Ph.D. Thesis

Toward Discovering Causal Relations from Manufacturing Data: Heteroscedasticity and Variable Groups

(Date of conferring)
January 2024

Name Genta Kikuchi

supervisor Shohei Shimizu
supervisor Hidetoshi Matsui

Graduate School of Data Science
Department of Data Science
(Doctor Course)
SHIGA UNIVERSITY

To my family

Abstract

Discovering causal relationships between quantities of interest is fundamental in many scientific disciplines. This thesis focuses on the field of manufacturing, where data-driven quality improvements are attracting increasing attention because of the more diverse data accumulated in the wake of Industry 4.0 and digital transformation. Understanding the causal relations among the various measurements, such as those of product qualities, machine parameters, and manufacturing environment, is crucial for data-driven quality improvement activities.

Although controlled experiments are the recommended approach to infer cause–effect relations, such experiments can be unethical, technically challenging, or too expensive. For example, manufacturing a set of defective products during mass production is unrealistic, as it decreases overall equipment effectiveness and might affect subsequent products. Numerous methods have been developed to estimate causal relationships from observational data, termed causal discovery, to tackle this issue.

Research that applies causal discovery methods to manufacturing data assumes that the data exhibit non-linearity, temporal dependencies, or both. However, they overlook a typical characteristic of manufacturing data, heteroscedasticity, which causes severe problems with many existing causal discovery methods. Another issue is handling groups of variables; when multiple measurements take similar values, selecting one of them or aggregating them by taking an average may impede the estimation performance. Several existing works on causal discovery address the aforementioned issues individually but not simultaneously.

This thesis addresses the problem of performing causal discovery on non-linear time-series data with heteroscedastic noise. We introduce an estimation method based on recently developed continuous optimization-based methods. Then, we extend the work to exploit the time structure and show that causal relationships can be uniquely recovered from data under specific assumptions. Furthermore, this thesis considers the problem of estimating causal relationships among multiple groups of variables where the functional relations are beyond linear. We propose a novel approach based on algebraic characterization of causal structure among multiple groups of variables that can be used as a constraint for the optimization problem on existing continuous optimization-based methods.

Acknowledgements

First and foremost, I would like to thank my supervisor, Shohei Shimizu, who has always been supportive and understanding. This thesis would not be possible without the freedom allowed, considering my full-time job and the guidance provided when I was facing difficulties. I would also like to thank my semi-supervisor, Hidetoshi Matsui, for providing insightful feedback on my research during my Ph.D. course and committee members Taichi Okumura and Tsuyoshi Esaki for their valuable comments on this manuscript.

Many thanks to the members of the DENSO F-IoT data analytics team. First, I would like to thank Mutsumi Yoshino, who encouraged me for this opportunity. Also, thanks to Toshikuni Shinohara, who supported me with fruitful advice based on his strong knowledge of both manufacturing and data analytics. Thanks as well to Tatsunori Kojo, Sho Takahashi, and Takero Arakawa for supporting me when I was too busy conducting my research. I would not have been able to complete this thesis without the exceptional freedom they allowed.

I am grateful to the members of the seminar of the causal discovery lab, who listened to my presentation and gave me many comments and questions about my research. A big thanks goes to Ibuki Hoshina for guiding me with fundamental ideas, such as how to think as a researcher and pursue my research, which encouraged me in the correct direction.

Lastly and most importantly, I would like to thank my whole family for supporting me during this long journey. Special thanks to my wife Satomi, my son Masaki, and my daughter Rika for being there with love and smiles the whole time, which was the most critical driving force for me to complete this thesis.

Table of contents

List of figures

List of tables

Nomenclature

Abbreviations

ANM	Additive Noise Model
CNN	Convolutional Neural Network
DAG	Directed Acyclic Graph
DoE	Design of Experiments
FCI	Fast Causal Inference
FCM	Functional Causal Model
GRCI	Generalized Root Causal Inference
HNM	Heteroscedastic Noise Model
IFMOC	Identifiable Functional Model Class
LiNGAM	Linear Non-Gaussian Acyclic Model
LSNM	Location-Scale Noise Model
MAD	Mean Absolute Deviation
PC	Peter-Clark
RCT	Randomized Controlled Trials
RESIT	Regression with a subsequent independence test
SVAR	Structural Vector Autoregressive Models
TiMINo	Time-series Model with Independent Noise

TS-LSNM	Time-Series Location-Scale Noise Models

Symbols

$\boldsymbol{E}$	Set of edges in a graph
$\boldsymbol{V}$	Set of vertices in a graph
$\boldsymbol{X}$	Set of variables
$\mathrm{E}[X], \mu_X$	Expectation of X
$\perp\!\!\!\perp$	Statistical independence
$\mathbf{x}$	Multidimensional random variable
$\mathscr{G}$	Graph
$\mathscr{N}$	Gaussian distribution
$\mathrm{PA_i}$	Set of parents of vertex V_i for variable X_i
PA_j^t	Set of parents of X_j^t at time t
$\rho_{X,Y}$	(Pearson) correlation coefficient of X and Y
Σ	Covariance matrix
$P(X)$	Probability distribution of X
cov	Covariance
$\mathrm{pa}(j)$	Index set of parents of a variable X_j
U	Uniform distribution
$\mathrm{Var}[X], \sigma_X^2$	Variance of X
B	Binary adjacency matrix
$B^{'}$	Group adjacency matrix
K	Set of index sets for each group of variables
L	Maximum time lag of time-series
M	Number of groups

N	Number of observations
N_j	Noise term corresponding to variable X_j
P	Number of variables
S	Number of convolutional kernels
W	Weighted adjacency matrix
$W^{'}$	Weighted group adjacency matrix
X	Random variable

Chapter 1

Introduction

Identifying the underlying causal relations and laws that govern a phenomenon (e.g., human behavior, economics, and physical phenomenon) is critical in many scientific disciplines (Spirtes et al., 2000). For instance, in neuroscience, causality among brain recordings is extracted to understand the complex behavior of brain networks (Smith et al., 2011). In biomedical and health informatics, inferring causality may facilitate disease diagnosis and decision-making regarding clinical treatments (Mani and Cooper, 2000; Shen et al., 2020). In many cases, including business scenarios, the knowledge of causal relations helps develop policies to control a quantity of interest, such as process control (Li and Shi, 2007), climate action (Addo et al., 2021), and root-cause analysis (Budhathoki et al., 2022). Therefore, it is crucial to understand the cause of the outcome and how the outcome changes if the cause is controlled to a specific value.

Collecting *experimental data* by performing controlled experiments is widely used to obtain high-level evidence of causal relationships. Randomized controlled trials (RCT) are widely used to draw causal inferences (Rubin, 1974). For example, in clinical research, each patient is randomly assigned to receive a treatment (*intervention*), and the difference between the outcomes is measured. In the manufacturing domain, the design of experiments (DoE) is often performed (Fisher, 1936; Kirk, 2009). For example, possible factors related to an outcome are examined for each level of factors, and the factorial analysis of variance is performed to understand the effect of the factors on the outcome. However, performing controlled experiments is difficult in practice. The experiments may be unethical (e.g., forcing a person to smoke to measure the effect of smoking), technically difficult (e.g., too many variables to perform an experiment), or cost-consuming (e.g., producing a large number of test products to check the causal connection between machine parameter and product failure). Thus, developing computational methods to infer causal relationships from *observational data* and not experimental data is essential.

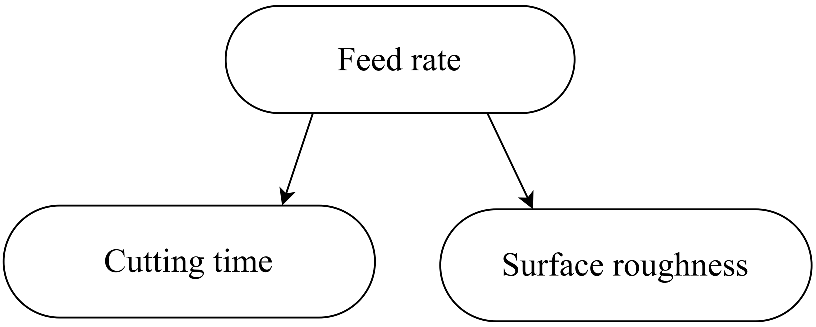


Fig. 1.1 Graph representing causal structure between variables of a cutting process. $x \rightarrow y$ represents a causal effect from x to y.

One method to infer causal relations is to leverage the supervised learning techniques of machine learning, a powerful tool for modeling the dependency between the outcome and other quantities (Mahesh, 2020). For example, machine learning models (e.g., LightGBM (Ke et al., 2017)) can be trained to accurately predict an outcome based on a set of explanatory variables that likely contain the cause of the outcome and then observe the importance of the variable or use model interpretation techniques, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017; Lundberg et al., 2020), to gain insights into the relation between explanatory variables and the outcome. However, they do not necessarily reflect the causal relationships as they ignore the causal data structure (Frye et al., 2020; Ma and Tourani, 2020). Furthermore, supervised learning techniques extract statistical dependencies such as correlations, which are insufficient to infer causal relationships (Pearl, 2009); hence, they fail to predict the outcome for the data generated after the intervention.

A causal structure among the variables of a cutting process in a manufacturing line is shown in Figure 1.1 as an example of when correlation is not equivalent to causation. Feed rate is the distance at which the cutting tool moves during one revolution, cutting time is the time taken to finish cutting, and surface roughness measures the unevenness of a surface. The feed rate is directly related to cutting time and surface roughness; hence, there is a correlation between the feed rate and the other two variables. Additionally, we observe a correlation between surface roughness and cutting time: if the surface roughness is high, the cutting time is small. This correlation is called *spurious* as it is due to the common cause between the variables, not by the direct causal effect (Pearl, 2009). However, intervening on the surface roughness (e.g., by changing the cutting tool) does not necessarily change the cutting time. Therefore, we need a deeper understanding of the underlying causal mechanism to control a quantity of interest.

