducting the causal inference between multiple continuous covariates  $\mathbf{X}$  and the single continuous variable  $Y$ , it does not make sense to establish a reverse causal model. Thus, entropy methods are applied to measure the appropriateness of causal model  $\mathbf{X} \rightarrow Y$ . That is, estimate the entropy information of residuals  $E_Y = Y - f(\mathbf{X})$ . Shannon entropy for continuous variable is defined as  $H(X) = - \int dx \mu(x) \log \mu(x)$  (Kraskov 2008). Since the density function of  $X$  in  $H(X)$  is usually unknown, the estimator in a simplified form is  $H(X) \approx \frac{1}{N-1} \sum_{i=1}^{N-1} \log(x_{i+1} - x_i) - \psi(1) + \psi(N)$ .

### 1.1.5 Structural Equation Model

In network analysis, multiple nodes can be regarded as variables, so a combination optimization problem on multiple variables can be introduced to causal network analysis. A standard form of integer linear programming (Xiong 2018) is

$$\min A^T X$$

$$BX \leq c$$

$$X \geq 0$$

$$X \in Z^n$$

where  $A \in R^n$ ,  $b \in R^m$ ,  $B \in R^{m \times n}$  and  $Z = \{0, 1, 2, \dots\}$ .

Considering whether a variable should be included as cause or not, the value of  $X$  is set as 0 or 1, so the above linear programming can be converted to 0-1 integer programming:

$$\min A^T X$$

$$BX \leq c$$

$$X \geq 0$$

$$X \in E^n$$

where  $E = \{0, 1\}$ .

The score function  $A^T X$  is just a simple example. In causal network analysis, the score function is more complicated to reflect the true relation between multiple variables. We can define a DAG as  $G=(V, E)$  where  $V$  refers to the nodes and  $E$  edges. The set of causal variables of  $v \in V$  defined as  $C_v$ , and the DAG can be illustrated by the causal variables sets  $C=\{C_1, C_2, \dots, C_p\}$ . The way to determine a directed acyclic graph (DAG)  $D$  is to select the case of having the optimal score of the summation of score functions. That is,  $S(D) = \sum_{v \in V} S(v, C_v)$ . Then the task is to find a DAG that minimizes the global score  $S(D)$  over all possible DAGs, i.e.  $\min_D \sum_{v \in V, C_v \in D} S(v, C_v)$ .

## 1.2 Public Health Significance

Association analysis has been used as a major tool for dissecting genetic architecture and unraveling mechanisms of complex diseases for more than a century (Fisher 1918; Timpson et al. 2017). Although significant progress in dissecting the genetic architecture of complex diseases by genome-wide association studies (GWAS) has been made, the overall contribution of the new identified genetic variants to the diseases is small and a large fraction



of disease risk genetic variants is still hidden. Understanding the etiology and causal chain of mechanism underlying many complex diseases remains elusive. The current approach to uncovering hidden genetic variants is (1) to increase sample sizes, (2) to study association of rare variants by next-generation sequencing and (3) to perform multi-omic analysis. Association and correlation analysis are the current paradigm of analysis for all these approaches. However, association analysis cannot identify causal signals that are quite different from the association signals, and cannot infer direct cause-effect relations. Thus, insistence on association analysis tends to hamper the theoretical development of genomic science and its application in practice. Causal inference coupled with multiple omics, imaging, physiological and phenotypic data is an essential component for the discovery of disease mechanisms. It is time to develop a new generation of genetic analysis for shifting the current paradigm of genetic analysis from shallow association analysis to deep causal inference.

Typical methods for unraveling cause-and-effect relationships are interventions and controlled experiments. Unfortunately, the experiments are sometimes unethical, expensive and technically impossible, especially in the field of public health. Thus, progress in causal inference on observational data will definitely benefit the research in public health.

### 1.3 Specific Aims

As described above, causal study is in high demand for genetic study as well as other fields of healthcare. When randomization controls trials are not available, it is really essential to make inference based on observational data. To fulfill the high demands in healthcare research, I propose several tests and models on causal inference. Then, bivariate causal test on continuous variables, bivariate causal test on discrete variables, the algorithm to detect confounders, the causal test for multiple causes and the causal network based on structured equations model can be combined as a system to deal with causal inference problems.

Aim 1(a): To develop bivariate causal inference test for continuous variables. This test will be proposed to deal with two continuous variables, applied to KEGG pathway analysis. Three types of simulations, independence, having both association and causation and having association without causation will be conducted to check the type I error rates and power of this test.

Aim 1(b): To develop bivariate causal inference test for discrete variables. This test will be proposed to deal with two discrete variables, applied to genome-wide causal study. Four types of simulations, no association and no causation, no association but having causation, having both association and causation, and having association without causation will be conducted to check the type I error rates and power of this test. Further, I will check the performance of the test in the case of linkage disequilibrium.

Aim 2: To develop nonlinear functional models for causal inference on genomic variables with measured confounders. One reason for having association without causation is the effect of confounder that causes spurious association. Machine learning scientist has developed an algorithm to detect confounders and given several examples. I will develop the algorithm into software packages available for statistician, and further conduct simulations to check its performance. If any defects found, I will improve the algorithm and develop a new method.

Aim 3(a): To develop the causal inference test for multiple causes and one effect. It is common that several factors cause a phenomenon. Since causes are multivariate and the effect one variable, the bivariate causal tests cannot be applied in this case. Thus, I will propose a test to detect the causal relationship between multiple factors and the effect.

Aim 3(b): To develop the causal network model for high-dimensional data. A linear structural equations model (Wang 2016) has been developed to deal with causal inference on high-dimensional data, which uses linear model to mimic causal relation between a specific node and its parents. However, in real world, causal relationships are usually nonlinear, so we need to extend it to nonlinear models.

## **2 Methods**

### **2.1 Overall Study Design**

#### **2.1.1 Simulation Studies**

At each of the five aims (sub-aims included), simulation studies will be conducted. Since in each simulation, data structures are different, so the simulated datasets will be generated separately. In aim 1a, I will generate simulated pairs of continuous data with three types of relationship, independence, association without causation and causation without association. Then causal test will be applied to the gene expression data and methylation data. In aim 1b, I will generate simulated genotype data mimicking 100 common SNPs in 1000 Genome. Then causal test will be applied to the simulated genotype data and phenotype data. In aim 2, two continuous variables with and without a hidden confounder will be generated. In aim 3a, multiple independent variables mimicking causes and the effect variable will be generated. In aim 3b, high-dimensional data will be generated from a pre-determined causal network. The simulations will be kept consistent to the paper on linear structural equations model (Wang 2016) to evaluate this performance.

#### **2.1.2 Real Data Application**

After simulation studies, I will apply the proposed methods to real datasets for each of the aim. In aim 1, I will use CATIE-MGS-SWD schizophrenia study dataset with 8,421,111 common SNPs typed in 13,557 individuals (Bergen 2012, Shi 2009, Stroup 2003). In aim 2 and 3, I will use a gene expression dataset with 51,060 genes and 432 samples from Rush

University Medical Center. I will use Kegg pathway as reference to check the performance of methodologies in aim 2 and 3 applied to real datasets.

## 2.2 Methods for Aim 1(a): To develop bivariate causal inference test for continuous variables

### 2.2.1 Statistical Modeling

Assume no confounding, no selection bias and no feedback. Consider a bivariate additive noise model  $X \rightarrow Y$  where  $Y$  is a nonlinear function of  $X$  and independent additive noise  $E_Y$ :

$$\begin{aligned} Y &= f_Y(X) + E_Y \\ X &\sim P_X, E_Y \sim P_{E_Y}, \end{aligned} \tag{2.1}$$

where  $X$  and  $E_Y$  are independent. Then, the density  $P_{X,Y}$  is said to be induced by the additive noise model (ANM) from  $X$  to  $Y$  (Mooij et al. 2016). The alternative additive noise model between  $X$  and  $Y$  is the additive noise model  $Y \rightarrow X$ :

$$\begin{aligned} X &= f_X(Y) + E_X \\ Y &\sim P_Y, E_X \sim P_{E_X}, \end{aligned} \tag{2.2}$$

where  $Y$  and  $E_X$  are independent.

*The general procedure for bivariate causal discovery is given as follows (Mooij et al. 2016):*

Step 1: Divide a data set into a training data set  $D_{train} = \{Y_n, X_n\}$  for fitting the model and a test data set  $D_{test} = \{\tilde{Y}_m, \tilde{X}_m\}$  for testing the independence.

Step 2: Use the training data set and nonparametric regression methods

(a) Regress  $Y$  on  $X$  :  $Y = f_Y(X) + E_Y$  and

(b) Regress  $X$  on  $Y$  :  $X = f_X(Y) + E_X$ .

Step 3: Use the test data set and estimated nonparametric regression model that fits the training data set  $D_{train} = \{Y_n, X_n\}$  to predict residuals:

(a)  $\hat{E}_{Y_X} = \tilde{Y} - \hat{f}_Y(\tilde{X})$

(b)  $\hat{E}_{X_Y} = \tilde{X} - \hat{f}_X(\tilde{Y})$

Step 4: Calculate the dependence measures  $HSIC^2(E_Y, X)$  and  $HSIC^2(E_X, Y)$ .

Step 5: Infer causal direction:

$$X \rightarrow Y \text{ if } HSIC^2(E_Y, X) < HSIC^2(E_X, Y); \quad (2.3)$$

$$Y \rightarrow X \text{ if } HSIC^2(E_Y, X) > HSIC^2(E_X, Y). \quad (2.4)$$

If  $HSIC^2(E_Y, X) = HSIC^2(E_X, Y)$ , then causal direction is undecided.

$H_0$ : no causations  $X \rightarrow Y$  and  $Y \rightarrow X$  (Both  $X$  and  $E_Y$  are dependent, and  $Y$  and  $E_X$  are dependent).

Calculate the test statistic:

$$T_C = |HSIC^2(E_Y, X) - HSIC^2(E_X, Y)|. \quad (2.5)$$

Assume that the total number of permutations is  $n_p$ . For each permutation, we fix  $x_i$ ,

$i = 1, \dots, m$  and randomly permute  $y_i, i = 1, \dots, m$ . Then, fit the ANMs and calculate the residuals  $E_X(i), E_Y(i), i = 1, \dots, m$  and test statistic  $T_C$ . Repeat  $n_p$  times. The P-values are defined as the proportions of the statistic  $\tilde{T}_C$  (computed on the permuted data) greater than or equal to  $\hat{T}_C$  (computed on the original data  $D_{TE}$ ). After cause is identified, we then use equations (4) and (5) to infer causal directions  $X \rightarrow Y$  or  $Y \rightarrow X$ .

### 2.2.2 Simulation Settings

There are three parts of simulation studies: 1) Independence. 2) Association exists. No causation. 3) Both association and causation exist.

#### 2.2.2.1 Independence



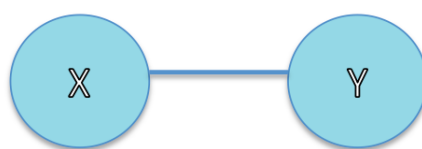
**Figure 2.1** Two independent continuous variables

We first generated the data with 100,000 subjects from the model:  $X \sim N(0,1), Y \sim N(0,1)$  and  $X, Y$  are independent. Then we randomly picked up 500, 1000 and 2000 samples from the population to calculate and compare their Type I error rates of causal tests.

**Table 2.1** Type I error rates of the ANMs between two continuous variables, assuming independence.

	Number of Samples		
Significance Level	500	1000	2000
0.05	0.033	0.051	0.043
0.01	0.006	0.005	0.01

2.2.2.2 Association exists. No causation.



**Figure 2.2** Two associated continuous variables without causation

**Table 2.2** Type 1 error rates of the ANMs between two continuous variables in the presence of association.

	Number of Samples		
Significance Level	500	1000	2000
0.05	0.044	0.048	0.050
0.01	0.011	0.011	0.011

We generated the data with 100,000 subjects from the model:  $X \sim N(0,1)$ ,  $Y \sim N(0,1)$ ,  $X$  and  $Y$



were associated, but without causation. Then we randomly picked up 500, 1000, and 2000 samples from the population to calculate and compare their Type I error rates of causation and association tests respectively.

### 2.2.2.3 Both association and causation exist.



**Figure 2.3** Two independent continuous variables

**Table 2.3** Power of the ANMs between two continuous variables.

Significance Level	Number of Samples				
	200	500	1000	2000	5000
0.05	0.3616	0.4833	0.5629	0.5997	0.6412
0.01	0.2066	0.3556	0.4382	0.4762	0.5241

We generated data with 100,000 subjects from the causal model:

$$Y = f(X) + N,$$

where  $f(x) = \sum_{j=1}^3 w_j \times \exp(-\gamma(x - x_j)^2)$ ,  $\gamma \sim N(0,1)$ ,  $x_j \sim N(0,1)$ ,  $X \sim N(0,1)$  and

$N \sim N(0, \sigma^2 = 0.01)$ .  $X$  and  $N$  are independent, and  $w_j$ 's are randomly-generated weights

from the uniform distribution. Then we randomly picked up 200, 500, 1000, 2000 and 5000 samples from the population to calculate and compare their powers.

## **2.3 Methods for Aim 1(b): To develop bivariate causal inference test for discrete variables**

### **2.3.1 Statistical Modeling**

Let  $X=(x_1, x_2, \dots, x_n)$  denote SNP data, and  $Y=(y_1, y_2, \dots, y_n)$  traits. Note that  $X$  is regarded as non-cyclic variable, since a person's SNP data cannot change.  $Y$  is regarded as cyclic variable if the patient can recover from a specific disease. Otherwise, it is also regarded as non-cyclic variable. The variable type will affect the regression model in the following.

I will first fit a non-parametric regression model from  $X$  to  $Y$ . That is  $Y = f(X) + N_1$ . Since  $X$  and  $Y$  are discrete variables, so non-parametric regression methods such as smoothing spline regression (Wang 2011) and Gaussian process regression are not applicable here. Here we use discrete regression with dependence minimization to solve it (Peters 2011).

Step 1: For each  $x_i$ ,  $f^{(0)}(x_i)$  is given the value  $y$  which maximizes probability mass function  $\hat{P}(X = x_i, Y = y)$ . Let  $\varepsilon_i = y_i - f^{(0)}(x_i)$ . If  $p^{(0)}$ , the p-value of association test between residuals  $E_1$  and  $X$  is less than significance level, go to Step 2. Otherwise, output  $f^{(0)}$  as  $\hat{f}$  and  $p^{(0)}$  as  $p_{X \rightarrow Y}$ .

Step 2: Given fitted regression  $f^{(j-1)}$ , for each  $x_i$  in a random ordering,  $f^{(j)}(x_i)$  is given the value  $y$  which maximize the p-value of association test between  $X$  and  $Y - f_{x_i \rightarrow y}^{(j-1)}(X)$ . If  $f^{(j)}$  is constant, choose the value  $y$  that results in the second largest p-value of association test.

Step 3: If  $p^{(0)}$ , the p-value is larger than significance level, output  $f^{(j)}$  as  $\hat{f}$  and  $p^{(0)}$  as  $p_{X \rightarrow Y}$ . Otherwise, go back to Step 2 until  $K$  iterations.

**Table 2.4** Rules to determine causal direction.

$P_c$	$p_{X \rightarrow Y}$	$p_{Y \rightarrow X}$	Causal Direction
$\geq \alpha$	X	X	No causation
$< \alpha$	$< \alpha$	$< \alpha$	No causation
$< \alpha$	$p_{X \rightarrow Y} < p_{Y \rightarrow X} \ \& \ p_{Y \rightarrow X} \geq \alpha$		$Y \rightarrow X$
$< \alpha$	$p_{Y \rightarrow X} < p_{X \rightarrow Y} \ \& \ p_{X \rightarrow Y} \geq \alpha$		$X \rightarrow Y$

It is the same to fit a regression model from  $Y$  to  $X$ ,  $X = g(Y) + N_2$ , and get the  $p_{Y \rightarrow X}$ . Then let  $\Delta_{obs} = |p_{X \rightarrow Y} - p_{Y \rightarrow X}|$ . If  $\Delta_{obs}$  is significantly large, I infer that there exists causal relationship because of this asymmetric phenomenon. To measure how large  $\Delta_{obs}$  is, I will use permutation test to generate a causation p-value,  $P_c$ . That is, resample  $Y$  as  $Y^{(k)}$  and get

$$\Delta_{permutation}^{(k)} = \left| p_{X \rightarrow Y}^{(k)} - p_{Y \rightarrow X}^{(k)} \right|, \text{ where } k=1, \dots, N. \text{ Then,}$$

$$P_c = P\left(\Delta_{obs} < \Delta_{permutation}^{(k)}\right) / N$$

Given significance level  $\alpha$ , the following table illustrates how to determine causal direction.

### 2.3.2 Simulation Settings

I will randomly select 100 common SNPs on gene TEKT4P2 in 1000 Genome datasets. The population is Caucasian. Then I will use resampling to generate the simulated genotypes of a 100,000 population. The phenotype data is generated either randomly or by the genotype data.

There are five parts of simulation studies: 1) No association. No causation. 2) No association. Causation exists. 3) Association exists. No causation. 4) Both association and causation exist. 5) The effect of linkage disequilibrium.

#### 2.3.2.1 No association. No causation.



**Figure 2.4** Two independent discrete variables

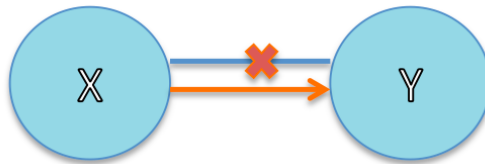
X refers to SNP and Y phenotype. I will first generate 100,000 subjects with X and Y independent of each other. Then I will randomly pick up 500, 1000, 2000, and 5000 samples

from the population to calculate and compare their Type I error rates of causation and association tests.

**Table 2.5** Type 1 error rates of the ANMs between two discrete variables, assuming no association and no causation.

Type I Error Rate	X non-cyclic			
Significance Level	N=500	N=1,000	N=2,000	N=5,000
0.05	0.044	0.046	0.048	0.051
0.01	0.005	0.006	0.007	0.009

#### 2.3.2.2 No association. Causation exists.



**Figure 2.5** Two causal discrete variables that do not show association

Here I will NOT generate 100,000-individual population, since the rare cases of no association but having causation only appear when their probability mass functions satisfy some criterion. That is,

$$\frac{a_1}{b_1} = \frac{b_2}{a_2} = \frac{a_3}{b_3} = k \quad (2.6)$$

**Table 2.6** Probability mass functions for X and Y

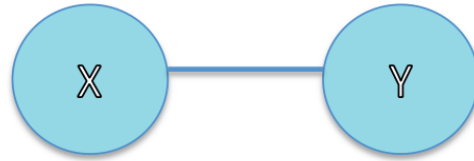
	Y=0	Y=1
X=0	$a_1$	$b_1$
X=1	$a_2$	$b_2$
X=2	$a_3$	$b_3$

Experiences show that when I simulate 100, 000 population and take a subgroup, the sub data will tend to present no causation. To show this rare case of no association but having causation, I will instead simulate the data with 500, 1000, 2000, 5000, 10000 and 20000 samples separately based on relationship very close to equation (2.6), and then test its power.

**Table 2.7** Power of the ANMs between two discrete variables, assuming no association but having causation.

Power		Number of samples					
		500	1,000	2,000	5,000	10,000	20,000
Significance	0.05	0.9032	0.9573	0.9935	0.9999	1	1
level	0.01	0.1058	0.2178	0.4096	0.7974	0.9631	0.9990

2.3.2.3 Association exists. No causation.



**Figure 2.6** Two associated discrete variables without causation

I will first generate 100,000 population with significantly associated X and Y. Then I will randomly pick up 500, 1000, 2000, and 5000 samples from the population to calculate and compare their Type I error rates of causation and association tests respectively.

**Table 2.8** Type 1 error rates of the ANMs between two discrete variables in the presence of association.

Type I Error Rate	X non-cyclic			
Significance Level	N=500	N=1,000	N=2,000	N=5,000
0.05	0.042	0.046	0.047	0.046
0.01	0.005	0.007	0.007	0.008

2.3.2.4 Both association and causation exist.



**Figure 2.7** Two causal discrete variables

X is simulated from 100 common SNPs from 1000 Genome with range  $\{0, 1, 2\}$ . N is a randomly generated discrete variable with range  $\{0, 1\}$ .

Data was generated by formula

$$Y = f(X) + N,$$

where  $f$  is a non-constant random function from  $\{0, 1, 2\}$  to  $\{0, 1\}$ .

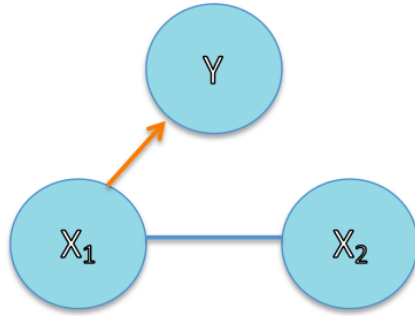
In the 100, 000-individual population, pairs with association p-value larger than significance level have been removed. Then I will randomly pick up 500, 1000, 2000, 5000, 10000, and 20000 samples from the population to calculate and compare their powers.

**Table 2.9** Power of the ANMs between two discrete variables in the presence of association.

Power		Number of samples					
		500	1,000	2,000	5,000	10,000	20,000
Significance level	0.05	0.553	0.662	0.751	0.837	0.885	0.921
	0.01	0.435	0.565	0.677	0.785	0.845	0.892

#### 2.3.2.5 Linkage disequilibrium





**Figure 2.8** One casual SNP and its neighboring associated SNP

$X_1$  is simulated from 100 common SNPs from 1000 Genome with range  $\{0, 1, 2\}$ .  $N$  is a randomly generated discrete variable with range  $\{0, 1\}$ .

