Dependencies in Learning Graphical Models

By

Md. Ashraful Islam

March 2025

Dependencies in Learning Graphical Models

A dissertation submitted to

THE GRADUATE SCHOOL OF ENGINEERING SCIENCE
OSAKA UNIVERSITY

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ENGINEERING

By

Md. Ashraful Islam

March 2025

Abstract

This thesis introduces two novel methodologies that significantly advance mutual information estimation for mixed-type variables and the identification of non-linear causal relationships. These contributions address critical challenges in data analysis and causal inference in complex settings, offering powerful tools for researchers across various scientific disciplines.

The first part focuses on estimating mutual information in datasets containing both discrete and continuous variables. Extending the Chow-Liu algorithm, our method constructs a forest that captures probabilistic dependencies among mixed-type variables. Using copula-based joint density estimation and the Watanabe Bayesian Information Criterion (WBIC) for computing free energies, our approach enables more accurate mutual information estimation, surpassing conventional likelihood-based methods. This method has been effectively applied to link genomic expression with Single Nucleotide Polymorphism (SNP) data in genome expression studies.

The second part introduces a method for learning non-linear causal structures by integrating Generalized Additive Models (GAMs) with the Hilbert-Schmidt Independence Criterion (HSIC). This approach addresses the challenges of estimating additive noise models without prior knowledge of the underlying non-linear relationships. Leveraging the adaptability of GAMs, our method models diverse non-linear dependencies without imposing strict parametric assumptions. We also provide a theoretical analysis demonstrating consistency in causal order identification, and our experimental results highlight the superior performance of the proposed approach in identifying causal relationships compared to existing methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the era of big data and complex systems, the ability to accurately analyze relationships between variables and infer causal structures has become paramount across various scientific disciplines. This thesis presents two novel methodologies that significantly advance our capabilities in mutual information estimation for mixed-type variables and the identification of non-linear causal relationships. These advancements address critical challenges in data analysis and causal inference, offering powerful tools for researchers across diverse fields.

1.1 Mutual Information Estimation for Mixed-Type Variables

Mutual information is a commonly employed metric for identifying dependencies between variables in a dataset. It measures how much information is shared between two random variables, figuring out how much two variables depend on each other. Mutual information is used in many areas of machine learning and information theory, for example, in feature selection, clustering, and classification (Shannon, 1948). It reduces the uncertainty associated with one variable when the other is known.

In the realm of graph theory, trees and forests offer a mathematical framework for representing variable interdependencies. An undirected acyclic graph, also known as a forest, is a group of trees that are not connected. It can visually show conditional independence (CI) relationships between variables (Bender and Williamson, 2010). The Chow-Liu algorithm uses mutual information metrics to build these trees or forests, allowing us to model the likely connections between variables using data. These graphs have been useful in complicated data, like gene differential analysis, where computational speed is very important (Suzuki, 2017).

As we navigate the complex world of mutual information estimation, it is essential to address the critical challenges, especially when the variables are of mixed types, with one being Gaussian and the others being discrete (Edwards et al., 2010). The computational landscape becomes significantly more intricate when examining graphical models that follow sequences like Y-X-Z, where X is Gaussian while Y and Z are discrete variables. In such models, optimizing mutual information estimates is far from trivial due to the intricate interdependencies among variables.

Previous approaches to this challenge include the Bayesian methodology proposed by Suzuki (2015), which focuses on constructing histograms and performing Bayesian computations. However, this approach has limitations, relying on unbounded histograms and facing difficulties with uneven sample sizes across datasets. Similarly, Suzuki (2017) proposed a hierarchical meshing mechanism to estimate mutual information, but this method struggles with small sample numbers and requires more data for proper quantification.

Our work introduces a new estimator of mutual information capable of dealing with variables of mixed types, including continuous and discrete variables. Initially, we use copula density estimation techniques to determine the joint density of mixed types of variables (Schmidt, 2007). We then employ a Bayesian approach to find the normalized constant needed to determine the minimum mutual information (Barron and Cover, 1991). We use the Watanabe Bayesian Information Criterion (WBIC) to obtain free energies, a method commonly used for model selection (Watanabe, 2013, 2021; Suzuki, 2023).

This novel approach overcomes the limitations of previous methods, allowing for more accurate estimation of mutual information in complex, mixed-type datasets. It is particularly useful in genomic data analysis, where we consider datasets containing information on gene expression in addition to Single Nucleotide Polymorphism (SNP) data.

1.2 Non-Linear Causal Inference from Data

Causal inference, the process of identifying cause-and-effect relationships from observational data, poses a significant challenge in various scientific domains, including economics, biology, and social sciences (Pearl, 2009). Understanding causal structures is crucial for the advancement of scientific knowledge, the informing of policy decisions, and the prediction of outcomes of interventions (Spirtes et al., 2000).

Traditional causal inference methods have been largely based on linear models or specific parametric assumptions. Techniques like the Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu et al., 2006) and its derivatives assume linear relationships or use independence tests that may not effectively capture non-linear dependencies. Similarly, structural equation models typically assume linear interactions among variables (Bollen, 1989). Although these methodologies have been successful in numerous applications, they often fall short of accurately modeling the complex nonlinear relationships prevalent in many real-world systems.

The recognition of the necessity for causal inference methods that can handle non-linear relationships has been growing, driven by the complexity observed in various fields. For example, ecological interactions and responses to environmental factors often exhibit non-linear patterns (Sugihara et al., 2012). Economic systems often display non-linear behaviors, such as the relationship between inflation and unemployment (Barnett et al., 2015). Neuroscience studies reveal brain connectivity patterns that involve complex non-linear interactions (Friston et al., 2003), while climate systems are marked by non-linear feedback and tipping points (Lenton et al., 2008).

Previous work in this area includes the approach by Hoyer et al. (2008), which addressed the issue of nonlinear causal inference but assumed prior knowledge of the underlying nonlinear function. Zhang and Hyvärinen (2009) proposed the use of nonlinear additive noise models (ANMs) for causal inference. Kernel-based methods, such as the kernel conditional independence test (KCI) (Zhang et al., 2011) and the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005), have gained attention due to their effectiveness in modeling complex, nonlinear relationships.

Our work builds upon these foundational ideas while extending the applicability of the LiNGAM framework to non-linear scenarios. We propose a novel method for learning nonlinear causal structures that combines generalized additive models (GAMs) with the Hilbert-Schmidt Independence Criterion (HSIC). This approach addresses several challenges in the field. Using the flexibility of GAMs, we can model a wide range of non-linear relationships without imposing strict parametric assumptions (Hastie and Tibshirani, 1990). Furthermore, incorporating HSIC improves the detection of intricate statistical dependencies that traditional correlation-based methods might overlook.

1.3 Thesis Structure

The remainder of this thesis is structured as follows:

- **Chapter 2** provides a comprehensive background and literature review for both research areas.
- **Chapter 3** details our novel methodology for mutual information estimation in mixed-type datasets, including its theoretical foundations.

- **Chapter 4** presents our approach to non-linear causal structure learning, outlining the integration of GAMs and HSIC, along with a rigorous theoretical analysis.
- **Chapter 5** demonstrates the application of our methods to both simulated and real-world datasets, showcasing their effectiveness and superiority over existing techniques.
- **Chapter 6** concludes the thesis with a discussion of our findings, their implications, and directions for future research.

Through this work, we aim to provide researchers and practitioners with robust, flexible, and interpretable tools for advanced data analysis and causal inference, contributing to the advancement of knowledge discovery in complex, non-linear systems.

Chapter 2

Background

This chapter provides the theoretical foundation and context for our research on mutual information estimation for mixed-type variables and non-linear causal structure learning. We begin by discussing the Chow-Liu algorithm and its relationship with mutual information, followed by an exploration of causal graphical models and structural equation models. We then delve into the challenges of causal discovery, particularly in non-linear settings.

2.1 Mutual Information and the Chow-Liu Algorithm

Mutual information, a concept rooted in information theory, quantifies the amount of information shared between two random variables (Shannon, 1948). For discrete random variables X and Y taking values in sets $\mathcal{X}$ and $\mathcal{Y}$ respectively, mutual information is defined as:

$$I(X,Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \tag{2.1}$$

where P_{XY}, P_X, and P_Y are the associated probability mass functions. Mutual information is non-negative and equals zero when X and Y are independent.

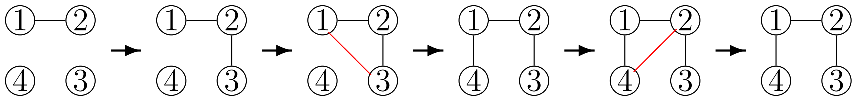

Figure 2.1: The Chow-Liu algorithm maximizes the sum of the mutual information values: $I(1,2) > I(2,3) > I(1,3) > I(1,4) > I(2,4) > I(3,4) \geq 0$.

The Chow-Liu algorithm (Chow and Liu, 1968) leverages mutual information to construct tree-structured graphical models. It approximates the joint probability distribution of N discrete variables $X^{(1)}, ..., X^{(N)}$ using a product of pairwise and univariate distributions:

$$Q(X^{(1)}, ..., X^{(N)}) = \prod_{k \in V} P(X^{(k)}) \prod_{\{i,j\} \in E} \frac{P(X^{(i)}, X^{(j)})}{P(X^{(i)})P(X^{(j)})} \tag{2.2}$$

where $V = \{1, ..., N\}$ is the set of vertices and E is the edge set of the forest $G = (V, E)$. The algorithm aims to maximize the sum of mutual information values across the edges of the forest.

Figure 2.1 illustrates the Chow-Liu algorithm's process of constructing a tree by selecting edges based on mutual information values.

2.2 Estimating Mutual Information from Data

Maximum likelihood estimators for mutual information tend to overfit, particularly for small sample sizes (Suzuki, 1993). To address this, alternative Bayesian estimates have been proposed. Suppose that the probabilities $P_X(x|\theta)$, $P_Y(y|\theta)$, and $P_{XY}(x, y|\theta)$ are indexed by a parameter $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^d$ is a parameter space. Let Q_X, Q_Y, and Q_{XY} be the associated marginal likelihood:

$$Q_X := \int_\Theta \prod_{i=1}^{n} P_X(x_i|\theta)\varphi(\theta)d\theta \tag{2.3}$$

and Q_Y, Q_{XY} are defined similarly, where $\varphi(\theta)$ is a prior probability of $\theta \in \Theta$. Suzuki (2012) proposed the following mutual information estimates:

$$J_n := \frac{1}{n} \log \frac{Q_{XY}}{Q_X Q_Y} \tag{2.4}$$

For discrete variables with $|\mathcal{X}| = \alpha_X$ and $|\mathcal{Y}| = \alpha_Y$, and using Jeffreys' prior, we have:

$$J_n = I_n - \frac{1}{2n}(\alpha_X - 1)(\alpha_Y - 1) \log n \tag{2.5}$$

where I_n is the maximum likelihood estimate. For Gaussian variables, the estimate takes the form:

$$J_n = I_n - \frac{1}{2n} \log n \tag{2.6}$$

These estimates provide more robust detection of independence, especially for large sample sizes (Suzuki, 1993).

2.3 Challenges with Mixed-Type Variables

Estimating mutual information becomes particularly challenging when dealing with a mixture of discrete and Gaussian variables (Edwards et al., 2010). Consider a scenario where X is Gaussian, while Y and Z are discrete. While it's relatively straightforward to compute mutual information for X-Y and X-Z pairs, the computation becomes complex for a Y-X-Z sequence.

Figure 2.2 illustrates the limitations of traditional methods when dealing with mixed-type variables in graphical models. Discrete vertices cannot be separated by a Gaussian vertex, significantly constraining the possible graph structures.

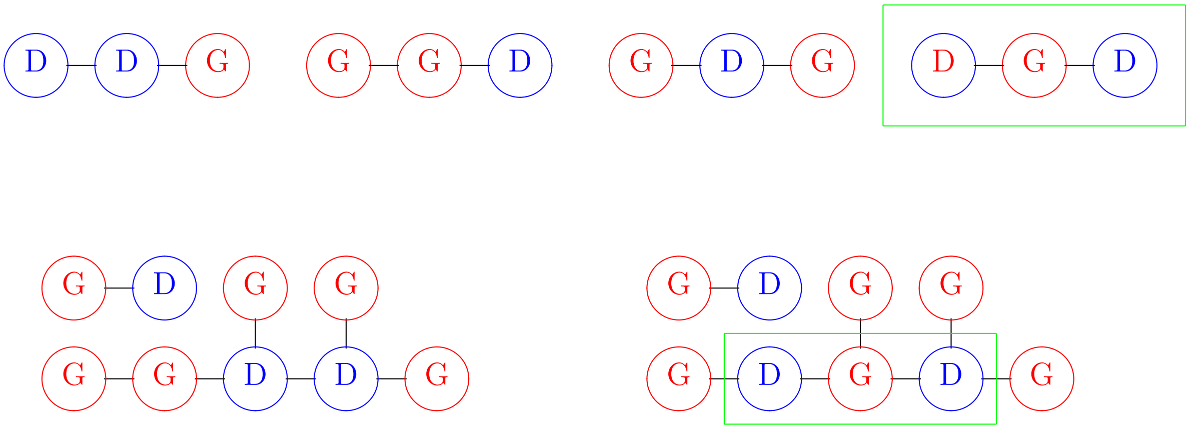

Figure 2.2: Upper: The labels ”D” and ”G” represent discrete and Gaussian variables, respectively. Out of the four cases presented, only the ”D-G-D” configuration is not permissible. Lower: The forest depicted on the right cannot be expressed by Edwards et al. (2010) due to the presence of the green rectangle.

2.4 Causal Graphical Models and Structural Equation Models

Causal graphical models provide a formalized approach to representing causal relationships (Pearl, 2000). Within this framework, causal structures are illustrated using directed acyclic graphs (DAGs), where the nodes correspond to random variables and the edges represent direct causal influences. For a set of p random variables, $\mathbf{X} = (X_1, \ldots, X_p)$, the causal relationships between these variables are described by a DAG $\mathcal{G} = (V, E)$, where $V = \{1, \ldots, p\}$ denotes the set of nodes corresponding to the variables, and $E \subseteq V \times V$ represents the set of directed edges that indicate direct causal relationships.

A DAG $\mathcal{G}$ is characterized by two essential properties: it is directed, which means that each edge has a direction from cause to effect, and it is acyclic, meaning that no variable can causally influence itself, directly or indirectly (Pearl, 2009).

To model the data-generating process, we use Structural Equation Models (SEMs) as described by Hoyle (2012). For the purpose of non-linear causal discovery, we focus on non-linear SEMs, which can be expressed as:

$$X_i = f_i(\text{PA}_i, \varepsilon_i), \quad i = 1, \ldots, p \tag{2.7}$$

In this expression, X_i denotes the i-th variable, and PA_i represents the set of parent variables that directly influence X_i in the graph $\mathcal{G}$, defined by $\text{PA}_i = \{X_j : (j, i) \in E\}$. The function f_i is an unspecified non-linear function that captures the relationship between X_i and its parent variables, while ε_i represents an independent noise term. This approach allows for the modeling of complex, non-linear interactions between causes and effects, extending the capabilities of traditional linear SEMs (Bollen, 1989).

2.5 The Causal Discovery Problem

Given observational data $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$, where each $\mathbf{x}^{(j)} = (x_1^{(j)}, \ldots, x_p^{(j)})$ is a realization of $\mathbf{X}$, our objective is to infer the underlying causal graph $\mathcal{G}$. This involves identifying the presence or absence of edges in $\mathcal{G}$ and determining the direction of these edges (Spirtes et al., 2000).