This thesis investigates *causal discovery*, a research field in which the causal structure among variables is recovered from data (Spirtes et al., 2000). Particularly, we address how to learn the causal structure from *observational data*, which is substantially more accessible in various applications, as opposed to experimental data. The causal discovery algorithm inputs observed data (and possibly prior knowledge) and output a *graph* that represents the causal structure with variables as vertices and cause–effect relationship as arrows ($\rightarrow$) that can be used to estimate the causal effects. Though estimating causal structure from purely observational data seems unachievable, causal discovery techniques address this problem by making assumptions about the data-generating process.

One important approach to infer the causal structure among variables is the *constraint-based* approach, which makes and leverages assumptions regarding the conditional independence of the data and the underlying graph structure (Spirtes et al., 2000). Constraint-based methods include well-known techniques such as the Peter–Clark (PC) algorithm and fast causal inference (FCI) algorithm (Spirtes et al., 2000); these methods are widely applied to various data types and distributions such as mixed data types (Tsagris et al., 2018), cyclic relations (Strobl, 2019), and time-series (Runge et al., 2019). However, they cannot uniquely identify the causal structure as they fail to distinguish causal relationships with the same sets of conditional independencies. For instance, when distinguishing cause and effect in two variable cases where a conditional independence relationship is unavailable, the constraint-based approach cannot determine the causal direction.

Another line of research is the *score-based* approach that maximizes the fitness measure of a graph for the observed data (Chickering, 2002). Typical methods involve the greedy equivalence search (GES) (Chickering, 2002), which starts from an empty graph and recursively adds and removes edges until the scoring criterion converges. Criteria such as the Akaike information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz, 1978) are used for scoring a graph. However, one needs to assume the data distribution, for instance, the Gaussian distribution for continuous variables (Chickering, 2020). Despite additional assumptions for data distribution, GES is limited to identifying a set of graphs that induce the same sets of conditional independencies as the constraint approach.

Then, the question remains: how can we fully identify the underlying causal structure? It requires capturing the asymmetries between the observed data of cases where the cause generates the effect and vice versa. Therefore, we need a framework to mathematically model the data-generating process and computational methods to exploit the asymmetry and distinguish the cause and effect. The major framework to represent the data-generating process is the *structural equation modeling* (SEM) (Pearl, 2009). However, many classical methods cannot distinguish the cause and effect, as they often assume that the data follow the Gaussian distribution (Pearl, 2009; Spirtes et al., 2000).

Recently, the *functional causal model* (FCM) approach for causal discovery (Glymour et al., 2019) has been attracting considerable attention. FCMs leverage SEM to represent the data-generating process with additional assumptions to the functional class, such as non-linearity. Thus, it does not assume a specific form of the probability distribution, provided with an estimation procedure to uniquely identify the causal structure among variables. The typical model is the *linear non-Gaussian acyclic model* (LiNGAM), which assumes that the causal relationship is linear with no cyclic relations, and the data follow non-Gaussian distributions (Shimizu et al., 2006). Later, various models that relax the assumptions of LiNGAM were proposed offering data characteristics such as non-linearity (Hoyer et al., 2008a), cyclic relations (Lacerda et al., 2008), unmeasured variables (Hoyer et al., 2008b), temporal dependencies (Hyvärinen et al., 2010; Peters et al., 2013), measurement error (Zhang and Hyvärinen, 2009, 2010), and location-scale noise (Immer et al., 2023; Strobl and Lasko, 2023; Xu et al., 2022).

Estimation of FCMs is typically done by recursively applying regression and evaluating the independence between regression residuals and explanatory variables (Peters et al., 2014; Shimizu et al., 2011; Strobl and Lasko, 2023). This induces combinatorial search space, exponentially growing with the number of variables. Recently, the combinatorial optimization problem for searching for the causal structure was converted to a continuous optimization problem, termed NOTEARS (Zheng et al., 2018). NOTEARS was originally designed for data with linear relationships and later was extended to nonlinear (Zheng et al., 2020) and time-series data (Sun et al., 2021). However, the least squares loss used in the optimization is equivalent to assuming standard Gaussian noise (Cai et al., 2021; Kaiser and Sipos, 2021), which hinders the estimation when the assumption does not hold. The violation can easily occur by scaling the variables; this issue is referred to as *scale sensitivity*, which remains an open problem (Reisach et al., 2021).

This thesis concentrates on the FCM approach for causal discovery (Glymour et al., 2019). The major reason is the identifiability of FCMs, where one can fully identify the underlying causal structure from the observed data using an appropriate estimation method if the true data-generating process satisfies their assumptions. However, it raises the following question: Does the model sufficiently represent the characteristics of the real-world data? The answer depends on the data one is planning to apply.

1.1 Causal Discovery in Manufacturing: Difficulties

This thesis applies causal discovery to the data acquired from a manufacturing process. In manufacturing, large volumes of data are collected owing to Industry 4.0 (Rüßmann et al., 2015; Vaidya et al., 2018) and the digital transformation (Albukhitan, 2020; Stolterman and Fors, 2004). The accumulated data are used for quality improvement activities, such as quality analysis (what affects the quality?), quality monitoring (how can we reduce the variability of the quality?), and process control (how can we maintain quality?) (Köksal et al., 2011). These activities seek to improve the quality of products and can be enhanced using causal relations learned from the data (Marazopoulou et al., 2016; Vuković and Thalmann, 2022).

There are various applications of causal discovery methods in manufacturing, such as quality analysis (Marazopoulou et al., 2016), root-cause analysis (Landman et al., 2014; Wunderlich and Niggemann, 2017) and process control (Li and Shi, 2007). However, to the best of our knowledge, although these methods often consider the typical characteristics of the manufacturing data, such as temporal dependencies and non-linearity, they overlook *heteroscedasticity*. Here, heteroscedasticity indicates that another quantity modulates the variance of a quantity; this phenomenon is a fundamental factor when considering quality improvements. Well-recognized traditional methods for quality improvements such as Six Sigma (Brady and Allen, 2006), statistical process control (Montgomery, 2019), and the Taguchi method (Karna et al., 2012) seek to assess and reduce the variance of quality. Heteroscedasticity can also be induced by measurement error of sensors, from the uncertainty related to readings and temperature (meettechniek.info, 2013; Weschler Instruments, 2020). Nonetheless, existing research on applying causal discovery to manufacturing data does not consider heteroscedasticity, which hinders the estimation performance of existing causal discovery methods. Recently, FCM capable of heteroscedasticity has been proposed (Immer et al., 2023); however, the corresponding estimation method for more than two variables is not sufficiently studied.

Another concern is handling *groups* of variables, where multiple variables take similar values. For example, measurements from the same machine can exhibit high correlation (Marazopoulou et al., 2016); hence, it is reasonable to aggregate multiple variables in a group. Typical procedures involve selecting one variable per group (Marazopoulou et al., 2016) or calculating the sum or average of the measurements in the same group (Scheines and Spirtes, 2008). Despite the reduced computational time by these methods, the performance of causal discovery methods is hindered due to changes in the conditional dependencies of the data (Scheines and Spirtes, 2008) or the cancellation of dependence (Wahl et al., 2023). Therefore, computational methods to infer the causal relations between groups of variables are essential, though existing FCM-based methods only handle linear relationships (Entner and Hoyer, 2012; Kawahara et al., 2010).

1.2 Overview of the Thesis

This thesis seeks to develop a causal discovery method for nonlinear time-series data with heteroscedasticity, considering the issues mentioned earlier. Furthermore, it introduces a novel approach to estimating causal structure between groups of variables, which is capable of data beyond linear functional relations.

First, a continuous optimization-based estimation method for nonlinear data with heteroscedasticity is proposed. The method estimates the conditional variance and the conditional expectation of each variable using multilayer perceptrons (MLP) to exploit the heteroscedasticity. We leverage the approximation of the log probability during the optimization, which does not assume a specific probability distribution. This bridges the gap between the continuous optimization-based methods and FCMs, mitigating scale sensitivity as FCMs do not require assumptions of probability distributions. Next, we propose an FCM capable of capturing the temporal dependencies and showing that the model is identifiable. In addition, we propose an estimation method using the convolutional neural network (CNN). Moreover, a novel approach for estimating the causal structure among groups of variables based on the optimization constraints is presented. Finally, the proposed and conventional methods are compared using synthetic and real-world data from a ceramic substrate manufacturing line.

In summary, this thesis introduces an FCM and corresponding estimation method capable of generating typical characteristics of manufacturing data. This leads to more accurate causal discovery, accelerating the realization of data-driven quality improvement activities.

The remainder of this thesis is structured as follows. Chapter 2 provides the mathematical background necessary to understand the subsequent chapters. Chapter 3 introduces how causality is mathematically defined and how the causal relation is modeled. Common assumptions regarding causal modeling are also introduced. Chapter 4 describes existing FCMs relevant to this thesis and the causal discovery method for nonlinear time series data with heteroscedasticity, which is the first contribution of the original research articles. In Chapter 5, we present the problem setting and existing works of performing causal discovery on groups of variables and the second contribution of the original research articles, which can be applied to data beyond linear relations. After that, a result of a numerical experiment on real-world data collected from a manufacturing process is described in Chapter 6. Finally, Chapter 7 concludes the thesis.

1.3 Publications

This thesis is based on the research presented in the following articles. The contributions of Articles I and II are presented in Chapters 4 and 5. An overview of model classes and existing works introduced in Chapters 4 and 5, and the positioning of each article are shown in Figure 1.2.

Article I Genta Kikuchi. Differentiable Causal Discovery under Heteroscedastic Noise. In Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part I, pages 284–295. Springer, 2023.

Article II Genta Kikuchi and Shohei Shimizu. Structure Learning for Group of Variables with Nonlinear Time-Series Data with Location-Scale Noise, Causal Analysis Workshop Series 2023, to be published in the Proceedings of Machine Learning Research Volume 223.

The idea of estimating the causal relationship among the groups of variables was suggested by Prof. Shohei Shimizu. The present author formulated and implemented the method, performed experiments, and drafted the article. Prof. Shohei Shimizu supported the present author by commenting on several stages of the draft.