Data was generated by formula

$$Y = f(X_1) + N,$$

where  $f$  is a non-constant random function from  $\{0, 1, 2\}$  to  $\{0, 1\}$ . Remove pairs with  $P_{\text{association}}$  larger than significance level. Then randomly generate  $X_2$  which should be associated with  $X_1$ .

Here I will simulate the data with 500, 1000, 2000 and 5000 samples separately, and then test the association and causation of  $X_2$  and  $Y$ .

**Table 2.10** Type 1 error rates of the ANMs between two discrete variables in the presence of linkage disequilibrium

Type I Error Rate	X non-cyclic			
Level $\alpha$	N=500	N=1,000	N=2,000	N=5,000
0.05	0.048	0.036	0.026	0.021
0.01	0.025	0.017	0.013	0.009

## 2.4 Methods for Aim 2: To develop nonlinear functional models for causal inference on genomic variables with measured confounders

### 2.4.1 Statistical Modeling

(Janzing 2009) introduces an algorithm to detect confounder based on additive noise model. Given the pair of datasets  $(X_1, Y_1), \dots, (X_n, Y_n)$ , we will first use Isomap algorithm to get an initial guess  $\hat{T}_k$  of the possible confounder. By Gaussian process regression, we have two functions  $\hat{u}$  and  $\hat{v}$  as the regression from  $X$  on  $\hat{T}$  and  $Y$  on  $\hat{T}$ , respectively. Then we re-choose  $T_k$  which minimizes  $\|(\hat{u}(T_k), \hat{v}(T_k)) - (X_k, Y_k)\|_{l_2}$ , which will be iterated for 5 times by default.

In the second step, we will minimize the dependence criterion. That is  $\hat{T}^{(j)} = \min_T \{HSIC(\hat{N}_X, \hat{N}_Y) + HSIC(\hat{N}_X, T) + HSIC(\hat{N}_Y, T)\}$ , where  $\hat{N}_X = X - \hat{u}(\hat{T}^{(j-1)})$  and  $\hat{N}_Y = Y - \hat{v}(\hat{T}^{(j-1)})$ . If  $\hat{N}_X \perp \hat{N}_Y$ ,  $\hat{N}_X \perp \hat{T}^{(j)}$  and  $\hat{N}_Y \perp \hat{T}^{(j)}$ , then we get the confounder  $\hat{T}^{(j)}$ . Otherwise, keep the iteration till maximum allowed steps.

### 2.4.2 Simulation Settings

A randomly generated nonlinear model will mimic the causal relation between two continuous variables. For example, Radial Basis Functions (RBF). That is,

$$b_k(x) = e^{-\frac{(x-c_k)^2}{2\sigma^2}}$$
$$f(x) = \sum w_k b_k(x)$$

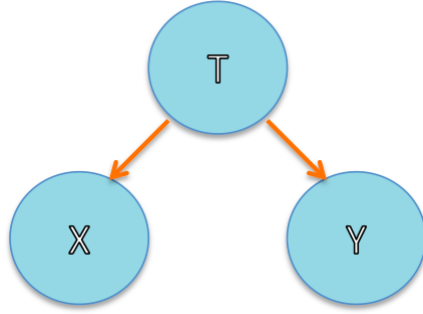
#### 2.4.2.1 No confounder exists.



**Figure 2.9** No confounder exists.

In each simulation,  $X$  is a randomly generated normal variable and  $Y = f(X) + N$ . I will simulate 1000 times, and the number of samples is still to be determined, since computational time increases with number of samples. Since there is no confounder in this system, Type I error rates will be calculated for the algorithm.

#### 2.4.2.2 Confounder exists.



**Figure 2.10** Confounder exists.

In each simulation, confounder  $T$  is a randomly generated normal variable, and then  $X = f(T) + N_1, Y = g(T) + N_2$ . I will simulate 1000 times. Since there exists confounder in this system, powers for each number of samples will be calculated for the algorithm.

## **2.5 Methods for Aim 3(a): To develop the causal inference test for multiple causes and one effect**

### **2.5.1 Statistical Modeling**

Suppose the number of sample is  $n$ . Given the effect variable  $Y$  and its parents set  $P = \{X_1, X_2, \dots, X_p\}$ , we consider an additive noise model (ANM):

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon$$

Let  $L$  be a continuous linear function defined on the Reproducing Kernel Hilbert Space (RKHS)  $H$ . Then,

$$Y^i = L_i f_d + \varepsilon_i, i = 1, \dots, n \quad (2.7)$$

where  $Y^i$  is the  $i^{th}$  sample of  $Y$ ,  $L_i$  is a continuous functional and  $\varepsilon_i$ 's are independent random errors with zero-mean and variance of  $\sigma_e^2$ .

Given  $p$  variables  $\{X_1, X_2, \dots, X_p\}$ , we have the tensor product  $H = H^{(1)} \otimes H^{(2)} \otimes \dots \otimes H^{(p)}$  on domain  $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_p$ , where  $H^{(k)} = H_0^{(k)} \oplus H_1^{(k)} \oplus \dots \oplus H_{r_k-1}^{(k)} \oplus H_{*1}^{(k)}$ ,  $k = 1, \dots, p$ . We further have

$$\begin{aligned} H &= H_0^* \oplus H_1^* \\ &= H^0 \oplus \{H^1 \oplus \dots \oplus H^q\} \end{aligned}$$

where  $H_0^* = H^0 = \sum_{j_1=0}^{r_1-1} \dots \sum_{j_p=0}^{r_p-1} H_{j_1}^{(1)} \otimes \dots \otimes H_{j_p}^{(p)}$ , a finite dimensional space including all functions that will not be penalized.  $H^1, \dots, H^q$  are orthogonal RKHS's with RKs  $R^1, \dots, R^q$ . The RK for the  $H_1^*$  is defined as  $R_1^* = \sum_{j=1}^q \theta_j R^j$ , where  $R^j$  is the RK for  $H^j$ .

Let  $\varphi_0^{(k)}, \dots, \varphi_{r_k-1}^{(k)}$  be the set of basis functions for space  $H_0^{(k)}$ . Then, all possible combinations of the basis functions for  $H^0$  are:

$$\left\{ \varphi_1^{(1)} \oplus \dots \oplus \varphi_{r_1}^{(1)} \right\} \dots \left\{ \varphi_1^{(p)} \oplus \dots \oplus \varphi_{r_p}^{(p)} \right\} = \sum_{j_1=0}^{r_1-1} \dots \sum_{j_p=0}^{r_p-1} \varphi_{j_1}^{(1)} \dots \varphi_{j_p}^{(p)} = \phi_1 + \dots + \phi_r,$$

where  $r = r_1 \dots r_p$  and  $\phi_v \in \left\{ \varphi_{j_1}^{(1)}, \dots, \varphi_{j_p}^{(p)} \right\}$ .

For simplicity, I consider only cubic spline, so  $r_1 = \dots = r_p = 2$  and then  $r = 2^p$ .  $q = 3^p - 2^p$ . When  $p=3$ , the basis functions are given by

$$\begin{aligned}
\phi_1(x_1, x_2, x_3) &= 1, \phi_2(x_1, x_2, x_3) = x_1 - 0.5, \phi_3(x_1, x_2, x_3) = x_2 - 0.5, \\
\phi_4(x_1, x_2, x_3) &= x_3 - 0.5, \phi_5(x_1, x_2, x_3) = (x_1 - 0.5)(x_2 - 0.5), \\
\phi_6(x_1, x_2, x_3) &= (x_1 - 0.5)(x_3 - 0.5), \phi_7(x_1, x_2, x_3) = (x_2 - 0.5)(x_3 - 0.5), \\
\phi_8(x_1, x_2, x_3) &= (x_1 - 0.5)(x_2 - 0.5)(x_3 - 0.5)
\end{aligned}$$

On the other hand, RKs  $R^j$  is the product of RKs of  $H_{j_1}^{(1)}, \dots, H_{j_p}^{(p)}$ , where  $R_0^{(l)} = 1, R_1^{(l)} = k_1(x_l)k_1(z_l)$  and  $R_2^{(l)} = k_2(x_l)k_2(z_l) - k_4(|x_l - z_l|)$  are RKs of  $H_0^{(l)}, H_1^{(l)}$  and  $H_2^{(l)}$  respectively.

To estimate smoothing splines regression in (2.7), we will minimize

$$\frac{1}{n} \sum_{i=1}^n (Y^i - L_i f_d)^2 + \sum_{j=1}^q \lambda_j \|P_j f\|^2$$

where  $P_j$ 's are the orthogonal projects of the function onto the RKHS  $H^j, j = 0, 1, \dots, q$ .

Once we have the fitted value  $\hat{Y}$  and thus the residuals  $\epsilon_i$  ( $i = 1, 2, \dots, n$ ), I will calculate and note down the entropy of residuals  $En_{obs}$ . If  $En_{obs}$  is significantly small, I infer that there exists causal relationship from multiple candidate factors to the effect Y. To measure how small  $En_{obs}$  is, I will use permutation test to generate a causation p-value,  $P_c$ . That is, resample Y as  $Y^{(k)}$  and get  $En_{permutation}^{(k)}$ , where  $k=1, \dots, N$ . Then,

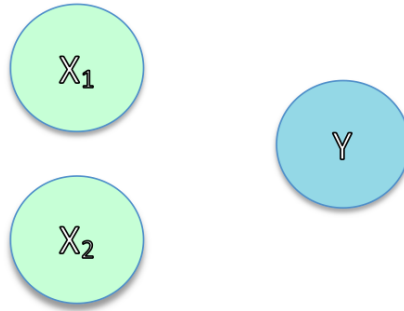
$$P_c = \frac{P\left(En_{obs} > En_{permutation}^{(k)}\right)}{N}$$

### 2.5.2 Simulation Settings

The causal relation between multiple factors and the effect will be mimicked by a randomly generated nonlinear regression based on Radial Basis Functions (RBF). That is,

$$b_k(\mathbf{x}) = e^{-\frac{\|\mathbf{x}-\mathbf{c}_k\|^2}{2\sigma^2}}$$
$$f(\mathbf{x}) = \sum w_k b_k(\mathbf{x})$$

#### 2.5.2.1 Independent Case



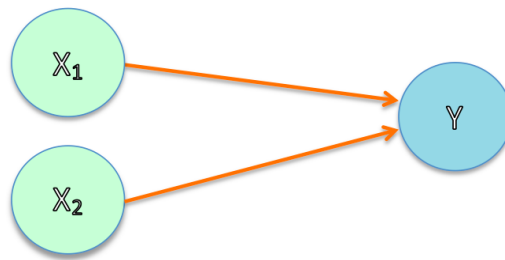
**Figure 2.11** Multiple covariates are independent of the response

$X_1$ ,  $X_2$  and  $Y$  are three randomly generated normal variables. I first generated 100,000 subjects and then randomly picked up 500 and 1000 samples from the population to calculate and compare their Type I error rates.

**Table 2.11** Type I error rates of causal inference test on multiple causes and one effect

Type I Error Rate	Number of Samples	
Significance level	N=500	N=1000
0.05	.048	.044
0.01	.012	.008

#### 2.5.2.2 Causal Case



**Figure 2.12** Multiple covariates causes the response

$X_1$ ,  $X_2$  and  $N$  are three randomly generated normal variables. Data was generated by formula

$$Y = f(X) + N = \sum_{k=1}^K w_k \times \exp(-\gamma \| \mathbf{X} - \mathbf{c}_k \|^2) + N$$

where  $\gamma, c_k \sim N(0, 0.01)$ ,  $w_k$ 's are randomly generated weights,  $k=1, 2, \dots, K$ . Then I will randomly pick up 500 and 1000 samples from the population to calculate and compare their powers.

**Table 2.12** Powers of causal inference test on multiple causes and one effect



Power	Number of Samples	
Significance level	N=500	N=1000
0.05	.988	.993
0.01	.966	.986

## 2.6 Methods for Aim 3(b): To develop the causal network model for high-dimensional data

### 2.6.1 Statistical Modeling

I will convert the causal network construction to a combinatorial optimization problem based on Integer programming. A standard form of integer linear programming is

$$\min c^T X$$

$$AX \leq b$$

$$X \in Z^n$$

where  $c \in R^n, b \in R^m, A \in R^{m \times n}$  and  $Z = \{0,1,2, \dots\}$

If all variables are restricted to the values from  $B=\{0,1\}$ , we have a 0-1-integer linear programming:

$$\min c^T X$$

$$AX \leq b$$

$$X \in B^n$$

where  $c \in R^n, b \in R^m, A \in R^{m \times n}$  and  $B = \{0,1\}$

Suppose the causal network is a directed acyclic graph (DAG), the optimization question can be converted to (Xiong 2018):

$$\begin{aligned} \min \quad & \sum_{v=1}^p \sum_{j_v=1}^{J_v} C(v, W_{j_v}) x(W_{j_v} \rightarrow v) \\ \text{constraints:} \quad & \sum_{j_v=1}^{J_v} x(W_{j_v} \rightarrow v) = 1, v = 1, \dots, p \end{aligned}$$

$$\forall C \subseteq V: \sum_{v \in C} \sum_{W_{j_v}: |W_{j_v} \cap C| < k, j_v=1, \dots, J_v} x(W_{j_v} \rightarrow v) \geq k, \forall k, 1 \leq k \leq |C| \quad (2.8)$$

$$x(W_{j_v} \rightarrow v) = 0 \text{ or } 1.$$

Here  $v$  refers to a specific node  $v$ , and  $W_{j_v}$  refers to one of the possible parent set of  $v$ .  $C(v, W_{j_v})$  denotes the score function for the pair of node  $v$  and its parent set  $W_{j_v}$ .  $x(W_{j_v} \rightarrow v) = 1$  if and only if  $W_{j_v}$  is the parent set for the node  $v$ . The constraint (2.8) is to ensure that there is no cycle in the DAG.

In real world, the causation relation from the parent set  $W_{j_v}$  to  $v$  is usually nonlinear, so I use a nonlinear score to represent  $C(v, W_{j_v})$ . The following is the algorithm (Xiong 2018) to get the score by the nonlinear regression:

Step 1: Select the penalty parameter  $\lambda$ . Define the node  $v$  as  $Y = [y_1, y_2, \dots, y_n]^T$  and the variables in  $W_{j_v}$  as  $x$ .

Step 2: Compute the matrices

$$T = \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_r(x_1) \\ \vdots & \vdots & \vdots \\ \phi_1(x_n) & \cdots & \phi_r(x_n) \end{bmatrix} \text{ and } \Sigma_j = \begin{bmatrix} R^j(x_1, z_1) & \cdots & R^j(x_1, z_n) \\ \vdots & \vdots & \vdots \\ R^j(x_n, z_1) & \cdots & R^j(x_n, z_n) \end{bmatrix}$$

$$\Sigma_\theta = \theta_1 \Sigma_1 + \cdots + \theta_q \Sigma_q, \text{ where } \theta_1, \dots, \theta_q \text{ are pre-determined weights.}$$

Step 3: Perform QR decomposition of the matrix  $T$ :

$$T = [Q_1 \ Q_2] \begin{pmatrix} R \\ 0 \end{pmatrix}$$

Step 4: Compute coefficients of the smoothing spline regression

$$\hat{a} = R^{-1} Q_1^T [I - M Q_2 (Q_2^T M Q_2)^{-1} Q_2^T] Y \text{ and } \hat{b} = Q_2 (Q_2^T M Q_2)^{-1} Q_2^T Y,$$

where  $M = \Sigma + n\lambda I$ .

Step 5: Compute the smoothing spline regression function

$$\hat{f}(x) = \sum_{j=1}^r \hat{a}_j \phi_j(x) + \sum_{v=1}^n \hat{b}_v \sum_{j=1}^q \theta_j L_{v(z)} R^j(x, z)$$

Step 6: Compute the fitted value:

$$\hat{f} = H(\lambda)Y, \text{ where } H(\lambda) = I - n\lambda Q_2 (Q_2^T M Q_2)^{-1} Q_2^T.$$

Step 7: Calculate the nonlinear score of the node  $v$ :

$$C(v, W_{j_v}) = \frac{1}{n} \|Y - T\hat{a} - \Sigma_\theta \hat{b}\|^2 + \hat{b} \Sigma_\theta \hat{b}.$$

### 2.6.2 Simulation Settings

First, randomly generated a DAG. Set nodes without parents as random normal distribution. Then use nonlinear/linear functions to generate the nodes in the next layer.

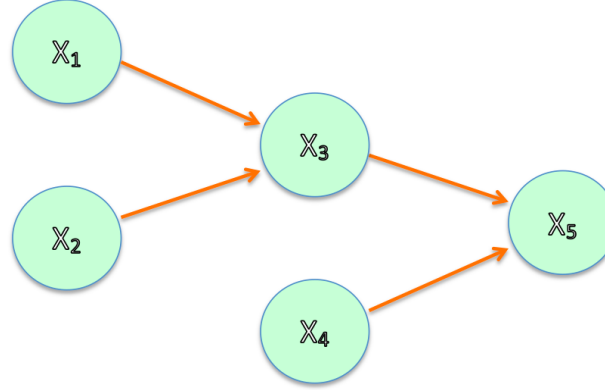


Figure 2.13 Causal network

For example, based on the above causal network, I will first generate three independent nodes  $X_1$ ,  $X_2$  and  $X_3$  that have normal distributions. Then randomly generate functions  $f$  and  $g$ , and generate  $X_3$  and  $X_5$  by  $X_3 = f(X_1, X_2) + N_1$  and  $X_5 = f(X_3, X_4) + N_2$ .

I randomly generated a DAG and the network data for 1,000 times. Let  $N_t$  be the total number of edges among 1,000 networks,  $N_o$  the total number of edges that do not appear in 1,000 networks,  $N_{True}$  the total number of edges detected by the algorithm and  $N_{False}$  the false edges directed among  $N_o$ . Then the false discovery rate (FDR) is defined by  $\frac{N_{False}}{N_o}$  and

power of detection (PD)  $\frac{N_{true}}{N_t}$ . The proposed numbers of sample is 500, 1000 and 2000, but

they may be changed according to computational resources and statistical performance.

## **2.7 Methods for Real Data Application for Proposed Aims**

Here I will summarize the methods of real data application for three proposed aims. I will apply the developed novel methods for both Aims 1 (a) and 2 to the gene expression dataset with 51,060 genes and 432 subjects from Rush University Medical Center. The gene variable is regarded as possible cause, effect or confounder in testing the performance of the novel methods.

For Aim 1 (b), I will apply the method to CATIE-MGS-SWD schizophrenia study dataset with 8,421,111 common SNPs typed in 13,557 participants and UK Biobank dataset with over 90 million SNPs and half million participants. For the schizophrenia dataset, I will use the common variants imputed by Shapeit and Impute2. I will follow conventional quality control (QC) criteria for GWAS. For example, the MAF of any SNP should be greater than 5%. In UK Biobank dataset, I will exclusively use the white British samples (n = 429, 512). Specifically, I will select coronary artery disease as the case (n=9,771) and samples with no recorded diseases as control.

In Aim 3, I will define the gene pathway by public pathway resources like KEGG (Ogata et al., 1999). The selected database is DNA Methylation and gene expression datasets with 448

subjects from Religious Orders Study or Rush Memory and Aging Project (ROSMAP). I will consider the mid-size pathways, for example, the pathways with 20 to 100 genes.

Via the real data applications, I expect to verify known risky genes and/or SNPs related to cardiovascular disease, schizophrenia or Alzheimer's disease, in order to validate my proposed methods. I also look forward to identify the novel risky genes and/or SNPs, providing novel valuable information to the disease research consortium.

## **2.8 Declaration on Human Subjects**

This dissertation study focuses on statistical method development. I used the Rush Alzheimer's Disease dataset and UK Biobank dataset for method demonstration purpose. I used the imaging, genotype, gene expression and methylation data in the Rush Alzheimer's Disease dataset and UK Biobank dataset. All data are pre-existing and de-identified. The IRB approval for the use of Rush Alzheimer's Disease dataset and UK Biobank dataset in my dissertation research was obtained by my dissertation advisor, Dr. Momiao Xiong, under UTHealth IRB approval (HSC-SPH-18-0819).

### **3. Results**

#### **3.1 Bivariate causal discovery for continuous variables in genetic and imaging data analysis**

##### **3.1.1 Introduction**

Despite significant progress in dissecting the genetic architecture of complex diseases by association analysis, understanding the etiology and mechanism of complex diseases remains elusive. Using association analysis and machine learning systems that operate, almost exclusively, in a statistical, or model-free modes as a major analytic platform for genetic studies of complex diseases is a key issue that hampers the discovery of mechanisms underlying complex traits (Pearl 2018).

As an alternative to association analysis, causal inference may provide tools for unraveling principles underlying complex traits. Power of causal inference is its ability to predict effects of actions on the system (Mooij et al. 2016). Typical methods for unraveling cause-and-effect relationships are interventions and controlled experiments. Unfortunately, the experiments in human genetics are unethical and technically impossible. Next generation genomic, epigenomic, sensing and image technologies produce ever deeper multiple omic, physiological, imaging, environmental and phenotypic data with millions of features. These data are almost all “observational”, which have not been randomized or otherwise experimentally controlled (Glymour 2015). In the past decades, a variety of statistical methods and computational algorithms for causal inference that attempts to abstract causal knowledge from purely

observational data, referred to as causal discovery, have been developed (Zhang et al. 2018). Causal inference is one of the most useful tools developed in the past century. The classical causal inference theory explores conditional independence relationships in the data to discover causal structures. The PC algorithms and the fast causal inference (FCI) algorithms developed at Carnegie Mellon University by Peter Spirtes and Clark Glymour are often used for cause discovery (Le et al. 2016). Despite its fundamental role in science, engineering and biomedicine, the conditional independence-based classical causal inference methods can only identify the graph up to its Markov equivalence class, which consists of all DAGs satisfying the same conditional independence distributions via the causal Markov conditions (Nowzohour and Bühlmann 2016). For example, consider three simple DAGs:  $x \rightarrow y \rightarrow z$ ,  $x \leftarrow y \leftarrow z$  and  $x \leftarrow y \rightarrow z$ . Three variables  $x, y$  and  $z$  in all three DAGs satisfy the same causal Markov condition:  $x$  and  $z$  are independent, given  $y$ . This indicates that these three DAGs form a Markov equivalence class. However, these three DAGs represent three different causal relationships among variables  $x, y$  and  $z$ , which prohibits unique causal identification. These non-unique causal solutions seriously limit their translational application.