The task of causal discovery is complex due to several factors. Nonlinearity poses a significant challenge, as functions f_i can be arbitrarily non-linear, complicating the distinction between cause and effect based on simple statistical associations (Hoyer et al., 2008). The high-dimensional nature of the problem further complicates the task, as the number of possible DAGs grows super-exponentially with the number of variables p, rendering an exhaustive search infeasible (Chickering et al., 2004). Furthermore, finite sample sizes make it difficult to distinguish true causal relationships from spurious correlations (Kalisch and Bühlmann, 2007). The faithfulness assumption, which posits that all conditional independence relationships in the distribution of $\mathbf{X}$ are reflected in the graph structure, is another critical factor; violations of this assumption can lead to errors in causal discovery (Spirtes et al., 2000). Finally, the presence of confounders, or unmeasured common causes, can result in spurious associations between variables (Pearl, 2009).

2.6 Assumptions in Causal Discovery

To make the causal discovery problem tractable, several assumptions are adopted in our approach. We assume causal sufficiency, which means that there are no unmeasured common causes of the observed variables. This allows us to focus solely on the relationships among observed variables without considering latent confounders (Spirtes et al., 2000). We also assume the causal Markov condition, which states that the joint distribution of the variables factors according to the causal graph. Formally, each variable X_i is conditionally independent of its non-descendants given its parents in the graph (Pearl, 2000). The faithfulness assumption is another critical component, positing that all conditional independence relationships in the distribution are mirrored in the graph structure and vice versa (Spirtes et al., 2000). Additionally, we assume acyclicity, meaning that the true causal structure does not contain feedback loops or reciprocal causation (Pearl, 2009). Finally, we assume that the noise terms ε_i are non-Gaussian, mutually independent, and independent of the causal parents PA_i for each variable X_i (Peters et al., 2017).

2.7 Challenges in Nonlinear Settings

Nonlinear causal discovery introduces several additional challenges that require careful consideration. One major challenge is identifiability, as causal directions are generally not identifiable from observational data alone without additional assumptions. In nonlinear contexts, while identifiability is theoretically possible (Hoyer et al., 2008), it often requires specific conditions on functional relationships and noise distributions. Model flexibility is another important consideration. The chosen model class for the functions f_i must be sufficiently flexible to capture complex nonlinear relationships, yet sufficiently constrained to avoid overfitting and ensure identifiability (Mooij et al., 2016). Nonlinear models also introduce increased computational complexity, as they often require more sophisticated estimation procedures and hypothesis tests, which significantly increase the computational burden, especially for high-dimensional problems (Heinze-Deml et al., 2018). Lastly, interpretability is a crucial factor. While nonlinear models can capture more complex relationships, they may be less interpretable compared to their linear counterparts, complicating the communication and validation of discovered causal structures (Molnar, 2020).

In the subsequent chapters, we present novel methodologies that address these challenges in both mutual information estimation for mixed-type variables and non-linear causal structure learning. Our approaches combine the strengths of various techniques, including Generalized Additive Models (GAMs) and the Hilbert-Schmidt Independence Criterion (HSIC), to provide robust, flexible, and interpretable tools for advanced data analysis and causal inference.

Chapter 3

Mutual Information Estimation in Mixed-Type Variables

3.1 Introduction

This chapter presents our novel approach to estimating mutual information in datasets comprising both discrete and continuous variables. We address the limitations of existing methods, particularly when dealing with combinations of Gaussian and discrete variables. Our methodology overcomes the challenges associated with mixed-type data analysis by employing copula-based joint density modeling, which offers greater flexibility in handling the collaboration between discrete and continuous variables.

3.2 Copula-Based Approaches for Joint Density Estimation

3.2.1 Theoretical Foundation

Copulas are fundamental tools in multivariate statistical modeling, enabling us to separate the dependency structure among random variables from their marginal distributions. A copula C is a function that links univariate marginal distribution functions to form a multivariate distribution function. Specifically, for a d-dimensional random vector $\mathbf{U} = (U_1, U_2, \ldots, U_d)$, where each U_i is uniformly distributed over the interval $[0, 1]$, the copula is defined as:

$$C(u_1, u_2, \ldots, u_d) = \Pr(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d).$$

This framework allows us to model the dependency structure independently of the marginal distributions (Schmidt, 2007). The Probability Integral Transform (PIT) facilitates this approach by mapping any random variable X with cumulative distribution function (CDF) $F_X(x)$ to a uniform random variable $U = F_X(X)$ on $[0, 1]$. Sklar's theorem (Sklar, 1959) provides the theoretical underpinning for copula modeling by expressing any multivariate joint CDF $H(x_1, x_2, \ldots, x_d)$ in terms of its marginals and a copula:

$$H(x_1, x_2, \ldots, x_d) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)),$$

where $F_i(x_i)$ are the marginal CDFs. This theorem also allows us to represent the joint probability density function (PDF) $f_{X_1,X_2,\ldots,X_d}(x_1, x_2, \ldots, x_d)$ in terms of the copula density $c(u_1, u_2, \ldots, u_d)$ and the marginal PDFs $f_{X_i}(x_i)$:

$$f_{X_1,X_2,\ldots,X_d}(x_1, x_2, \ldots, x_d) = c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \prod_{i=1}^{d} f_{X_i}(x_i), \tag{3.1}$$

where the copula density $c(u_1, u_2, \ldots, u_d)$ is the mixed partial derivative of the copula function C:

$$c(u_1, u_2, \ldots, u_d) = \frac{\partial^d C(u_1, u_2, \ldots, u_d)}{\partial u_1 \partial u_2 \ldots \partial u_d}.$$

3.2.2 Copula Families for Modeling Dependencies

Selecting an appropriate copula family is crucial when modeling joint distributions, especially with mixed data types like discrete and continuous variables (Aas et al., 2009; Kojadinovic and Yan, 2010). In our study, we focus on the Gaussian and Clayton copulas to model the joint distribution of a Bernoulli-distributed variable X and a normally distributed variable Y.

3.2.2.1 Gaussian Copula

The Gaussian copula utilizes the standard normal CDF Φ and its inverse Φ^{-1}. The copula density function $c_{\text{Gaussian}}(u, v; \rho)$ is defined as:

$$c_{\text{Gaussian}}(u, v; \rho) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{z_u^2 - 2\rho z_u z_v + z_v^2}{2(1-\rho^2)} \right),$$

where $z_u = \Phi^{-1}(u), z_v = \Phi^{-1}(v)$, and ρ is the correlation parameter.

The joint PDF for X and Y using the Gaussian copula is:

$$f_{X,Y}(x, y; p, \mu, \sigma^2, \rho) = P(X = x) \cdot f_Y(y) \cdot c_{\text{Gaussian}}(F_X(x), F_Y(y); \rho), \tag{3.2}$$

where $P(X = x)$ is the probability mass function (PMF) of the Bernoulli variable X with parameter p, and $f_Y(y)$ is the PDF of the normal variable Y with mean μ and variance σ^2.

3.2.2.2 Clayton Copula

The Clayton copula, suitable for modeling asymmetric dependencies, has the density function:

$$c_{\text{Clayton}}(u, v; \theta) = (\theta + 1)u^{-\theta-1}v^{-\theta-1}\left(u^{-\theta} + v^{-\theta} - 1\right)^{-\theta-2},$$

where $\theta > 0$ is the dependence parameter. The corresponding joint PDF is:

$$f_{X,Y}(x, y; p, \mu, \sigma^2, \theta) = P(X = x) \cdot f_Y(y) \cdot c_{\text{Clayton}}(F_X(x), F_Y(y); \theta). \tag{3.3}$$

3.2.3 Bayesian Marginal and Joint Distributions

In the Bayesian framework, we incorporate prior distributions over the model parameters to account for uncertainty.

Marginal Distribution for X:

$$Q_X = \int_0^1 \prod_{i=1}^{n} P(X_i = x_i \mid p)\pi(p)dp,$$

where $\pi(p)$ is the prior for the Bernoulli parameter p, and n is the sample size.

Marginal Distribution for Y:

$$Q_Y = \int_{-\infty}^{\infty}\int_0^{\infty} \prod_{i=1}^{n} f_Y(y_i \mid \mu, \sigma^2)\pi(\mu, \sigma^2)d\mu d\sigma^2,$$

where $\pi(\mu, \sigma^2)$ is the joint prior over the mean μ and variance σ^2.

Joint Distribution Using Gaussian Copula:

$$Q_{XY} = \int\int\int\int \prod_{i=1}^{n} f_{X,Y}(x_i, y_i \mid p, \mu, \sigma^2, \rho)\pi(p, \mu, \sigma^2, \rho)dpd\mu d\sigma^2 d\rho,$$

where $\pi(p, \mu, \sigma^2, \rho)$ is the prior over all parameters.

3.3 Proposed Framework for Mutual Information Estimation

3.3.1 Bayesian Marginal and Joint Distributions

For a Bernoulli-distributed variable X with parameter p, the Bayesian marginal distribution Q_X is defined as:

$$Q_X = \int_0^1 \prod_{i=1}^{n} P(X = x_i)\varphi(p)dp$$

For a Gaussian-distributed variable Y with parameters μ and σ^2, the Bayesian marginal distribution Q_Y is:

$$Q_Y = \int_{-\infty}^{\infty} \int_0^{\infty} \prod_{i=1}^{n} f_Y(y_i)\varphi(\mu, \sigma^2)d\sigma^2 d\mu$$

Utilizing the Gaussian copula, the Bayesian joint distribution Q_{XY} becomes:

$$Q_{XY} = \int\int\int\int \prod_{i=1}^{n} P(X = x_i) \times f_Y(y_i) \times c_{\text{Gaussian}}(u_i, v_i; \rho)\varphi(p, \mu, \sigma^2, \rho)dpd\mu d\sigma^2 d\rho$$

3.3.2 Mutual Information Estimation

The mutual information of the mixture of discrete and continuous variables J_n incorporating the copula-based joint distribution is:

$$J_n = \frac{1}{n}\log\left(\frac{\int_\Theta \prod_{i=1}^{n} f_{XY}(x_i, y_i|\theta)\varphi(\theta)d\theta}{\left(\int_\Theta \prod_{i=1}^{n} f_X(x_i \mid \theta)\varphi(\theta)d\theta\right)\left(\int_\Theta \prod_{i=1}^{n} f_Y(y_i|\theta)\varphi(\theta)d\theta\right)}\right) \tag{3.4}$$

3.3.3 Watanabe Bayesian Information Criterion (WBIC)

In Bayesian inference, calculating the free energy is a key step for model selection and comparison. The free energy F_n serves as a measure that evaluates the fit of a model to observed data while incorporating a regularization effect from the prior distribution $\varphi(\theta)$ over the parameters θ. It is defined as:

$$F_n = -\log \int_\theta \prod_{i=1}^{n} P(x_i \mid \theta)\varphi(\theta)d\theta$$

However, solving this high-dimensional integral directly requires substantial computational resources, especially as the sample size n grows. To address this challenge, we employ the Watanabe Bayesian Information Criterion (WBIC), which offers an approximation to free energy while reducing computational effort. WBIC was introduced by Watanabe (Watanabe, 2013), as an extension of the traditional Bayesian Information Criterion (BIC), and it is particularly useful for dealing with non-regular models. WBIC is expressed as:

$$\text{WBIC} = \mathbb{E}_{1/\log n}\left[-\sum_{i=1}^{n} \log P(x_i \mid \theta)\right] = F_n + O_P(\sqrt{\log n})$$

This expression allows us to compute an approximation of F_n in a more tractable manner, which can be implemented via probabilistic programming frameworks like `Stan`.

3.3.3.1 Posterior Distribution and WBIC Approximation

The key idea behind WBIC is to modify the posterior distribution by introducing a *reverse temperature coefficient* $\beta > 0$, which adjusts the influence of the likelihood on the posterior distribution:

$$p(\theta \mid X^n, Y^n, \beta) = \frac{1}{Z(\beta)}\varphi(\theta)\prod_{i=1}^{n} P(X_i, Y_i \mid \theta)^\beta$$

Here, $Z(\beta)$ is the normalizing constant ensuring that the posterior integrates to 1, and β is a temperature-like parameter that controls the weight of the likelihood function. For WBIC, we use $\beta = \frac{1}{\log n}$, which asymptotically approaches zero as $n \to \infty$, making the approximation tighter.

By employing WBIC, we can compute the free energy approximation more efficiently, avoiding the direct evaluation of the high-dimensional integral. The approximation to free energy can be written as:

$$\mathbb{E}_\beta[Y(\cdot)] = \frac{\int_\theta Y(\theta) \prod_{i=1}^n P(x_i \mid \theta)^\beta \varphi(\theta) d\theta}{\int_\theta \prod_{i=1}^n P(x_i \mid \theta)^\beta \varphi(\theta) d\theta}, \quad \beta > 0$$

The log loss function $L_n(\theta)$ is used as the measure of fit:

$$L_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log P(X_i, Y_i \mid \theta)$$

where (X_i, Y_i) represent the observed data points. This log loss function reflects the predictive performance of the model, and WBIC balances the model fit against complexity, preventing overfitting.

3.3.3.2 Implementation via Stan

WBIC can be implemented efficiently using `Stan`, a probabilistic programming language that facilitates sampling from the posterior distribution. `Stan` uses Monte Carlo methods to approximate the integrals involved in the WBIC computation, making it computationally feasible for large datasets. Through this method, we can obtain an accurate and computationally efficient approximation of the free energy, F_n (Suzuki, 2020, 2021).

3.3.4 Free Energy and Mutual Information

In this framework, the free energy for the joint distribution $F_n^{XY}(x, y)$ can be expressed as:

$$F_n^{XY}(x, y) = -\log \int_w \prod_{i=1}^n f_{XY}(x_i, y_i \mid w) \varphi(w) dw = -\log Q_{XY}.$$

This leads us to the final formula for the mutual information J_n:

$$J_n = \frac{1}{n} \log \frac{Q_{XY}}{Q_X Q_Y} = \frac{1}{n} \log \frac{e^{-F_{XY}}}{e^{-F_X} \cdot e^{-F_Y}} = \frac{1}{n}(F_X + F_Y - F_{XY}).$$

This equation establishes a direct relationship between mutual information and the difference between the free energies of the marginal distributions and the joint distribution. It shows that mutual information can be viewed as the gain in free energy obtained by knowing the joint distribution of X and Y instead of their marginal distributions separately.

3.4 Evaluating Marginal Likelihood Estimates

We conducted a comparative study of marginal probability estimations using both direct approaches and our Stan-based approach. This analysis focused on Bernoulli, Gaussian, and pairwise distributions across diverse sample sizes ranging from 100 to 4000.

3.4.1 Experimental Setup

We generated synthetic datasets: Bernoulli data using a binomial process with a probability of success p, and Gaussian data with a specified mean μ and standard deviation σ. For pairwise distributions, we merged datasets consisting of Bernoulli and Gaussian variables, assuming independence between these variables.

Our Stan model, employing a Gaussian copula, was designed to model the relationship between Bernoulli and Gaussian variables. It includes priors for the mean (μ) and standard deviation (σ) of the Gaussian variable, along with a correlation parameter (ρ).

3.4.2 Results and Analysis

To visualize the performance of our marginal likelihood estimation methods, we plotted the results for Bernoulli, Gaussian, and pairwise distributions across various sample sizes, as shown in Figure 3.1.