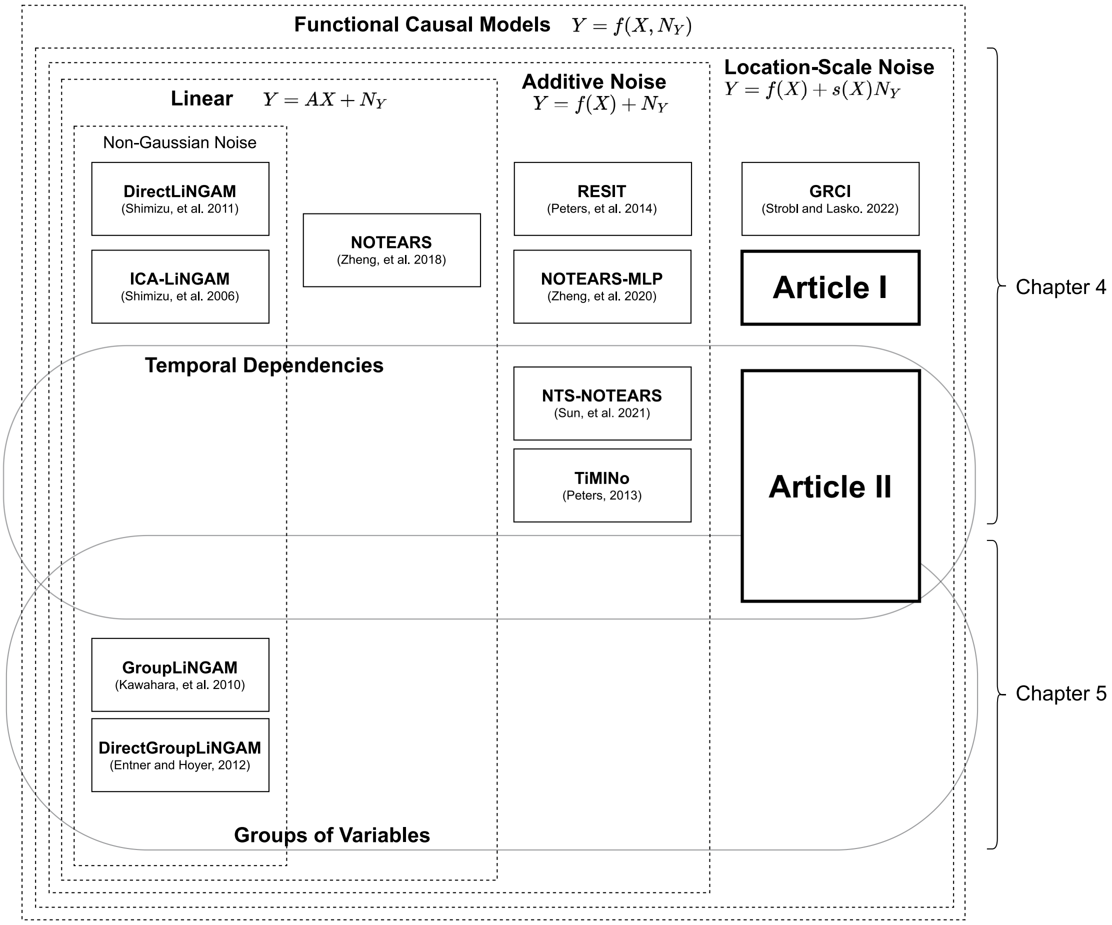

Functional Causal Models $Y = f(X, N_Y)$
Linear $Y = AX + N_Y$
Additive Noise $Y = f(X) + N_Y$
Location-Scale Noise $Y = f(X) + s(X)N_Y$
Non-Gaussian Noise
DirectLiNGAM (Shimizu, et al. 2011)
ICA-LiNGAM (Shimizu, et al. 2006)
NOTEARS (Zheng, et al. 2018)
RESIT (Peters, et al. 2014)
NOTEARS-MLP (Zheng, et al. 2020)
GRCI (Strobl and Lasko. 2022)
Article I
Temporal Dependencies
NTS-NOTEARS (Sun, et al. 2021)
TiMINo (Peters, 2013)
Article II
GroupLiNGAM (Kawahara, et al. 2010)
DirectGroupLiNGAM (Entner and Hoyer, 2012)
Groups of Variables
Chapter 4
Chapter 5

Fig. 1.2 Overview of model classes and relevant works (estimation methods for the corresponding model class) introduced in Chapter 4 and Chapter 5. The positioning of each article listed in Section 1.3 is also depicted. Equations of the bivariate case $X \to Y$ are shown for brevity.

Chapter 2

Background

This chapter presents the mathematical background to understand subsequent chapters. We first introduce the necessary principles of probability and statistics. Then, graph terminologies used to represent causal relationships are summarized.

2.1 Probability and Statistics

We follow the notation of (Spirtes et al., 2000). Given a *random variable* X, the *probability distribution* modeling the stochastic uncertainty of X is given as $P(X)$. Throughout this thesis, we consider only continuous random variables, where the probability of X is represented using a *probability density function* $p()$. A P-dimensional random variable is represented as $\mathbf{x} = (X_1, X_2, ..., X_P)^T$, and the *joint probability distribution* (or joint distribution for short) of $\mathbf{x}$ is denoted as $P(\mathbf{x}) = P(X_1, ..., X_P)$.

Suppose we have two random variables, X and Y. When $P(Y) > 0$, we define the *conditional probability distribution* of X given Y as

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}. \tag{2.1}$$

We can also define the conditional probability distribution of Y given X as

$$P(Y|X) = \frac{P(X,Y)}{P(X)}. \tag{2.2}$$

From Equations (2.1) and (2.2), we can write $P(X,Y)$ in two ways:

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X). \tag{2.3}$$

The *expectation* and *variance* of X are given as

$$\mathrm{E}\left[X\right] = \mu_X = \int_{-\infty}^{\infty} xp(x)dx,$$
$$\mathrm{Var}\left[X\right] = \sigma_X^2 = \mathrm{E}\left[(X - \mathrm{E}\left[X\right])^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2.$$

Using the conditional probability distribution, the conditional expectation and variance are calculated as

$$\mathrm{E}\left[X|Y=y\right] = \int_{-\infty}^{\infty} xp(x|y)dx,$$
$$\mathrm{Var}\left[X|Y=y\right] = \mathrm{E}\left[(X - \mathrm{E}\left[X|Y\right])^2|Y\right] = \mathrm{E}\left[X^2|Y\right] - \mathrm{E}\left[X|Y\right]^2.$$

The *covariance* of X and Y is

$$\mathrm{cov}\left[X,Y\right] = \mathrm{E}\left[(X - \mathrm{E}\left[X\right])(Y - \mathrm{E}\left[Y\right])\right] = \mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right],$$

and *(Pearson) correlation coefficient* is defined as

$$\rho_{X,Y} = \frac{\mathrm{cov}\left[X,Y\right]}{\sqrt{\mathrm{Var}\left[X\right]\mathrm{Var}\left[Y\right]}},$$

where the value measures the "linear" correlation between variables. Throughout the thesis, we call the correlation coefficient as correlation for brevity.

2.1.1 Probability Distributions

Here, we introduce some probability distributions relevant to this thesis. The shape of each distribution with several parameter settings is depicted in Figure 2.1. The most widely used probability distribution for continuous variables is the (multivariate) *Gaussian distribution*, where the probability density function is defined by

$$p(X_1,...,X_P) = \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^P \mathrm{det}(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad (2.4)$$

where $\boldsymbol{\mu}$ is a *mean vector*; the expectation for each dimension of $\mathbf{x}$ is stored in each element. Σ is a $P \times P$ *covariance matrix*, with covariance $\mathrm{cov}\left[X_i,X_j\right]$ in the non-diagonal elements and variance $\mathrm{Var}\left[X_i\right]$ in the diagonal elements.

In causal discovery, *non-Gaussianity* of a probability distribution is important in identifying causal relationships (Shimizu et al., 2006). One of the non-Gaussian probability distributions is the *uniform distribution*

$$p(X) = \mathrm{U}(X; a, b) = \begin{cases} \frac{1}{b-a} & \text{if} \quad a \leq X \leq b \\ 0 & \text{if} \quad X < a \text{ or } X > b \end{cases},$$

where a and b define the support of the probability distribution. X values in the range $[a, b]$ are equally likely to occur, showing a flat shape. Another well-known non-Gaussian distribution is the Laplace distribution:

$$p(X) = \mathrm{Laplace}(X; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|X-\mu|}{b}\right),$$

where μ and b are location and scale parameters, respectively. Compared to the Gaussian distribution, the Laplace distribution shows sharper peaks and heavier tails. The probability distributions introduced so far are all symmetric. An example of the asymmetric probability distributions is the *Gumbel distribution*

$$p(X) = \mathrm{Gumbel}(X; \mu, \beta) = \frac{1}{\beta} \exp\left(-\left(z + \exp^{-z}\right)\right),$$

where $z = \frac{X-\mu}{\beta}$, and μ is a location parameter which equal to the mode of the distribution, and $\beta > 0$ is a scale parameter that controls the expectation.

Probability distributions with heavier tails than the Gaussian distribution are called *super-Gaussian* distributions. In contrast, probability distributions with more decay of the tail are called *sub-Gaussian* distributions. For instance, the Laplace distribution is a super-Gaussian distribution, and the uniform distribution is a sub-Gaussian distribution (Hyvärinen et al., 2001).

2.1.2 Statistical Independence

Two random variables X and Y are statistically *independent*, denoted as $X \perp\!\!\!\perp Y$ (Dawid, 1979), if and only if

$$P(X, Y) = P(X)P(Y). \tag{2.5}$$

Equation (2.5) is equivalent to Equation (2.3) if $P(X|Y) = P(X)$ and $P(Y|X) = P(Y)$. Intuitively, this means that the variable X has no information of Y; thus, observing X does not benefit estimating Y, and vice versa.