In the past decade, causal inference theory is undergoing exciting and profound changes from discovering only up to the Markov equivalent class to identify unique causal structure (Peters et al. 2011; Peters and Bühlman, 2014). A class of powerful algorithms for finding a unique causal solution are based on properly defined functional causal models (FCMs).



They include the linear, non-Gaussian, acyclic model (LiNGAM) (Zhang et al. 2018; Shimizu et al. 2006), the additive noise model (ANM) (Hoyer et al. 2009; Peters et al. 2014), and the post-nonlinear (PNL) causal model (Zhang and Hyvärinen 2009).

In genomic and epi-genomic data analysis, we usually consider four types of associations: association of discrete variables (DNA variation) with continuous variables (phenotypes, gene expressions, methylations, imaging signals and physiological traits), association of continuous variables (expressions, methylations and imaging signals) with continuous variables (gene expressions, imaging signals, phenotypes and physiological traits), association of discrete variables (DNA variation) with binary trait (disease status) and association of continuous variables (gene expressions, methylations, phenotypes and imaging signals) with binary trait (disease status). All these four types of associations can be extended to four types of causations. This paper focuses on studying causal relationships between two continuous variables.

The many causal inference algorithms using observational data require that two variables being considered as cause-effect relationships are part of a larger set of observational variables (Mooij et al. 2016). Similar to genome-wide association studies where only two variables are considered, we mainly investigate bivariate causal discovery to infer cause-effect relationships between two observed variables. To simplify the cause discovery studies, we assume no selection bias, no feedback and no confounding. We first introduce

the basic principle underlying the modern causal theory. It assumes that nature consists of autonomous and independent causal generating process modules and attempts to replace causal faithfulness by the assumption of Independence of Cause and Mechanism (ICM) (Peters et al. 2017; Besserve et al. 2017; Schölkopf et al. 2012; Janzing et al. 2010; Lemeire et al. 2012). Then, we will present ANM as a major tool for causal discovery between two continuous variables. We will investigate properties of ANM for causal discovery. Finally, the ANM will be applied to gene expression data to infer gene regulatory networks and longitudinal phenotype-imaging data to identify brain regions affected by intermediate phenotypes.

### 3.1.2 Data and Notation

Assume no confounding, no selection bias and no feedback. Consider a bivariate additive noise model  $X \rightarrow Y$  where  $Y$  is a nonlinear function of  $X$  and independent additive noise  $E_Y$ :

$$\begin{aligned} Y &= f_Y(X) + E_Y \\ X &\sim P_X, E_Y \sim P_{E_Y}, \end{aligned} \tag{3.1}$$

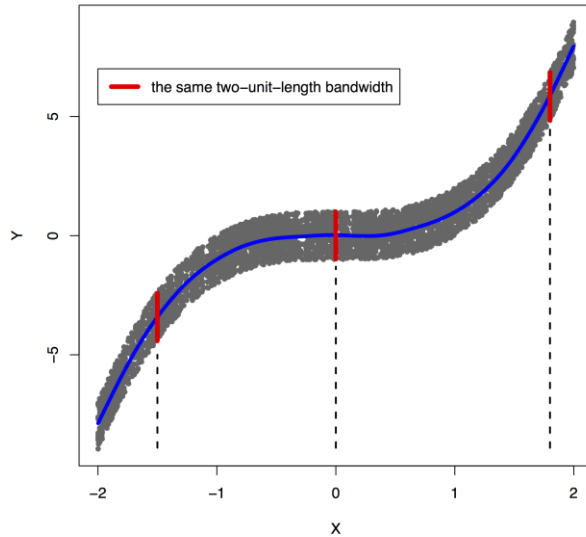
where  $X$  and  $E_Y$  are independent. Then, the density  $P_{X,Y}$  is said to be induced by the additive noise model (ANM) from  $X$  to  $Y$  (Mooij et al. 2016). The alternative additive noise model between  $X$  and  $Y$  is the additive noise model  $Y \rightarrow X$ :

$$\begin{aligned} X &= f_X(Y) + E_X \\ Y &\sim P_Y, E_X \sim P_{E_X}, \end{aligned} \tag{3.2}$$

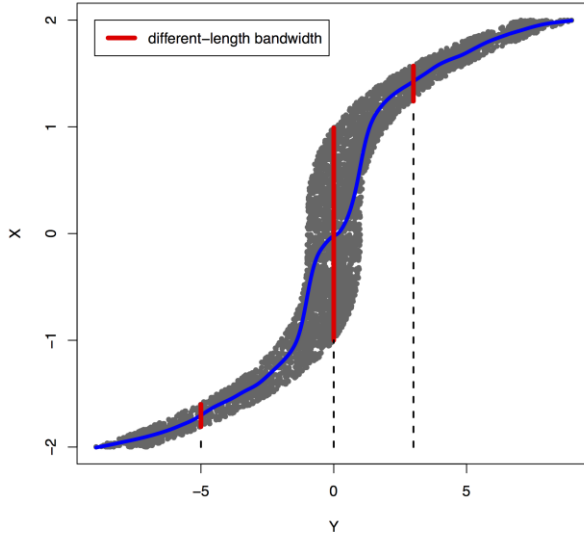
where  $Y$  and  $E_X$  are independent.

If the density  $P_{X,Y}$  is induced by the ANM  $X \rightarrow Y$ , but not by the ANM  $Y \rightarrow X$ , then the ANM  $X \rightarrow Y$  is identifiable. To illustrate application of the algorithmic mutual information, we show that independence of cause and mechanism will imply that the cause  $X$  and error  $E_Y$  in the nonlinear function model (3.1) are independent.

Peters et al. (2017) showed that a joint distribution  $P_{X,Y}$  does not admit an ANM in both directions at the same time under some quite generic conditions. To illustrate that ANMs are generally identifiable, i.e., a joint distribution only admits an ANM in one direction, we plotted Figures 3.1 and 3.2. The data in Figures 3.1 and 3.2 were generated by  $Y = X^3 + E_Y$ , where  $E_Y$  is uniformly distributed in  $[-1, 1]$ .



**Figure 3.1** An example of joint distribution  $p(x, y)$  generated by  $Y := f(X) + E_Y$ , where  $f(X) = X^3$  and  $E_Y$  is uniformly distributed in  $[-1, 1]$ . The interval of the red line represents the bandwidth of the conditional distribution  $p_{Y|X}$ .



**Figure 3.2** An example of joint distribution  $p(x, y)$  generated by  $Y := f(X) + E_Y$ , where  $f(X) = X^3$  and  $E_Y$  is uniformly distributed in  $[-1, 1]$ . The interval of the red line represents the bandwidth of the conditional distribution  $p_{X|Y}$ .

The joint distribution satisfied an ANM  $X \rightarrow Y$ , but did not admit an ANM  $Y \rightarrow X$ . We plotted Figures 3.1 and 3.2 in which red lines indicated the bandwidth of the conditional distribution. Figure 3.1 showed that all bandwidth of the conditional distribution  $P_{Y|X}$  represented by the red line was two units. This clearly demonstrated that conditional distribution  $P_{Y|X}$  did not depend on the cause  $X$ . However, Figure 3.2 showed that the bandwidth of the conditional distribution  $P_{X|Y}$ , represented by the red line varied as  $Y$  changed. This demonstrated that the conditional distribution  $P_{X|Y}$ , indeed, depended on  $Y$ . In other words, it violated the principal of independence of cause and mechanism. The joint distribution in this example only admitted an ANM in only one direction  $X \rightarrow Y$ .

The ANMs should assume that the functions  $f_X$  and  $g_Y$  are nonlinear. If the functions are linear, then additional assumptions for identifiability should be made. In other words, for the

linear functions, if at least one of the distributions of the cause and noise is non-Gaussian (e.g., linear non-Gaussian acyclic model (LiNGAM)), then the linear model is identifiable. Otherwise, the linear model is not identifiable (Moneta et al. 2013; Shimizu et al. 2011). In this scenario, we cannot get different bandwidths. The limitation of the ANMs is that it cannot be applied to linear case if both distributions of cause and noise are Gaussian.

Empirically, if the ANM  $X \rightarrow Y$  fits the data, then we infer that  $X$  causes  $Y$ , or if the ANM  $Y \rightarrow X$  fits the data, then  $Y$  causes  $X$  will be concluded. Although this statement cannot be rigorously proved, in practice, this principle will provide the basis for bivariate cause discovery (Mooij et al. 2016). To implement this principal, we need to develop statistical methods for assessing whether the additive noise model fits the data or not.

Now we summarize procedures for using ANM to assess causal relationships between two variables. Two variables can be two gene expressions, or one gene expression and one methylation level of CpG site, or an imaging signal of one brain region and a functional principal score of gene. Divide the dataset into a training data set by specifying  $D_{train} = \{Y_n, X_n\}$ ,  $Y_n = [y_1, \dots, y_n]^T$ ,  $X_n = [x_1, \dots, x_n]^T$  for fitting the model and a test data set  $D_{test} = \{\tilde{Y}_m, \tilde{X}_m\}$ ,  $\tilde{Y}_m = [\tilde{y}_1, \dots, \tilde{y}_m]^T$ ,  $\tilde{X}_m = [\tilde{x}_1, \dots, \tilde{x}_m]^T$  for testing the independence, where  $n$  is not necessarily equal to  $m$ .

*Algorithm for causal discovery with two continuous variables is given below.*

Step 1: Regress  $Y$  on  $X$  using the training dataset  $D_{train}$  and non-parametric regression methods:

$$Y = \hat{f}(X) + E_Y. \quad (3.3)$$

Step 2: Calculate residual  $\hat{E}_Y = Y - \hat{f}(X)$  using the test dataset  $D_{test}$  and test whether the residual  $\hat{E}_Y$  is independent of causal  $X$  to assess the ANM  $X \rightarrow Y$ .

Step 3: Repeat the procedure to assess the ANM  $Y \rightarrow X$ .

Step 4: If the ANM in one direction is accepted and the ANM in the other is rejected, then the former is inferred as the causal direction.

There are many non-parametric methods that can be used to regress  $Y$  on  $X$  or regress  $X$  on  $Y$ . For example, we can use smoothing spline regression methods (Wang 2011), B-spline (Wang 2017) and local polynomial regression (LOESS, see Cleveland, 2012).

Covariance can be used to measure association, but cannot be used to test independence between two variables. A covariance operator can measure the magnitude of dependence, and is a useful tool for assessing dependence between variables. Specifically, we will use the Hilbert-Schmidt norm of the cross-covariance operator or its approximation, the Hilbert-Schmidt independence criterion (HSIC) to measure the degree of dependence between the residuals and potential causal variable (Gretton et al. 2005; Mooij et al. 2016).

*Calculation of the HSIC consists of the following steps.*

Step 1: Use test data set to compute

$$y_i = \hat{f}(x_i) + E_Y(i), i = 1, \dots, m.$$

Step 2: Compute the residuals:

$$\varepsilon_i = E_Y(i) = y_i - \hat{f}(x_i), i = 1, \dots, m.$$

Step 3: Select two kernel functions  $k_E(\varepsilon_i, \varepsilon_j)$  and  $k_x(x_i, x_j)$ . In practice, we often use the

Gaussian kernel function. Compute the Kernel matrices:

$$K_{E_Y} = \begin{bmatrix} k_E(\varepsilon_1, \varepsilon_1) & \cdots & k_E(\varepsilon_1, \varepsilon_m) \\ \vdots & & \vdots \\ k_E(\varepsilon_m, \varepsilon_1) & \cdots & k_E(\varepsilon_m, \varepsilon_m) \end{bmatrix}, K_x = \begin{bmatrix} k_x(x_1, x_1) & \cdots & k_x(x_1, x_m) \\ \vdots & & \vdots \\ k_x(x_m, x_1) & \cdots & k_x(x_m, x_m) \end{bmatrix}.$$

Step 4: compute the HSCI for measuring dependence between the residuals and potential causal variable.

$$HSIC^2(E_Y, X) = \frac{1}{m^2} \text{Tr}(K_{E_Y} H K_X H),$$

where  $H = I - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T$ ,  $\mathbf{1}_m = [1, 1, \dots, 1]^T$  and  $\text{Tr}$  denotes the trace of the matrix.

*In summary, the general procedure for bivariate causal discovery is given as follows (Mooij et al. 2016):*

Step 1: Divide a data set into a training data set  $D_{train} = \{Y_n, X_n\}$  for fitting the model and a test data set  $D_{test} = \{\tilde{Y}_m, \tilde{X}_m\}$  for testing the independence.

Step 2: Use the training data set and nonparametric regression methods

(c) Regress  $Y$  on  $X$  :  $Y = f_Y(X) + E_Y$  and

(d) Regress  $X$  on  $Y$  :  $X = f_X(X) + E_X$ .

Step 3: Use the test data set and estimated nonparametric regression model that fits the training data set  $D_{train} = \{Y_n, X_n\}$  to predict residuals:

$$(c) \hat{E}_{Y_X} = \tilde{Y} - \hat{f}_Y(\tilde{X})$$

$$(d) \hat{E}_{X_Y} = \tilde{X} - \hat{f}_X(\tilde{Y}).$$

Step 4: Calculate the dependence measures  $HSIC^2(E_Y, X)$  and  $HSIC^2(E_X, Y)$ .

Step 5: Infer causal direction:

$$X \rightarrow Y \text{ if } HSIC^2(E_Y, X) < HSIC^2(E_X, Y); \quad (3.4)$$

$$Y \rightarrow X \text{ if } HSIC^2(E_Y, X) > HSIC^2(E_X, Y). \quad (3.5)$$

If  $HSIC^2(E_Y, X) = HSIC^2(E_X, Y)$ , then causal direction is undecided.

We do not have closed analytical forms for the asymptotic null distribution of the HSIC and hence it is difficult to calculate the P-values of the independence tests. To overcome these limitations, the permutation/bootstrap approach can be used to calculate the P-values of the causal test statistics. The null hypothesis is

$H_0$ : no causations  $X \rightarrow Y$  and  $Y \rightarrow X$  (Both  $X$  and  $E_Y$  are dependent, and  $Y$  and  $E_X$  are dependent).

Calculate the test statistic:



$$T_C = |HSIC^2(E_Y, X) - HSIC^2(E_X, Y)|. \quad (3.6)$$

Assume that the total number of permutations is  $n_p$ . For each permutation, we fix  $x_i, i = 1, \dots, m$  and randomly permute  $y_i, i = 1, \dots, m$ . Then, fit the ANMs and calculate the residuals  $E_X(i), E_Y(i), i = 1, \dots, m$  and test statistic  $T_C$ . Repeat  $n_p$  times. The P-values are defined as the proportions of the statistic  $\tilde{T}_C$  (computed on the permuted data) greater than or equal to  $\hat{T}_C$  (computed on the original data  $D_{TE}$ ). After cause is identified, we then use equations (3.4) and (3.5) to infer causal directions  $X \rightarrow Y$  or  $Y \rightarrow X$ .

### 3.1.3 Linear Correlation and Causation

In everyday language, correlation and association are used interchangeably. However, correlation and association are different terminologies. Pear correlation coefficient is defined as  $\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y}$  from covariance, Spearman correlation coefficient is defined as measuring increasing or decreasing trends. Association characterizes dependence between two variables (Altman and Krzywinski 2015). In this paper, association is equivalent to Pearson linear correlation. We will focus on linear correlation. We investigate the relationships between causation and correlation. The correlation between two continuous variables can be investigated by a linear regression model:

$$Y = \beta X + \varepsilon, \quad (3.7)$$

where  $\beta \neq 0$ .

The causation  $X \rightarrow Y$  is identified by the ANM:

$$Y = f(X) + \varepsilon, \quad X \perp\!\!\!\perp \varepsilon. \quad (3.8)$$

In classical statistics, if we assume that both variables  $X$  and  $\varepsilon$  follow a normal distribution, then  $cov(X, \varepsilon) = 0$  if and only if  $X$  and  $\varepsilon$  are independent. If  $X$  and  $\varepsilon$  are not normal variables, this statement will not hold. For general distribution, we extend the concept of covariance to cross covariance operator  $\tilde{C}_{X\varepsilon}$  (Zhang et al. 2017). It is shown that for the general distributions of  $X$  and  $\varepsilon$ ,  $\tilde{C}_{X\varepsilon} = 0$  if and only if  $X$  and  $Y$  are independent (Mooij et al. 2016).

Let  $h$  and  $g$  be any two nonlinear functions.  $\tilde{C}_{X\varepsilon} = 0$  is equivalent to (Gretton et al. 2005)

$$\max cov(h(X), g(\varepsilon)) = \max cov(h(X), g(Y - f(X))) = 0, \quad (3.9)$$

Subject to  $\|h\| = 1, \|g\| = 1$ .

Now we give examples of a pair of random variables to illustrate existence of three cases: a) both linear correlation and causation  $X \rightarrow Y$ , b) causation  $X \rightarrow Y$ , but no linear correlation and c) linear correlation, but no causation  $X \rightarrow Y$ .

*a) Both linear correlation and causation  $X \rightarrow Y$ .*

We consider a special case:  $Y = f(X)$ . When  $Y = f(X)$ , equation (3.9) holds, which implies  $X \rightarrow Y$ . If we assume that  $h(X) = X$  and  $g(Y - f(X)) = Y - f(X)$ , then equation (3.9) holds and implies that

$$\text{cov}(X, Y) = \text{cov}(X, f(X)). \quad (3.10)$$

If we further assume  $f(X) = \beta X$ , then equation (3.10) implies

$$\beta = \frac{\text{cov}(X, Y)}{\text{Var}(X)}. \quad (3.11)$$

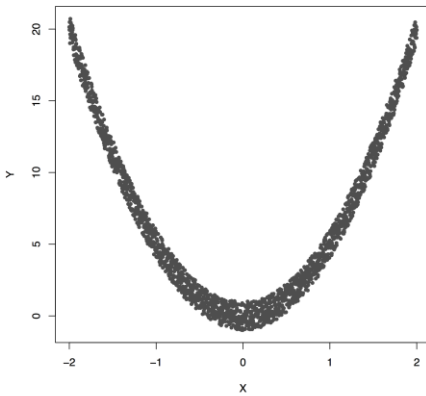
This is estimation of linear regression coefficient.

*b) Causation  $X \rightarrow Y$ , but no linear correlation*

Consider the model:

$$Y = 5X^2 + \varepsilon,$$

where  $X$  follows a uniform distribution between  $-2$  and  $2$  and  $\varepsilon$  follows a uniform distribution between  $-1$  and  $1$ .



**Figure 3.3** The data generated by  $Y = 5X^2 + \varepsilon$ , where  $X$  follows a uniform distribution between  $-2$  and  $2$  and  $\varepsilon$  follows a uniform distribution between  $-1$  and  $1$ .

Figure 3.3 plotted functions  $Y = 5X^2 + \varepsilon$ . Assume that 2,000 subjects were sampled. Permutation was used to calculate P-value for testing causation. We found that the Pearson correlation was  $-0.00070$  and P-value for testing causation  $X \rightarrow Y$  was  $10^{-5}$ . This example showed the presence of causation, but lack of linear correlation (Pearson correlation was near zero).

*c) Linear correlation, but no causation  $X \rightarrow Y$ .*

Consider the model:

$$X = Z + \varepsilon_1,$$

$$Y = Z + \varepsilon_2,$$

and  $Z \sim N(0,2)$ ,  $\varepsilon_1 \sim N(0,1)$ ,  $\varepsilon_2 \sim N(0,1)$ ,  $Z, \varepsilon_1, \varepsilon_2$  are independent.

The model can be rewritten as

$$Y = X + \varepsilon_2 - \varepsilon_1.$$

First we show that linear correlation between  $Y$  and  $X$  exists. In fact,

$$\text{cov}(Y, X) = \text{cov}(Z + \varepsilon_2, Z + \varepsilon_1) = \text{var}(Z) = 2, \text{Var}(Y) = 3, \text{Var}(X) = 3.$$

Thus, the Pearson linear correlation coefficient is equal to  $\rho = \frac{2}{3}$ . Thus, linear correlation between  $Y$  and  $X$  exists.

Next we show that  $X$  and  $\varepsilon_2 - \varepsilon_1$  are not independent. Note that  $cov(X, \varepsilon_2 - \varepsilon_1) = -var(\varepsilon_1) = -1$  and  $X, \varepsilon_2 - \varepsilon_1$  follow normal distribution. Since the covariance between  $X$  and  $\varepsilon_2 - \varepsilon_1$  is not equal to zero, this implies that  $X$  and  $\varepsilon_2 - \varepsilon_1$  are not independent. The conditional distribution  $P_{Y|X}$  is the distribution of  $\varepsilon_2 - \varepsilon_1$ . But, we show that the normal variables  $X$  and  $\varepsilon_2 - \varepsilon_1$  are not independent. This implies that the distribution  $P(X)$  and  $P_{Y|X}$  are not independent. Therefore, we finally show that there is no causation  $X \rightarrow Y$ . Similar conclusions hold for  $Y \rightarrow X$ .

#### *ANMs with Different Nonlinear Functions*

To investigate their feasibility for causal inference, the ANMs were applied to simulation data. Similar to Nowzohour and Bühlmann (2016), we considered three nonlinear functions: quadratic, exponential and logarithm functions and two random noise variables: normal and  $t$  distribution. We assumed that the cause  $X$  follows a normal distribution  $N(0,1)$ .