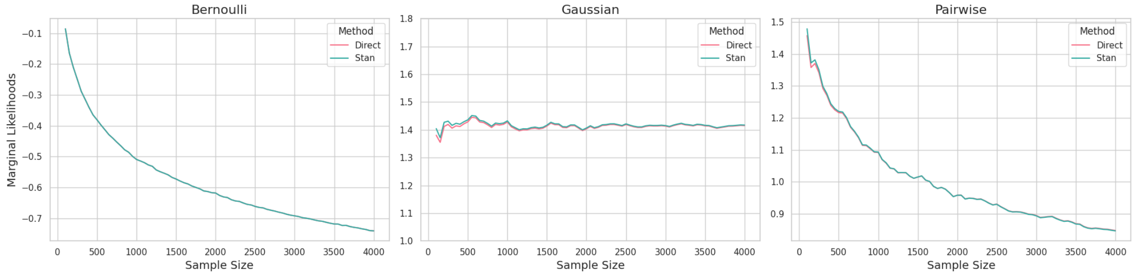

Figure 3.1: Marginal Likelihood Estimation Analysis for Diverse Sample Sizes within Bernoulli, Gaussian, and Pairwise Models. This visualization contrasts the outcomes from conventional direct estimation techniques (highlighted in red) with those obtained through our Stan-based methodology (depicted in green), showcasing the relationship between estimation methods and sample size on the precision of marginal likelihood calculations.

Figure 3.1 provides a comprehensive view of how our Stan-based approach compares to direct estimation methods across different distribution types and sample sizes. Our findings indicate:

1. For the Bernoulli distribution, marginal likelihood estimates from both direct and Stan-based approaches were closely aligned across all sample sizes. This is evident from the overlapping red and green lines in the top panel of the figure.

2. For the Gaussian distribution (middle panel), both techniques provided estimates in close agreement, with discrepancies becoming negligible as the sample size increased. The convergence of the red and green lines as we move right on the x-axis illustrates this trend.

3. For the pairwise distribution (bottom panel), integrating Bernoulli and Gaussian data, we observed consistent estimates from both approaches. The parallel

trajectories of the red and green lines demonstrate this consistency, underscoring their capability to effectively model mixed distribution models.

These visual results reinforce our conclusion that the direct and Stan-based approaches provide similar and consistent marginal likelihood estimates. Our Stan methodology, in particular, shows robustness across various statistical models, affirming its utility as a reliable tool for Bayesian model comparisons, especially in cases where direct methods might be computationally challenging or impractical. The consistency observed across different sample sizes and distribution types provides strong evidence for the reliability and versatility of our proposed method. This graphical analysis complements our numerical findings and offers intuitive insight into the performance of our methodology across a range of scenarios.

3.5 Summary

Our novel methodology for estimating mutual information in mixed-type variables addresses the challenges associated with analyzing datasets comprising both discrete and continuous variables. By leveraging copula-based joint density modeling and employing the Watanabe Bayesian Information Criterion, we provide a robust and flexible approach that overcomes the limitations of traditional methods. The comparative analysis of marginal likelihood estimates further validates the effectiveness of our approach across various distribution types and sample sizes. This methodology opens new avenues for analyzing complex, multidimensional datasets with increased precision and comprehensiveness.

Chapter 4

Non-Linear Causal Inference from Data

4.1 Introduction

This chapter presents a detailed description of our proposed method for nonlinear causal inference. We introduce a novel approach that integrates Generalized Additive Models (GAMs) with the Hilbert-Schmidt Independence Criterion (HSIC) to perform non-linear causal discovery. This method addresses the challenges associated with estimating additive noise models from data without prior knowledge of the nonlinear function, a situation where overfitting often hinders the detection of noise independence.

We begin by discussing the foundational components of our approach: Generalized Additive Models (GAMs) and the Hilbert-Schmidt Independence Criterion (HSIC). We then provide a comprehensive derivation of the objective function, followed by a detailed description of the algorithm. The chapter concludes with a theoretical analysis of our method and a discussion of practical considerations for implementation.

4.2 Generalized Additive Models (GAMs)

Generalized Additive Models (GAMs), introduced by Hastie and Tibshirani (1990), provide an extension of the Generalized Linear Models (GLMs) by allowing non-linear relationships between predictors and the response variable. GLMs assume that the effects of predictors on the response are linear, which often limits their applicability in real-world scenarios. GAMs relax this assumption by modeling the effect of each predictor as a smooth, non-linear function. This added flexibility enables GAMs to capture complex data patterns while preserving the interpretability of the model.

In many real-world systems, the assumption of linearity between predictors and a response variable is unrealistic. For example, in biological systems, the relationship between drug dosage and patient outcomes may be non-linear, with small doses having little effect, medium doses providing significant benefits, and large doses leading to harmful side effects. Linear models, which assume a constant effect of predictors, cannot capture such complex dynamics. GAMs provide a solution by allowing each predictor to have its own non-linear function, which can vary across the range of the predictor, offering a flexible, data-driven approach. Furthermore, despite the increased flexibility, GAMs retain the interpretability of traditional linear models because the effect of each predictor can be examined individually.

4.2.1 Mathematical Formulation of GAMs

The formulation of a GAM for a response variable Y and predictors $X_1, X_2, \ldots, X_p$ is given as:

$$g(E[Y \mid X_1, \ldots, X_p]) = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p) \tag{4.1}$$

Here, $g(\cdot)$ is a link function that transforms the expected value of Y into a scale appropriate for additive modeling. The term $E[Y \mid X_1, \ldots, X_p]$ represents the conditional expectation of Y given the predictors. The link function $g(\cdot)$ depends on the type of response variable. For continuous data, the identity link $g(\mu) = \mu$ is often used, while for binary responses, the logit link $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ is more common. For count data, the log link $g(\mu) = \log(\mu)$ is typically applied. The smooth functions $f_j(X_j)$ allow each predictor to have a non-linear effect on the response variable. These functions are estimated non-parametrically, typically using techniques such as splines or kernel methods, allowing for data-driven estimation of the functional form.

4.2.2 Estimation of GAMs

The parameters of a GAM are estimated by maximizing a penalized likelihood, which incorporates both the likelihood of the data and a penalty that ensures the smoothness of the estimated functions. The penalized log-likelihood function is expressed as:

$$L(\beta_0, f_1, \dots, f_p) = -l(Y, \eta) + \sum_{j=1}^{p} \lambda_j \int [f_j''(x)]^2 dx \tag{4.2}$$

In this formulation, $l(Y, \eta)$ represents the log-likelihood function for the data, and $\eta = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$ is the additive predictor. The second term, $\int [f_j''(x)]^2 dx$, penalizes the roughness of each smooth function $f_j(X_j)$ by incorporating the second derivative of $f_j(X_j)$. This ensures that the estimated functions remain smooth and do not overfit the data. The smoothing parameters λ_j control the trade-off between fitting the data closely and ensuring that the functions $f_j(X_j)$ are smooth. Larger values of λ_j enforce smoother functions, while smaller values allow for more flexibility.

4.2.2.1 Smoothing Parameter Selection

The selection of smoothing parameters λ_j is crucial to the performance of a GAM. If the smoothing parameters are too small, the model may overfit the data by capturing noise in the observations. Conversely, if the smoothing parameters are too large, the model may underfit the data, leading to overly smooth functions that fail to capture important patterns. To select optimal smoothing parameters, Generalized Cross-Validation (GCV) is commonly used (Craven and Wahba, 1978). The GCV score is defined as:

$$\mathrm{GCV}(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\left(1 - \frac{\mathrm{tr}(S)}{n}\right)^2} \tag{4.3}$$

In this equation, the numerator represents the mean squared error (MSE) between the observed values y_i and the predicted values $\hat{y}_i$, while the denominator adjusts for the complexity of the model, accounting for the effective degrees of freedom through the trace of the smoother matrix S. By minimizing the GCV score, the smoothing parameters are chosen in a way that balances model complexity and goodness of fit.

4.3 Hilbert-Schmidt Independence Criterion (HSIC)

The Hilbert-Schmidt Independence Criterion (HSIC), introduced by Gretton et al. (2005), is a kernel-based method for measuring the statistical independence between two random variables. Unlike traditional methods such as Pearson correlation, which can only detect linear dependencies, HSIC is capable of detecting both linear and non-linear relationships. HSIC is based on embedding the random variables into reproducing kernel Hilbert spaces (RKHS), allowing for a more flexible and powerful measure of dependence.

4.3.1 Definition

HSIC measures the dependence between two random variables X and Y by evaluating the Hilbert-Schmidt norm of the cross-covariance operator between their respective RKHS embeddings. Mathematically, HSIC is defined as:

$$\text{HSIC}(X, Y) = \|C_{XY}\|_{\text{HS}}^2 \tag{4.4}$$

Here, C_{XY} represents the cross-covariance operator between the RKHSs associated with X and Y, and $\| \cdot \|_{\text{HS}}$ denotes the Hilbert-Schmidt norm. HSIC evaluates how far the joint distribution of X and Y deviates from the product of their marginal distributions. If X and Y are independent, their joint distribution will equal the product of their marginals, and HSIC will be close to zero. If X and Y are dependent, HSIC will be positive, indicating the presence of statistical dependence.

4.3.2 Empirical Estimation of HSIC

In practice, HSIC is estimated from finite samples of the random variables. Given a sample of size n, the empirical estimate of HSIC is computed as:

$$\text{HSIC}(X, Y) = \frac{1}{(n-1)^2}\text{tr}(KHLH) \tag{4.5}$$

In this equation, K and L are the kernel matrices for X and Y, respectively, where each entry in the kernel matrices is computed as $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$. The centering matrix $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ ensures that the kernel matrices are mean-centered in the RKHS. The trace operator $\text{tr}(\cdot)$ computes the sum of the diagonal elements of the matrix product $KHLH$, which measures the dependence between X and Y.

4.4 Proposed Methodology

We propose a method for non-linear causal discovery that integrates the flexibility of Generalized Additive Models (GAMs) with the robust independence testing capabilities of the Hilbert-Schmidt Independence Criterion (HSIC). Our approach combines the power of GAMs to model complex, non-linear relationships with the sensitivity of HSIC to detect dependencies in the residuals, allowing us to uncover underlying causal structures.

4.4.1 Problem Formulation

We consider a set of variables $X = \{X_1, \ldots, X_p\}$, where the causal relationships between these variables are represented by a Directed Acyclic Graph (DAG) G. Each variable X_i is assumed to be generated as a function of its parents in the DAG, denoted as PA_i, and an independent noise term ε_i. The data generation process is assumed to follow the form:

$$X_i = f_i(\mathrm{PA}_i) + \varepsilon_i \tag{4.6}$$

Here, $f_i(\cdot)$ represents an unknown, non-linear function, and ε_i is a non-Gaussian noise term that is independent of the parent variables PA_i. The objective is to estimate the causal structure by fitting GAMs to model the relationships between the variables and using HSIC to test for statistical independence in the residuals.

4.4.2 Objective Function

The objective function in the context of Generalized Additive Models (GAMs) combined with Hilbert-Schmidt Independence Criterion (HSIC) is fundamental for estimating the relationships between variables while maintaining the smoothness of the functional forms and avoiding overfitting.

For each variable X_i, we model the influence of the other variables X_j, where $j \neq i$, through a set of smooth functions $f_{ij}(X_j)$. These functions capture the non-linear effects of predictor X_j on the response X_i. The general form of the GAM for each variable X_i is:

$$X_i = \beta_0 + \sum_{j \neq i} f_{ij}(X_j) + \varepsilon_i$$

Here:

- X_i is the dependent variable.
- β_0 is the intercept.
- $f_{ij}(X_j)$ is a smooth, non-linear function representing the effect of X_j on X_i.
- ε_i is the noise or residual term, assumed to be independent and non-Gaussian.

The estimation of the smooth functions $f_{ij}(X_j)$ involves minimizing a penalized least squares criterion. The penalization ensures that the functions $f_{ij}(X_j)$ do not overfit the data by being overly flexible. The objective function for estimating the smooth functions f_{ij} is expressed as:

$$\hat{\beta}_0, \hat{f}_{ij} = \arg \min_{\beta_0, f_{ij}} \sum_{i=1}^{p} \left[\sum_{k=1}^{n} \left(X_{i,k} - \beta_0 - \sum_{j \neq i} f_{ij}(X_{j,k}) \right)^2 + \sum_{j \neq i} \lambda_j \int \left(f''_{ij}(X_j) \right)^2 dX_j \right]$$

This objective function consists of two main components:

1. **Least Squares Error**: The term $\sum_{k=1}^{n} \left(X_{i,k} - \beta_0 - \sum_{j \neq i} f_{ij}(X_{j,k}) \right)^2$ represents the sum of squared residuals, where $X_{i,k}$ is the observed value of X_i for the k-th observation, and $\beta_0 + \sum_{j \neq i} f_{ij}(X_j)$ is the additive model fitted to the data. This term ensures that the model fits the data accurately.

2. **Smoothness Penalty**: The term $\sum_{j \neq i} \lambda_j \int \left(f''_{ij}(X_j) \right)^2 dX_j$ penalizes the roughness of the functions $f_{ij}(X_j)$ by incorporating their second derivatives. This

smoothness penalty controls the flexibility of the functions, with larger values of the smoothing parameters λ_j enforcing smoother (i.e., less wiggly) functions. The integration $\int \left(f''_{ij}(X_j)\right)^2 dX_j$ ensures that functions with large second derivatives (which correspond to rapid changes) are penalized, preventing overfitting.

The smoothing parameters λ_j play a crucial role in balancing the trade-off between model fit and smoothness. Smaller values of λ_j allow the functions $f_{ij}(X_j)$ to capture more complex patterns in the data, while larger values force the functions to be smoother and avoid capturing noise.

4.4.2.1 Interpretation of the Objective Function

The objective function minimizes the total error across all variables X_i, with each predictor X_j contributing to the additive model for X_i. The penalization ensures that the estimated functions are smooth and interpretable. The second derivative penalty is a key feature of non-parametric regression techniques, such as splines, that are commonly used in GAMs.

In the context of causal discovery, this objective function allows us to model the non-linear dependencies between variables while ensuring that the estimated relationships are smooth and generalize well to unseen data.

4.4.3 Determining Causal Order

The threshold τ is derived from the asymptotic distribution of the normalized HSIC (nHSIC) statistic. Under the null hypothesis of independence, it has been shown that as the sample size n increases, nHSIC converges in distribution to a weighted sum of squared normal variables, i.e.,

$$\tau \leftarrow \lim_{n\to\infty} \sum_{l=1}^{\infty} \lambda_l z_l^2,$$

where $\{\lambda_l\}$ are the eigenvalues of the kernel integral operator and $\{z_l\}$ are independent standard normal variables, i.e., $z_l \sim \mathcal{N}(0, 1)$. This asymptotic result provides a theoretically justified threshold for assessing statistical independence. For a detailed derivation and discussion of these properties, please refer to Gretton et al. (2005, 2008)

4.4.4 Algorithm for Non-Linear Causal Discovery

This section describes the steps in the proposed non-linear causal discovery method, which integrates Generalized Additive Models (GAMs) and the Hilbert-Schmidt Independence Criterion (HSIC). The method estimates causal relationships between variables by modeling non-linear dependencies using GAMs and applying HSIC to test for independence in the residuals. The final output is a causal graph representing the discovered causal structure.