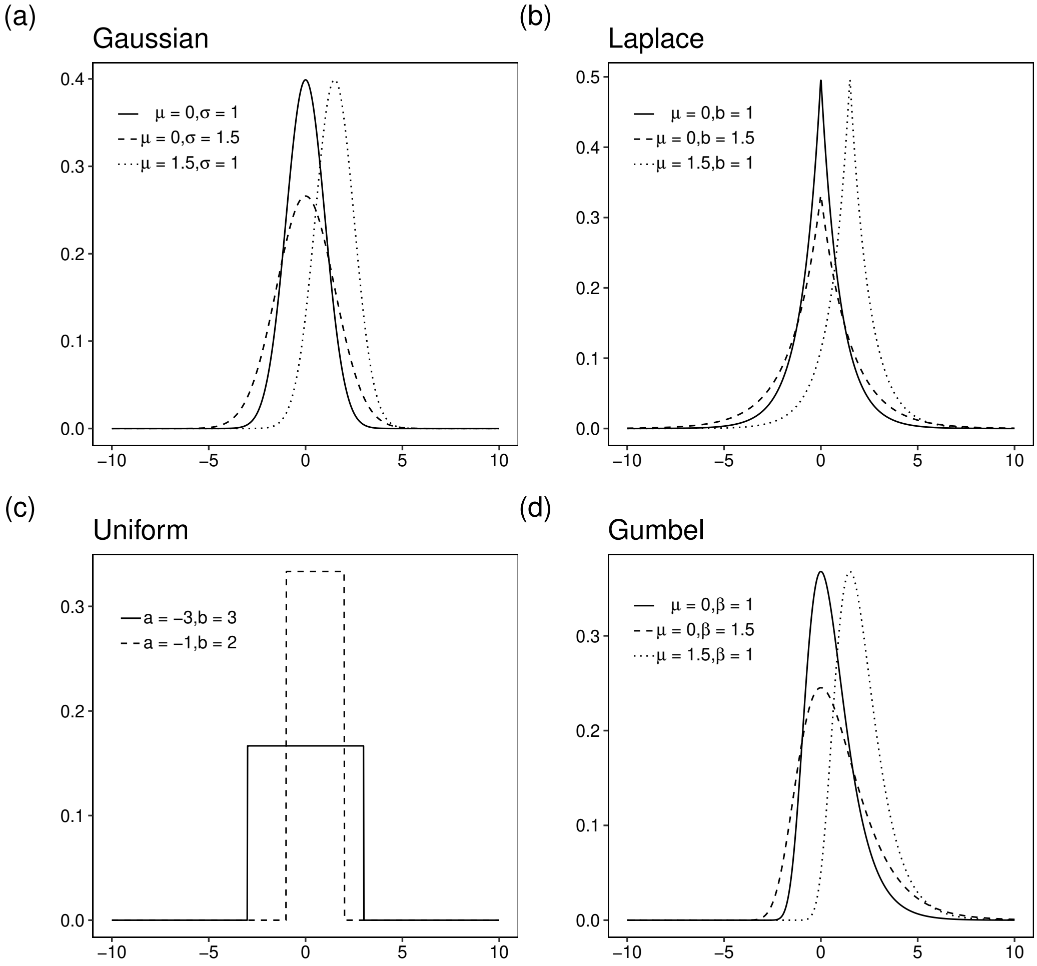


Fig. 2.1 Probability distributions. (a) The Gaussian distribution, (b) the Laplace distribution, (c) the uniform distribution, and (d) the Gumbel distribution. Different parameter sets are shown for each probability distribution.

A *conditional independence* of random variables X and Y given Z is defined as

$$X \perp\!\!\!\perp Y|Z \quad \text{if and only if} \quad P(X,Y|Z) = P(X|Z)P(Y|Z).$$

For example, suppose two persons, A and B, independently visit a supermarket simultaneously. Let X and Y be events of whether person A and B purchase an ice cream, respectively. A and B will likely purchase ice creams if the outside temperature is high; hence, X and Y more or less exhibit dependence. However, given the outside temperature Z, whether one purchases ice cream cannot be inferred from whether the other purchases it.

If X and Y are independent, we get

$$\begin{aligned} \mathrm{E}[XY] &= \mathrm{E}[X]\,\mathrm{E}[Y], \\ \mathrm{cov}[X,Y] &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] = \mathrm{E}[X]\,\mathrm{E}[Y] - \mathrm{E}[X]\,\mathrm{E}[Y] = 0, \end{aligned}$$

which shows no statistical correlation. However, if X and Y are uncorrelated, it does not mean that X and Y are independent. An example of non-independent data with no correlation is shown in Figure 2.2 [1]. A special case is when the variables follow the multivariate Gaussian distribution; no correlation is equivalent to independence. Recall the probability distribution of the multivariate Gaussian distribution given in Equation (2.4). If there is no correlation between the variables, the non-diagonal elements of the covariance matrix Σ is zero; then, we can rewrite Equation (2.4) as

$$\begin{aligned} p(X_1,\ldots,X_P) &= \frac{1}{\sqrt{(2\pi)^P \prod_{i=1}^{P} \sigma_i}} \exp\left(-\frac{1}{2}\sum_{i=1}^{P}\frac{(X_i-\mu_i)^2}{\sigma_i}\right) \\ &= \prod_{i=1}^{P} \frac{1}{\sqrt{(2\pi)^P \sigma_i}} \exp\left(-\frac{1}{2}\frac{(X_i-\mu_i)^2}{\sigma_i}\right), \end{aligned}$$

where the product of the univariate Gaussian distribution satisfies the Equation (2.5). We used the term σ_i as σ_{X_i} for brevity. However, in practice, real-world data deviates from the Gaussian distribution; hence, focusing only on the statistical measurements may lead to mistakes in inferring the relationship between the variables.

[1] The data is generated with a R package *datasauRus* (https://github.com/jumpingrivers/datasauRus). The original data are generated using simulated annealing (Matejka and Fitzmaurice, 2017).

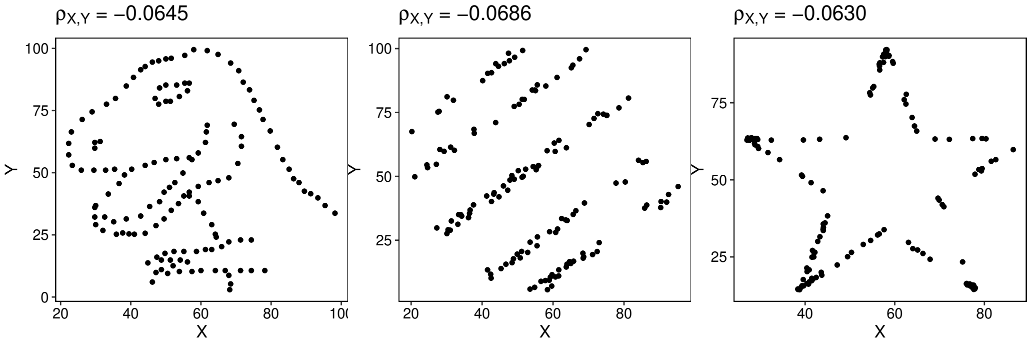


Fig. 2.2 Data with no statistical correlation, though not independent. They also share nearly equal average and standard deviation.

$$B = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

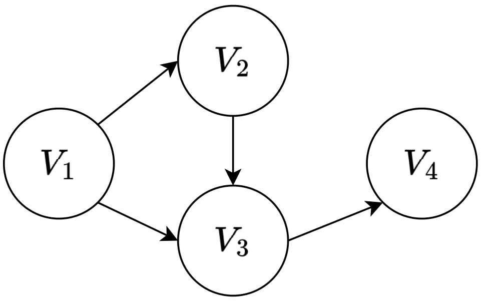


Fig. 2.3 Adjacency matrix and corresponding directed graph. The non-zero element of the adjacency matrix B indicates the graph edges.

2.2 Graphs

Causal discovery uses graphs to represent the causal data structure. This section introduces some graph terminologies relevant to this thesis.

A *graph* $\mathscr{G}$ is defined by a tuple $(\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} \in \{V_1, ..., V_P\}$ is a set of *vertices* (nodes), and $\boldsymbol{E}$ is a set of *edges*. An edge is defined by an ordered pair of vertices $< V_i, V_j >$, written as $V_i \to V_j$. A sequence of edges, for example $V_i \to ... \to V_s \to ... \to V_j$ is called a *path* from V_i to V_j. A path is called *cyclic* if it goes through any vertex more than once. Edges consisting of directed arrows are called *directed edges*, and paths with directed edges pointing from the first to the second vertex of every pair along the sequence of vertices are called *directed paths*. A *directed graph* is a graph with only directed edges, and a *skeleton* of a graph $\mathscr{G}$ is a graph with all directed edges of $\mathscr{G}$ replaced with *undirected edges* V_i–V_j.

A directed graph can be represented by an *adjacency matrix* $B \in \{0,1\}^{P \times P}$, which is a binary matrix with connections from V_i to V_j encoded in the (i, j)-th element. An example of an adjacency matrix and corresponding directed graph is given in Figure 2.3.

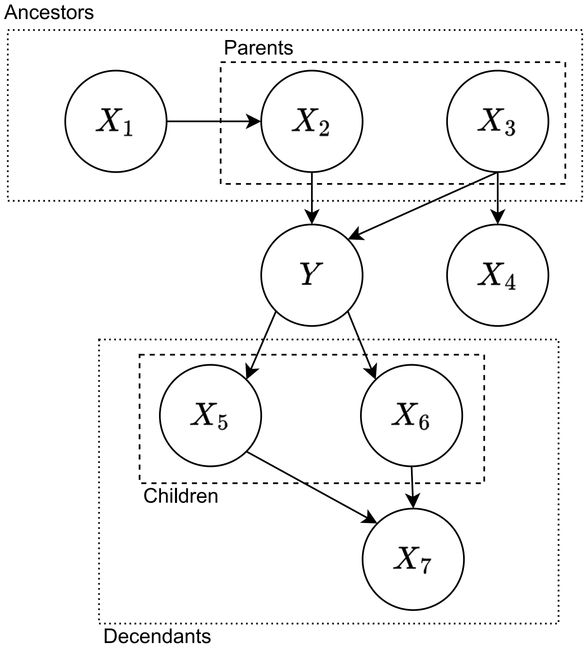


Fig. 2.4 Terminology of the vertices in a DAG. Ancestors, parents, children, and descendants are the vertices with respect to Y. Furthermore, X_1 and X_3 are the exogenous vertices and X_4 and X_7 are the sink.

A directed graph containing no cycles is called a *directed acyclic graph* (DAG). A *child* of a vertex V_i is a vertex that has a direct arrow from V_i. A vertex with a direct arrow to V_i is called a *parent* of V_i; we denote a set of parents of V_i as PA_i. If there is a directed path from a vertex to V_i, that vertex is called an *ancestor* of V_i, and if there is any directed path from V_i, the vertex is called a *descendant* of V_i. A vertex that has no parents is called *exogenous* or a *root*, and a vertex that has no children is called a *sink*.