First we consider two models with a quadratic function and two types of random noise variables, normal  $N(0,1)$  and  $t$  distribution with 5 degrees of freedom:

Model 1:

$$Y = X + b \cdot X^2 + \varepsilon_1,$$

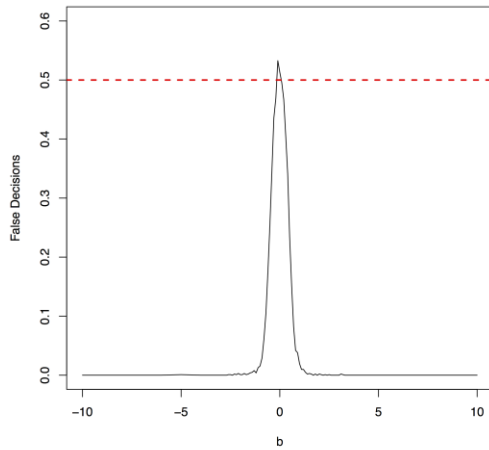
where the parameter  $b$  ranges from -10 to 10 and  $\varepsilon_1$  is distributed as  $N(0,1)$ .

Model 2:

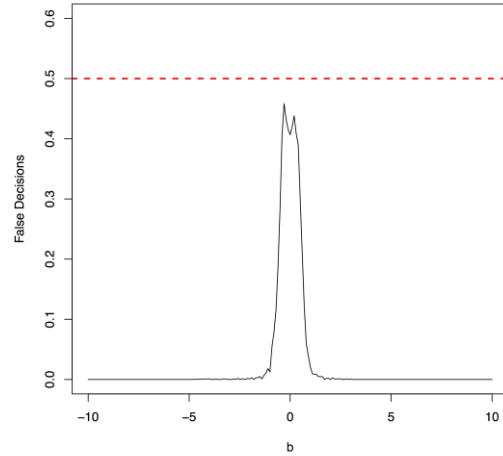
$$Y = X + b \cdot X^2 + \varepsilon_2,$$

where the parameter  $b$  is defined as before and  $\varepsilon_2$  is distributed as  $t$  distribution with 5 degrees of freedom.

The parameter space  $b \in [-10, 10]$  was discretized. For each grid point, 1,000 simulations were repeated. For each simulation, 500 samples were generated. The ANMs were applied to the generated data. Smoothing spline is used to fit the functional model. The true causal direction is the forward model:  $X \rightarrow Y$ . The false decision rate was defined as the proportion of times when the backward model  $Y \rightarrow X$  is wrongly chosen by the ANMs. Figures 3.4 and 3.5 presented false decision rate as a function of the parameter  $b$  for the models 1 and 2, respectively. We observed from Figures 3.4 and 3.5 that the false decision rate reached its maximum 0.5 when  $b = 0$ . This showed that when the model is close to linear, the ANMs



**Figure 3.4** False decision rates as a function of the parameter  $b$  for the model 1.



**Figure 3.5** False decision rates as a function of the parameter  $b$  for the model 2.

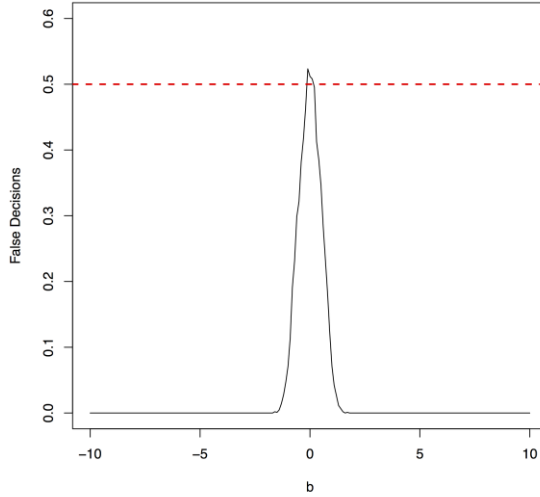
could not identify the true causal direction. However, when  $b$  moved away from 0, the false decision rates approached 0 quickly. This showed that when the data was generated by nonlinear models, with high probability, we can accurately identify the true causal directions. To further confirm these observations, we consider another two nonlinear functions.

Model 3:

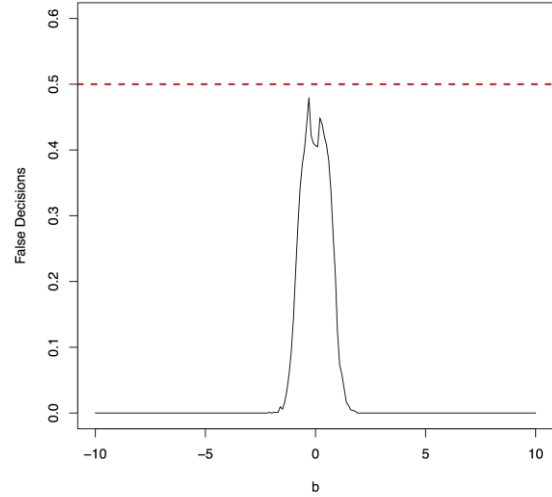
$$Y = X + b \log(|X|) + \varepsilon_1,$$

Model 4:

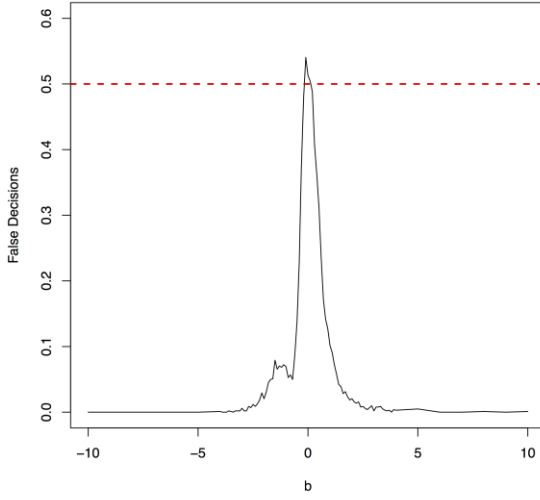
$$Y = X + b \log(|X|) + \varepsilon_2,$$



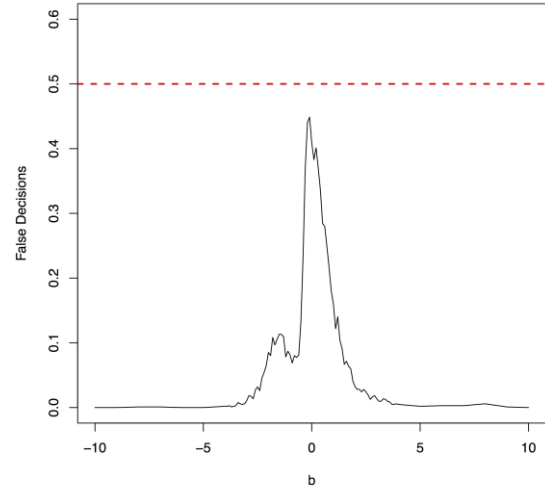
**Figure 3.6** The false decision rates of the ANMs for detecting the true causal direction  $X \rightarrow Y$  for the model 3.



**Figure 3.7** The false decision rates of the ANMs for detecting the true causal direction  $X \rightarrow Y$  for the model 4.



**Figure 3.8** The false decision rates of the ANMs for detecting the true causal direction  $X \rightarrow Y$  for the model 5.



**Figure 3.9** The false decision rates of the ANMs for detecting the true causal direction  $X \rightarrow Y$  for the model 6.

Model 5:

$$Y = X + b \cdot e^X + \varepsilon_1,$$

Model 6:

$$Y = X + b \cdot e^X + \varepsilon_2,$$

where the parameter  $b$  and the noise variables  $\varepsilon_1$  and  $\varepsilon_2$  were defined previously.

The false decision rates of the ANMs for detecting the true causal direction  $X \rightarrow Y$  for the models 3, 4, 5 and 6 were presented in Figures 3.6, 3.7, 3.8 and 3.9, respectively. Again, the observations for the models 1 and 2 still held for the models 3, 4, 5 and 6. When the data were generated by nonlinear models, we can accurately identify the true causal directions. However, when the data were generated by linear models, the false decision rates reached



0.5, which was equivalent to random guess.

### *Type I Error Rates*

To evaluate the performance of the ANMs for bivariate cause discovery, we calculate the type I error rates. We consider two scenarios: (a) no association, (b) presence of association.

#### (a) No Association

We first generated the data with 100,000 subjects from the model:  $X \sim N(0,1), Y \sim N(0,1)$  and  $X, Y$  are independent. Number of permutations was 500. Number of replication of tests was 1,000. The sampled subjects from the generated population for type I error rate calculations were 500, 1,000 and 2,000 respectively. The test statistic  $T_c$  and permutations were used to test for causation between two variables  $X$  and  $Y$ . Table 3.1 summarized type 1 error rates of the ANMs for testing causation, assuming no association.

**Table 3.1** Type 1 error rates of the ANMs for testing causation, assuming no association.

	Number of Samples		
Nominal Levels	500	1000	2000
0.05	0.033	0.051	0.043
0.01	0.006	0.005	0.01

#### (b) Presence of Association

Then, we generated the data with 100,000 subjects from the model:  $X \sim N(0,1), Y \sim N(0,1)$ ,  $X$  and  $Y$  were associated, but without causation. Number of

permutations was 500. Number of replication of tests was 1,000. The sampled subjects from the generated population for type I error rate calculations were 500, 1,000 and 2,000 respectively. The test statistic  $T_c$  and permutations were used to test for causation between two variables  $X$  and  $Y$ . Table 3.2 summarized type I error rates of the ANMs for testing causation in the presence of association.

**Table 3.2** Type 1 error rates of the ANMs for testing causation in the presence of association.

	Number of Samples		
Nominal Levels	500	1000	2000
0.05	0.044	0.048	0.050
0.01	0.011	0.011	0.011

In summary, Tables 3.1 and 3.2 showed that type I error rates of the ANM based on permutation even in the presence of association were not significantly deviated from nominal levels.

### *Power Simulations*

To further evaluate the performance of the ANMs for bivariate cause discovery, we used simulated data to estimate their power to detect causation. We generated data with 100,000 subjects from the causal model:

$$Y = f(X) + N,$$

where  $f(x) = \sum_{j=1}^3 w_j \times \exp\left(-\gamma(x - x_j)^2\right)$ ,  $\gamma \sim N(0,1)$ ,  $x_j \sim N(0,1)$ ,  $X \sim N(0,1)$  and  $N \sim N(0, \sigma^2 = 0.01)$ .  $X$  and  $N$  are independent, and  $w_j$ 's are randomly generated weights

from the uniform distribution. Number of permutations was 500. Number of replication of tests was 1,000. The sampled subjects from the population were 200, 500, 1,000, 2,000 and 5,000 respectively. The test statistic  $T_c$  and permutations were used to test for causation between two variables  $X$  and  $Y$ . Table 3.3 summarized the power of the ANMs for detecting causation between two variables.

**Table 3.3** Power of the ANMs for detecting causation between two variables.

Significance Level	Number of Samples				
	200	500	1000	2000	5000
0.05	0.3616	0.4833	0.5629	0.5997	0.6412
0.01	0.2066	0.3556	0.4382	0.4762	0.5241

### 3.1.4 Application to KEGG Pathway

Regulation of gene expression is a complex biological process. Large-scale regulatory network inference provides a general framework for comprehensively learning regulatory interactions, understanding the biological activity, devising effective therapeutics, identifying drug targets of complex diseases and discovering the novel pathways. Uncovering and modeling gene regulatory networks are one of the long-standing challenges in genomics and computational biology. Various statistical methods and computational algorithms for network inference have been developed. The ANMs can also be applied to inferring gene regulatory networks using gene expression data. Similar to co-gene expression networks where correlations are often used to measure dependence between two gene expressions, the ANMs can be used to infer regulation direction, i.e., whether changes in expression of gene  $X$  causes

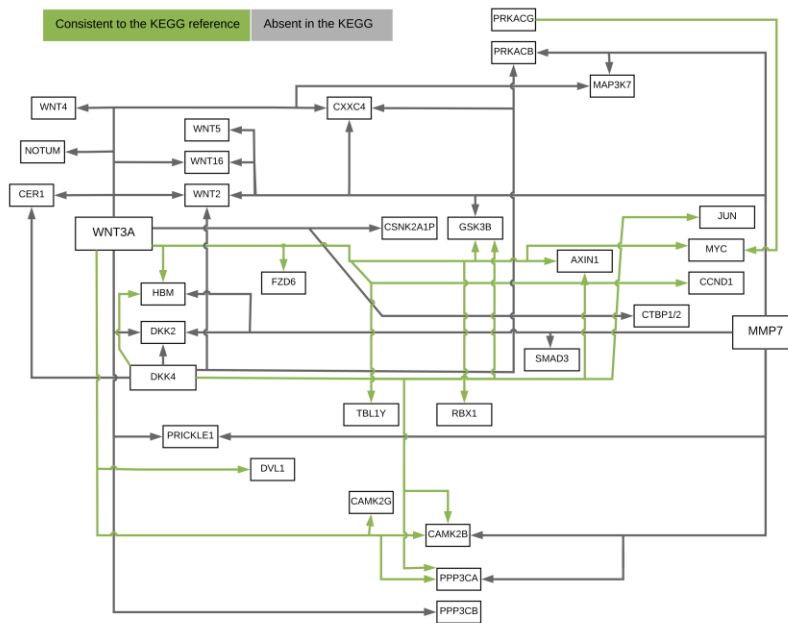
changes in expression of gene  $Y$  or vice versa changes in expression of gene  $Y$  causes changes in expression of gene  $X$ .

The ANMs were applied to Wnt signaling pathway with RNA-Seq of 79 genes measured in 447 tissue samples in the ROSMAP dataset (White et al. 2017). For comparisons, the structural equation models (SEMs) integrating with integer programming (Xiong 2018), causal additive model (CAM) (Bühlmann et al. 2014), PC algorithm (Tan et al. 2011), random network, glasso (Friedman et al. 2008), and Weighted Correlation Network Analysis (WGCNA) (Langfelder and Horvath, 2008) were also included in the analysis. We

**Table 3.4** Accuracy of the ANMs and other six methods for inferring Wnt pathway.

Wnt Pathway	Directed Paths			Undirected Paths Included		
Top Selected Edge Number	40	50	60	40	50	60
Pairwise ANM	37.50%	38%	35%	47.50%	46%	41.70%
CAM	17.50%	16%	13.30%	25%	24%	25%
SEM	22.50%	20%	15%	32.50%	26%	25%
Random Network	25.80%	25.40%	25.40%	31%	30.60%	30.50%
PC Algorithm	19.50%	21.60%	16.40%	36.60%	39.20%	27.90%
WGCNA Association	X	X	X	25%	22%	23.30%
Glasso	X	X	X	25%	28%	26.70%

ranked directed edges according to the values of the test statistics for the ANMs. The results for top 40, 50 and 60 edges were included in comparison. The results were summarized in Table 3.4. True directed path was defined as the paths that matched KEGG paths with directions. True undirected path was defined as the paths that matched KEGG paths with or without directions. Detection accuracy was defined as the proportion of the number of true paths detected over the number of all paths detected.



**Figure 3.10** The ANM-inferred network structure of the Wnt pathway. The green lines represented the inferred paths consistent to the KEGG while the gray ones represented the inferred edges absent in the KEGG.

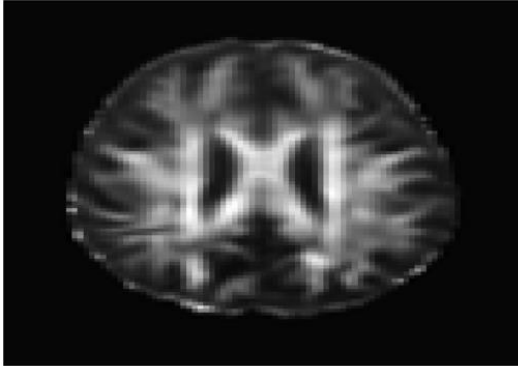
Figure 3.10 presented the ANM-inferred network structure of the Wnt pathway. The green lines represented the inferred paths consistent to the KEGG while the gray ones represented

the inferred edges absent in the KEGG. The ANM, CAM, SEM, PC, and random network methods inferred directed networks, and Glasso and WGCNA association methods inferred undirected networks. We took the structure of Wnt in the KEGG as the true structure of the Wnt in nature. We observed from Table 3.4 that the ANM more accurately inferred the network structure of the Wnt than the other six statistical and computational methods for identifying directed or undirected networks. Table 3.4 also showed that the accuracy of widely used Glasso and WGCNA algorithms for identifying the structure of Wnt was even lower than that of random networks, however, the accuracy of the ANM was much higher than that of random networks. The causal network with 50 selected top edges identified by the ANMs reached the highest accuracy. Varying the number of selected edges in the network will affect accuracy, but their accuracies were not largely different for the ANMs. This observation may not be true for other methods.

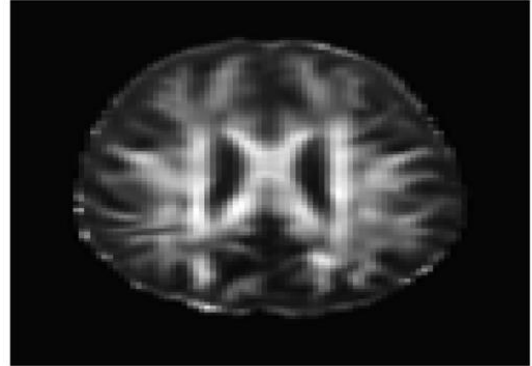
### **3.1.5 Application to Imaging Analysis**

To evaluate the performance for causal inference, the ANMs were applied to the Alzheimer’s Disease Neuroimaging Initiative (ADNI) data with 91 individuals with Diffusion Tensor Imaging (DTI) and cholesterol phenotypes measured at four time points: baseline, 6 months, 12 months and 24 months. After normalization and image registration, the dimension of a single DTI image is  $91 \times 109 \times 91$ . Three dimensional functional principal component analysis (3D-FPC) was used to summarize imaging signals in the brain region (Lin et al. 2015), because of the technical difficulty and operational cost, only 44 of the 91

individuals have all the DTI imaging data at all the four data points. Based on our own analysis experience, usually the first one or two 3D-FPC scores can explain more than 95% of the variation of the imaging signals in the region. To evaluate the performance of 3D-FPC for imaging signal feature extraction, we present Figures 3.11(A) and 3.11(B). Figure 3.11(A) is a layer of the FA map of the DTI image from a single individual and the dimension of this image is  $91 \times 109$ . A total of 91 images were used to calculate the 3D-FPC scores. Figure 3.11(B) was the reconstruction of the same layer of the FA map of the DTI image from the same individual in Figure 3.11(A) using 5 FPC scores. Comparing Figure 3.11(A) with Figure 3.11(B), we can see that these two images are very similar indicating that the 3D-FPC score is an effective tool to represent the image features.



**(A)** A slice of the FA map from a single individual's DTI data.

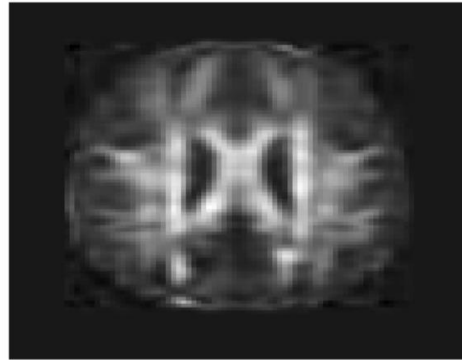


**(B)** FA map reconstruction with the first two 3D-FPC scores.

**Figure 3.11(A)** A slice of the FA map from a single individual's DTI data.

**Figure 3.11(B)** FA map reconstruction with the first two 3D-FPC scores.

To investigate feasibility of image imputation by using a mixed strategy of 3D-FPC scores and matrix completion, we used the DTI image of the 44 individuals who have measurement at all four time points as the investigation dataset. Since at baseline, the DTI image of all individuals was available, we did not have missing value problems. We only need to impute images at 6, 12 and 24 months for some individuals. We randomly sampled 20 individuals assuming that their imaging data were missing. Matrix completion methods were used to impute missing images (Thung et al. 2018). To perform 3D FPCA, all missing imaging signals at 6, 12 and 24 months of the individuals were replaced by their imaging signals at the baseline. Then, 3D FPCA was performed on the original images and replaced images of 44 individuals at all time points (base line, 6, 12 and 24 months). The FPC scores of 22 individuals without missing images were used for matrix completion. The imputed FPC score



**Figure 3.12** Imputed FA map in Figure 3.11(A) using 3D-FPC scores and matrix completion.

were then used to form reconstruction of the DTI images. To evaluate performance of the above image imputation, we presented Figure 3.12 that was the reconstruction of the DTI



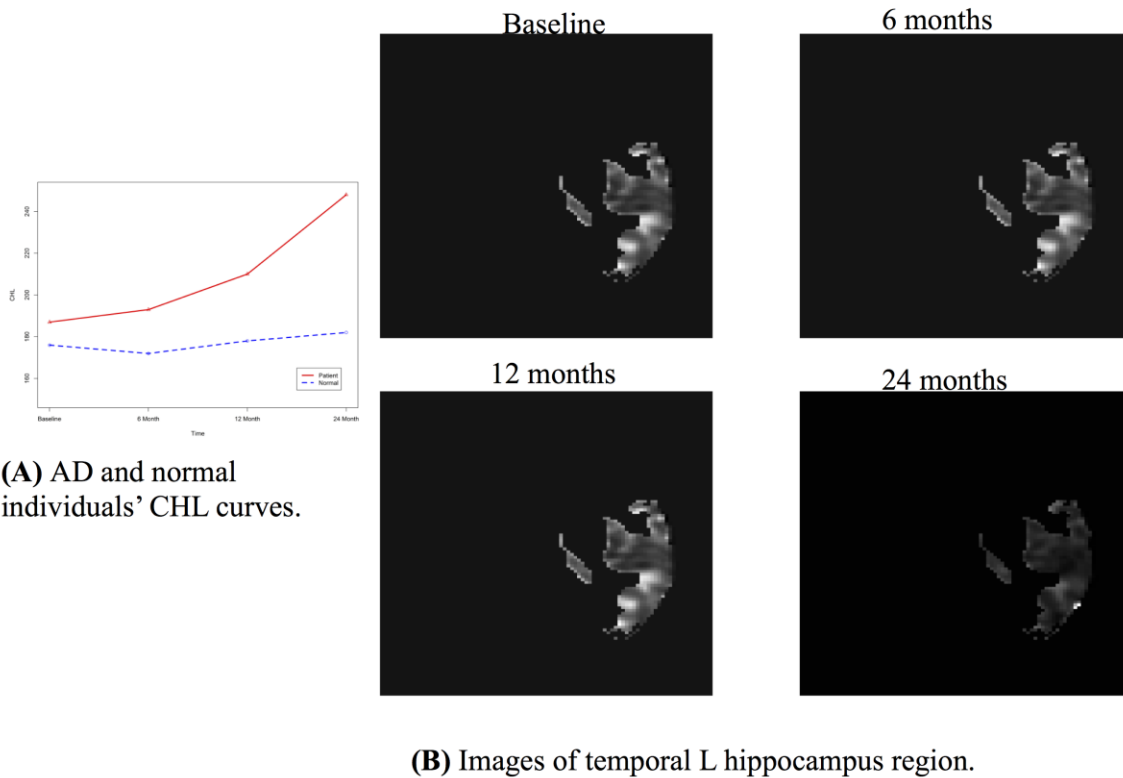
image in Figure 3.11(A). We observed from these figures that the imputed image captured the majority of the information in the original DTI image data.