Algorithm 1 Proposed Algorithm for Non-linear Causal Discovery

```
Require: Dataset $D = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ where each $\mathbf{x}^{(k)} = \left(x_1^{(k)}, \dots, x_p^{(k)}\right)$
Ensure: A causal graph $G = (V, E)$
Standardize each variable: $X'_i = \frac{X_i - \mu_i}{\sigma_i}$ for $i = 1$ to $p$
for $i = 1$ to $p$ do
    Fit GAM: $X'_i \approx \beta_{0,i} + \sum_{j \neq i} f_{ij}(X'_j)$
    Minimize penalized loss:
        $$\mathcal{L}(\beta_{0,i}, \{f_{ij}\}_{j \neq i}) = \sum_{k=1}^{n} \Big( X'_{i,k} - \beta_{0,i} - \sum_{j \neq i} f_{ij}(X'_{j,k}) \Big)^2 + \sum_{j \neq i} \lambda_{ij} \int \Big[ f''_{ij}(x) \Big]^2 dx,$$
    choosing $\lambda_{ij}$ via GCV or REML.
    Compute residuals: $r_i^{(k)} = X'_{i,k} - \Big( \beta_{0,i} + \sum_{j \neq i} f_{ij}(X'_{j,k}) \Big)$
end for
Initialize matrix $M \in \mathbb{R}^{p \times p}$
for $i = 1$ to $p$ do
    for $j = 1$ to $p$ where $j \neq i$ do
        Compute HSIC between $r_i$ and $X'_j$:
            $$M_{ij} = \mathrm{HSIC}(r_i, X'_j) = \frac{1}{(n-1)^2} \operatorname{tr}(K_i H L_j H),$$
    where $K_i$ is the kernel matrix for residuals $r_i$, $L_j$ is the kernel matrix for $X'_j$,
    and $H$ is the centering matrix.
    end for
end for
for each pair $(i, j)$ where $i \neq j$ do
    if $M_{ij} < \tau$ then
        Infer edge $X'_j \to X'_i$
    else if $M_{ij} > \tau$ then
        Infer edge $X'_i \to X'_j$
    end if
end for
Construct initial graph $G = (V, E)$ using the directed edges determined
    above.
while $G$ contains at least one directed cycle do
    Identify a cycle $C$ in $G$
    Remove edge $e^* = \arg\min_{(u,v) \in C} M_{uv}$
    $E \leftarrow E \setminus \{e^*\}$
end while
return The resulting acyclic graph $G = (V, E)$
```

4.5 Theoretical Analysis

Our method is supported by several theoretical results, which we state here. The detailed proofs can be found in the appendix.

Lemma 4.1 (Consistency of GAM Estimation)**.** *Let $\mathcal{X}$ be the domain of the input variables, $f : \mathcal{X} \to \mathbb{R}$ be the true underlying function, and $\hat{f}_n : \mathcal{X} \to \mathbb{R}$ be the GAM estimator based on n samples. Under suitable regularity conditions, the GAM estimators are consistent, that is, as the sample size $n \to \infty$, the estimated functions converge to the true functions in probability:*

$$\sup_{x \in \mathcal{X}} |\hat{f}_n(x) - f(x)| \xrightarrow{p} 0 \tag{4.7}$$

where $|\cdot|$ denotes the absolute value.

The proof of this lemma is provided in Appendix A.1.

Proposition 4.1 (HSIC and Independence)**.** *For characteristic kernels k and l, HSIC$(X, Y) = 0$ if and only if X and Y are independent (Gretton et al., 2005).*

The proof of this proposition is given in Appendix A.1.

Theorem 4.1 (Consistency of Proposed Method)**.** *Under the assumptions of Lemma 1 and Proposition 1, and assuming that the true data-generating process follows a non-linear additive noise model with non-Gaussian noise terms, the proposed method consistently recovers the true causal structure, including both the presence of causal relationships and their directions, as the sample size approaches infinity.*

The detailed proof of this theorem can be found in Appendix A.1.

These theoretical results provide a strong foundation for our method, ensuring its consistency and effectiveness in recovering true causal structures in non-linear settings with non-Gaussian noise.

4.6 Computational Considerations

The computational complexity of the proposed method is $O(n^3 + n^2p^3)$, where n is the number of samples and p is the number of variables. This complexity arises from the $O(n^3)$ operations needed for computing HSIC and the $O(n^2p^3)$ required for fitting GAMs when using cubic spline basis functions.

To enhance scalability and efficiency, we implemented several optimizations:

1. Utilization of the `mgcv` package in R (Wood, 2011) for fast restricted maximum likelihood estimation for GAMs.
2. Leveraging parallel processing to handle independent tasks.
3. Employment of sparse matrix representations and operations in high-dimensional settings.
4. Implementation of iterative methods for GAM fitting and HSIC computation for extremely large datasets.

These optimizations enable the proposed method to efficiently address moderate to large-scale causal discovery tasks.

4.7 Summary

The proposed method presented in this chapter offers a novel approach to nonlinear causal discovery by combining the flexibility of Generalized Additive Models with the independence testing power of the Hilbert-Schmidt Independence Criterion. By addressing the challenges of non-linear causal discovery, the proposed method provides a powerful tool for uncovering causal structures in complex systems across various domains, including economics, biology, and the social sciences.

In the subsequent chapters, we will present empirical evaluations of the proposed method on simulated data, demonstrating its effectiveness in practical applications, and comparing its performance to existing state-of-the-art causal discovery methods.

Chapter 5

Results and Discussion

5.1 Introduction

This chapter presents a comprehensive analysis of our novel methodologies: (1) estimating mutual information in mixed-type variables, and (2) non-linear causal structure learning. We evaluate these methods using both simulated and real-world datasets, comparing their performance against existing approaches. Our analysis employs various statistical techniques, including Bayesian modeling with Stan and rigorous performance metrics.

The exploration of accurate causal inference in complex, non-linear systems is a fundamental challenge in various scientific disciplines, including economics, biology, and social sciences (Pearl, 2009). Our research addresses this critical need by proposing novel methods that overcome limitations of traditional approaches in capturing intricate relationships within data.

5.2 Mutual Information Estimation in Mixed-Type Variables

5.2.1 Experimental Setup

Our analysis focused on three distinct datasets: synthetic data, a heart disease dataset, and a gene expression dataset. Each dataset was selected based on its unique characteristics, representing a variety of real-world scenarios that allow us to evaluate the robustness and flexibility of our proposed method. For all datasets, we employed the Stan programming language, which is widely used for Bayesian statistical modeling. We interfaced Stan with R version 4.3.2 through the `rstan` package to enable efficient and reproducible analyses.

For each dataset, we executed Stan models using four independent Markov Chain Monte Carlo (MCMC) chains, each running for a total of 2000 iterations. This process included a burn-in period of 1000 iterations to allow the chains to reach stationarity before the final sample was collected. By standardizing the MCMC configuration across datasets, we ensured consistent methodological rigor and facilitated coherent comparison of results between datasets. The datasets and scripts used for these analyses can be accessed via the link below:

https://github.com/ash141886/Forest-based-on-WBIC.

5.2.2 Model Performance with Simulated Data

We began our evaluation by applying our proposed method to a synthetic dataset containing 151 variables, including 75 discrete variables and 76 continuous variables. The aim of this analysis was to validate the model's performance under controlled conditions, allowing us to observe how effectively it can handle a mix of variable types and model the relationships between them. We utilized the Stan statistical model to examine these relationships, producing a series of diagnostic graphs and posterior distribution plots to assess model convergence and reliability.

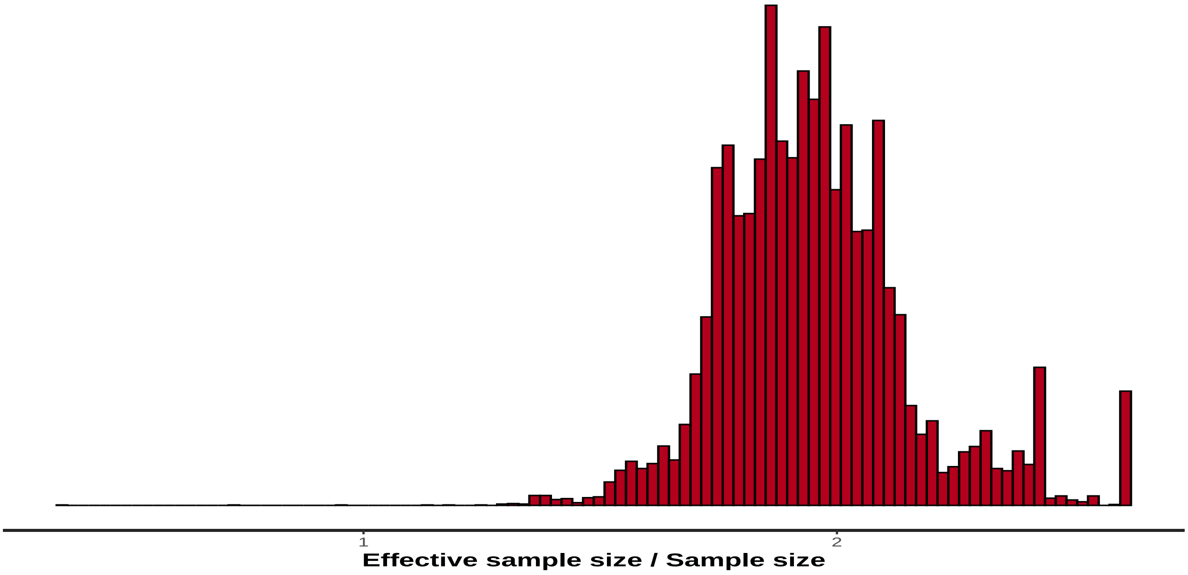

Figure 5.1: Distribution of the Effective Sample Size (ESS) over various parameters. A higher ESS indicates better sampling efficiency, meaning that the MCMC chains are effectively mixing, with minimal autocorrelation between samples.

In Figure 5.1, we observe the distribution of the Effective Sample Size (ESS) ratio across all parameters. The ESS is a critical diagnostic measure, as it quantifies the degree of independence between sampled values. A higher ESS ratio indicates more efficient sampling and ensures that the posterior samples are representative of the true distribution. The peak observed near a ratio of 1 suggests that the MCMC chains mixed well for most parameters, resulting in an accurate representation of the posterior distribution (Betancourt, 2017).

The results of the Gelman-Rubin potential scale reduction factor (R-hat) analysis are presented in Figure 5.2. The R-hat diagnostic is a widely used metric for assessing MCMC convergence, where values close to 1 suggest that the chains have converged to a stationary distribution. In this case, most R-hat values cluster near 1, indicating that our MCMC chains converged effectively across all parameters, providing confidence in the reliability of the posterior estimates (Gelman and Rubin, 1992).

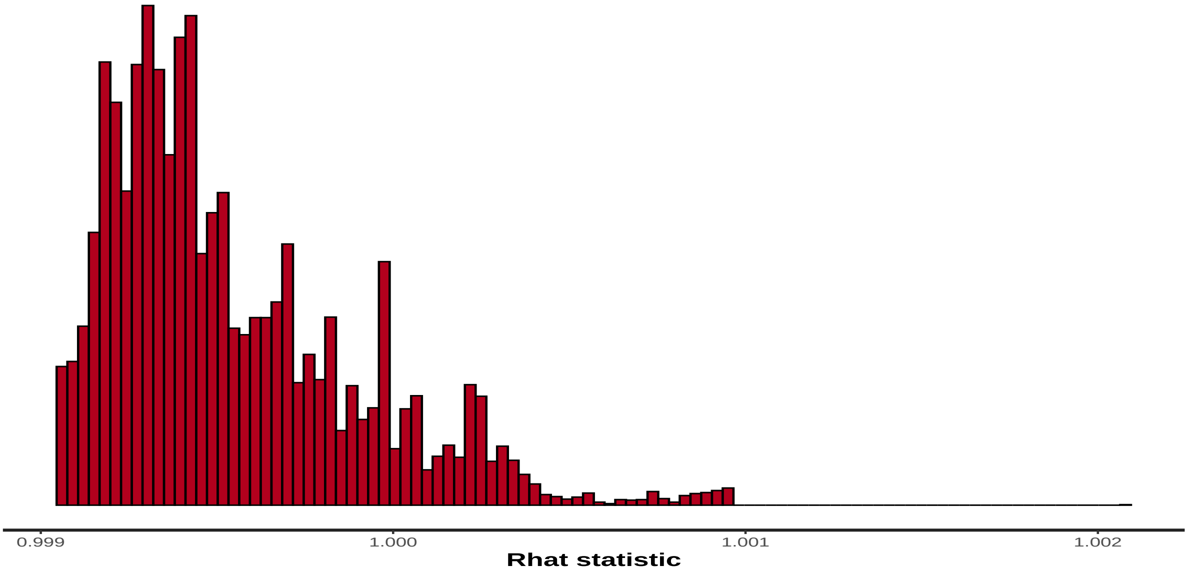

Figure 5.2: Gelman-Rubin R-hat diagnostic values for model parameters. R-hat values close to 1 indicate convergence across multiple MCMC chains, confirming the reliability of the posterior estimates.

Figure 5.3 illustrates the distribution of the model's Log Posterior values and the Mean Metropolis Acceptance rates. The concentration of Log Posterior values around a threshold of 560 reflects the model's consistent likelihood across iterations. Additionally, the high acceptance rates in the Metropolis algorithm indicate that the proposed samples were largely accepted, suggesting efficient exploration of the parameter space (Gelman, 2011). However, these high acceptance rates should be interpreted with caution, as they may indicate that the model is not thoroughly exploring the entire parameter space.

The trace plots and posterior distributions for selected parameters, presented in Figure 5.4, offer further insights into the behavior of the MCMC chains. On the left, the trace plots for parameters $p[2]$, $\mu[4]$, and $\sigma[3]$ show overlapping paths across the four chains, indicating good mixing and convergence to a stationary distribution. On the right, the posterior distributions exhibit bell-shaped curves, implying that the parameters are well-estimated and that the likelihood function is smooth and unimodal. These findings confirm that the model is capable of accurately estimat-

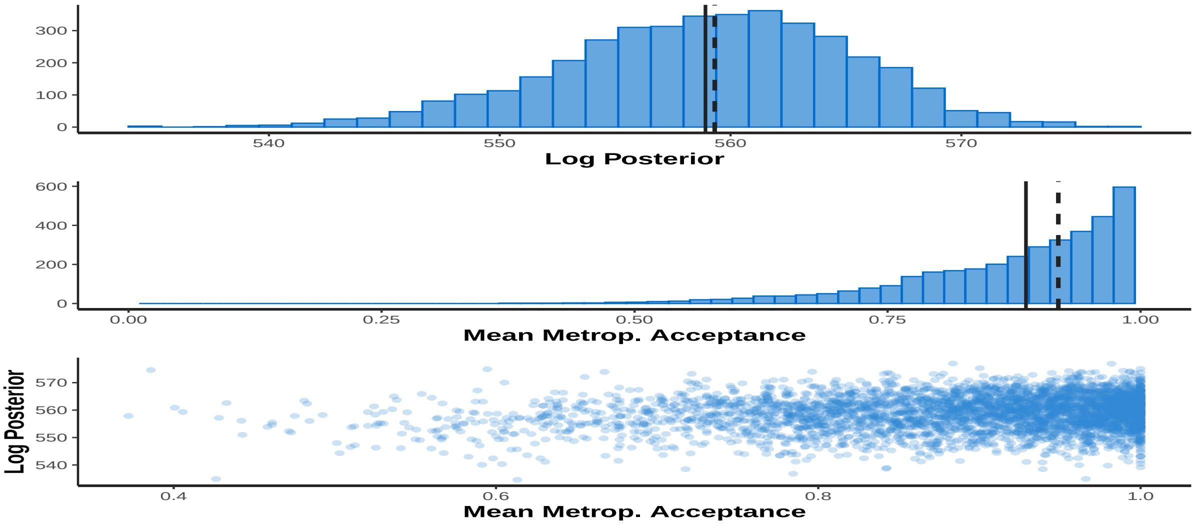

Figure 5.3: Distribution of the model's Log Posterior values and Mean Metropolis Acceptance rates. The concentration around 560 suggests a stable model likelihood, while high acceptance rates indicate efficient sampling.