An example of a DAG with 8 nodes $\boldsymbol{V} = \{Y, X_1, ..., X_7\}$ is shown in Figure 2.4. With respect to Y, $\{X_1, X_2, X_3\}$ are the ancestors, $\{X_2, X_3\}$ are the parents, $\{X_5, X_6\}$ are the children, and $\{X_5, X_6, X_7\}$ are the descendants. Furthermore, $\{X_1, X_3\}$ are the exogenous vertices and $\{X_4, X_7\}$ are the sink.

Given a DAG $\mathscr{G}$, we can obtain a *topological* or *causal order* of the vertices $\{V_1, ..., V_P\}$, which is a sequence of indices $1, .., P$. For every $i = 1, ..., P$, V_i precedes V_j in the ordering if V_i is not a descendant of V_j in $\mathscr{G}$. For example, the causal order of the vertices in the DAG of Figure 2.3 is $(1, 2, 3, 4)$.

Chapter 3

Causal Modeling

In this chapter, we describe mathematically modeling causal relationships. To explain what we refer to as *causality* in this thesis, we start with the potential outcome framework (Neyman, 1923; Rubin, 1974), also referred to as the counterfactual model, a major framework for describing causation between variables. Next, we introduce causal Bayesian networks (CBN), the most common framework to mathematically model the causal relationships between variables. After that, we explain the framework of structural equation models (SEM) that model the functional relationships between variables. We show that by using the framework of structural causal models (SCM), also referred to as functional causal models (FCM), which uses SEMs to describe the data-generating process, we can represent the population-level causal relationships defined by the potential outcome model. Finally, common assumptions regarding causal modeling relevant to this thesis are introduced.

3.1 Potential Outcome Framework

How can we say that *a variable X has a causal effect on a variable Y*? Intuitively, we can say that X caused Y if the value of Y changes with changes in X. For example, suppose we want to know the effect of teaching material in helping students pass the examination. In this case, $X = \{0,1\}$ is a binary variable representing whether a student uses the teaching material (0: did not use, 1: used), and $Y = \{0,1\}$ is a binary variable representing whether that student passed the examination (0: did not pass, 1: passed). If the student did not pass the examination when not using the teaching material ($X = 0, Y = 0$) and passed when using the material ($X = 1, Y = 1$), we can say that the teaching material caused the student to pass the examination.

More formally, let u be an individual *unit* (in this case, the student), $Y_0(u)$ and $Y_1(u)$ be the value of the *outcome* Y (result of the examination) when *treatment* X (usage of the teaching material) is 0 and 1, respectively. Then, $Y_1(u) - Y_0(u)$ is the causal effect of using the teaching material. In practice, however, if a student uses the teaching materials and takes the examination, the results when the student does not use the teaching materials, which is *counterfactual*, cannot be known. This problem is called the *fundamental problem of causal inference* (Holland, 1986).

At the population level, for instance, suppose there are multiple students in a class, and we want to know the effect of the teaching material for that population. Consider estimating *average treatment effect* (ATE), an average of the effect of the teaching material given by

$$\mathrm{ATE} = \mathrm{E}\left[Y_1 - Y_0\right].$$

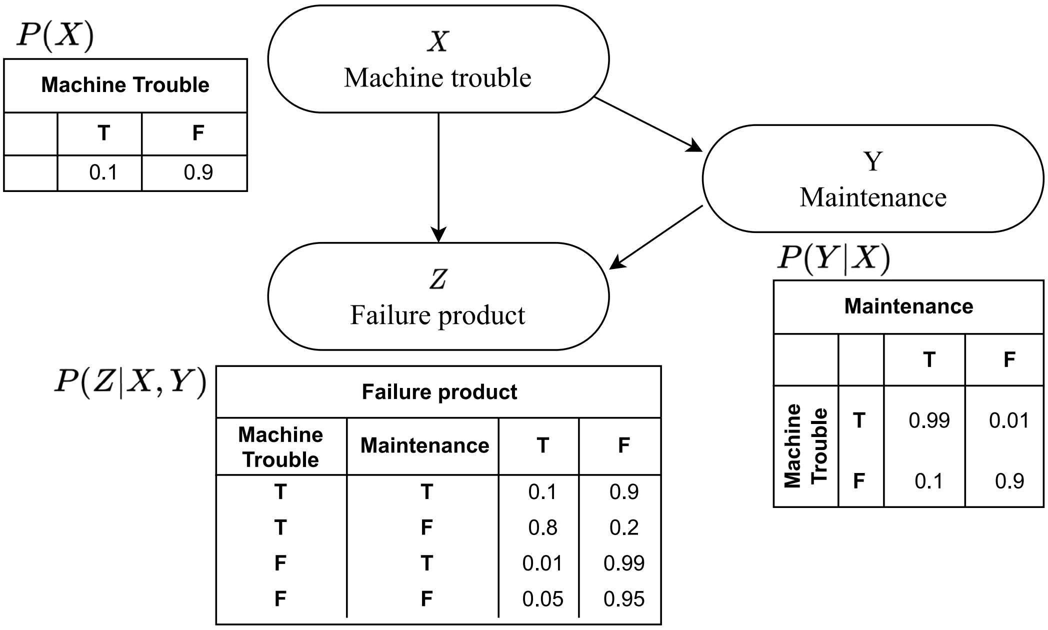
In this case, we also suffer from the problem as in the individual case: If all students use the teaching material, observing the counterfactual from the examination results is impossible. However, if we assign teaching materials *randomly* (Rubin, 1974) to each student, we can obtain a causal effect of the teaching material on the population. Suppose we assign teaching material $X(u)$ to student u. As assignment X is determined randomly, we have $X \perp\!\!\!\perp Y_0, Y_1$. Then, the difference between the average of the outcomes is equivalent to ATE.

$$\begin{aligned}
\hat{\mathrm{ATE}} &= \mathrm{E}\left[Y_1 | X = 1\right] - \mathrm{E}\left[Y_0 | X = 0\right] \\
&= \mathrm{E}\left[Y_1\right] - \mathrm{E}\left[Y_0\right] \\
&= \mathrm{E}\left[Y_1 - Y_0\right] \\
&= \mathrm{ATE}.
\end{aligned}$$

Note that the potential outcome framework does not explicitly model the data-generating process. In Section 3.4, we explain that SEMs also can represent population-level causation through a set of equations (Pearl, 2009).

3.2 Causal Bayesian Networks

Next, we introduce frameworks to model the causal relations. A causal Bayesian network (CBN) is a framework defined on a directed acyclic graph $\mathscr{G} = (\boldsymbol{V}, \boldsymbol{E})$ with random variables used as the vertices $\boldsymbol{V} = \{X_1, ..., X_P\}$; it models the conditional dependence over the variables (Pearl, 2009). CBN defines the joint probability distribution as a product of conditional probability distributions.



$P(X)$

Machine Trouble		
	T	F
	0.1	0.9

$P(Y|X)$

		Maintenance	
		T	F
Machine Trouble	T	0.99	0.01
	F	0.1	0.9

$P(Z|X,Y)$

		Failure product	
Machine Trouble	Maintenance	T	F
T	T	0.1	0.9
T	F	0.8	0.2
F	T	0.01	0.99
F	F	0.05	0.95

Fig. 3.1 A simple example of a causal Bayesian network (CBN), represents a relationship between the occurrence of the machine trouble, whether maintenance is performed, and the failure of a product.

Definition 3.1 (Causal Bayesian Networks). *A causal Bayesian network of P variables $\{X_1,...,X_P\}$ defines the joint probability distribution:*

$$P(X_1,...,X_P) = \prod_{j=1}^{P} P(X_j|PA_j), \tag{3.1}$$

where the joint probability is factorized with the conditional probability distributions.

Equation (3.1) indicates that the variables are mutually independent if conditioned with their parents. Intuitively, this means that the values of each variable are determined only by their parents. Formally, parents PA_j is the smallest subset of variables $\boldsymbol{V} \setminus X_j$ that satisfies the following equation:

$$P(X_j|PA_j) = P(X_j|\boldsymbol{V} \setminus X_j).$$

An example of a CBN is given in Figure 3.1. Machine trouble increases the probability of performing maintenance and whether the maintenance affects the possibility of producing a failed product.

3.3 Structural Equation Models

Another well-known and most important framework for this thesis is SEM. SEMs were first introduced from genetics and econometrics (Haavelmo, 1943; Wright, 1921). Generally, CBNs are used for modeling discrete random variables, whereas SEMs are used for modeling continuous random variables. We consider the definitions of SEM given in (Pearl, 2009) and follow the notations in (Peters et al., 2014).

Definition 3.2 (Structural Equation Models)**.** *SEM of P variables $\{X_1,...,X_P\}$ is a collection of P equations:*

$$X_j = f_j(\mathrm{PA}_j, N_j), \quad j = 1,...,P \tag{3.2}$$

where N_j is a noise term that is mutually independent, and f_j is a function that maps the parents PA_j and noise N_j to X_j.

Equation (3.2) is a non-linear and non-parametric generalization of *linear SEMs*:

$$X_j = \sum_{i \neq j} b_{ji} X_i + N_j, \quad j = 1,...,P \tag{3.3}$$

where coefficient b_{ji} is a scalar representing a direct effect from X_i to X_j; hence, the parents of X_j is X_i having b_{ji} with non-zero values in this setting. Generally, the functional relationship is not deterministic, which is modeled by noise N_j. From Equation (3.3), we obtain

$$N_j = X_j - \sum_{i \neq j} b_{ji} X_i. \quad j = 1,...,P$$

When the probability distribution of N_j is $P_j(N_j)$, the joint distribution over $\{X_1,...,X_P\}$ is

$$\begin{aligned} P(X_1,...,X_P) &= \prod_{j=1}^{P} P(X_j|\mathrm{PA}_j) \\ &= \prod_{j=1}^{P} P_j(N_j) \\ &= \prod_{j=1}^{P} P_j\left(X_j - \sum_{i \neq j} b_{ji} X_i\right). \end{aligned}$$

Therefore, if we assume a probability distribution for noise N_j, the linear SEM can be seen as a CBN, with conditional probabilities $P(X_j|\mathrm{PA}_j) = P_j\left(X_j - \sum_{i \neq j} b_{ji} X_i\right)$. Generally, P_j is assumed to be the Gaussian distribution. In particular, a linear SEM with an assumption of Gaussian noise, we call the linear SEM a *linear Gaussian SEM*.