After image imputation, DTI images at all four points and cholesterol and working memory of 91 individuals were available. The DTI images were segmented into 19 brain regions using the Super-voxel method (Achanta et al. 2012). Three-dimensional functional principal component analysis was used to summarize imaging signals in the brain region (Lin et al.

Table 3.5 P-values for assessing association and causal relationships between the cholesterol and brain region.

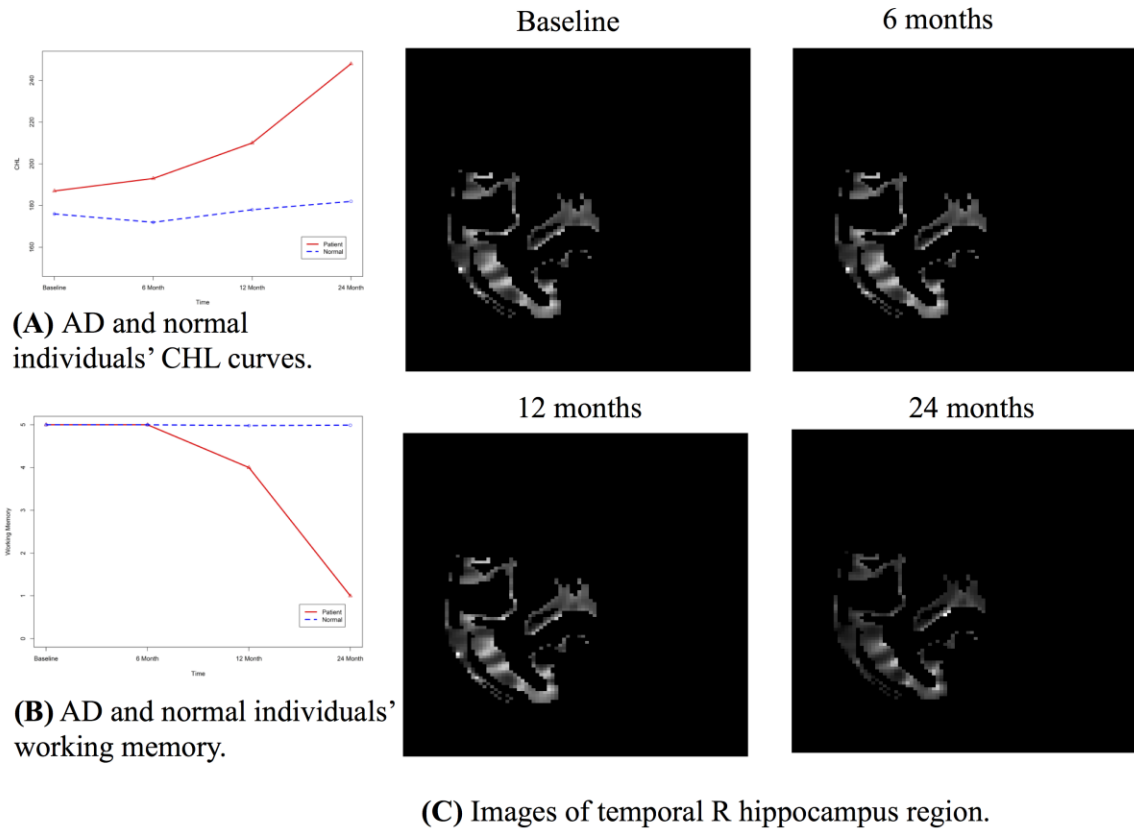
	Baseline		6 Months		12 Months		24 Months	
	Causal	Association	Causal	Association	Causal	Association	Causal	Association
Frontal_Inf_R	0.5699	0.4318	0.2927	0.9390	0.2169	0.7145	0.6624	0.1580
Frontal_Sup_Mid_L	0.4061	0.5539	0.0203	0.0301	0.6905	0.8670	0.3316	0.9664
Insula_L	0.9274	0.4602	0.2766	0.3102	0.5396	0.2724	0.7734	0.6819
Fusiform_L	0.3253	0.6601	0.8358	0.1778	0.5720	0.6238	0.8411	0.4510
Insula_R	0.3853	0.2367	0.6093	0.8874	0.0109	0.1218	0.2575	0.1832
Temporal_R	0.3740	0.7487	0.2997	0.3214	0.2813	0.8856	0.0165	0.0044
Occipital_Mid	0.7275	0.3344	0.8082	0.4159	0.6794	0.0003	0.1922	0.00004
Temporal_L	0.1455	0.4873	0.5384	0.9752	0.5262	0.0038	0.0001	0.0001
Frontal_L_R	0.1673	0.9822	0.8928	0.9269	0.3784	0.4762	0.5832	0.8093
Frontal & Temp_L	0.6067	0.4698	0.9643	0.3847	0.2945	0.9249	0.5057	0.1937
Lingual	0.2625	0.5307	0.8354	0.0834	0.7238	0.8036	0.2230	0.5510
Cingulum	0.6232	0.6483	0.3061	0.1381	0.0587	0.7611	0.3581	0.6024
Precentral_R	0.7113	0.4946	0.7263	0.0948	0.1565	0.6969	0.5169	0.6388
Frontal_Inf_L	0.9167	0.9260	0.5886	0.0138	0.3091	0.0929	0.3568	0.7203
Occipital	0.2444	0.3753	0.0782	0.9927	0.8490	0.2909	0.7388	0.4617
Precuneus	0.8480	0.2492	0.4183	0.9418	0.7208	0.5096	0.9071	0.8899
SMP	0.9866	0.1630	0.4416	0.6642	0.1175	0.3797	0.9788	0.3388
Precentral_L	0.6825	0.7937	0.4142	0.0759	0.9402	0.5150	0.5254	0.9770
Precentral_R	0.0488	0.4103	0.9759	0.9831	0.7251	0.9000	0.5008	0.0105

2015). The ANMs were used to infer causal relationships between cholesterol, or working memory and image where only first FPC score (accounting for more than 95% of the imaging signal variation in the segmented region) was used to present the imaging signals in the segmented region. Table 3.5 presented P-values for testing causation (cholesterol  $\rightarrow$  image variation) and association of cholesterol with images of 19 brain regions where the canonical correlation method was used to test association (Lin et al. 2017). Two remarkable



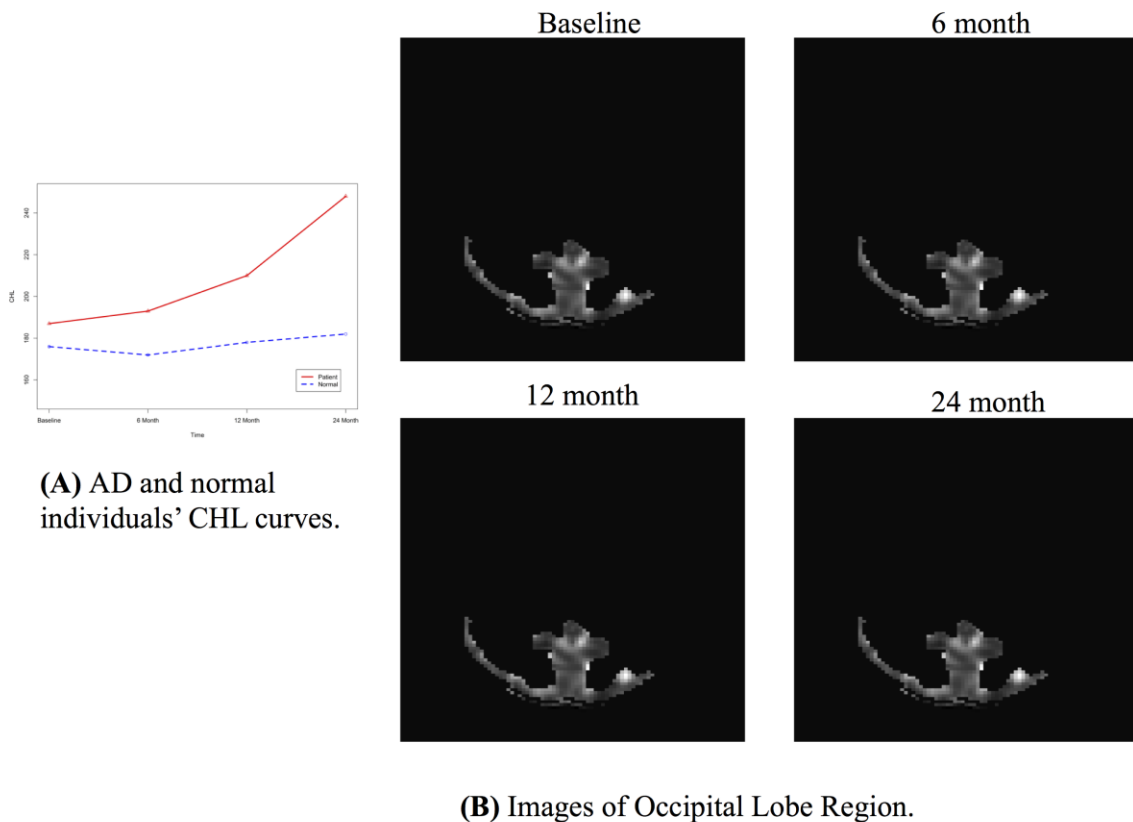
**Figure 3.13(A)** AD and normal individuals' CHL curves. **Figure 3.13(B)** Images of temporal L hippocampus region.

features emerged. First, we observed both causation and association of cholesterol with imaging signal variation at 24 months in the temporal L hippocampus (P-value for causation  $< 0.0001$ , P-value for association  $< 0.0001$ ) and temporal R hippocampus regions (P-value for causation  $< 0.0165$ , P-value for association  $< 0.0044$ ), and only association of cholesterol with imaging signal variation at 12 months in the temporal L region (P-value for causation  $< 0.5262$ , P-value for association  $< 0.0038$ ). Figures 3.13(A) and 3.13(B) presented the curves of cholesterol level of an AD patient and average cholesterol level of normal individuals, and



**Figure 3.14(A)** AD and normal individuals' CHL curves. **Figure 3.14(B)** AD and normal individuals' working memory. **Figure 3.14(C)** Images of temporal R hippocampus region.

images at baseline, 6 months, 12 months and 24 months of the temporal L hippocampus of an individual with AD diagnosed at 24 months time point, respectively. Figures 3.14(A) and 3.14(C) presented the curves of cholesterol level of an individual with AD diagnosed at 24 months' time point and average cholesterol levels of normal individuals, and images at variation at 12 and 24 months in the Occipital\_Mid brain region (P-value < 0.0003 at 12 months), P-value at baseline, 6 months, 12 months and 24 months of the Temporal R regions



**Figure 3.15(A)** AD and normal individuals' CHL curves. **Figure 3.15(B)** Images of Occipital Lobe Region.

of an individual with AD diagnosed at 24 months' time point, respectively. Figures 3.13 and 3.14 showed that images of the temporal L hippocampus and Temporal R regions at 24 months became black, which indicated that temporal L hippocampus and temporal R regions were affected by the high cholesterol. Second, we observed only association of cholesterol with imaging signal  $< 0.00004$  at 24 months), but no causation (P-value  $< 0.6794$  at 12 months, P-value  $< 0.1922$  at 24 months). Figure 3.15 showed images of the occipital lobe region. We observed that there was no significant imaging signal variation in the occipital

Table 3.6 P-values for assessing association and causal relationships between the working memory and brain region.

	Baseline		6 Months		12 Months		24 Months	
	Causal	Association	Causal	Association	Causal	Association	Causal	Association
Frontal_Inf_R	0.7515	0.6348	0.4857	0.5088	0.3709	0.5807	0.5028	0.0572
Frontal_Sup_Mid_L	0.2022	0.2877	0.0187	0.8929	0.2355	0.8327	0.4114	0.7976
Insula_L	0.0300	0.5539	0.4928	0.1057	0.8959	0.5846	0.6212	0.0332
Fusiform_L	0.3244	0.5135	0.0931	0.0503	0.0617	0.9162	0.6927	0.0741
Insula_R	0.2212	0.9885	0.7729	0.6777	0.5171	0.1434	0.7416	0.4923
Temporal_R	0.9042	0.5224	0.9641	0.6987	0.2813	0.0939	0.0001	0.5904
Occipital_Mid	0.8350	0.4884	0.0309	0.7277	0.6280	0.9993	0.2067	0.4716
Temporal_L	0.9491	0.8716	0.1052	0.4597	0.0001	0.0006	0.0001	0.5836
Frontal_L_R	0.8957	0.0212	0.2522	0.5165	0.2658	0.7134	0.1474	0.1720
Frontal & Temp_L	0.9189	0.3919	0.7792	0.1148	0.3951	0.3585	0.7691	0.7355
Lingual	0.4241	0.3219	0.4952	0.5941	0.1707	0.8981	0.8382	0.6736
Cingulum	0.5063	0.5778	0.0383	0.9534	0.5947	0.3123	0.1482	0.6307
Precentral_R	0.1398	0.2945	0.9875	0.5693	0.3247	0.7966	0.7323	0.7358
Frontal_Inf_L	0.8985	0.0989	0.2982	0.3727	0.8644	0.0363	0.9291	0.9581
Occipital	0.3828	0.8736	0.5267	0.8378	0.4624	0.1352	0.6937	0.1991
Precuneus	0.7215	0.8909	0.1169	0.5417	0.0406	0.6599	0.0429	0.9704
SMP	0.0900	0.7818	0.9407	0.6380	0.4428	0.3417	0.3151	0.8178
Precentral_L	0.9660	0.7217	0.6289	0.6630	0.8759	0.5526	0.8848	0.1713
Precentral_R	0.4051	0.3829	0.4783	0.5286	0.6365	0.0569	0.9260	0.5996

lobe region. This strongly demonstrates that association may not provide information on unraveling mechanism of complex phenotypes.

In our phenotype-image studies, we also identified causal relationships between working memory and activities of the temporal R (hippocampus) at 24 months with P-value < 0.00014) (image  $\rightarrow$  working memory), but identified no association of working memory with imaging signal variation in the temporal R (hippocampus) region (P-value < 0.5904) (Table 3.6). Figure 3.14 (C) showed the weak imaging signal or decreased neural activities in the temporal R (hippocampus) region at 24 months and Figure 3.14 (B) showed lower working memory measure of an AD patient than the average working memory measurements of normal individuals at 24 months. This demonstrated that the decreased neural activities in the temporal R (hippocampus) region deteriorated working memory of the AD patient. This result provided evidence that causation may be identified in the absence of association signals. These observations can be confirmed from the literature. It was reported that cholesterol level impacted the brain white matter connectivity in the temporal gyrus (Haltia et al. 2007) and was related to AD (Sjogren et al. 2005; Teipel et al. 2006). Abnormality in working memory was observed in patients with temporal lobe epilepsy (Stretton et al. 2013). Next we investigate two examples from the gold-standard data set in (Mooij et al 2016) to evaluate performance. The first dataset was collected at 349 weather stations in Germany from 1961 to 1990. Let X be altitude and Y be temperature. Meteorology assumes that places with higher altitude tend to be colder than those with lower altitude (roughly 1

centigrade per 100 meter). There is no doubt that altitude is the cause and temperature the effect, so ground truth is  $X \rightarrow Y$ . P-value of using the ANMs and permutation test for detecting the causation was 0.001.

The second dataset was Old Faithful geyser data. Old Faithful is a hydrothermal geyser in Yellowstone National Park in the state of Wyoming, USA. Each observation corresponds to a single eruption. The data consists of 194 samples, and was collected in a single continuous measurement from August 1 to August 15, 1985. Let  $X$  be duration of eruption in minutes and  $Y$  be time to the next eruption in minutes. It is commonly accepted that the time interval between the current and the next eruption is an effect of the duration of the current eruption, so ground truth is  $X \rightarrow Y$ . P-value of using the ANMs and permutation test for detecting the causation was 0.003. Both examples demonstrated that the ANMs and permutation test were able to detect causation between two variables.

## **3.2 Bivariate Causal Discovery for Discrete Variables**

### **3.2.1 Simulation Results**

To examine the validity of statistics  $T_C$  for testing the causal relationships between a SNP and disease, we performed a series of simulation studies to compare their empirical levels with the nominal ones. We consider two scenarios: (1) no causation in the absence of association and (2) no causation in the presence of association. We selected top 100 common SNPs (MAF between 0.19 and 0.49) from gene TEKT4P2 on chromosome 21 from 1,000

Genome Project. In scenario 1, a binary trait  $Y$  is randomly generated and independent of indicator variables  $X$  for genotypes of SNPs. In scenario 2, we first randomly generated  $X$  and  $Y$ , and then picked up the associated pairs of data as our dataset  $(X, Y)$ .

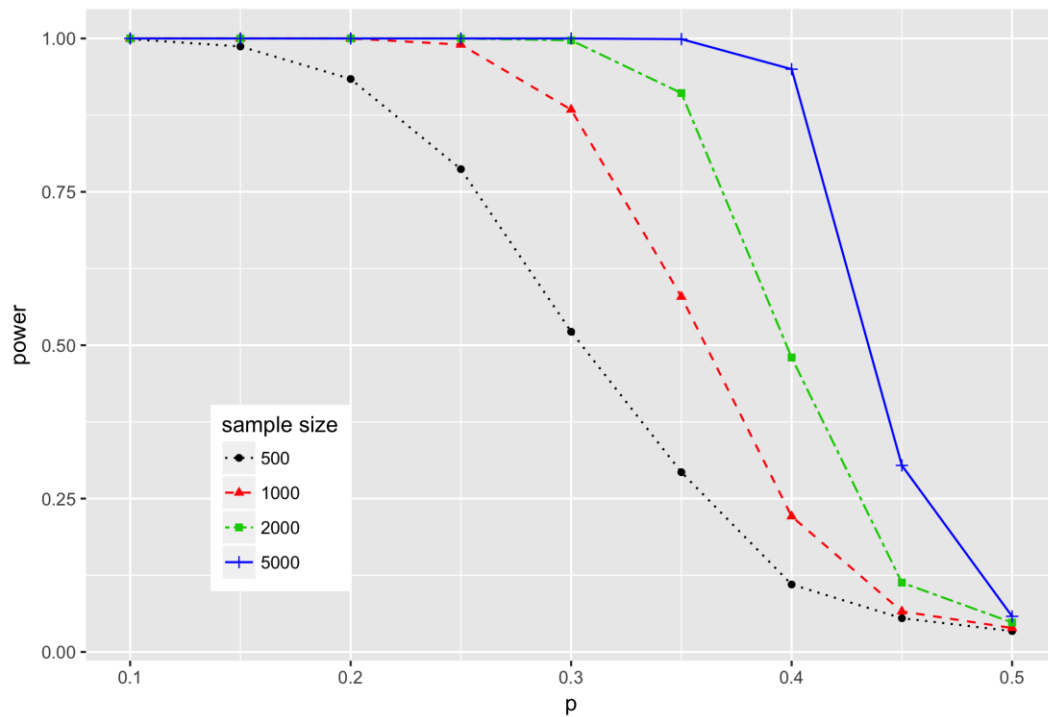
We generated the data with 100,000 subjects by resampling from the 99-individual CEU population in 1,000 Gnome Project. Number of permutations was 1,000, Number of replication of tests was 1,000. The sampled subjects from the generated population for type I error rate calculations were 500, 1,000, 2,000 and 5,000 respectively. We first consider scenario 1. Table 2 summarized the average type I error rates of the test statistics for testing the causal relationships between SNP and disease in the absence of association between SNP and disease over all 100 SNPs at the nominal levels  $\alpha = 0.05$  and  $\alpha = 0.01$  respectively. To ensure no association in the data, we also presented Table 3 that summarized average type I error rates of the association test over 100 SNPs. These tables showed that the type I error rates of the test statistics for testing the causal relationships between SNPs and disease were not appreciably different from the nominal levels. Next we consider scenario 2. Table 3.7 presented the average type I error rates of the test statistics for testing the causal relationships between SNP and disease in the presence of association between SNP and disease over all 100 SNPs at the nominal levels  $\alpha = 0.05$  and  $\alpha = 0.01$ , respectively. Again, these results demonstrated even in the presence of association the type I error rates of the test statistics for testing the causal relationships between SNPs and disease were not much appreciably different from the nominal levels.