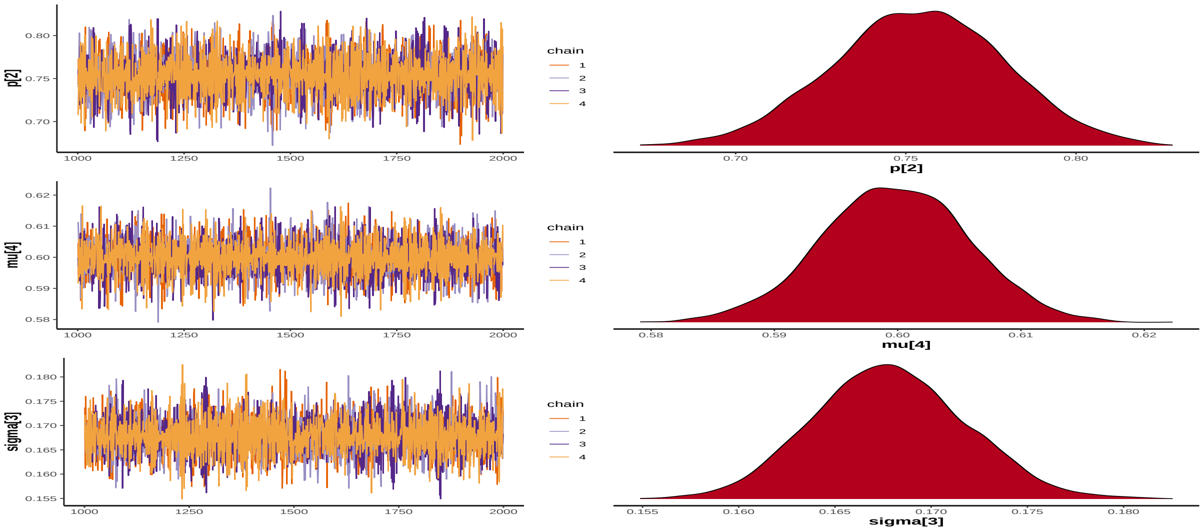

Figure 5.4: Trace plots (left) and posterior distributions (right) for parameters $p[2]$, $\mu[4]$, and $\sigma[3]$. The trace plots show good mixing and convergence across chains, while the posterior distributions indicate well-defined parameter estimates.

ing the parameters, even in the presence of both discrete and continuous variables (McElreath, 2020).

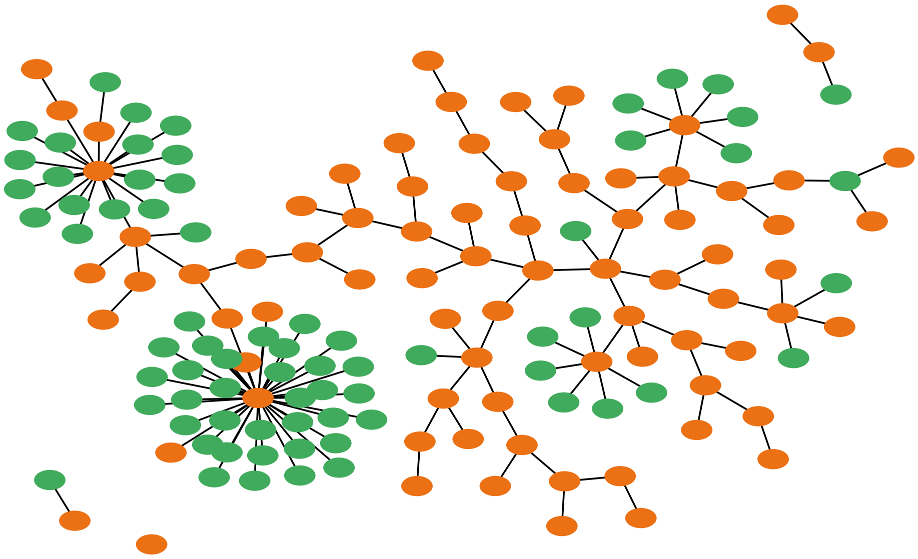

Figure 5.5: Forest graph constructed from the simulated dataset. This graph includes 75 discrete variables (orange nodes) and 76 continuous variables (green nodes), demonstrating the method's ability to identify independent and dependent structures.

The forest graph produced using the Chow-Liu algorithm, shown in Figure 5.5, illustrates the relationships between the 75 discrete and 76 continuous variables in the simulated dataset. Unlike traditional spanning tree models, which impose a single-parent constraint, this forest structure reveals the model's ability to detect independent variables while preserving the complexity of the dataset. This structure demonstrates the robustness of our proposed method in identifying both dependencies and independencies in the data.

5.2.3 Analysis of Heart Disease Dataset

Following the simulated data analysis, we applied our method to a real-world heart disease dataset comprising 303 patient records, each containing 14 attributes. These attributes represent a combination of demographic, physiological, and clinical factors related to cardiovascular health (Janosi et al., 1988). This dataset presents a more complex scenario than the synthetic data, as it includes real-world variability and interdependencies between factors.

Table 5.1: Summary of Markov Chain Monte Carlo (MCMC) Simulation Results for Heart Disease Dataset.

Parameter	Mean	SD	25%	50%	95%	n_eff	Rhat
p[1]	0.27	0.03	0.25	0.27	0.31	8294	1
p[2]	0.78	0.02	0.76	0.78	0.81	10191	1
p[3]	0.42	0.03	0.40	0.42	0.46	8774	1
p[4]	0.91	0.02	0.90	0.91	0.94	9896	1
p[5]	0.93	0.01	0.92	0.93	0.96	10349	1
p[6]	0.06	0.01	0.05	0.06	0.08	10554	1
p[7]	0.52	0.03	0.50	0.52	0.57	7749	1
p[8]	0.87	0.02	0.86	0.87	0.90	10059	1
p[9]	0.57	0.03	0.55	0.57	0.62	9872	1
mu[1]	1.54	0.03	1.52	1.54	1.60	8815	1
mu[2]	-0.56	0.03	-0.59	-0.56	-0.51	9717	1
mu[3]	-0.82	0.03	-0.84	-0.82	-0.77	10202	1
mu[4]	-1.55	0.03	-1.57	-1.55	-1.50	9435	1
sigma[1]	0.98	0.02	0.96	0.97	1.01	8562	1
sigma[2]	1.06	0.02	1.05	1.06	1.10	9903	1
sigma[3]	0.93	0.02	0.91	0.93	0.96	9785	1
sigma[4]	0.98	0.02	0.96	0.98	1.02	8224	1
sigma[5]	1.01	0.02	0.99	1.01	1.05	10058	1
lp__	-7417.46	6.08	-7421.36	-7417.12	-7408.12	1213	1

Table 5.1 presents the results of the MCMC simulations for the heart disease dataset. The Rhat values, which converge to 1 for all parameters, indicate that the multiple chains reached a consistent distribution, affirming the stability and reliability of the parameter estimates. The effective sample sizes (n_eff) are also substantial relative to the total number of iterations, indicating minimal autocorrelation in the samples, which further enhances the robustness of the model's estimates.

Figure 5.6 depicts the graphical model constructed for the heart disease dataset. The spanning tree structure highlights the interrelated nature of the variables, suggesting that the risk factors for heart disease are not independent, but rather interconnected in complex ways. This provides valuable insights into the underlying mechanisms contributing to cardiovascular disease.

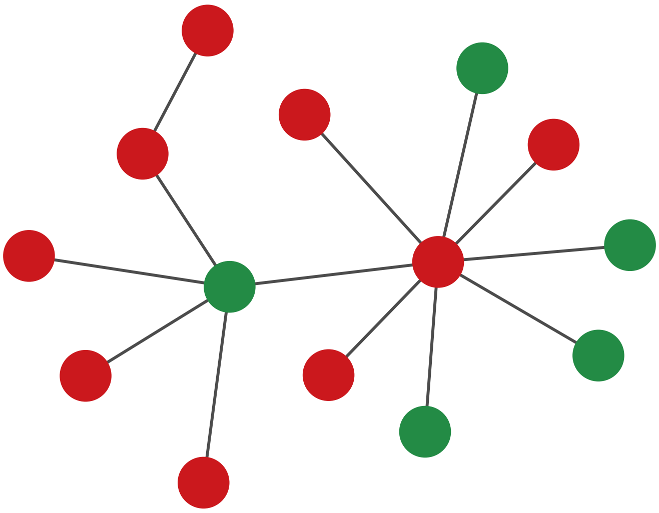

Figure 5.6: Graphical representation of the heart disease dataset, with discrete variables marked in red and continuous variables in green. This structure illustrates the interconnectedness of the variables and their potential relationships.

5.2.4 Analysis of Gene Expression Data

We further applied our method to a gene expression dataset from the HapMap project (Miller et al., 2005; Gamazon et al., 2010). This dataset includes genetic variants and gene expression levels for 90 Utah individuals from the CEU population. Our analysis focused on the first 200 polymorphic SNPs and 200 expression values to examine the associations between genetic variants and gene expression.

The forest graph shown in Figure 5.7 visualizes the relationships between 336 variables, including genomic expressions, SNPs, and gender information. The graph provides a detailed overview of the complex relationships within the dataset, with nodes representing different variable types. The presence of independent components in the graph demonstrates our method's ability to accurately identify truly independent variables, thus offering a deeper understanding of the underlying genetic mechanisms.

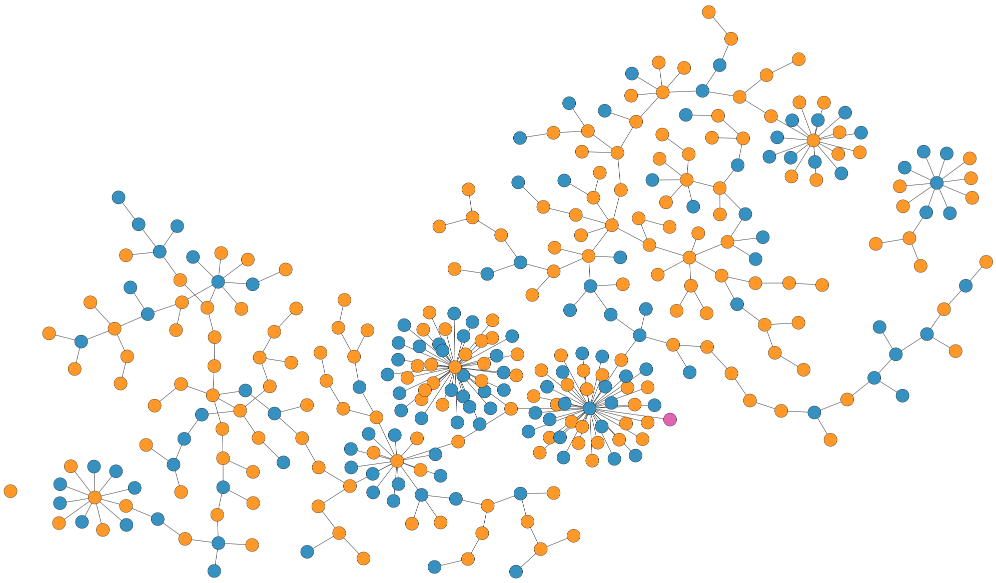

Figure 5.7: Forest graph illustrating relationships between 336 variables, including genomic expressions, SNPs, and gender information. The structure reveals the complex relationships between these variables, with distinct nodes highlighting independent variables.

5.3 Non-Linear Causal Inference from Data

5.3.1 Performance in Linear Scenarios

We start our analysis by evaluating the performance of our proposed method under linear conditions, providing a baseline for comparison with LiNGAM (Shimizu et al., 2006). To investigate the effect of sample size on model performance, we generated datasets with sample sizes of 400, 800, 1200, and 1500. The generated data followed a predefined causal structure consisting of five variables: x_1, x_2, x_3, x_4, and x_5. The causal relationships were structured such that x_1 influenced x_2, which subsequently affected x_4 and x_5, with x_5 further influencing x_3.

Figure 5.8 provides an in-depth comparison of the performance of our proposed method versus LiNGAM, across four important metrics: Mean Root Mean Square Error (RMSE), Mean Kolmogorov-Smirnov (KS) statistics, Mean KS p-values, and Mean t-test p-values. The results are evaluated over various sample sizes to assess

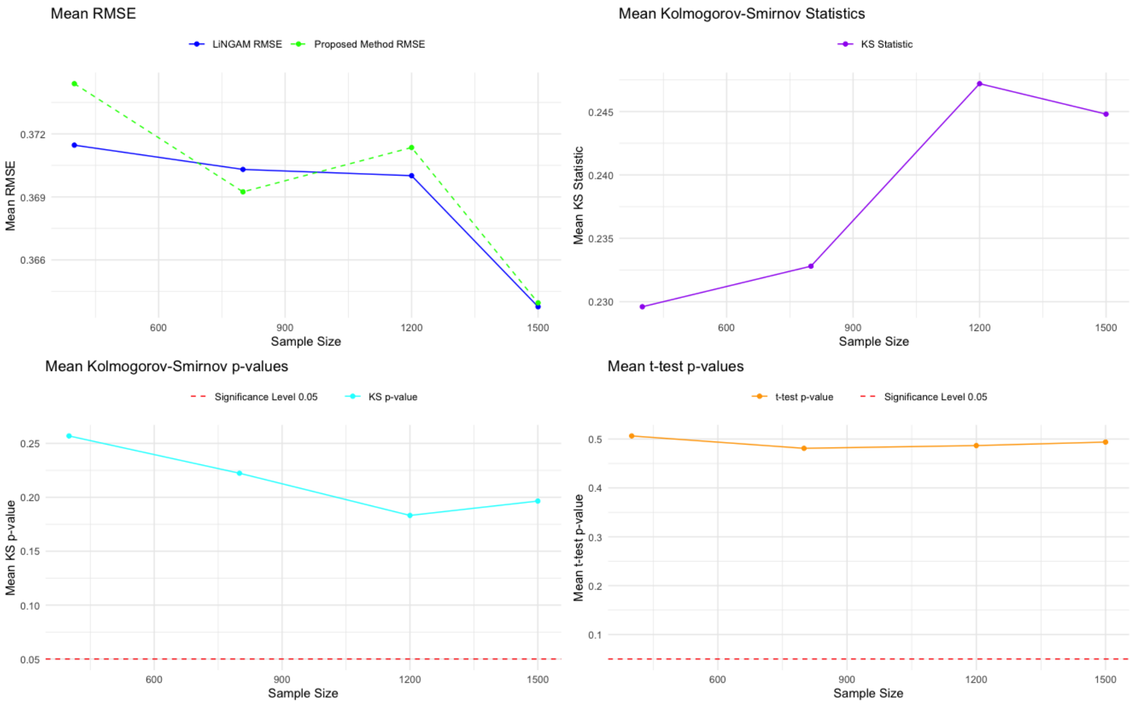

Figure 5.8: Performance comparison between LiNGAM and our method in linear scenarios. The plots show Mean RMSE, Mean KS statistics, Mean KS p-values, and Mean t-test p-values across different sample sizes.

how both methods behave under different data conditions.

The upper left subplot shows the mean RMSE for both methods. RMSE is used to measure the average deviation between the predicted and actual values, where lower values indicate better model performance. Although LiNGAM consistently achieves slightly lower RMSE values, suggesting a marginally better fit to the data, our method demonstrates nearly equivalent performance. The gap between the two methods remains small, indicating that despite LiNGAM's specialized design for linear models, our method effectively captures linear relationships with only a minor increase in error magnitude (Shimizu et al., 2006).

The upper right subplot illustrates the Mean Kolmogorov-Smirnov (KS) statistics, which quantify the maximum distance between the empirical cumulative distribution functions (ECDFs) of the predicted residuals and the true residuals. Higher KS statistics imply greater divergence between the predicted and actual distributions.