3.4 Describing Causation with Structural Equation Models

This section explains representing population-level causation using SEMs (Pearl, 2009). The framework using SEMs to represent the data-generating process is called *structural causal models* (SCM), also termed as *functional causal models* (FCM). In the potential outcome framework, the random assignment of a treatment to a unit is given by $X(u)$. For instance, $X(u) = 1$ means that we force the treatment to be 1 independent of other variables; this manipulation of holding a variable to a certain value is also called *intervention*. Suppose we have the following SEM:

$$\begin{aligned} X &= N_X \\ Y &= f_Y(X, N_Y). \end{aligned} \tag{3.4}$$

We consider intervening on X as a constant $X = 1$, denoted as the *do-operator* as $do(X = 1)$. To represent an intervention on a variable with SEM, we replace Equation (3.4) with the constant value and obtain a modified set of equations that represent the behavior under the intervention $do(X = 1)$ (Pearl, 2009):

$$\begin{aligned} X &= 1 \\ Y &= f_Y(X, N_Y). \end{aligned}$$

Intervening on a variable gives *interventional distribution* on Y: $P(Y|do(X = 1))$, induced by the modified set of equations. Therefore, if the interventional distribution changes according to X, we can say that X affects Y:

$$P(Y|do(X = 1)) \neq P(Y|do(X = 0)).$$

The difference between the expectation of the interventional distributions can be used to quantify the causal effect of X on Y (Pearl, 2009), referred to as the average causal effect:

$$\mathrm{E}\left[Y|do(X = 1)\right] - \mathrm{E}\left[Y|do(X = 0)\right].$$

Generally, if the causal structure is known, we can identify the average causal effect (Pearl, 2009; Shpitser and Pearl, 2008). However, in most cases, the true causal structure is unknown; hence, developing mathematical methods to estimate the causal structure from observational data is essential. The choice of the functional form and the noise distributions are the central factors to exploit causal relationships from observational data (Peters et al., 2011). In Section 4.1, we introduce some of the structures used in the context of the causal discovery.

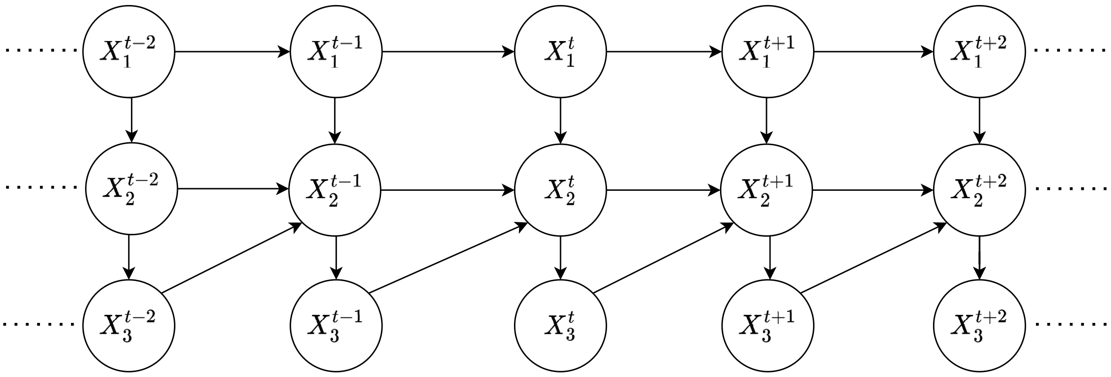


Fig. 3.2 An example of a full time graph for 3-dimensional time-series data.

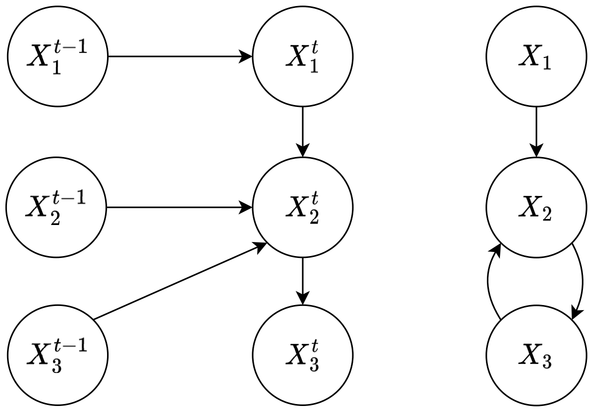
3.5 Time Series Models

In many cases, variables show temporal dependencies. For instance, temperatures measured at multiple machine positions can influence the temperatures in the following observations. To model the time-series data, we consider the case where each observation is collected in fixed time intervals, and each variable is generated from the past and present variables. We first explain several graph representations for time-series data and then introduce the time-series model.

Suppose we have a set of three time-series $\boldsymbol{X} = \{X_1, X_2, X_3\}$, and let $\boldsymbol{X^t} = \{X_1^t, X_2^t, X_3^t\}$ be a set of variables for a fixed time t. Then, a graph that represents the full (dynamic) data-generating structure is called *full time graph* (Peters et al., 2013) or *full time causal graph* (Assaad et al., 2022). Figure 3.2 shows an example of a full time graph of $\boldsymbol{X}$. The past effects are called *lagged effects*, and effects on the same time frame are called *instantaneous* or *contemporaneous effects*.

Recovering the full time graph from observational data is difficult as only a single observation is available for each time frame. Therefore, a common assumption for the time series is stationarity (Runge, 2020), which states that for a causal relation $X_i^{t-\tau} \to X_j^t$, the causal relation $X_i^{t'-\tau} \to X_j^{t'}$ is observed for all $t' \neq t$. Intuitively, stationarity states that the conditional independencies among the variables do not change over time.

Assuming stationarity, we simplify the full time graph and obtain a *window causal graph* using the variable set from the present time frame t to the maximum lag of the causal relations as vertices (Assaad et al., 2022). Furthermore, we can create a more simplified *summary causal graph* (Assaad et al., 2022), referred to as *summary time graph* (Peters et al., 2013), comprising variables in $\boldsymbol{X}$ as vertices, by drawing an arrow $X_i \to X_j$ if any $X_i^{t-\tau} \to X_j^t$ exist. Figure 3.3 (a) and (b) depict examples of window causal and summary time graphs, respectively.



(a) Window causal graph (b) Summary time graph

Fig. 3.3 Example of window causal and summary time graphs for three-dimensional time-series data.

A major model for representing multivariate time-series relevant to this thesis is the *structural vector autoregressive model* (SVAR) (Lütkepohl, 2005), which includes linear SEM as a special case.

Definition 3.3 (Structural Vector Autoregressive Model)**.** *SVAR with maximum lag L, denoted as the L-th order SVAR or SVAR(L) of P time-series $\{X_1^t, ..., X_P^t\}$ is defined as*

$$X_j^t = \sum_{i \neq j} b_{ji}^t X_i^t + \sum_{\tau=1}^{L} \sum_{i=1}^{P} b_{ji}^{t-\tau} X_i^{t-\tau} + N_j^t, \quad j = 1, ..., P \tag{3.5}$$

where b_{ji}^t and $b_{ji}^{t-\tau}$ are the connection strengths of instantaneous and lagged effects, respectively, and N_j^t is the noise term, which is mutually independent over j and identically distributed over t. The instantaneous effects are assumed to be acyclic. By setting $\boldsymbol{X}^t = (X_1^t, ..., X_P^t)^T$ and $\boldsymbol{N}^t = (N_1^t, ..., N_P^t)^T$, Equation (3.5) can be represented in the vector form:

$$\boldsymbol{X}^t = B^t \boldsymbol{X}^t + \sum_{\tau=1}^{L} B^{t-\tau} \boldsymbol{X}^{t-\tau} + \boldsymbol{N}^t,$$

where $B^t, B^{t-\tau}$ are $P \times P$ matrix of the connection strengths, where B^t can be permuted to a strictly lower triangular matrix because of the acyclicity of the instantaneous effects by simultaneously permuting row and column with the topological order of the variables.

Estimation of SVAR is typically done by maximum likelihood estimation (Lütkepohl, 2005). However, many estimation methods assume Gaussian noise, hence suffer from the identifiability (Hyvärinen et al., 2010).

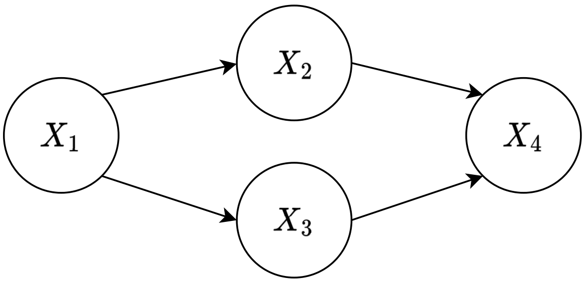


Fig. 3.4 Causal Markov condition. The graph shows that $X_2 \perp\!\!\!\perp X_3 | X_1$ and $X_1 \perp\!\!\!\perp X_4 | X_2, X_3$.

3.6 Common Assumptions for Causal Models

In addition to the functional form and probability distribution, reasonable assumptions about the data-generating process are required to model the underlying (unknown) causal relationships from data. Here, we introduce some of the common assumptions of the data-generating process.