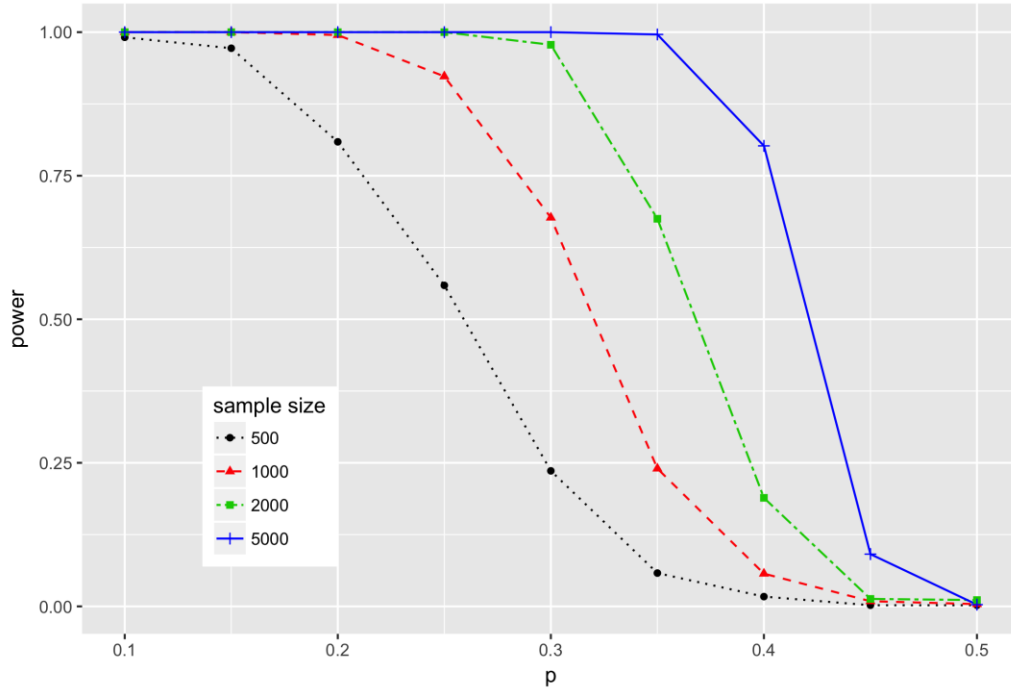
**Table 3.7** Average type 1 error rates of the statistics for testing causal relationships between SNP and disease in the presence of association

Nominal Level	500	1,000	2,000	5,000
0.05	0.042	0.046	0.047	0.046
0.01	0.005	0.007	0.007	0.008

To evaluate the performance of the ANMs for assessing the causal relationships between SNP and disease, simulated data were used to estimate their power to detect a true causation. First,



**Figures 3.16** The power curves of the causation test as a function of the parameter  $p$  with significance levels  $\alpha = 0.05$



**Figures 3.17** The power curves of the causation test as a function of the parameter  $p$  with significance levels  $\alpha = 0.01$

we investigate the power as a function of sample sizes with fixed causal measure parameter.

The data were generated by the following cyclic model:

$$Y = f(X) + N_Y, N_Y \perp\!\!\!\perp X, \quad (3.12)$$

where  $Y = \{0, 1\}$  was a binary trait and generated by the model (43),  $X = \{0, 1, 2\}$  was an indicator function for genotype of a SNP selected from 1,000 Genome Project, the minor

allele frequency of the SNP was 0.1,  $f$  was an integer function:  $f(0) = 0, f(1) = 0, f(2) = 1$ ,  $N_Y = \{0, 1\}$  was a noise distributed as a binomial with probability parameter  $P$ . We used the model (43) to generate the population of 100,000 individuals with  $Y$  and  $X$ . A set of 500, 1,000, 2,000, 5,000, 10,000 and 20,000 individuals were sampled from the population. A total of 1,000 simulations were repeated for the power calculation. Three factors: the probability parameter  $P$  in the binomial distribution, significance level  $\alpha$  and sample sizes affect the power of the ANMs for testing causation. We first fixed the parameter  $P$  and significance level  $\alpha$ . Figure 3.16 plotted the power curves as a function of sample sizes where four scenarios: (1)  $P = 0.2, \alpha = 0.05$ ; (2)  $P = 0.2, \alpha = 0.01$ ; (3)  $P = 0.4, \alpha = 0.05$  and (4)  $P = 0.4, \alpha = 0.01$  were considered. We can observe from Figure 1 that for  $P = 0.2, \alpha = 0.01$ , we can reach power 81% even when sample sizes were only 500 and for  $P = 0.4, \alpha = 0.01$ , we still can reach power 80% when sample sizes were 5,000.

We then fixed sample sizes  $n$  and significance level  $\alpha$ . Figures 3.16 and 3.17 showed the power curves of the causation test as a function of the parameter  $P$  with significance levels  $\alpha = 0.05$  and  $\alpha = 0.01$ , respectively. We observed that when the parameter  $P$  increased, the power of the causal tests decreased. In deed, the parameter  $P$  determined the value of the residual  $N_Y$ , which in turn, influenced the causality measure. When the parameter  $P$  was small, the values of the response variable  $Y$  were mainly determined by causal  $X$ . As the parameter  $P$  increased, the impact of the noise  $N_Y$  on  $Y$  increased and hence causality measure decreased, in turn, the power of the causal tests decreased. Finally, when  $P = 0.5$ ,

with the equal probability, the noise  $N_Y$  produced values 1 and 0,  $Y$  was mainly determined by noise  $N_Y$ , the ANMs had almost no power to detect causation.

### 3.2.2 Application to Genome-wide Causal Study of Schizophrenia

To further evaluate its performance, the ANMs for testing causation was applied to CATIE-MGS-SWD schizophrenia (SCZ) study dataset with 8,421,111 common SNPs typed in 13,557 individuals.

Both GWAS and GWCS where  $\chi^2$  test was used for association analysis were conducted. For the clarity of view, in the Manhattan plot of GWAS and GWCS, we only showed P-values of causal analysis (in green color) and association analysis (in black and grey colors) of all SNPs with P-values  $< 10^{-5}$ . We observed that associated SNPs were quite uniformly distributed across the genome, but the causal SNPs concentrated only on some genome regions. This may indicate that the Causal SNPs contained more information than the associated SNPs.

Due to computational time limitation of permutations, a P-value for declaring significant causation was  $10^{-6}$ . In total, 245 SNPs in 29 genes showed significant causations with SCZ. Selected top 15 causal SNPs were listed in Table 3.8. Among them, 62 causal SNPs can be confirmed from the literature and four of them were on the typical 108 schizophrenia-associated genetic loci (Nature, 511 (2014), pp. 421-427; Sullivan et al. 2007; Fatemi et al.

2011; Lei et al. 2013; Costas et al. 2013; Athanasiu et al. 2013; Misztak et al. 2018; Ren et al. 2011; Suzuki et al. 2003; Cho et al. 2015; Ide and Lewis 2010). We also conducted GWAS for this dataset. A total of 5,917 SNPs are associated with SCZ at the significance level of  $10^{-6}$  and only 89 of them showed causation.

These results showed several remarkable features. First, we can observe some SNPs that showed both significant causation and association. For example, four SNPs: rs1324544, rs2829725, rs9931378 and rs12057989 showed both strong causation and association (Table

**Table 3.8** P-values of top 15 SNPs that had significant causal relationships with schizophrenia

RS Number	Chr	Position	Gene	Related Disease	P-values	
					Causation	Association
rs1324544	6	9181479			<E-06	3.14E-12
rs2829725	21	26764027			<E-06	4.53E-11
rs9931378	16	5783022			<E-06	1.23E-09
rs12057989	1	144617251			<E-06	1.28E-08
rs7110863	11	112843138	NCAM1	Schizophrenia	<E-06	4.34E-08
rs1420643	7	35874928	SEPT7	Schizophrenia	<E-06	2.02E-07
rs1534440	6	145017328	UTRN	Schizophrenia	<E-06	2.36E-07
rs228768	17	42191893	HDAC5	Mental Depression	<E-06	3.76E-06
rs1940713	11	112906285	NCAM1	Schizophrenia	<E-06	4.57E-06
rs1940714	11	112906391	NCAM1	Schizophrenia	<E-06	4.57E-06
rs12739344	1	243791312	AKT3	Schizophrenia	<E-06	8.95E-06
rs876983	8	18407858	PSD3	Schizophrenia	<E-06	1.42E-05
rs10075211	5	147839537	HTR4	Schizophrenia	<E-06	2.24E-05
rs725515	16	82854696	CDH13	Mental Depression	<E-06	3.80E-05
rs10986439	9	101262400	GABBR2	Major Depressive Disorder	<E-06	0.000457917

3.8). Second, the number of causal SNPs was much smaller than the number of associated SNPs. Third, highly significantly associated SNPs may show no significant causation. Forth, the SNPs that showed strong causation signals may not demonstrate association. For example, SNP rs12739344 in gene AKT3 showed strong causation ( $P\text{-value} < 10^{-6}$ ), but did not reach threshold  $P\text{-value}$  for association ( $P\text{-value}$  for association is  $8.95 \times 10^{-6}$ ). It is well known that the genetic variation in the gene AKT3 is a top risk signal in schizophrenia and network analysis identified that AKT3 contributes to four of the pathways involved in SCZ (Howell et al. 2017). SNP rs10986439 in gene GABBR2 showed significant causation ( $P\text{-value} < 10^{-6}$ ), but no association with SCZ ( $P\text{-value}$  is 0.000458). Genetic-imaging analysis showed that gene GABBR2 was in neuron development, synapse organization and axon pathways which could affect cognition in schizophrenia (Luo et al. 2018). Fifth, proportion of SNPs showed both causation and association was small (36.3% of causal SNPs showed association and only 0.98% of associated SNPs showed causation).

### **3.2.3 Application to Disease Prediction**

Genomic predictors and risk estimates for a large number of diseases can be constructed from SNPs. The traditional methods for developing genomic risk scores (GRS) utilize small numbers of SNPs, typically those identified as genome-wide significant association (Abraham and Inouye 2015). To evaluate the predictive ability of causal SNPs and associated SNPs, we selected top 245 causal SNPs (all  $P\text{-values} < 10^{-6}$ ) and top 245 associated SNPs for SCZ risk prediction. Logistic regression and 10 fold cross validation were used to

calculate prediction accuracy. Table 3.9 listed ten-fold cross-validated accuracy for prediction of SCZ. Table 3.9 showed that using the same number of SNPs, all the sets of SNPs selected by causal analysis had higher prediction accuracy than the set of SNPs selected by association analysis. Specifically, the prediction accuracy of 245 top causal SNPs was about 3% higher than that of 245 top SNPs selected by association analysis. This may imply that the causal SNPs contain more biological information than associated SNPs.

**Table 3.9** Ten-fold cross-validated accuracy and AUC for SCZ risk prediction of using top 15 causal SNPs and association SNPs.

Number of SNPs	11	12	13	14	15	245
Accuracy of Causal SNPs	0.5542	0.554	0.5534	0.5531	0.5521	0.5737
AUC of Causal SNPs	0.5344	0.5342	0.5336	0.5333	0.5324	0.5491
Accuracy of Associated SNPs	0.5415	0.541	0.5404	0.5401	0.5395	0.5430
AUC of Associated SNPs	0.5178	0.5173	0.5168	0.5163	0.5158	0.5249

### 3.2.4 Application to Linkage Disequilibrium

In this section, we investigate the impact of linkage disequilibrium (LD) on the causal analysis. It is well known that linkage disequilibrium has big impact on the association analysis. For the convenience of presentation, we first consider the true linear model for a quantitative trait (Xiong 2018):

$$Y = \mu + X\alpha + N_Y, X \perp\!\!\!\perp N_Y, \quad (3.13)$$

where  $X$  is an indicator variable for the genotype at the true causal locus and distribution of  $N_Y$  is not normal.

Suppose that  $X^m$  is an indicator variable for the genotype at a marker locus with marker allele frequencies  $P_M$  and  $P_m$  and LD  $D_m$  between the marker and true causal loci. Then, we have the following linear regression model for the marker locus:

$$Y = \mu + X^m \alpha_m + N_Y^m . \quad (3.14)$$

Then, we can show (Xiong 2018) that

$$\alpha_m \xrightarrow{a.s} \frac{D_m}{P_M P_m} \alpha . \quad (3.15)$$

Equation (3.15) implies that in the presence of LD, the marker locus still shows some association with genetic additive effect  $\frac{D_m}{P_M P_m} \alpha$  approximately.

Now we investigate the impact of LD on causal inference. Substituting equation (3.13) into equation (3.14), we obtain

$$N_Y^m = N_Y + X\alpha - X^m \alpha_m . \quad (3.16)$$

Define

$$\Delta = X\alpha - X^m \alpha_m \approx (X - \frac{D_m}{P_M P_m} X^m) \alpha . \quad (3.17)$$

When  $\Delta \neq 0$ , distance covariance  $dCov^2(X^m, N_Y^m)$  is equal to

$$0 \leq dCov^2(X^m, N_Y^m) = dCov^2(X + X^m - X, N_Y + \Delta)$$



$$\begin{aligned}
&\leq dCov^2(X, N_Y) + dCov^2(X^m - X, \Delta) \\
&= dCov^2(X^m - X, \Delta).
\end{aligned} \tag{3.18}$$

$X^m \rightarrow Y$  must imply that  $\Delta = 0$  (Sze'kely and Rizzo, 2009) or

$$X = \frac{D_m}{P_M P_m} X^m. \tag{3.19}$$

Equation (3.19) indicates that  $X^m \rightarrow X$ . However, in general, SNPs do not have causal relationships. Therefore,  $dCov^2(X^m, N_Y^m) \neq 0$  and  $X^m, N_Y^m$  are not independent, which implies that  $X^m$  does not cause  $Y$ .

Now we calculate the causal measure. Let  $C_{X \rightarrow Y} = 1 - R(X, N_Y)$  be the causal measure of the causal SNP  $X$ . Then, the causal measure of the marker  $X^m$  is given by

$$C_{X^m \rightarrow Y} = C_{X \rightarrow Y} - R(X^m - X, X - \frac{D_m}{P_M P_m} X^m). \tag{3.20}$$

$$1 \geq R(X^m - X, X - \frac{D_m}{P_M P_m} X^m) \geq 0 \text{ implies}$$

$$C_{X \rightarrow Y} \geq C_{X^m \rightarrow Y} \geq 0. \tag{3.21}$$

Causation measure  $C_{X^m \rightarrow Y}$  depends on the distance correlation between  $X^m - X$  and  $X -$

$$\frac{D_m}{P_M P_m} X^m.$$

For qualitative trait, we can use logistic integer function as a nonlinear function. After some algebraic operations, we can have the model:

$$Y = \frac{e^{X\alpha}}{1+e^{X\alpha}} + N_Y \quad (3.22)$$

or

$$Y = f(X\alpha) + N_Y, \quad (3.23)$$

where  $f(X\alpha)$  is a nonlinear function. When  $f(X\alpha) \geq 0.5$  then  $\frac{e^{X\alpha}}{1+e^{X\alpha}} = 1$ ; when  $f(X\alpha) < 0.5$ , we set  $\frac{e^{X\alpha}}{1+e^{X\alpha}} = 0$ .

Equation (3.23) can be approximated by

$$Y = f(0) + f'(0)X\alpha + N_Y. \quad (3.24)$$

Thus, the model (3.23) is reduced to model (3.13). Using the same arguments for the model (3.13), we can define the causality measure for marker  $X^m$ :

$$C_{X^m \rightarrow Y} = C_{X \rightarrow Y} - R(X^m - X, X - \frac{f'(0)D_m}{P_M P_m} X^m). \quad (3.25)$$

For the discrete ANMs, we cannot find  $f'(0)$ , the causal measure for the marker may simply be written as

$$C_{X^m \rightarrow Y} = C_{X \rightarrow Y} - R(X^m - X, X - \frac{\gamma D_m}{P_M P_m} X^m), \quad (3.26)$$

where  $\gamma$  is a appropriate constant.

Next we use simulations to investigate the impact of LD on the causation analysis. Data for

two markers: rs150012736 and rs376953511 were taken from 1000 Genome Project. In the 1000 Genome Project dataset, LD ( $r^2$ ) between rs150012736 and rs376953511 was calculated as 0.5. Assume that SNP1 was a causal SNP. We did not make assumption about whether or not SNP2 was a causal SNP. The trait values was generated by the discrete cyclic ANMs:

$$Y = f_Y(X) + N_Y, \quad (3.27)$$

where  $f_Y$  is a specified nonlinear integer function and  $N_Y$  is a binomial variable. We fitted the ANMs to the data  $(Y, X^m)$  where  $X^m$  represented the indicator variable for genotypes of SNP2. The results of causation and association tests were summarized in Tables 3.10 and 3.11, and Tables 3.12 and 3.13. Tables 3.10 and 3.11 showed that we can detect both association and causation between SNP1 and Disease with high power when sample sizes were larger than 2,000. Table 3.12 showed that Type 1 error rates of test to detect causation between SNP2 and disease was not very high and decreased when sample sizes increased. In other words, we did not detect causation at SNP2. However, Table 3.13 showed that association test detected association of SNP2 with disease with high power. The simulation results showed that the impact of LD on the causal tests was much smaller than on the association tests.

**Table 3.10** Power to detect association between SNP1 and Disease

Sample Sizes	500	1000	2000	5000
0.05	0.999	1	1	0.999
0.01	0.992	0.992	0.993	0.992

**Table 3.11** Power to detect causation between SNP1 and disease

Sample Sizes	500	1000	2000	5000
0.05	0.684	0.888	0.949	0.949
0.01	0.418	0.701	0.936	0.948

**Table 3.12** Type I error rates of causal test between SNP2 and disease

Significance Level	500	1,000	2,000	5,000
0.05	0.183	0.159	0.142	0.104
0.01	0.105	0.118	0.105	0.093

**Table 3.13** Power of test for association between SNP2 and disease

Significance Level	500	1,000	2,000	5,000
0.05	0.918	0.979	0.992	0.994
0.01	0.860	0.957	0.990	0.992

To further evaluate the impact of LD on causation test by real data analysis. From the results of GWCS of SCZ, we selected SNP rs6578689 that had P-values  $< 10^{-6}$  and  $2.82 \times 10^{-7}$  for causation and association tests, respectively. Then, we selected 20 neighboring SNPs of causal SNP rs6578689. We tested their causation and association with SCZ. Table 3.14 summarized the results of the causation and association tests. These results showed that neighboring SNPs that had  $r^2 > 0.44$  demonstrated no causation with SCZ, but strong associations with small P-values  $< 4.59 \times 10^{-9}$  with SCZ. These results of real data analysis demonstrated that LD had small impact on causation analysis, but large impact on association tests.

**Table 3.14** P-values for causation and association tests of 20 neighboring SNPs of causal SNP rs6578689

SNPs	Chr	P-values		Neighbor SNPs	Position	$r^2$	P-values	
		Causation	Association				Causation	Association
rs6578689	11	<E-06	2.82E-07	rs10742794	5826464	0.7196	0.03	9.95E-10
rs6578689	11			rs11039135	5836787	0.63226	0.39	2.65E-09
rs6578689	11			rs7115498	5831847	0.53094	0.94	9.31E-10
rs6578689	11			rs10838661	5830617	0.53093	0.96	1.03E-09
rs6578689	11			rs35898746	5830823	0.53093	0.96	1.03E-09
rs6578689	11			rs11039085	5823651	0.53034	0.93	6.03E-10
rs6578689	11			rs10742791	5819152	0.5272	0.94	4.03E-10
rs6578689	11			rs12226188	5837141	0.52658	0.9	6.80E-10
rs6578689	11			rs10838674	5836857	0.52634	0.93	9.01E-10
rs6578689	11			rs35271555	5833707	0.5233	0.9	7.99E-10
rs6578689	11			rs6578687	5813985	0.52136	0.95	5.00E-10
rs6578689	11			rs7114690	5814376	0.51743	0.88	3.83E-10
rs6578689	11			rs80316576	5827945	0.44329	0.37	3.01E-09
rs6578689	11			rs73390385	5809052	0.44286	0.42	3.96E-09
rs6578689	11			rs73392251	5821745	0.44191	0.44	4.59E-09
rs6578689	11			rs73392254	5822797	0.44191	0.54	4.59E-09
rs6578689	11			rs73390383	5808495	0.44143	0.48	3.90E-09
rs6578689	11			rs73392222	5817732	0.44136	0.47	2.80E-09
rs6578689	11			rs73392226	5817797	0.44136	0.46	2.80E-09
rs6578689	11			rs77107630	5818487	0.44136	0.47	2.80E-09

### 3.3 Confounder Detection

#### 3.3.1 Data and Notation

If there exists a variable  $T$  and let  $X = u(T) + N_X$  and  $Y = v(T) + N_Y$ , then  $T$  is regarded as the hidden confounder. We here apply the Algorithm 1 (Janzing et al. 2009) for confounder detection based on additive noise model.

Algorithm 1 Identifying Confounders using Additive Noise Models (ICAN)

Input:  $(X_1, Y_1), \dots, (X_n, Y_n)$  (normalized)

Initialization:

Fit a curve  $\hat{\mathbf{s}}$  to the data that minimizes  $l_2$  distance:  $\hat{\mathbf{s}} := \operatorname{argmin}_{\mathbf{s} \in S} \sum_{k=1}^n \operatorname{dist}(\mathbf{s}, (X_k, Y_k))$ .

repeat

Projection:

$$\begin{aligned}\hat{T} &:= \operatorname{argmin}_T DEP(\hat{N}_X, \hat{N}_Y) + DEP(\hat{N}_X, T) + DEP(\hat{N}_Y, T) \text{ with } (\hat{N}_{X,k}, \hat{N}_{Y,k}) \\ &= (X_k, Y_k) - \hat{\mathbf{s}}(T_k)\end{aligned}$$

if  $\hat{N}_X \perp \hat{N}_Y$  and  $\hat{N}_X \perp \hat{T}$  and  $\hat{N}_Y \perp \hat{T}$  then

Output:  $(\hat{T}_1, \dots, \hat{T}_n)$ ,  $\hat{u} = \hat{s}_1$ ,  $\hat{v} = \hat{s}_2$ , and  $\frac{\operatorname{var} \hat{N}_X}{\operatorname{var} \hat{N}_Y}$ .