Both methods show comparable KS values, but as the sample size increases, there is a slight rise in the KS statistic for our proposed method. This trend may indicate that as our model gains flexibility with larger datasets, it begins to capture more subtle patterns, which can lead to a slight deviation from the true residual distribution.

The bottom left subplot depicts the mean KS p-values, which test the null hypothesis that there is no significant difference between the distributions of the predicted residuals and the actual residuals. Higher p-values suggest that the null hypothesis cannot be rejected, implying that the distributions are statistically indistinguishable. Both LiNGAM and our method yield p-values well above the conventional significance level of 0.05 for all sample sizes, indicating that neither method produces residuals significantly different from the true residuals. This result reinforces the validity of both methods in accurately modeling the residual distributions.

Finally, the bottom right subplot presents the mean p-values from the t-test, which assesses whether the mean of the predicted residuals differs significantly from the mean of the true residuals. A higher p-value suggests no significant difference between the two means. Both methods consistently achieve high t-test p-values across all sample sizes, showing that the residuals produced by our method are statistically comparable to those generated by LiNGAM, particularly in terms of mean accuracy. This further supports the robustness of our method in handling linear scenarios.

In summary, although LiNGAM achieves slightly lower RMSE values, the performance of our method is comparable, especially when looking at residual distribution metrics. The KS statistics and p-values, as well as the t-test results, show that our method performs similarly to LiNGAM in terms of residual accuracy, particularly as sample sizes increase. These findings suggest that our method is highly capable of modeling linear relationships, offering robust and reliable results across different evaluation criteria.

5.3.2 Performance in Non-Linear Scenarios

The primary objective of this study is to develop a method that effectively identifies causal relationships in complex nonlinear systems, where conventional linear techniques such as LiNGAM (Shimizu et al., 2006) may be inadequate. To evaluate the performance of the proposed method, synthetic datasets were created with sample sizes of 400, 800, 1200, and 1500, as well as with different numbers of variables (4, 8, 12, 15), thus allowing the method to be tested under various levels of complexity. Nonlinear transformations, such as sine, cosine, and exponential functions, were randomly applied to selected parent variables, with a sparsity parameter to regulate the introduction of nonlinearity in the data. Furthermore, non-Gaussian noise, an important component of LiNGAM's assumptions, was incorporated and extended to a nonlinear context. This noise was introduced into each variable using random samples from several non-Gaussian distributions, such as exponential, chi-squared, or t-distribution, and then scaled by a specified noise level. This approach ensured that the datasets featured both nonlinear relationships and non-Gaussian noise, allowing for a thorough evaluation of the method's effectiveness in identifying complex causal structures, particularly when the non-Gaussianity assumption is critical for establishing causal directions. The results, presented in Figures 5.9 to 5.13, provide a comprehensive comparison between the proposed method and LiNGAM, demonstrating the benefits of the proposed method in nonlinear contexts, especially regarding the utilization of non-Gaussian noise for causal discovery.

However, as illustrated in Figure 5.13, the improvements in accuracy and structural precision come with the trade-off of increased computational time. As the number of variables increases, the computational cost associated with the proposed method rises, reflecting the greater complexity and resources required for accurately modeling non-linear relationships. While the computational demand is higher, the significant gains in accuracy and structural fidelity justify this cost in many real-world applications where precision is paramount (Chickering, 2002).

In conclusion, the results demonstrate that the proposed method offers clear advantages over LiNGAM in non-linear settings. The method consistently outperforms LiNGAM in key metrics such as accuracy, F1 score, MSE, and SHD, highlighting its robustness in capturing complex non-linear causal relationships. This makes it a highly valuable tool for researchers working on causal discovery in fields where such non-linearities are prevalent (Peters et al., 2014; Mooij et al., 2016; Hastie et al., 2009).

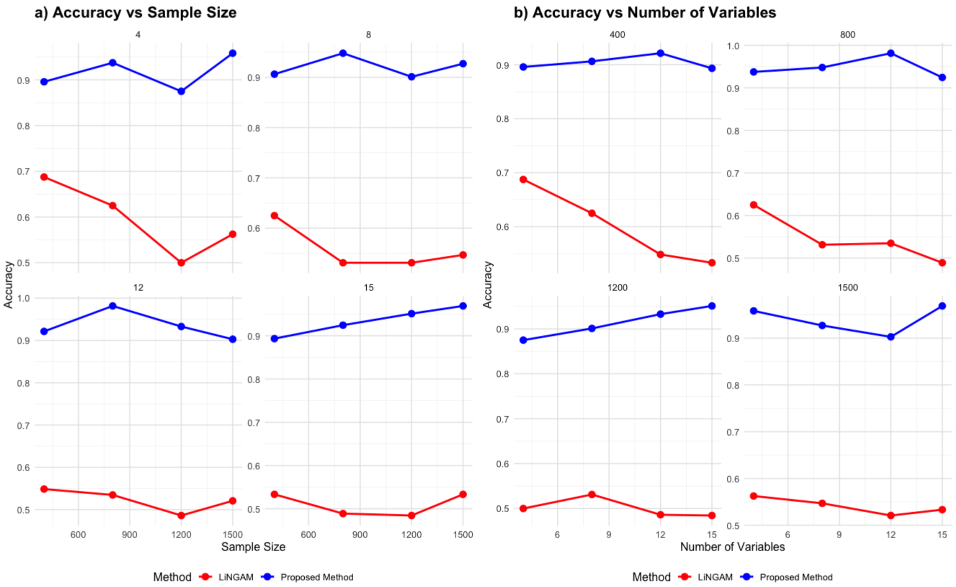

Figure 5.9: Accuracy comparison between LiNGAM and our method in non-linear scenarios. Our method consistently outperforms LiNGAM, particularly as the number of variables increases.

The accuracy results depicted in Figure 5.9 clearly indicate that the proposed method consistently surpasses the performance of LiNGAM across different sample sizes and numbers of variables. This consistent trend highlights the robustness of the method and its capability to model complex non-linear dependencies, where LiNGAM's effectiveness is notably diminished (Shimizu et al., 2006). The results demonstrate that our approach is better equipped to capture the intricacies of non-linear relationships, making it a more suitable choice in such scenarios.

Additionally, the F1 scores illustrated in Figure 5.10 further validate the proposed method's efficacy, particularly in non-linear environments. As the number of vari-

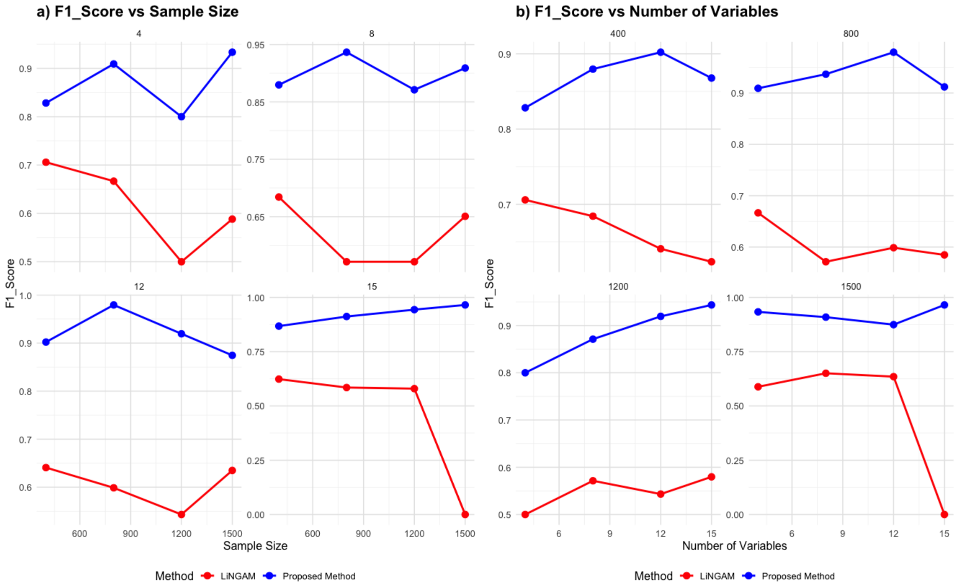

Figure 5.10: F1 score comparison between our method and LiNGAM. Our method achieves higher F1 scores, especially in datasets with a greater number of variables.

ables increases, our method shows a more balanced trade-off between precision and recall, reflecting its ability to accurately identify causal relationships while minimizing false positives and negatives. This improved balance is crucial in the context of causal discovery, especially when the underlying relationships become more complex (Hoyer et al., 2008).

Figure 5.11 presents a comparison of the mean squared error (MSE) between the proposed method and LiNGAM. Across all conditions, the proposed method consistently achieves lower MSE values, demonstrating its superior prediction accuracy in non-linear contexts. The lower MSE values suggest that the predictions made by our method are more reliable, even when faced with intricate non-linear structures in the data (Zhang and Hyvärinen, 2012).

The Structural Hamming Distance (SHD) results, shown in Figure 5.12, underscore the enhanced structural accuracy of our method when reconstructing the true causal graphs. The consistently lower SHD values indicate that our method more effectively captures the true underlying causal structures compared to LiNGAM, particularly in scenarios involving non-linear relationships. This is a crucial metric in causal

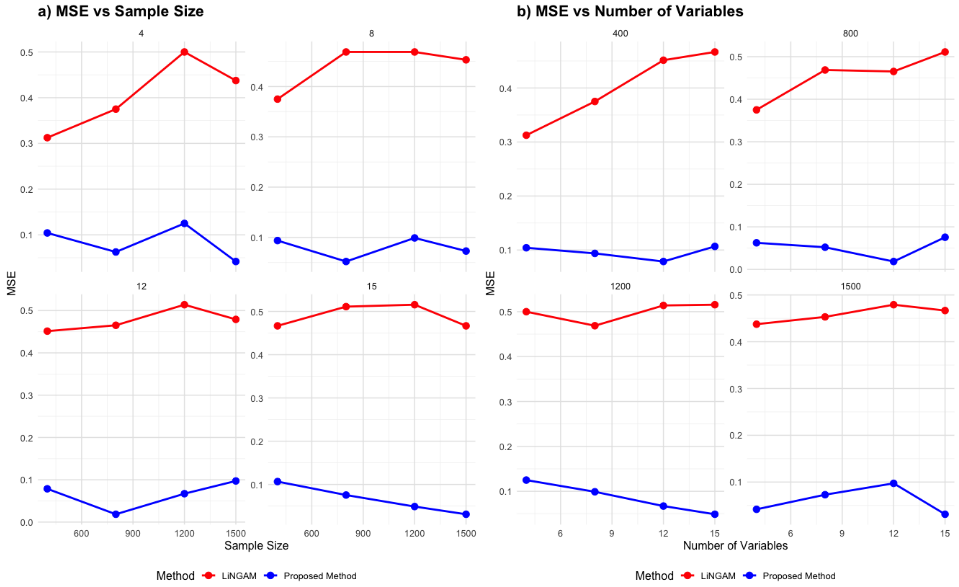

Figure 5.11: Comparison of Mean Squared Error (MSE) values between LiNGAM and our method. Our method consistently achieves lower MSE values, particularly as system complexity increases.

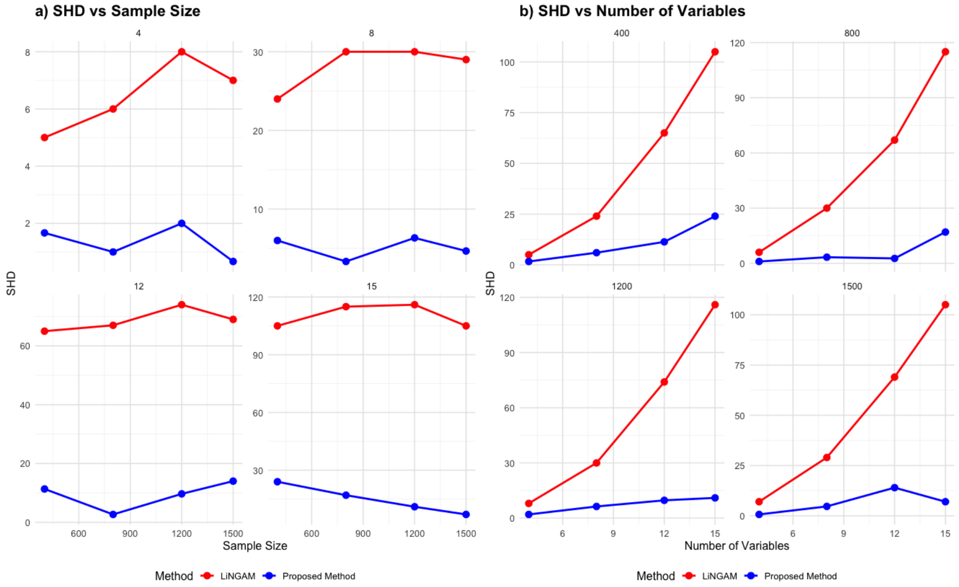

Figure 5.12: Comparison of Structural Hamming Distance (SHD) between our method and LiNGAM. Our method shows lower SHD values, indicating better reconstruction of causal structures.

discovery, as a lower SHD implies greater alignment between the estimated and true causal networks (Peters et al., 2017).

However, as illustrated in Figure 5.13, the improvements in accuracy and structural precision come with the trade-off of increased computational time. As the

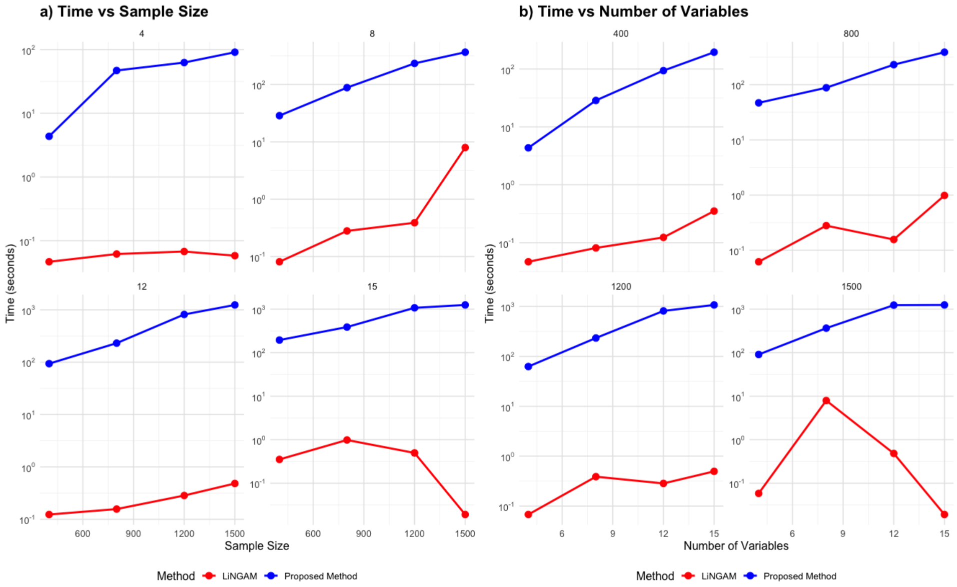

Figure 5.13: Comparison of computational time required by LiNGAM and our method. Our method requires more computational resources, particularly as the number of variables grows.

number of variables increases, the computational cost associated with the proposed method rises, reflecting the greater complexity and resources required for accurately modeling non-linear relationships. While the computational demand is higher, the significant gains in accuracy and structural fidelity justify this cost in many real-world applications where precision is paramount (Chickering, 2002).