3.6.1 Causal Markov Condition

The causal Bayesian network (Definition 3.1) defines the joint probability distribution as a product of conditional probability distributions of each variable X_j given its parents PA$_j$. Therefore, we assume that the data-generating process follows this formulation. This is given by the *causal Markov condition*, which connects the graphs with the probability distribution as follows: (Spirtes et al., 2000):

Definition 3.4 (Causal Markov Condition). *Let $\mathscr{G} = (\boldsymbol{V}, \boldsymbol{E})$ be a graph with vertices $\boldsymbol{V} = \{X_1, ..., X_P\}$ inducing a joint probability $P(X_1, ..., X_P)$ over the vertices. $\mathscr{G}$ and $P(X_1, ..., X_P)$ satisfy the* causal Markov condition *if and only if the following condition holds for all $W \subseteq \boldsymbol{V}$:*

$$W \perp\!\!\!\perp \boldsymbol{V} \setminus (Descendants(W) \cup Parents(W)) \, | Parents(W).$$

Thus, if the causal Markov condition holds, each variable X_j is independent of its non-descendants $\boldsymbol{V} \setminus \big(\text{Descendants}(X_j) \cup \text{Parents}(X_j)\big)$ given its parents $\text{Parents}(X_j)$. An example is given in Figure 3.4.

3.6.2 Causal Minimality Condition

The causal Markov condition states the sufficient condition for conditional independence. However, it cannot tell whether a pair of variables is dependent even if they are directly

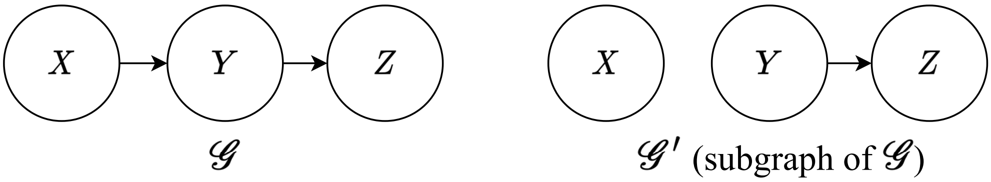


Fig. 3.5 Example of a graph and its subgraph. If we consider a joint distribution $P(X,Y,Z)$ with $X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z$, $\mathscr{G}$ and $P(X,Y,Z)$ do not satisfy the causal minimality condition, because its subgraph $\mathscr{G}'$ and $P(X,Y,Z)$ also satisfies the causal Markov condition.

connected in the graph. The *causal minimality condition* states an additional assumption regarding conditional dependencies:

Definition 3.5 (Causal Minimality Condition)**.** *Let $\mathscr{G} = (\boldsymbol{V}, \boldsymbol{E})$ be a graph with vertices $\boldsymbol{V} = \{X_1, ..., X_P\}$ inducing a joint probability $P(X_1, ..., X_P)$ over the vertices. $\mathscr{G}$ and $P(X_1, ..., X_P)$ satisfy the* causal minimality condition *if and only if $\mathscr{G}$ and $P(X_1, ..., X_P)$ satisfy the causal Markov condition, but not for every proper subgraph of $\mathscr{G}$.*

Suppose we have a graph $\mathscr{G}$ that consists of three vertices $\{X,Y,Z\}$ shown in Figure 3.5. From $\mathscr{G}$, causal Markov condition entails conditional independence relations $X \perp\!\!\!\perp Z|Y$. Then, assume we have a joint probability $P(X,Y,Z)$ that shows $X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z$, which satisfies the causal Markov condition $X \perp\!\!\!\perp Z|Y$. Next, consider graph $\mathscr{G}'$, which is a subgraph of $\mathscr{G}$ obtained by removing an arrow from X to Y. We can see that $P(X,Y,Z)$ also satisfies the causal Markov condition with respect to $\mathscr{G}'$. Therefore, $P(X,Y,Z)$ that exhibit $X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z$ does not satisfy the causal minimality condition with respect to $\mathscr{G}$.

3.6.3 Faithfulness Condition

When inferring a causal relationship based on conditional independence relationships, it will be problematic if the conditional independence disappears because of the specific choice of model parameters (e.g., b_{ji} of the linear SEMs). The *faithfulness condition* requires that the conditional independence relationship of the joint probability distribution $P(X_1, ..., X_P)$ is via the causal Markov condition:

Definition 3.6 (Faithfulness Condition)**.** *Given a graph $\mathscr{G} = (\boldsymbol{V}, \boldsymbol{E})$ inducing a joint probability $P(X_1, ..., X_P)$ over the vertices $\boldsymbol{V} = \{X_1, ..., X_P\}$, $\mathscr{G}$ and $P(X_1, ..., X_P)$ satisfy the* faithfulness condition*, or we say $P(X_1, ..., X_P)$ is* faithful *to $\mathscr{G}$ if and only if every conditional independence relationship in $P(X_1, ..., X_P)$ is entailed by the causal Markov condition.*

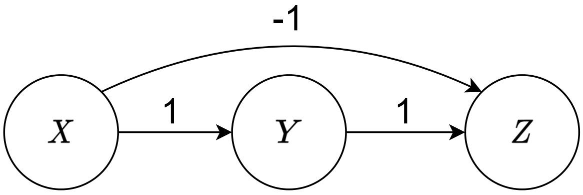


Fig. 3.6 Linear SEM which is not faithful to the graph. Coefficients are shown on each arrow. If the noise terms follow the Gaussian distribution, we observe $X \perp\!\!\!\perp Z$ despite the given causal structure.

An example of a linear SEM that satisfies the causal Markov condition but violates the faithfulness condition is shown in Figure 3.6. The corresponding linear SEM is

$$X = N_X, \tag{3.6}$$

$$Y = X + N_Y, \tag{3.7}$$

$$Z = -X + Y + N_Z. \tag{3.8}$$

Let the independent noise terms N_X, N_Y, and N_Z follow the standard Gaussian distribution. From Equations (3.6)–(3.8), we obtain

$$Z = -N_X + N_X + N_Y + N_Z = N_Y + N_Z. \tag{3.9}$$

From Equations (3.8) and (3.9), we calculate $\mathrm{cov}\,[X,Z]$ as

$$\begin{aligned}
\mathrm{cov}\,[X,Z] &= \mathrm{E}\,[XZ] - \mathrm{E}\,[X]\,\mathrm{E}\,[Z] \\
&= \mathrm{E}\,[N_X N_Y + N_X N_Z] - \mathrm{E}\,[N_X]\,\mathrm{E}\,[N_Y + N_Z] \\
&= \mathrm{E}\,[N_X]\,\mathrm{E}\,[N_Y] + \mathrm{E}\,[N_X]\,\mathrm{E}\,[N_Z] - \mathrm{E}\,[N_X]\,\mathrm{E}\,[N_Y] - \mathrm{E}\,[N_X]\,\mathrm{E}\,[N_Z] \\
&= 0.
\end{aligned}$$

As X and Z follow the Gaussian distribution, no correlation is equivalent to the independence; therefore, $X \perp\!\!\!\perp Z$. This violates the faithfulness condition because the independence relationship between X and Z is not due to the causal Markov condition.

3.6.4 Causal Sufficiency

Another concern for inferring the causal structure is the *unobserved variable*. If an unobserved variable is a common effect of two or more variables, the variables are *confounded* and exhibit statistical dependency regardless of the true causal relationships. Unobserved variables that confound observed variables are called *latent confounders*.

Causal sufficiency requires no latent confounders of the vertices in the population.

Definition 3.7 (Causal Sufficiency)**.** *Let* $\mathscr{G} = (\boldsymbol{V}, \boldsymbol{E})$ *be a graph with vertices* $\boldsymbol{V} = \{X_1, ..., X_P\}$. $\boldsymbol{V}$ *is* causal sufficient *if and only if every common cause of any two or more variables is in* $\boldsymbol{V}$.

Chapter 4

Causal Discovery on Nonlinear Time-Series Data with Heteroscedasticity

In this chapter, we describe the contributions of a series of studies conducted on the original research Article I and part of Article II that propose a causal discovery method for data with nonlinearity, temporal dependency, and heteroscedasticity. Before explaining the proposed method, we first introduce relevant existing works that seek to estimate the underlying *causal structure* (Pearl, 2009) from observational data.

Definition 4.1 (Causal Structure)**.** *Let $\boldsymbol{X}$ be a set of variables. A* causal structure *of $\boldsymbol{X}$ is a directed acyclic graph (DAG) where each vertice represents a corresponding variable in $\boldsymbol{X}$, and edges represent the existence of direct functional relationships between the variables.*

This thesis uses SEM-based FCM. There are two main components for FCM: First is the model, which involves the design of the functional relationships and noise probability distribution, where the central issue is the *identifiability* of the model; that is, whether the causal structure can be uniquely recovered from the joint distribution of the observed data. Second is the corresponding estimation method to learn the functional relationships from the data. Throughout this thesis, we assume that the causal structure is acyclic, suggesting that the causal structure can be represented by a DAG.

In Section 4.1, we first introduce existing FCMs relevant to this thesis. Then, we explain the estimation methods based on continuous optimization in Section 4.2. Section 4.3 describes the proposed estimation method for nonlinear data with heteroscedasticity. After that, we introduce the proposed FCM and corresponding estimation method that extends the former work to be capable of temporal dependencies in Section 4.4. Finally, the results of numerical experiments conducted to assess the effectiveness of the proposed method are shown in Section 4.5. Other causal discovery methods include the constraint-based methods

(Spirtes et al., 2000) and score-based methods (Chickering, 2002); however, these methods are out of the scope of this thesis.

4.1 Functional Causal Models

This section introduces the existing FCMs that exploit causal asymmetries based on the functional relationship and noise distribution. A sufficient identifiability condition for bivariate and multivariate cases is given in (Peters et al., 2011) to show that an FCM is *identifiable*, where the underlying causal structure can be uniquely recovered from the joint distribution. Many existing works on FCMs leverage this observation; here, we briefly introduce the definitions for the *identifiable functional model class* (IFMOC).