Break.

end if

Regression:

Estimate  $\hat{\mathbf{s}}$  by regression  $(X, Y) = \hat{\mathbf{s}}(\hat{T}) + \hat{N}$ . Set  $\hat{u} = \hat{s}_1$ ,  $\hat{v} = \hat{s}_2$ .

Until K iterations

Output: Data cannot be fitted by a CAN model.

The algorithm has been realized by R and uploaded to Github:

[https://github.com/jiaorong007/Confounder-Detection/blob/master/conf\\_detection\\_0928.R](https://github.com/jiaorong007/Confounder-Detection/blob/master/conf_detection_0928.R)

### 3.3.2 Simulation Results

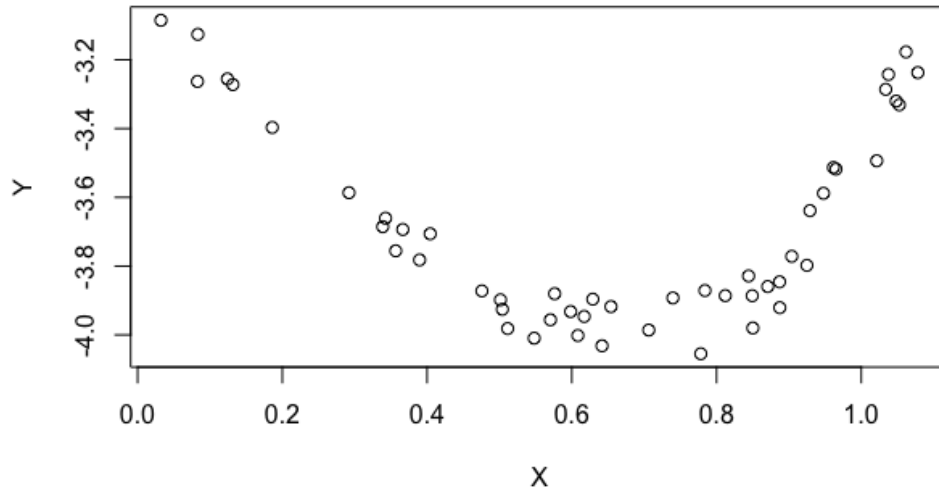
Using the Algorithm 1 (Janzing et al. 2009), we here show its performance on confounder detection for three kinds of datasets.

#### 3.3.2.1 Simulation A

Data were generated by

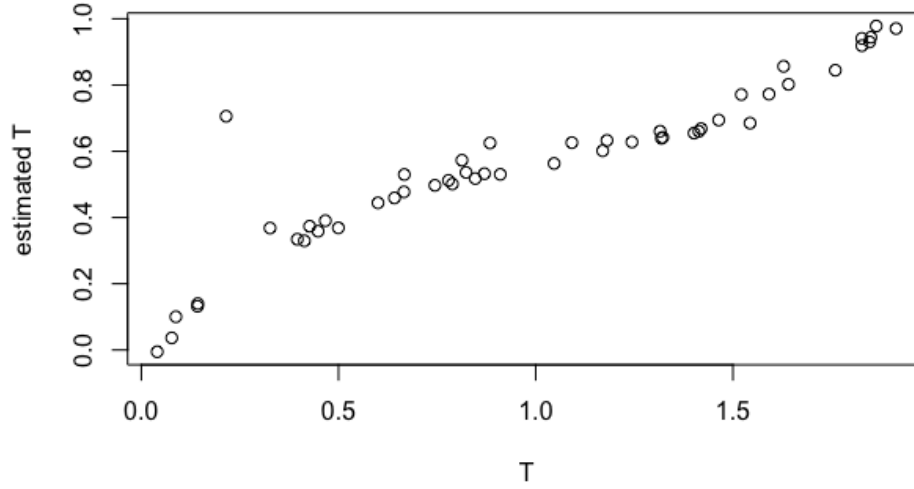
$$\begin{cases} X = \log(T + 1) + N_X \\ Y = (T - 1)^2 - 4 + N_Y \end{cases}$$

where  $T \sim \text{unif}(0, 2)$ ,  $N_X \sim \text{unif}(-0.01, 0.01)$ ,  $N_Y \sim \text{unif}(-0.1, 0.1)$ .



**Figure 3.18** Scatterplot of model  $X = \log(T + 1) + N_X$  and  $Y = (T - 1)^2 - 4 + N_Y$

Then we have  $p_{HSIC}(\hat{N}_X, \hat{N}_Y) = 0.72, p_{HSIC}(\hat{N}_X, \hat{T}) = 0.97, p_{HSIC}(\hat{N}_Y, \hat{T}) = 0.75$ , which means there exists a hidden confounder. Further, we notice that the mapping from  $T$  to  $\hat{T}$  is actually bijective transformation, which confirms the existence of hidden confounders.



**Figure 3.19** Scatterplot of  $\hat{T}$  (estimated T) vs. T

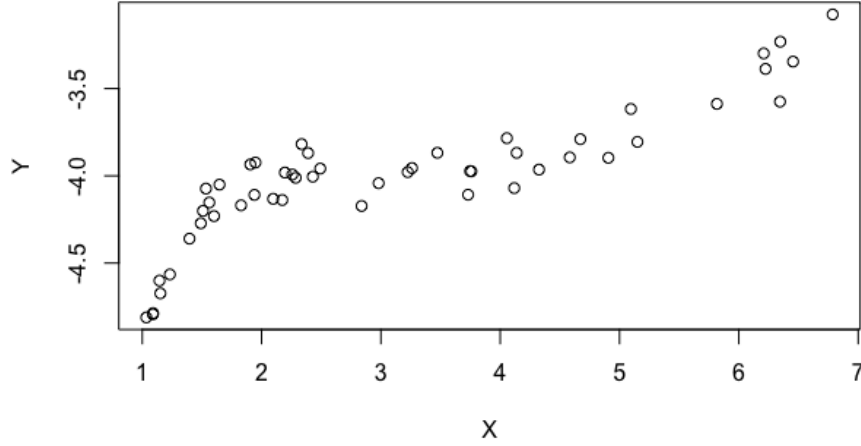
### 3.3.2.2 Simulation B

Data were generated by

$$\begin{cases} X = \log(T + 1) + N_X \\ Y = (T - 1)^2 - 4 + N_Y \end{cases}$$

where  $T \sim \text{unif}(0,2), N_X \sim \text{unif}(-0.01,0.01), N_Y \sim \text{unif}(-0.1,0.1)$ .





**Figure 3.20** Scatterplot of model  $X = e^T + N_Y$  and  $Y = (T - 1)^3 - 4 + N_X$

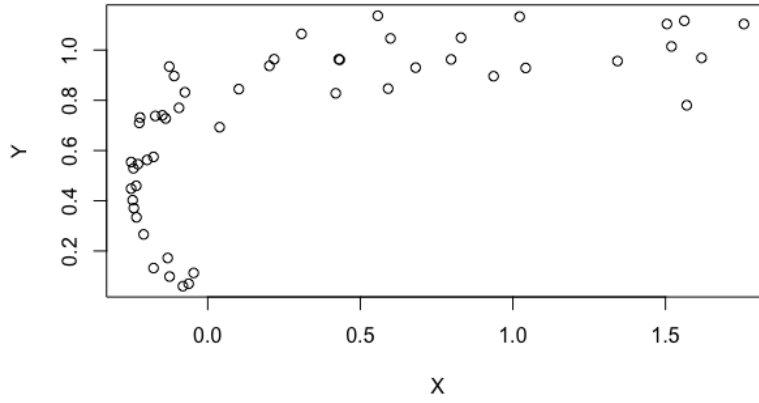
Then we have  $p_{HSIC}(\hat{N}_X, \hat{N}_Y) = 0.9996, p_{HSIC}(\hat{N}_X, \hat{T}) = 0.9979, p_{HSIC}(\hat{N}_Y, \hat{T}) = 0.9995$ , which means there exists a hidden confounder. Further, we notice that the mapping from  $T$  to  $\hat{T}$  is actually bijective transformation, which confirms the existence of hidden confounders.

### 3.3.2.3 Simulation C

Data were generated by

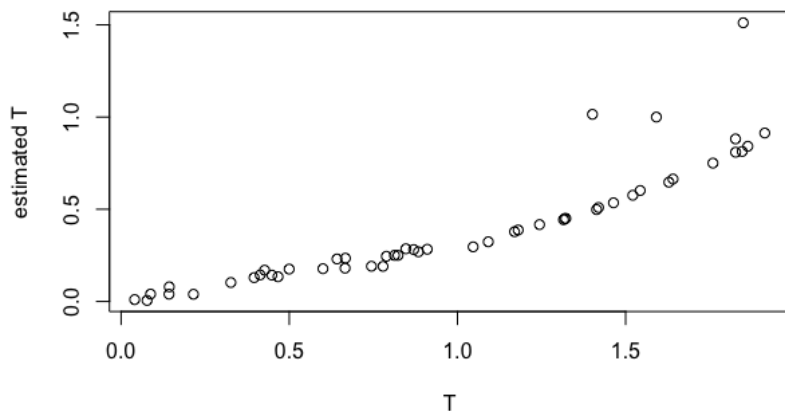
$$\begin{cases} X = T^2 - T + N_X \\ Y = \sin T + N_Y \end{cases}$$

where  $T \sim \text{unif}(0, 2), N_X \sim \text{unif}(-0.01, 0.01), N_Y \sim N(0, \sigma^2 = 0.01)$ .



**Figure 3.21 (A)** Scatterplot of model  $X = T^2 - T + N_X$  and  $Y = \sin T + N_Y$

Then we have  $p_{HSIC}(\hat{N}_X, \hat{N}_Y) = 0.78$ ,  $p_{HSIC}(\hat{N}_X, \hat{T}) = 0.60$ ,  $p_{HSIC}(\hat{N}_Y, \hat{T}) = 0.12$ , which means there exists a hidden confounder. Further, we notice that the mapping from  $T$  to  $\hat{T}$  is actually bijective transformation, which confirms the existence of hidden confounders.



**Figure 3.21 (B)** Scatterplot of  $\hat{T}$  (estimated  $T$ ) vs.  $T$

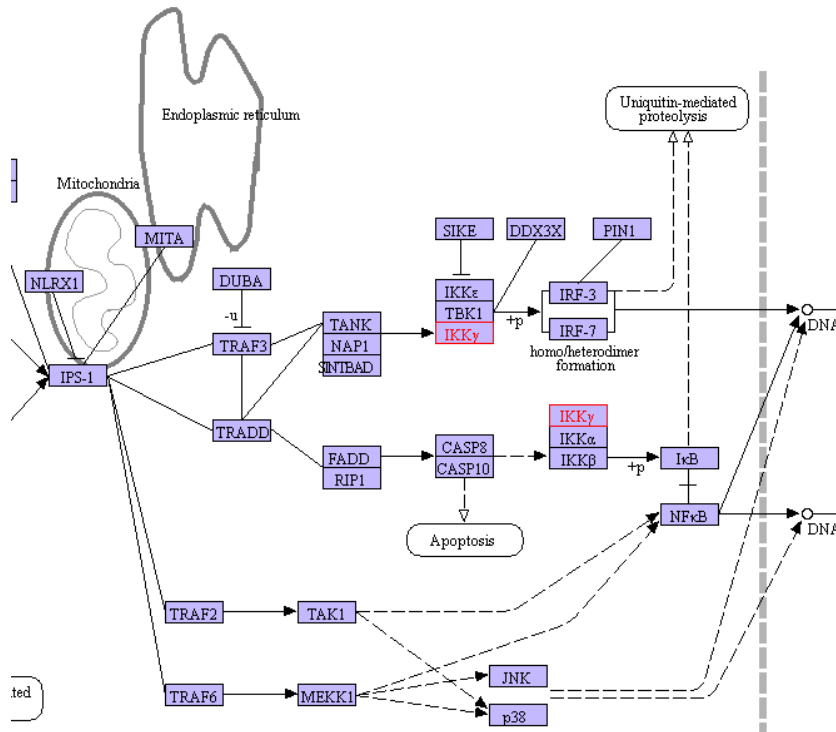
### 3.3.3 Application to Gene Expression Data

The confounder detection algorithm was applied to RIG-I-like receptor signaling pathway measured in 447 tissue samples in the ROSMAP dataset (White et al. 2017). Two variables were detected as having a hidden confounder.

**Table 3.15** Variables with possible hidden confounder in RIG-I-like receptor signaling pathway

Var1	Var2	pHSIC_Nx_Ny	pHSIC_Nx_T_est	pHSIC_Ny_T_est	P_association
IKBKG	TRADD	0.5562405	0.1096206	0.05367892	0.005994006

Based on the KEGG signaling pathway, there exists a causal pathway between TRADD and IKBKG, which demonstrate the significant association. However, there is no confounder shown in the reference pathway graph, so it leaves much space for scientists to explore.



**Figure 3.22** Part of RIG-I-like receptor signaling pathway

### 3.4 Nonlinear Causal Network

#### 3.4.1 Data and Notation

Suppose the causal network is a directed acyclic graph (DAG), the optimization question can be converted to (Xiong 2018):

$$\begin{aligned} \min \quad & \sum_{v=1}^p \sum_{j_v=1}^{J_v} C(v, W_{j_v}) x(W_{j_v} \rightarrow v) \\ \text{constraints:} \quad & \sum_{j_v=1}^{J_v} x(W_{j_v} \rightarrow v) = 1, v = 1, \dots, p \end{aligned}$$

$$\forall C \subseteq V: \sum_{v \in C} \sum_{W_{j_v}: |W_{j_v} \cap C| < k, j_v=1, \dots, J_v} x(W_{j_v} \rightarrow v) \geq k, \forall k, 1 \leq k \leq |C| \quad (3.28)$$

$$x(W_{j_v} \rightarrow v) = 0 \text{ or } 1.$$

Here  $v$  refers to a specific node  $v$ , and  $W_{j_v}$  refers to one of the possible parent set of  $v$ .

$C(v, W_{j_v})$  denotes the score function for the pair of node  $v$  and its parent set  $W_{j_v}$ .  $x(W_{j_v} \rightarrow v) = 1$  if and only if  $W_{j_v}$  is the parent set for the node  $v$ . The constraint (3.28) is to ensure that there is no cycle in the DAG.

In real world, the causation relation from the parent set  $W_{j_v}$  to  $v$  is usually nonlinear, so I use a nonlinear score to represent  $C(v, W_{j_v})$ . The following is the algorithm (Xiong 2018) to get the score by the nonlinear regression:

Step 1: Select the penalty parameter  $\lambda$ . Define the node  $v$  as  $Y = [y_1, y_2, \dots, y_n]^T$  and the variables in  $W_{j_v}$  as  $x$ .

Step 2: Compute the matrices

$$T = \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_r(x_1) \\ \vdots & \vdots & \vdots \\ \phi_1(x_n) & \cdots & \phi_r(x_n) \end{bmatrix} \text{ and } \Sigma_j = \begin{bmatrix} R^j(x_1, z_1) & \cdots & R^j(x_1, z_n) \\ \vdots & \vdots & \vdots \\ R^j(x_n, z_1) & \cdots & R^j(x_n, z_n) \end{bmatrix}$$

$$\Sigma_\theta = \theta_1 \Sigma_1 + \cdots + \theta_q \Sigma_q, \text{ where } \theta_1, \dots, \theta_q \text{ are pre-determined weights.}$$

Step 3: Perform QR decomposition of the matrix  $T$ :

$$T = [Q_1 \ Q_2] \begin{pmatrix} R \\ 0 \end{pmatrix}$$

Step 4: Compute coefficients of the smoothing spline regression

$$\hat{a} = R^{-1} Q_1^T [I - M Q_2 (Q_2^T M Q_2)^{-1} Q_2^T] Y \text{ and } \hat{b} = Q_2 (Q_2^T M Q_2)^{-1} Q_2^T Y,$$

where  $M = \Sigma + n\lambda I$ .

Step 5: Compute the smoothing spline regression function

$$\hat{f}(x) = \sum_{j=1}^r \hat{a}_j \phi_j(x) + \sum_{v=1}^n \hat{b}_v \sum_{j=1}^q \theta_j L_{v(z)} R^j(x, z)$$

Step 6: Compute the fitted value:

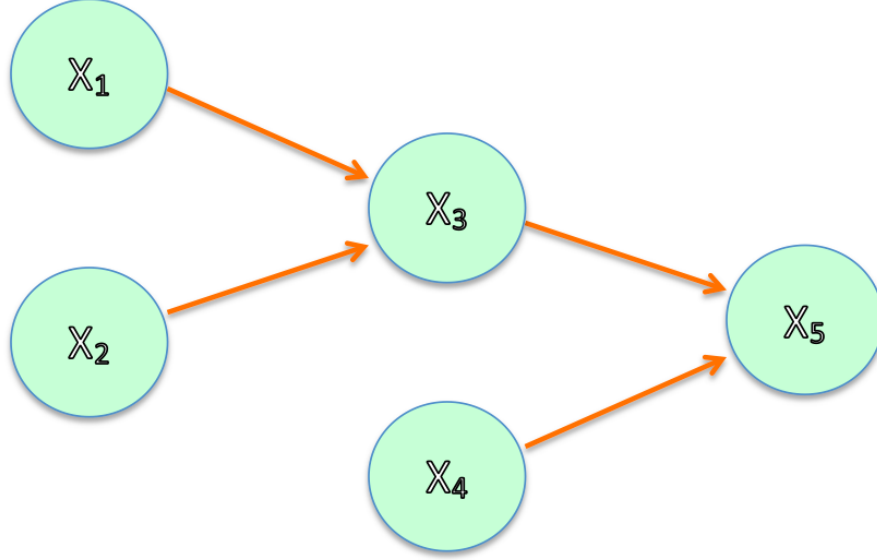
$$\hat{f} = H(\lambda) Y, \text{ where } H(\lambda) = I - n\lambda Q_2 (Q_2^T M Q_2)^{-1} Q_2^T.$$

Step 7: Calculate the nonlinear score of the node  $v$ :

$$C(v, W_{j_v}) = \frac{1}{n} \|Y - T\hat{a} - \Sigma_\theta \hat{b}\|^2 + \hat{b} \Sigma_\theta \hat{b}.$$

### 3.4.2 Simulation Results

On the simulations of nonlinear causal network, we randomly generated a DAG. Set nodes without parents as random normal distribution. Then use nonlinear/linear functions to generate the nodes in the next layer.



**Figure 3.23** Causal Network

Based on the above causal network, I first generated three independent nodes  $X_1$ ,  $X_2$  and  $X_3$  that have normal distributions. Then randomly generate functions  $f$  and  $g$ , and generate  $X_3$  and  $X_5$  by  $X_3 = f(X_1, X_2) + N_1$  and  $X_5 = f(X_3, X_4) + N_2$ .

I here randomly generated a DAG and the network data for 1,000 times. Let  $N_t$  be the total number of edges among 1,000 networks,  $N_o$  the total number of edges that do not appear in

1,000 networks,  $N_{True}$  the total number of edges detected by the algorithm and  $N_{False}$  the false edges directed among  $N_o$ . Then the false discovery rate (FDR) is defined by  $\frac{N_{False}}{N_o}$  and power of detection (PD)  $\frac{N_{True}}{N_t}$ . The numbers of sample are 100, 300 and 500.

**Table 3.16** Power and FDR of four methods in causal network analysis

	Power			FDR		
Nsample	100	300	500	100	300	500
Nonlinear+Integer programming	0.2533	0.3608	0.8218	0.1098	0.1140	0.0857
Only nonlinear	0.4469	0.6013	0.6601	0.1850	0.2126	0.2191
CAM	0.9122	0.9307	0.9373	0.3256	0.3230	0.3220
Linear + Integer programming	0.6954	0.7083	0.7146	0.1306	0.0841	0.0680

The method of nonlinear SEM with integer programming was compared with only nonlinear SEM, Causal Additive Models (CAM) and linear SEM with integer programming in power and FDR. Although CAM shows higher power than nonlinear SEM with integer programming, its FDR is also much higher.

### 3.4.3 Real Data Analysis

Four methods of causal network analysis were applied to Wnt signaling pathway with RNA-Seq of 79 genes measured in 447 tissue samples in the ROSMAP dataset (White et al. 2017), and RNA-Seq of 144 genes measured in 744 tissue samples in the ADNI dataset. We picked up 50 top significant edges and matched them to the reference. Since WGCNA is

based on association analysis without causal direction, we also gave the power of detection of the other three methods without consideration of causal direction.

**Table 3.17** Power of detection of four causal methods on two datasets

Methodology/Dataset	Directed		Undirected	
	ROSMAP	ADNI	ROSMAP	ADNI
CAM	16%	16%	24%	26%
Nonlinear SEM	24%	28%	28%	44%
SEM	20%	6%	26%	21%
WGCNA association	X	X	18%	6%



#### **4. Discussion**

The major purpose of this dissertation is to address several issues for shifting the paradigm of genetic analysis from association analysis to causal inference. The first issue is the basic principles for causal inference from observational data only. Typical methods for unraveling cause and effect relationships are interventions and controlled experiments. Unfortunately, the experiments in human genetics are unethical and technically impossible. In the past decade, the new principles for causal inference from pure observational data have been developed. The philosophical causal principle assumes that nature consists of autonomous and independent causal generating process modules and attempts to replace causal faithfulness by the assumption of Independence of Cause and Mechanism (ICM). In other words, causal generating processes of a system's variables are independent. If we consider two variables, the ICM states that distribution of cause and conditional distribution of effect given the cause are independent.

The second issue is how to measure independence (or dependence) between two distributions. Statistics only provides tools for measuring independence between two random variables. There are no measures or statistics to test independence between two distributions. Therefore, we introduce algorithmic information theory that can offer notion and mathematical formulation of independence between two distributions or independence of mechanisms. We use algorithmic mutual information to measure independence between two distributions which can be used to assess causal relationships between two variables.