In conclusion, the results demonstrate that the proposed method offers clear advantages over LiNGAM in non-linear settings. The method consistently outperforms LiNGAM in key metrics such as accuracy, F1 score, MSE, and SHD, highlighting its robustness in capturing complex non-linear causal relationships. This makes it a highly valuable tool for researchers working on causal discovery in fields where such non-linearities are prevalent (Peters et al., 2014; Mooij et al., 2016; Hastie et al., 2009).

5.4 Discussion and Implications

Our research makes substantial contributions to mutual information estimation and causal discovery by introducing methods that overcome previous limitations and offer key advancements. One major breakthrough lies in the integration of discrete and continuous variables. Unlike prior approaches that analyzed these variables separately, our methodology successfully combines them, leading to a more comprehensive understanding of complex systems, which is particularly relevant in contexts such as biological and social science datasets (Edwards et al., 2010). Additionally, our methods address scalability issues, which have often posed challenges to existing techniques. We demonstrate consistent performance across datasets of different sizes, which is crucial for real-world applications where the size and complexity of data can vary significantly (Suzuki, 2017).

Another key advancement is our use of forest-based models rather than traditional spanning trees to represent variable independence. This approach allows for a more nuanced depiction of the structure of the data, detecting complex interdependencies that are often overlooked by simpler models. This provides deeper insights into intricate relationships within the dataset. Moreover, our research advances the field of non-linear causal discovery, improving the ability to accurately identify causal relationships in complex systems characterized by non-linear interactions. This enhancement addresses a significant gap in existing methods (Peters et al., 2014; Mooij et al., 2016; Hastie et al., 2009) and opens new opportunities for understanding sophisticated phenomena across diverse scientific domains.

Furthermore, in cases of non-linear causal inference, the evaluation metrics, including accuracy, F1 score, mean squared error (MSE), and structural precision as measured by Structural Hamming Distance (SHD), consistently demonstrate the superior performance of our proposed method compared to LiNGAM. This improvement is particularly noticeable in scenarios involving complex non-linear relationships. The ability of our method to balance multiple aspects of model performance—such as minimizing residual errors, reducing false discoveries, and accurately reconstructing causal structures—has broad implications across several domains. In biomedical research, for example, the method's capacity to capture intricate non-linear interactions between genetic, environmental, and clinical factors could support the development of personalized treatment strategies and contribute to a deeper understanding of disease mechanisms (Peters et al., 2014; Hastie et al., 2009). In economics and the social sciences, where systems are often driven by mixed-type variables and complex dependencies, the enhanced accuracy and structural fidelity of our approach could lead to more effective modeling of economic behavior and social networks, which in turn could improve the design and evaluation of policies and interventions (Mooij et al., 2016; Peters et al., 2017). In climate science, the method's ability to model non-linear dynamics within climate systems offers potential improvements in the accuracy of climate models and long-term forecasting (Zhang and Hyvärinen, 2012). Finally, in neuroscience, this approach holds promise for advancing our understanding of brain connectivity, particularly by revealing the complex, non-linear interactions between brain regions and their connections to cognitive functions and neurological disorders (Chickering, 2002; Peters et al., 2017).

5.5 Summary

Our proposed methods for mutual information estimation involving mixed-type variables and for learning non-linear causal structures represent significant advancements within their respective areas of research. By effectively addressing challenges such as the integration of discrete and continuous variables as well as the identification of non-linear relationships, these methods become powerful tools for uncovering intricate structures in diverse real-world contexts.

Our non-linear causal discovery approach has demonstrated consistent superiority over LiNGAM in various metrics, indicating its robustness in identifying non-linear causal relationships. Although the increased computational cost presents a trade-off, the resulting gains in accuracy and the enhanced structural representation often justify this additional effort.

These encouraging results provide a solid foundation for future advancements in mutual information estimation and non-linear causal discovery. By offering more precise and comprehensive tools for analyzing complex systems, these methods have the potential to foster new insights across multiple scientific disciplines, including genomics, neuroscience, economics, and climate science.

Chapter 6

Conclusions and Future Works

6.1 Conclusion

This thesis introduces two key advancements in the fields of mutual information estimation and non-linear causal discovery, addressing fundamental challenges in analyzing complex, real-world datasets and determining causal relationships within non-linear systems. The first major contribution is the development of a robust mutual information estimator designed to work with hybrid datasets containing both discrete and continuous variables. By incorporating the Watanabe Bayesian Information Criterion (WBIC) and copula density estimation, this estimator effectively calculates joint densities and free energies, surpassing traditional limitations associated with data type constraints in mutual information calculations. Enhancements were made to the Chow-Liu algorithm, enabling it to construct multiple trees, thus providing a more nuanced representation of dependencies between variables. The practical success of this approach was demonstrated in genomics, where it efficiently uncovered gene expression-SNP relationships without the need to separate discrete from Gaussian components. This integrative method simplified the analysis and interpretation of complex genetic data, enriching the insights obtained and making significant strides in the domain of mutual information estimation for mixed-type datasets.

The second contribution of this thesis is a novel approach to non-linear causal discovery, integrating Generalized Additive Models (GAMs) with the Hilbert-Schmidt Independence Criterion (HSIC) to extend the applicability of the LiNGAM framework into non-linear contexts. This method tackles the problem of estimating additive noise models without requiring prior knowledge of the underlying non-linear functions, thus maintaining interpretability while accommodating a wide range of non-linear relationships. The use of GAMs provides flexibility, while HSIC enhances the capacity to detect complex statistical dependencies that are often missed by correlation-based techniques. Furthermore, the method assumes non-Gaussian noise, which is essential for correctly identifying causal directions in cases where Gaussian noise would otherwise hinder identifiability. Theoretical consistency in determining causal order, combined with superior experimental results across metrics such as accuracy, F1 score, mean squared error, and structural Hamming distance, makes this approach a valuable tool for researchers studying complex, non-linear systems in fields such as economics, biology, neuroscience, and climate science.

While the methods developed in this thesis represent significant advancements, several promising directions for future research remain. One potential area for exploration is improving the computational efficiency of both methods, especially for non-linear causal discovery, by focusing on algorithmic optimizations or leveraging advanced computational techniques to reduce runtime while retaining accuracy. The scalability of these methods to very large datasets should also be investigated, as real-world data continues to grow in size and complexity. This could involve developing distributed or parallel processing versions of the algorithms to handle larger datasets effectively. Another promising avenue is to explore how the integration of these techniques with deep learning models might enhance their performance, particularly when dealing with high-dimensional data or complex, non-linear relationships. Expanding the application of these methods beyond genomics to other domains, such as climate science, social networks, and financial markets, would help validate their versatility and generalizability.

6.2 Future research directions

Further research could also focus on extending the proposed methods to handle time series data, thereby capturing temporal causal relationships and expanding their applicability. Developing robust approaches for quantifying uncertainty in mutual information estimates and causal structures would also be beneficial, enhancing the interpretability and reliability of the findings. Addressing challenges related to latent confounders is another potential research direction, where extending the non-linear causal discovery method to account for hidden variables would greatly enhance its utility in real-world scenarios. Finally, further theoretical investigation into the identifiability conditions and convergence properties of the proposed methods would provide deeper insights, ensuring robustness and guiding future refinements. These future directions provide a foundation for advancing the contributions of this thesis, leading to the development of more powerful, efficient, and versatile tools for understanding complex systems and discovering causal relationships across a broad spectrum of real-world applications.

Appendix A

Appendix

A.1 Proof of Consistency of GAM Estimation

Lemma A.1 (Consistency of GAM Estimation)**.** *Let $\mathcal{X}$ be the domain of the input variables, $f : \mathcal{X} \to \mathbb{R}$ be the true underlying function, and $\hat{f}_n : \mathcal{X} \to \mathbb{R}$ be the GAM estimator based on n samples. Under suitable regularity conditions, the GAM estimators are consistent, that is, as the sample size $n \to \infty$, the estimated functions converge to the true functions in probability:*

$$\sup_{x \in \mathcal{X}} |\hat{f}_n(x) - f(x)| \xrightarrow{p} 0 \tag{A.1}$$

where $|\cdot|$ denotes the absolute value.

Proof. We consider the Generalized Additive Model (GAM) for each response variable X_i, modeled as:

$$X_i = \beta_{0i} + \sum_{j \neq i} f_{ij}(X_j) + \varepsilon_i, \quad i = 1, \ldots, p$$

where:

- β_{0i} is the intercept term.

- $f_{ij}(X_j)$ are unknown smooth functions of the predictors X_j, $j \neq i$.
- ε_i are error terms with $E[\varepsilon_i] = 0$ and $\text{Var}(\varepsilon_i) = \sigma_i^2 < \infty$.

Assumptions:

(A1) **Smoothness of True Functions:** Each true function f_{ij} belongs to a Sobolev space $W_2^m(\mathcal{X})$ of order $m \geq 2$, implying that f_{ij} has square-integrable derivatives up to order m.

(A2) **Choice of Smoothing Parameters:** The smoothing parameters satisfy $\lambda_{ij} \to 0$ and $n\lambda_{ij} \to \infty$ as $n \to \infty$.

(A3) **Design Density:** The predictor variables X_j have a joint density $p_X(x)$ that is bounded away from zero and infinity on a compact support $\mathcal{X} \subset \mathbb{R}^p$.

(A4) **Error Terms:** The error terms ε_i are independent and identically distributed (i.i.d.) with $E[\varepsilon_i] = 0$ and $\text{Var}(\varepsilon_i) = \sigma_i^2 < \infty$.

Our goal is to show that under these assumptions, the estimators $\hat{f}_{ij}$ converge uniformly in probability to the true functions f_{ij} as $n \to \infty$.

Step 1: Approximation by Finite Basis Expansion

We represent each function f_{ij} using a finite basis expansion:

$$f_{ij}(x) = \sum_{l=1}^{M_{ij}} \theta_{ijl}\phi_{ijl}(x)$$

where:

- $\{\phi_{ijl}(x)\}_{l=1}^{M_{ij}}$ are known basis functions (e.g., B-splines or truncated power series) forming a basis for $W_2^m(\mathcal{X})$.
- M_{ij} is the number of basis functions, which may increase with n.
- $\theta_{ij} = (\theta_{ijl})_{l=1}^{M_{ij}}$ are coefficients to be estimated.

This representation allows us to approximate f_{ij} to arbitrary accuracy as $M_{ij} \to \infty$.

Step 2: Penalized Least Squares Estimation

We estimate the coefficients θ_{ij} by minimizing the penalized least squares criterion:

$$\mathrm{PLS}_n(\theta_i) = \frac{1}{n}\sum_{k=1}^{n}\left(X_i^{(k)} - \beta_{0i} - \sum_{j\neq i}\sum_{l=1}^{M_{ij}} \theta_{ijl}\phi_{ijl}(X_j^{(k)})\right)^2 + \sum_{j\neq i}\lambda_{ij}\theta_{ij}^\top \mathbf{K}_{ij}\theta_{ij}$$

where:

- $\mathbf{K}_{ij}$ is a positive semi-definite matrix corresponding to the roughness penalty, typically derived from the integral of the squared second derivative of f_{ij}:

$$\theta_{ij}^\top \mathbf{K}_{ij}\theta_{ij} = \int \left(f_{ij}''(x)\right)^2 dx$$

Step 3: Consistency of Penalized Least Squares Estimators

Under assumptions (A1)-(A4), and by applying results from nonparametric regression theory (Newey, 1997; Huang, 2003), we have the following:

- The estimation error of $\hat{\theta}_{ij}$ satisfies:

$$\|\hat{\theta}_{ij} - \theta_{ij}^*\| = O_P\left(\left(\frac{M_{ij}}{n}\right)^{1/2} + \lambda_{ij}\right)$$

where θ_{ij}^* are the true coefficients of f_{ij} in the basis expansion.

- The approximation error due to using a finite basis is $O(M_{ij}^{-m})$, because the Sobolev space W_2^m functions can be approximated with error $O(M_{ij}^{-m})$ (Schumaker, 2007)).

Step 4: Uniform Convergence of Estimated Functions

Combining the estimation error and approximation error, we have:

$$\sup_{x\in\mathcal{X}}\left|\hat{f}_{ij}(x) - f_{ij}(x)\right| = O_P\left(\left(\frac{M_{ij}^2}{n}\right)^{1/2} + \lambda_{ij}M_{ij}^{1/2}\right)$$

Similarly, the approximation error is:

$$\sup_{x \in \mathcal{X}} \left| f_{ij}^{M_{ij}}(x) - f_{ij}(x) \right| = O(M_{ij}^{-m})$$

Step 5: Choice of M_{ij} and λ_{ij}

To balance the estimation and approximation errors, we select M_{ij} and λ_{ij} such that:

$$\left(\frac{M_{ij}^2}{n} \right)^{1/2} \approx M_{ij}^{-m}, \quad \lambda_{ij} M_{ij}^{1/2} \approx M_{ij}^{-m}$$

Let us choose $M_{ij} \propto n^{1/(2m+1)}$. Then:

$$\left(\frac{M_{ij}^2}{n} \right)^{1/2} = O\left(n^{-m/(2m+1)} \right)$$

$$M_{ij}^{-m} = O\left(n^{-m/(2m+1)} \right)$$

Similarly, choose $\lambda_{ij} \propto n^{-m/(2m+1)} M_{ij}^{-1/2}$.

Step 6: Total Error and Convergence

Combining the errors:

$$\sup_{x \in \mathcal{X}} \left| \hat{f}_{ij}(x) - f_{ij}(x) \right| = O_P\left(n^{-m/(2m+1)} \right)$$

As $n \to \infty$, $n^{-m/(2m+1)} \to 0$, so:

$$\sup_{x \in \mathcal{X}} \left| \hat{f}_{ij}(x) - f_{ij}(x) \right| \xrightarrow{P} 0$$

Step 7: Conclusion

By applying this result to each function f_{ij}, we have shown that under the regularity conditions (A1)-(A4), the estimated functions $\hat{f}_i(x)$ for the generalized additive model are consistent. Specifically:

$$\sup_{x\in\mathcal{X}}\left|\hat{f}_i(x)-f_i(x)\right|\leq\sum_{j\neq i}\sup_{x\in\mathcal{X}}\left|\hat{f}_{ij}(x)-f_{ij}(x)\right|\xrightarrow{P}0\quad\text{as}\quad n\to\infty$$

This completes the proof of the consistency of the generalized additive model estimators. □

Proposition A.1 (HSIC and Independence). *For characteristic kernels k and l, HSIC$(X,Y)=0$ if and only if X and Y are independent (Gretton et al., 2005).*

Proof. We will prove this proposition following the approach of Gretton et al. (2005):

Step 1: Define HSIC in terms of cross-covariance operators.

Let $\mathcal{F}$ and $\mathcal{G}$ be Reproducing Kernel Hilbert Spaces (RKHS) associated with kernels k and l, respectively. Let $\phi:\mathcal{X}\to\mathcal{F}$ and $\psi:\mathcal{Y}\to\mathcal{G}$ be the feature maps corresponding to k and l.

The mean elements in $\mathcal{F}$ and $\mathcal{G}$ are defined as:

$$\mu_X=\mathbb{E}_X[\phi(X)],\quad\mu_Y=\mathbb{E}_Y[\psi(Y)]$$

The cross-covariance operator $C_{XY}:\mathcal{G}\to\mathcal{F}$ is defined by:

$$C_{XY}=\mathbb{E}_{XY}[(\phi(X)-\mu_X)\otimes(\psi(Y)-\mu_Y)]$$

where $\otimes$ denotes the tensor product.

The Hilbert-Schmidt Independence Criterion (HSIC) is defined as the squared Hilbert-Schmidt norm of C_{XY}:

$$\text{HSIC}(X,Y)=\|C_{XY}\|_{\text{HS}}^2$$

Step 2: Show that HSIC is zero if and only if the cross-covariance operator is zero.

By the properties of the Hilbert-Schmidt norm, we have:

$$\|C_{XY}\|_{\text{HS}}^2 = 0 \quad \iff \quad C_{XY} = 0$$

Therefore:

$$\text{HSIC}(X, Y) = 0 \quad \iff \quad C_{XY} = 0$$

Step 3: Prove that a zero cross-covariance operator implies independence.

Assume that $C_{XY} = 0$. We need to show that X and Y are independent.

For any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, consider the covariance between $f(X)$ and $g(Y)$:

$$\begin{aligned}\text{Cov}(f(X), g(Y)) &= \mathbb{E}_{XY}[(f(X) - \mathbb{E}[f(X)])(g(Y) - \mathbb{E}[g(Y)])] \\ &= \mathbb{E}_{XY}[\langle f, \phi(X) - \mu_X\rangle_{\mathcal{F}}\langle \psi(Y) - \mu_Y, g\rangle_{\mathcal{G}}] \\ &= \langle f, C_{XY} g\rangle_{\mathcal{F}}\end{aligned}$$

Since $C_{XY} = 0$, it follows that:

$$\text{Cov}(f(X), g(Y)) = 0$$

for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$.

Because k and l are characteristic kernels, their corresponding RKHS $\mathcal{F}$ and $\mathcal{G}$ are dense in the space of continuous functions vanishing at infinity on $\mathcal{X}$ and $\mathcal{Y}$, respectively (see Fukumizu et al., 2008).

This implies that for all bounded continuous functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$:

$$\text{Cov}(f(X), g(Y)) = 0$$

Therefore, all such functions $f(X)$ and $g(Y)$ are uncorrelated.

However, uncorrelatedness of all bounded continuous functions implies independence of X and Y, because any joint distribution where all bounded continuous functions are uncorrelated must be the product of the marginals.

Conversely, assume that X and Y are independent.

Then:

$$\mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)] = \mathbb{E}_X[\phi(X)] \otimes \mathbb{E}_Y[\psi(Y)]$$

Therefore:

$$\begin{aligned} C_{XY} &= \mathbb{E}_{XY}[(\phi(X) - \mu_X) \otimes (\psi(Y) - \mu_Y)] \\ &= (\mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)]) - \mu_X \otimes \mu_Y \\ &\quad - \mu_X \otimes \mu_Y + \mu_X \otimes \mu_Y \\ &= (\mu_X \otimes \mu_Y) - \mu_X \otimes \mu_Y \\ &= 0 \end{aligned}$$

Thus, $\text{HSIC}(X, Y) = \|C_{XY}\|^2_{\text{HS}} = 0$.

Conclusion:

We have shown that:

$$\text{HSIC}(X, Y) = 0 \quad \Longleftrightarrow \quad X \perp Y$$

when using characteristic kernels k and l. This completes the proof.

□

A.2 Proof of Theorem

Theorem A.1 (Consistency of Proposed Method)**.** *Under the assumptions of Lemma 1 and Proposition 1, and assuming that the true data-generating process follows a non-linear additive noise model with non-Gaussian noise terms, the proposed method consistently recovers the true causal structure, including both the presence of causal relationships and their directions, as the sample size approaches infinity.*

Proof. To prove the consistency of our method, we will show that under suitable assumptions, the probability that our method correctly identifies the true causal graph approaches 1 as $n \to \infty$.

Assumptions:

(A1) **Data Generating Process:** The true model is a non-linear additive noise model (ANM):

$$X_i = f_i(\mathrm{PA}_i) + \varepsilon_i, \quad i = 1, \ldots, p$$

where:

- PA_i denotes the set of parent variables of X_i in the true causal graph G.
- f_i are unknown smooth, non-linear functions belonging to a Sobolev space $W_2^m(\mathcal{X})$ with $m \geq 2$.
- ε_i are mutually independent, non-Gaussian noise terms with $E[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_i^2 < \infty$.
- The errors ε_i are independent of their predictors PA_i.

(A2) **Observational Data:** We have an i.i.d. sample $\{X^{(k)} = (X_1^{(k)}, \ldots, X_p^{(k)})\}_{k=1}^n$ from the joint distribution of $(X_1, \ldots, X_p)$.

(A3) **Design Density:** The covariates X_j have a joint density $p_X(x)$ that is bounded away from zero and infinity on a compact support $\mathcal{X} \subset \mathbb{R}^p$.

(A4) **Smoothness Penalty Parameters:** The smoothing parameters λ_{ij} satisfy $\lambda_{ij} \to 0$ and $n\lambda_{ij} \to \infty$ as $n \to \infty$.

(A5) **Kernel Functions:** The kernels used in HSIC are characteristic, continuous, and bounded.

Our goal is to demonstrate that, under these assumptions, the estimated causal graph $\hat{G}$ converges to the true causal graph G with probability approaching 1 as $n \to \infty$.

Proof Outline:

1. **Consistency of Function Estimation:** Show that the estimated functions $\hat{f}_i$ converge uniformly to the true functions f_i.

2. **Convergence of Residuals:** Establish that the residuals $\hat{\varepsilon}_i = X_i - \hat{f}_i(\hat{\mathrm{PA}}_i)$ converge in distribution to the true noise terms ε_i.

3. **Independence Testing Using HSIC:** Use the HSIC to test for independence between residuals and predictors, correctly identifying parent and non-parent variables.

4. **Recovery of Causal Structure:** Combine the results to show that the estimated parent sets $\hat{\mathrm{PA}}_i$ converge to the true parent sets PA_i, thus recovering the true causal structure.

Detailed Proof:

Step 1: Consistency of Function Estimation

By Lemma A.1, under assumptions (A1)-(A4), the estimated functions $\hat{f}_i$ obtained from the penalized least squares criterion converge uniformly in probability to the true functions f_i:

$$\sup_{x \in \mathcal{X}} \left| \hat{f}_i(x) - f_i(x) \right| \xrightarrow{P} 0, \quad \forall i = 1, \ldots, p.$$

Step 2: Convergence of Residuals

Define the residuals:

$$\hat{\varepsilon}_i^{(k)} = X_i^{(k)} - \hat{f}_i(\hat{\mathrm{PA}}_i^{(k)}), \quad k = 1, \ldots, n.$$

Expressing $\hat{\varepsilon}_i^{(k)}$ in terms of the true model:

$$\hat{\varepsilon}_i^{(k)} = \varepsilon_i^{(k)} + \left[f_i(\mathrm{PA}_i^{(k)}) - \hat{f}_i(\hat{\mathrm{PA}}_i^{(k)}) \right].$$

Let $\delta_i^{(k)} = f_i(\mathrm{PA}_i^{(k)}) - \hat{f}_i(\hat{\mathrm{PA}}_i^{(k)})$.

Since $\hat{f}_i$ converges to f_i uniformly and $\hat{\text{PA}}_i$ will be shown to converge to PA_i, we have $\delta_i^{(k)} \xrightarrow{P} 0$.

Thus:

$$\hat{\varepsilon}_i^{(k)} = \varepsilon_i^{(k)} + \delta_i^{(k)} \xrightarrow{d} \varepsilon_i^{(k)}.$$

Step 3: Independence Testing Using HSIC

We employ Proposition A.1 which states that for characteristic kernels:

$$\text{HSIC}(X, Y) = 0 \quad \text{if and only if} \quad X \perp Y.$$

Testing Procedure:

For each variable X_i and each potential predictor X_j $(j \neq i)$:

1. **Fit a GAM Excluding X_j:** Estimate $\hat{f}_i$ without including X_j as a predictor.

2. **Compute Residuals:** Calculate $\hat{\varepsilon}_i^{(k)} = X_i^{(k)} - \hat{f}_i(\hat{\text{PA}}_i^{(k)})$.

3. **Compute HSIC:** Compute $\text{HSIC}_n(\hat{\varepsilon}_i, X_j)$ using the empirical estimate:

   $$\text{HSIC}_n(\hat{\varepsilon}_i, X_j) = \frac{1}{(n-1)^2} \operatorname{tr}(\mathbf{KHLH}),$$

   where $\mathbf{K}$ and $\mathbf{L}$ are kernel matrices for $\hat{\varepsilon}_i$ and X_j, and $\mathbf{H}$ is the centering matrix.

4. **Hypothesis Testing:** Test $H_0 : \hat{\varepsilon}_i \perp X_j$ versus $H_1 : \hat{\varepsilon}_i \not\perp X_j$.

5. **Decision Rule:** Reject H_0 if $\text{HSIC}_n(\hat{\varepsilon}_i, X_j) > \tau_n$, where τ_n is determined based on the asymptotic distribution under H_0 and the desired significance level α.

Justification:

- In **Case 1**, when $X_j \notin \text{PA}_i$, the noise term ε_i is independent of X_j. As a result, the estimated residuals $\hat{\varepsilon}_i$ converge in distribution to the true residuals, $\hat{\varepsilon}_i \xrightarrow{d}$

ε_i, meaning that asymptotically, $\hat{\varepsilon}_i$ becomes independent of X_j. Therefore, as the sample size increases, the Hilbert-Schmidt Independence Criterion (HSIC) between $\hat{\varepsilon}_i$ and X_j, denoted as $\mathrm{HSIC}_n(\hat{\varepsilon}_i, X_j)$, converges in probability to zero, $\mathrm{HSIC}_n(\hat{\varepsilon}_i, X_j) \xrightarrow{P} 0$.

- In **Case 2**, when $X_j \in \mathrm{PA}_i$, excluding X_j from the model results in misspecification. This misspecification leads to residuals $\hat{\varepsilon}_i$ that retain some dependence on X_j, meaning that $\hat{\varepsilon}_i \not\perp X_j$. Consequently, the HSIC value between $\hat{\varepsilon}_i$ and X_j, $\mathrm{HSIC}_n(\hat{\varepsilon}_i, X_j)$, converges in probability to a positive constant, $\mathrm{HSIC}_n(\hat{\varepsilon}_i, X_j) \xrightarrow{P} c > 0$.

Consistency of the Test:

The test is consistent because, under the null hypothesis H_0, which states that $\hat{\varepsilon}_i$ and X_j are independent, $\mathrm{HSIC}_n \xrightarrow{P} 0$. On the other hand, under the alternative hypothesis H_1, which posits dependence between $\hat{\varepsilon}_i$ and X_j, the HSIC statistic converges in probability to a positive value, $\mathrm{HSIC}_n \xrightarrow{P} c > 0$. As the sample size n tends to infinity, the test reliably distinguishes between independence and dependence with a probability approaching 1.

Step 4: Recovery of Causal Structure

By applying the testing procedure for all pairs (X_i, X_j), we construct estimated parent sets:

$$\hat{\mathrm{PA}}_i = \{X_j : \text{Reject } H_0 \text{ between } \hat{\varepsilon}_i \text{ and } X_j\}.$$

Since the tests are consistent, we have:

$$P(\hat{\mathrm{PA}}_i = \mathrm{PA}_i) \to 1 \quad \text{as } n \to \infty.$$

Conclusion:

The estimated causal graph $\hat{G}$, constructed from the estimated parent sets $\hat{\mathrm{PA}}_i$, converges to the true causal graph G with probability approaching 1:

$$P(\hat{G} = G) \to 1 \quad \text{as } n \to \infty.$$

Therefore, the proposed method is consistent.

□

List of Publications

- Md. Ashraful Islam, Joe Suzuki, ”Forest construction of Gaussian and discrete variables with the application of Watanabe Bayesian Information Criterion” *Behaviormetrika*, 2024. Available: https://link.springer.com/article/10.1007/s41237-024-00227-4
- Md. Ashraful Islam, Joe Suzuki, ”Exploring Non-linear Causal Inference in Observational Data.” (2024+), *Under Review.*

Acknowledgements

First and foremost, I am deeply grateful to my advisor, **Professor Joe Suzuki**, for his unwavering support, insightful guidance, and generous assistance throughout my PhD journey. His thoughtfulness, positivity, and constant encouragement have significantly impacted my academic growth and personal development. I am genuinely fortunate to have had such an exceptional mentor, whose support has made this journey not only enriching but also fulfilling and enjoyable.

I also want to extend my heartfelt appreciation to my friends and colleagues. Since moving to Japan, I have been fortunate to receive support from friends back home and to build lasting friendships within joe-lab. I am also thankful to the administrative staff, whose assistance has made my life easier during my studies.

Finally, I want to express my deepest gratitude to my family for their boundless love and unwavering support. To my beloved wife, **Sumaiya Afrin Smriti**, thank you for your patience and steadfast support during my PhD journey.

Bibliography

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198.

Barnett, W. A., Serletis, A., and Serletis, D. (2015). Nonlinear and complex dynamics in economics. *Macroeconomic Dynamics*, 19(8):1749–1779.

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054.

Bender, E. A. and Williamson, S. G. (2010). *Lists, decisions and graphs*. S. Gill Williamson.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Chickering, D. M., Heckerman, D., and Meek, C. (2004). Large-sample learning of bayesian networks is np-hard. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 124–133.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403.

Edwards, D., De Abreu, G. C., and Labouriau, R. (2010). Selecting high-dimensional mixed graphical models using minimal aic or bic forests. *BMC Bioinformatics*, 11(1):1–13.

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302.

Gamazon, E. R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E. O., Nicolae, D. L., Dolan, M. E., and Cox, N. J. (2010). Scan: Snp and copy number annotation. *Bioinformatics*, 26(2):259–262.

Gelman, A. (2011). Induction and deduction in bayesian data analysis. *Statistical Science*.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the Algorithmic Learning Theory Conference*, pages 63–77. Springer.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. *Advances in neural information processing systems*, 20.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models.* Chapman and Hall/CRC.

Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N. (2018). Causal inference in systems biology: a synthesis of computational approaches. *Bioinformatics*, 34(14):2477–2485.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21.

Hoyle, R. H. (2012). *Handbook of structural equation modeling.* Guilford press.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635.

Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R., and MD, M. (1988). Heart disease, uci machine learning repository, 1988.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636.

Kojadinovic, I. and Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula r package. *Journal of Statistical Software*, 34:1–20.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J. (2008). Tipping elements in the earth's climate system. *Proceedings of the national Academy of Sciences*, 105(6):1786–1793.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan.* CRC Press.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., et al. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences*, 102(38):13550–13555.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168.

Pearl, J. (2000). *Models, reasoning and inference*. Cambridge, UK: CambridgeUniversityPress.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053.

Schmidt, T. (2007). Coping with copulas. *Copulas: From Theory to Application in Finance*, 3:1–34.

Schumaker, L. L. (2007). *Spline functions: Basic theory*. Cambridge University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search.* MIT press.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, 338(6106):496–500.

Suzuki, J. (1993). A construction of bayesian networks from databases based on an mdl principle. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 266–273. Elsevier.

Suzuki, J. (2012). The bayesian chow-liu algorithm. In *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models*, pages 315–322. Springer.

Suzuki, J. (2015). Consistency of learning bayesian network structures with continuous variables: an information theoretic approach. *Entropy*, 17(8):5752–5770.

Suzuki, J. (2017). A novel chow–liu algorithm and its application to gene differential analysis. *International Journal of Approximate Reasoning*, 80:1–18.

Suzuki, J. (2020). *Statistical Learning with Math and R.* Springer.

Suzuki, J. (2021). *Sparse Estimation with Math and R: 100 Exercises for Building Logic.* Springer Nature.

Suzuki, J. (2023). *WAIC and WBIC with R Stan: 100 Exercises for Building Logic.* Springer Nature.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(27):867–897.

Watanabe, S. (2021). Waic and wbic for mixture models. *Behaviormetrika*, 48(1):5–21.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655.

Zhang, K. and Hyvärinen, A. (2012). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). On the kernel choice for unsupervised learning of causal relations. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 793–800.