Algorithmically independent conditional implies that the joint distribution has a shorter description in causal direction than in non-causal direction.

The third issue is to develop causal models that can easily assess algorithmic independent conditions. The algorithmic independent condition states that the direction with the lowest Kolmogorov complexity can be identified to be the most likely causal direction between two random variables. However, it is well known that the Kolmogorov complexity is not computable (Budhathoki and Vreeken, 2017). Although stochastic complexity was proposed to approximate Kolmogorov complexity via the Minimum Description Length (MDL) principle, it still needs heavy computations. The ANM was developed as practical causal inference methods to implement algorithmically independent conditions. We showed that algorithmic independence between the distribution of cause  $X$  and conditional distribution  $P_{Y|X}$  of effect given the cause is equivalent to the independence of two random variables  $X$  and  $E_Y$  in the ANM.

The fourth issue is the development of test statistics for bivariate causal discovery. The current ANM helps to break the symmetry between two variables  $X$  and  $Y$ . Its test statistics are designed to identify causal directions:  $X \rightarrow Y$  or  $Y \rightarrow X$ . Statistics and methods for calculation of P-values for testing the causation between two variables have not been developed. To address this issue, we have developed a new statistic to directly test for

causation between two variables and a permutation method for the calculation of P-value of the test.

The fifth issue is the power of the ANM. The challenge arising from bivariate causal discovery is whether the ANM has enough power to detect causation between two variables. To investigate their feasibility for causal inference, the ANMs were applied to simulation data. We considered three nonlinear functions: quadratic, exponential and logarithm functions and two random noise variables: normal and t distribution. We showed that the ANM had reasonable power to detect existence of causation between two variables. To further evaluate its performance, the ANM was also applied to reconstruction of the Wnt pathway using gene expression data. The results demonstrated that the ANM had higher power to infer gene regulatory networks than six other statistical methods using KEGG pathway database as gold standard.

The sixth issue is how to distinguish association from causation. In everyday language, correlation and association are used interchangeably. However, correlation and association are different terminologies. Correlation is to characterize the trend pattern between two variables, particularly; the Pearson correlation coefficient measures linear trends, while association characterizes the simultaneous occurrence of two variables. The widely used notion of association often indicates the linear correlation. When two variables are linearly correlated we say that there is association between them. Pearson correlation or its equivalent, linear regression is often used to assess association. Causation between two

variables is defined as independence between the distribution of cause and conditional distribution of the effect, given cause. In the nonlinear ANM, the causal relationship is assessed by testing independence between the cause variable and residual variable. We investigated the relationships between causation and association (linear correlation). Some theoretical analysis and real trait-imaging data analysis showed that there were three scenarios: (1) presence of both association and causation between two variables, (2) presence of association, while absence of causation and (3) presence of causation, while lack of association in causal analysis.

Finally, in real imaging data analysis, we showed that causal traits change the imaging signal variation in the brain regions. However, the traits that were associated with the imaging signal in the brain regions did not change imaging signals in the region at all.

The experiences in association analysis in the past several decades strongly demonstrate that association analysis is lack of power to discover the mechanisms of the diseases and provide powerful tools for medicine. It is time to shift the current paradigm of genetic analysis from shallow association analysis to more profound causal inference. Transition of analysis from association to causation raises great challenges. The results in this paper are considered preliminary. A large proportion of geneticists and epidemiologists have serious doubt about the feasibility of causal inference in genomic and epi-genomic research. Causal genetic analysis is in its infancy. The novel concepts and methods for causal analysis in genomics, epi-genomics and imaging data analysis should be developed in the genetic community.

Large-scale simulations and real data analysis for causal inference should be performed. We hope that our results will greatly increase the confidence in genetic causal analysis and stimulate discussion about whether the paradigm of genetic analysis should be changed from association to causation or not.

The results of confounder detection imply a possible confounder for two associated variables in gene expression data. Although the reference KEGG pathway does now show such a confounder, this finding implies that there may exist such confounder and shows a further direction for scientists to conduct experiments. In several other pathway datasets, however, no such hidden confounders were reported, which requires us to further improve the power of the algorithm.

## 5. Conclusion

Alternative to association analysis, the major goal of this dissertation is to propose a notion of causal analysis and to address several important issues for causal study. The standard approach to causal discovery is to use interventions or randomized experiments. Many genetic epidemiologists have always thought it impossible to detect causal SNPs using observational data. However, intervention or randomized experiments are unethical, time-consuming, expensive and infeasible in many cases. To address this critical barrier in GWCS, we focus on causal discovery methods developed for causal inference from observational data, not from interventional or randomized experiments and propose to use discrete ANMs as a major tool for GWCS. By large simulations and real data analysis we demonstrate the feasibility and limitations of the proposed GWCS as a new paradigm of genetic analysis.

Association is to measure dependent relationships and association analysis can be done from observational data. Causal inference is inductive reasoning (Causal inference in AI, 2019). In other words, causal inference is reasoned from the observed part to the unobserved general. The goal of causal inference is to learn the response of taking an action and is usually carried out from interventions. However, it may be expensive, infeasible and unethical to conduct intervention experiments. Modern causal theory attempts to learn outcome of an intervention from the observed data. Causation that can be inferred from observational data has been debated for more than one century. In this paper, we review great progress that have been made in causal inferences over the past several decades, and define causation as effect of

taking action in some system from observational data in terms of interventions or counterfactuals (Lattimore and Ong 2018). We also review three emerging major approaches to bivariate causal discovery: “do” action, counterfactuals and ICM and showed that these three approaches can be unified. The ANMs that are widely used algorithms to implement ICM are explored for GWCS. In GWCS, we assume that there are no confounding and selection bias. Methods for causation analysis with confounders will be presented elsewhere. Therefore, we lay down theoretic foundations for GWCS.

The original ANMs are used to distinguish cause-effect direction and do not provide P-value calculation for testing the causation of the SNP with disease. To overcome this limitation, we develop a test statistic and use permutations to calculate the P-value of statistics for testing the causation of the SNP with disease. This provides a practical approach to GWCS.

An essential problem for performing GWCS in practice is the type I error rates, power of the test statistics and feasibility of computations. We showed that type I error rates of the ANMs for testing the causation in both presence and absence of association were not significantly deviated from nominal level. In other words, large simulation results demonstrated that the ANMs for causation analysis of genetic variants were valid. Power of the ANMs depends on the probability parameter  $P$  in binomial distribution generating noise  $N_Y$ , sample sizes and significance levels. As we discussed in the text, probability parameter  $P$  determines the strength of causation. We showed that even for significance level  $\alpha = 0.01$  and  $P = 0.4$ ,

when sample sizes were 5,000, the power of the ANMs was close to 80%. If the parameter  $P \leq 0.15$ , using 500 sample sizes, we could ensure that the ANMs can reach power greater than 90% under both  $\alpha = 0.05$  and  $\alpha = 0.01$ . These results implied that the ANMs had high power to detect causation in many cases.

Distinguishing causation from association is an age-old problem. The most classical causal inference theory focuses on inferring causal relationships among more than three variables. Due to lack of methods for bivariate causal discovery, very few GWCS and very few results of significant causal genetic variants from GWCS have been reported. In the past decade, the rapid development in modern causal analysis theory has provided several efficient methods for bivariate causal discovery including ANMs. To promote application of causal inference to genetic analysis, we applied the ANMs to GWCS of SCZ. From the GWCS of SCZ, we have several important observations.

Causality is not only critical for us to understand disease mechanisms, but also particularly important for development of efficient treatment. Much of the failure of previous efforts of drug development was attributable to the insufficient understanding of disease mechanism.

The question whether we can infer causal relationships between genetic variants and disease from observational data has been debated for more than one century. Association and correlation analysis are the current paradigm of most genetic studies and have been used for



more than one century. Our study demonstrated that large proportions of causal loci couldn't be discovered by association analysis. Finding causal SNPs only via searching the set of associated SNPs may not be good enough for unraveling mechanisms of complex diseases. Causal analysis as an alternative to association analysis for genetic studies has never been systematically investigated. The main purpose of this paper is to stimulate discussion about causal analysis and association analysis, and both theoretical and practical researches in genomic causal analysis. We hope that our results will greatly increase confidence in applying causal inference to genetic analysis, more and more intelligent methods for causal inference will be developed, and more and more real causal analysis of complex diseases will be investigated.

## 6. References

- Abraham G, Inouye M. (2015). Genomic risk prediction of complex human disease and its clinical application. *Curr Opin Genet Dev.* 33:10-6.
- Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. (2012). SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 34: 2274-82.
- Altman N, Krzywinski M. (2015). Association, correlation and causation. *Nat Methods.* 12(10):899-900.
- Ashrafian H, Darzi A, Athanasiou T. (2015). A novel modification of the Turing test for artificial intelligence and robotics in healthcare. *Int J Med Robot* 11: 38-43.
- Athanasios L, Mattingsdal M, Kähler AK, Brown A, Gustafsson O, Agartz I, Giegling I, Muglia P, Cichon S, Rietschel M, et al. (2010). Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort. *J Psychiatr Res.* 44(12):748-53.
- Bausch J. (2012). On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. arXiv:1208.2691.
- Bergen SE, O'Dushlaine CT, Ripke S, et al. (2012). Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry* 17:880–886. doi: 10.1038/mp.2012.73
- Besserve, M., Shajarisales, N., Schölkopf, B., Janzing, D. (2017). Group invariance principles for causal generative models. *arXiv preprint arXiv:1705.02212*.

Boyle EA, Li YI, Pritchard JK. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 169, 1177-1186.

Budhathoki, K., and Vreeken, J. (2017). Causal inference by stochastic complexity. *arXiv:1702.06776*.

Callaway E. (2017). Genome studies attract criticism: Geneticists question ability of genome-wide association studies to find useful disease links. *Nature*. 546: 463.

Causal inference in AI. (2019). <https://www.allerin.com/blog/causal-inference-in-ai>.

Cho CH, Lee HJ, Woo HG, Choi JH, Greenwood TA, Kelsoe JR. (2015). CDH13 and HCRT2 may be associated with hypersomnia symptom of bipolar depression: A genome-wide functional enrichment pathway analysis. *Psychiatry Investig*. 12(3):402-7.

Clyde D. (2017). Disease genomics: Transitioning from association to causation with eQTLs. *Nat Rev Genet*. 18, 271

Costas J, Suárez-Rama JJ, Carrera N, Paz E, Páramo M, Agra S, Brenlla J, Ramos-Ríos R, Arrojo M. (2013). Role of DISC1 interacting proteins in schizophrenia risk from genome-wide analysis of missense SNPs. *Ann Hum Genet*. 77(6):504-12.

Elwert F. (2013). A Brief review of counterfactual causality. [https://www.ssc.wisc.edu/~felwert/causality/wp-content/uploads/2013/06/1-Elwert\\_Causal\\_Intro.pdf](https://www.ssc.wisc.edu/~felwert/causality/wp-content/uploads/2013/06/1-Elwert_Causal_Intro.pdf).

Fatemi SH, Folsom TD, Thuras PD. (2011). Deficits in GABA(B) receptor system in schizophrenia and mood disorders: a postmortem study. *Schizophr Res*. 128(1-3):37-43.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Proc. Roy. Soc. Edinburgh*. 52, 99-433.

- Friedman, J., Hastie, T., Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 9: 432–441.
- Glymour, C. (2015). The causal revolution, observational science and big data. *Lecture presented at Ohio University in the History and Philosophy of Science series, Athens, Ohio.*
- Greenwood V. (2018). Theory suggests that all genes affect every complex trait. <https://www.quantamagazine.org/omnigenic-model-suggests-that-all-genes-affect-every-complex-trait-20180620/>
- Gretton, A., Bousquet, O., Smola, A., and Schölkopf B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. In *Proceedings of the International Conference on Algorithmic Learning Theory*, 63–77.
- Grunwald, P. D., & Vitanyi, P. M. B. (2004). Shannon Information and Kolmogorov complexity. *IEEE Trans. Information Theory*, *arXiv:cs/0410002*.
- Haltia LT, Viljanen A, Parkkola R, Kemppainen N, Rinne JO, Nuutila P, et al. (2007). Brain White Matter Expansion in Human Obesity and the Recovering Effect of Dieting. *The Journal of Clinical Endocrinology & Metabolism*. 92(8): 3278-84. doi: 10.1210/jc.2006-2495.
- Holland PW. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 945–960.
- Howell KR, Floyd K, Law AJ. (2017). PKB $\gamma$ /AKT3 loss-of-function causes learning and memory deficits and deregulation of AKT/mTORC2 signaling: relevance for schizophrenia. *PLoS One*. 12(5):e0175993.

- Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*. 689-696.
- Ide M, Lewis DA. (2010). Altered cortical CDC42 signaling pathways in schizophrenia: implications for dendritic spine deficits. *Biol Psychiatry*. 68(1):25-32.
- Jaffe A. (2010). Correlation, causation, and association - What does it all mean?? <https://www.psychologytoday.com/us/blog/all-about-addiction/201003/correlation-causation-and-association-what-does-it-all-mean>.
- Janzing D, Schölkopf B. (2010). Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194.
- Jones MP, Walker MM, Attia JR. (2017). Understanding statistical principles in correlation, causation and moderation in human disease. *Med J Aust*. 207(3):104-106.
- Kahrilas IJ, Kahrilas PJ. (2019). Reflux disease and idiopathic lung fibrosis: association does not imply causation. *Chest*. 155(1):5-6.
- Kaplan D. (2018). Causal inference for observational studies. *The Journal of Infectious Diseases*. 219: 1-2.
- Langfelder P, Horvath S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 9, 559.
- Lattimore F, Ongv CS. (2018). A Primer on causal analysis. arXiv:1806.01488.
- Le, T., Hoang, T., Li, J., Liu, L., Liu, H. & Hu, S. (2016). A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. doi:10.1109/TCBB.2016.2591526.

- Lee HC, Melduni RM. (2018). Autoimmunity and cardiac arrhythmias in endemic pemphigus foliaceus - association, correlation or causation? *Heart Rhythm*. 15(5):732-733.
- Lei Y1, Zhu H, Duhon C, Yang W, Ross ME, Shaw GM, Finnell RH. (2013). Mutations in planar cell polarity gene SCRIB are associated with spina bifida. *PLoS One*. 8(7):e69262.
- Lemeire J and Janzing D. (2013). Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*. 23(2):227–249.
- Liu, F. and Chan, L. (2016). Causal inference on discrete data via estimating distance correlations. *Neur. Comp.*, vol. 28, no. 5, pp. 801–814.
- Lin N, Jiang J, Guo S, Xiong M. (2015). Functional Principal Component Analysis and Randomized Sparse Clustering Algorithm for Medical Image Analysis. *PLOS ONE*. 10(7): e0132945.
- Lin N, Zhu Y, Fan R and Xiong MM. (2017). A Quadratically Regularized Functional Canonical Correlation Analysis for Identifying the Global Structure of Pleiotropy with NGS Data. *PLOS Computational Biology*. 13(10): e1005788.
- Marsala T. (2015). Causality, correlation and artificial intelligence for rational decision making. *World Scientific*.
- Misztak P, Pańczyszyn-Trzewik P, Sowa-Kućma M. (2018). Histone deacetylases (HDACs) as therapeutic target for depressive disorders. *Pharmacol Rep*. 70(2):398-408.
- Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., & Schölkopf, B. (2016). Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research*. 17, 32:1-32:102.

- Nowzohour, C. and Bühlmann, P. (2016). Score-based causal learning in additive noise models. *Statistics*. 50, 471-485.
- Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27, 29–34.
- Ongen H, Brown AA, Delaneau O, Panousis NI, Nica AC, GTEx Consortium, Dermitzakis ET. (2017). Estimating the causal tissues for complex traits and diseases. *Nat Genet*. 49, 1676-1683.
- Orho-Melander M. (2015). Genetics of coronary heart disease: towards causal mechanisms, novel drug targets and more personalized prevention. *J Intern Med*. 278(5):433-46.
- Parascandolo, G., Rojas-Carulla, M., Kilbertus, N., Schölkopf, B. (2017). Learning Independent Causal Mechanisms. *arXiv preprint arXiv:1712.00961*.
- Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics* 6 DOI: 10.2202/1557–4679.1203
- Pearl J. (2000). Causality: models, reasoning and inference. MIT Press, Cambridge.
- Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*.
- Pearl J. (2019). The Seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*. 62: 54-60
- Peters, J., Bühlman, P. (2014). Identifiability of Gaussian Structural Equation Models with Equal Error Variances. *Biometrika*. 101, 219-228.
- Peters J, Janzing D, Schölkopf B. (2011). Causal inference on discrete data using additive noise models. *IEEE Trans Pattern Anal Mach Intell*. **33**, 2436-2350.

Peters, J., Janzing, D., Schölkopf, B. (2017). Elements of Causal Inference - Foundations and Learning Algorithms Adaptive Computation and Machine Learning Series. Cambridge, MA: The MIT Press.

Peters J, Mooij J, Janzing D, Schoelkopf B. (2011). Identifiability of Causal Graphs using Functional Models. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.

Peters, J., Mooij, J. M., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*. 15, 2009-2053.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Ren RJ1, Wang LL, Fang R, Liu LH, Wang Y, Tang HD, Deng YL, Xu W, Wang G, Chen SD. (2011). The MTHFD1L gene rs11754661 marker is associated with susceptibility to Alzheimer's disease in the Chinese Han population. *J Neurol Sci*. 308(1-2):32-4.

Rosenbaum PR, Rubin DB. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Ross SM. (1985). Introduction to probability models. Third Edition. Academic Press, Inc. London.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 511(7510):421-7.



- Schölkopf, B. and Janzing, D. and Peters, J. and Sgouritsa, E. and Zhang, K. and Mooij, J. (2012). On Causal and Anticausal Learning. *Proceedings of the 29th International Conference on Machine Learning*. 1255-1262.
- Shajarisales, N., Janzing, D., Schölkopf, B., Besserve, M. (2015). Telling cause from effect in deterministic linear dynamical systems. *Proceedings of the 32nd International Conference on Machine Learning*. 37, 285–294.
- Shi J, Levinson DF, Duan J, et al. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 460:753–757. doi: 10.1038/nature08192
- Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*. 7, 2003-2030.
- Sjogren M, Blennow K. (2005). The link between cholesterol and Alzheimer's disease. *World J Biol Psychiatry*. 6(2): 85-97.
- Spirtes P, Glymour C and Scheines R. (2000). Constructing Bayesian networks models of gene expression networks from microarray data. *In Proceedings of the Atlantic Symposium on Computational Biology*.
- Stretton J, Winston GP, Sidhu M, Bonelli S, Centeno M, Vollmar C, et al. (2013). Disrupted segregation of working memory networks in temporal lobe epilepsy. *Neuroimage Clin*. 2: 273-81. doi: 10.1016/j.nicl.2013.01.009.
- Stroup TS, McEvoy JP, Swartz MS, et al. (2003). The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophr Bull* 29:15–31

Sullivan PF, Keefe RS, Lange LA, Lange EM, Stroup TS, Lieberman J, Maness PF. (2007). NCAM1 and neurocognition in schizophrenia. *Biol Psychiatry*. 61(7):902-10.

Suzuki T, Iwata N, Kitamura Y, Kitajima T, Yamanouchi Y, Ikeda M, Nishiyama T, Kamatani N, Ozaki N. (2003). Association of a haplotype in the serotonin 5-HT<sub>4</sub> receptor gene (HTR4) with Japanese schizophrenia. *Am J Med Genet B Neuropsychiatr Genet*. 121B(1):7-13.

Székely,G.J., Rizzo, M.L., Bakirov,N.K. (2007). Measuring and testing independence by correlation of distances. *Annals of Statistics* 35(6), 2769–2794.

Székely,G.J., Rizzo, M.L. (2009). Brownian distance covariance. *Annals of Applied Statistics* 3(4),1236–1265.

Tan M, Alshalalfa M, Alhajj R, Polat F. (2011). Influence of prior knowledge in constraint-based learning of gene regulatory networks. *IEEE/ACM Trans Comput Biol Bioinform*. 8: 130-42.

Teipel SJ, Pruessner JC, Faltraco F, Born C, Rocha-Unold M, Evans A, et al. (2006). Comprehensive dissection of the medial temporal lobe in AD: measurement of hippocampus, amygdala, entorhinal, perirhinal and parahippocampal cortices using MRI. *J Neurol*. 253(6): 794-800. doi: 10.1007/s00415-006-0120-4.

Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D.J., Richards, J. B. (2017). Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet*. 19, 110-124. doi: 10.1038/nrg.2017.101.

Thung KH, Yap PT, Adeli E, Lee SW, Shen D; Alzheimer's Disease Neuroimaging Initiative. (2018). Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion. *Med Image Anal*. 45, 68-82.

- Wang, P. M., Rahman, L., Jin, and M., Xiong, (2016). A new statistical framework for genetic pleiotropic analysis of high dimensional phenotype data. *BMC Genomics* 17: 881.
- White, C. C. et al. (2017). Identification of genes associated with dissociation of cognitive performance and neuropathological burden: Multistep analysis of genetic, epigenetic, and transcriptional data. *PLoS Med.* 14, e1002287.
- Xiong MM. (2018). Big data in omics and image: integrated analysis and causal inference. CRC Press.
- Zakhari S, Hoek JB. (2018). Epidemiology of moderate alcohol consumption and breast cancer: association or causation? *Cancers (Basel)*. 10(10). pii: E349.
- Zenil H, Kiani NA, Zea AA, Tegnér J. (2019). Causal deconvolution by algorithmic generative models. *Nature Machine Intelligence* volume. 1: 58–66.
- Zhang, K., Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Zhang, K., Schölkopf, B., Spirtes, P., Glymour, C. (2018). Learning causality and causality-related learning. *National Science Review*. 5. 26-29. doi:10.1093/nsr/nwx137.