Definition 4.2 (Functional Model Class). *Recall the SEMs given in Definition 3.2:*

$$X_j = f_j(\mathrm{PA}_j, N_j). \quad j = 1, ..., P$$

An SEM with noise distributions $P(N_j)$ is called a functional model*, if the noise terms are mutually independent and the underlying graph is acyclic. Then, consider the following set of functions:*

$$\mathscr{F} \subset \{f | f : \mathbb{R}^m \to \mathbb{R} \quad \textit{for any} \quad 2 \leq m \leq P\}.$$

A functional model belongs to a functional model class with function class $\mathscr{F}$ *(*$\mathscr{F}$-FMOC*) if $f_j \in \mathscr{F}$ for all $j = 1, ..., P$ and induces joint probability distribution that all the probabilities are positive.*

Definition 4.3 (Bivariate Identifiable Set). *Suppose we observe a functional model with two variables X, Y with corresponding noise terms N_X and N_Y, and write $\mathscr{F}_{|2} := \{f | f : \mathbb{R}^2 \to \mathbb{R}\}$. For instance, $X = N_X$ and $Y = f(X, N_Y)$. Let $\mathscr{B}$ be a set of combinations of functions $f \in \mathscr{F}_{|2}$ and probability distributions $P(X)$, $P(N_Y)$:*

$$\mathscr{B} = \mathscr{F}_{|2} \times P_{\mathbb{R}} \times P_{\mathbb{R}}.$$

A set $\mathscr{B}$ is bivariate identifiable in $\mathscr{F}$ *when the following condition holds:*

$$\begin{aligned} &\textit{if} \quad (f, P(X), P(N_Y)) \in \mathscr{B} \quad \textit{and} \quad Y = f(X, N_Y),\ N_Y \perp\!\!\!\perp X \\ &\textit{then} \quad \nexists g \in \mathscr{F}_{|2} : \quad X = g(Y, N_X),\ N_X \perp\!\!\!\perp Y. \end{aligned}$$

Hence, we cannot define a function in both directions ($X \rightarrow Y, X \leftarrow Y$) that satisfies the independence of the noise terms ($N_Y \perp\!\!\!\perp X, N_X \perp\!\!\!\perp Y$). Moreover, the effect must not be independent of the cause: $Y = f(X, N_Y) \not\perp\!\!\!\perp X$ for all $(f, P(X), P(N_Y)) \in \mathscr{B}$ under $N_Y \perp\!\!\!\perp X$.

Given the bivariate identifiability of a functional model, by further constraining conditional distributions, the identifiability is also valid for the multivariate case.

Definition 4.4 (Identifiable Functional Model Class). *Let $\mathscr{B}$ be bivariate identifiable in $\mathscr{F}$; consider functional models with P variables:*

$$X_j = f_j(\mathrm{PA}_j, N_j). \quad j = 1, ..., P$$

An $\mathscr{F}$-FMOC *is called a* $(\mathscr{B}, \mathscr{F})$-identifiable functional model class *(IFMOC) if for all $j \in \{1, ..., P\}$, $i \in \mathrm{pa}(j)$ and for all $x_{\mathrm{pa}(j)\setminus i}$, we have*

$$f_j\big(x_{\mathrm{pa}(j)\setminus i}, \underbrace{\cdot}_{X_i}, \underbrace{\cdot}_{N_j}\big) \in \mathscr{F}_{|2},$$

where $\mathrm{pa}(j)$ *is an index set of* PA_j*, and the underbraces indicate the input component of f_j for X_i and N_j. Additionally, for all sets* $\mathbf{S}$ *with* $\mathrm{pa}(j)\setminus\{i\} \subseteq \mathbf{S} \subseteq \mathrm{nd}(j)\setminus\{i, j\}$*, there exists $x^*_{\mathbf{S}}$ with $P(x^*_{\mathbf{S}}) > 0$ and*

$$\Big(f_j\big(x_{\mathrm{pa}(j)\setminus i}, \underbrace{\cdot}_{X_i}, \underbrace{\cdot}_{N_j}\big), P(X_i | X_{\mathbf{S}} = x^*_{\mathbf{S}}), P(N_j)\Big) \in \mathscr{B},$$

where $\mathrm{nd}(j)$ *is an index set of non-descendants of X_j.*

One can uniquely recover the underlying causal structure if the data are generated by IFMOC (Peters et al., 2011). Intuitively, Definition 4.4 requires that when one obtains a bivariate model by fixing all arguments of the functions f_j, except for one parent (X_i), and noise variable (N_j), the bivariate identifiability remains. IFMOC assumes acyclicity of the causal structure; the noise terms are jointly independent, have positive densities, and imply causal minimality, which is a weaker assumption compared to faithfulness (Peters et al., 2011)

The remaining section introduces the relevant FCMs and estimation methods. We start with the linear non-Gaussian acyclic models (LiNGAM) (Shimizu et al., 2006), an FCM assuming linear causal relationships and non-Gaussian noise. Then, we describe the additive noise models (ANM) (Hoyer et al., 2008a) that leverages non-linearity to infer the causal direction. Next, the location-scale noise models (LSNM) (Immer et al., 2023), which can handle non-linearity and heteroscedasticity, are explained. Finally, we introduce non-linear time-series models with independent noise (TiMINo) (Peters et al., 2013), capable of nonlinearity and temporal dependency.

4.1.1 LiNGAM: Linear Non-Gaussian Acyclic Models

Model Definition

Suppose we observe a set of P variables $\boldsymbol{X} = \{X_1, ..., X_P\}$ generated from a system, represented by a directed acyclic graph. Let b_{ji} denote connection strength from X_i to X_j, and recall that the parent set of X_j is denoted as PA_j, where $b_{ji} \neq 0$ if $X_i \in PA_j$ and $b_{ji} = 0$ if $X_i \notin PA_j$. LiNGAM is a special case of SEM (Equation (3.2)), where each variable $X_j, (j = 1, ..., P)$ is generated from the following equation:

$$X_j = \sum_{i \in \mathrm{pa}(j)} b_{ji} X_i + N_j, \quad j = 1, ..., P$$

where $\mathrm{pa}(j)$ is an index set of PA_j, and noise term N_j is a continuous random variable with non-Gaussian densities of non-zero variance. N_j is assumed to be mutually independent; thus, there are no latent confounders. It is shown that one can recover true causal order if data strictly follow the given assumptions (Shimizu et al., 2006).

The principle of the identification of LiNGAM is shown in Figure 4.1. We consider two variable cases where the true causal direction is $X \rightarrow Y$ and noise terms are generated from the uniform distribution. Comparing Figures 4.1 (b) and (d) show that the predictor and the regression residuals are independent only when regressed in the true direction. LiNGAM leverages this asymmetry induced by the non-Gaussianity for determining the causal structure.

Estimation

Estimation for LiNGAM can be performed using the independent component analysis (Shimizu et al., 2006), called *ICA-LiNGAM*. Later, an estimation method called *DirectLiNGAM* (Shimizu et al., 2011) was proposed to improve the convergence and scale-sensitivity of ICA-LiNGAM. DirectLiNGAM recursively performs regression and independence tests between predictor and regression residuals. Let r_{ji} denote residuals obtained when X_j is regressed on X_i. Under the LiNGAM assumption, X_i is exogenous if and only if it is independent of its residuals r_{ji} for all $j \neq i$ (Shimizu et al., 2011). After DirectLiNGAM estimates an exogenous variable, its effect is removed from the remaining graph using ordinary least squares and repeated until full estimation of the causal order is obtained. Next, regularized

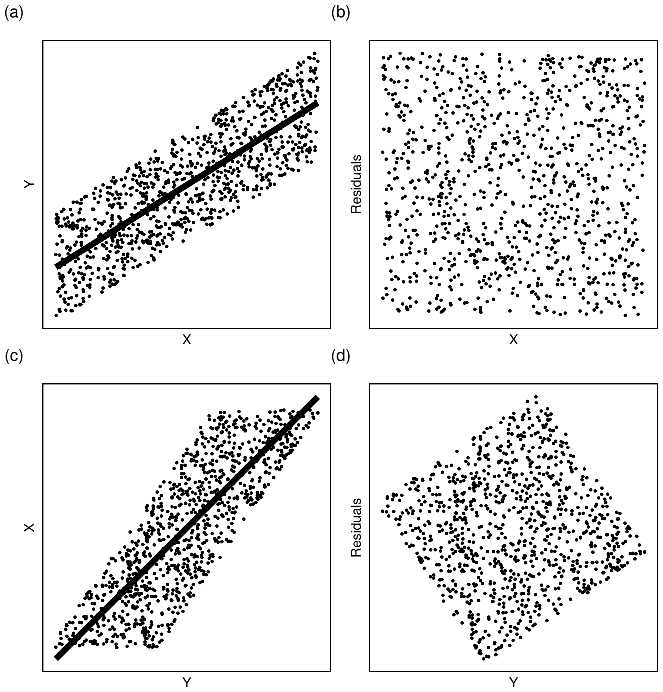


Fig. 4.1 Role of non-Gaussianity in LiNGAM. (a) Scatter plot and fitted linear regression model on true causal direction ($X \rightarrow Y$). (b) Predictor and residuals on the true causal direction. (c) Scatter plot and fitted linear regression model on anti-causal direction ($Y \rightarrow X$). (d) Predictor and residuals on the anti-causal direction. The predictor and residuals are independent in the true casual direction but not in the anti-causal direction.

regression such as adaptive lasso (Zou, 2006) is performed on each variable to estimate the causal strength b_{ji}, using the preceding variables in the causal order as predictors.

Measuring the independence between predictor and regression residuals may increase computational time. For practical estimation, the likelihood ratio, asymptotically equivalent to the difference in mutual information, is proposed (Hyvärinen and Smith, 2013), which can be computed efficiently for the linear case. Let $\tilde{X}_i, \tilde{X}_j$ denote variables X_i, X_j standardized to zero mean and unit variance, respectively. The difference between mutual information is given by

$$\begin{aligned} m(X_i, X_j) &= I(\tilde{X}_j, r_{ij}) - I(\tilde{X}_i, r_{ji}) \\ &= H(\tilde{X}_j) + H\left(\frac{r_{ij}}{\sigma_{r_{ij}}}\right) - H(\tilde{X}_i) - H\left(\frac{r_{ji}}{\sigma_{r_{ji}}}\right), \end{aligned} \tag{4.1}$$

where σ denotes standard deviation. Causal order is estimated as $X_i \rightarrow X_j$ and $X_i \leftarrow X_j$ when $m(X_i, X_j)$ is positive and negative, respectively. Entropies $H(\cdot)$ are estimated by maximum entropy approximation (Hyvärinen, 1998). A collection of mutual information is aggregated to find an exogenous variable and recover the causal order among multiple variables:

$$m_i = -\sum_j \min(0, [\boldsymbol{M}]_{i,j})^2, \tag{4.2}$$

where $\boldsymbol{M}$ is a $P \times P$ matrix, with the (i, j) th element $[\boldsymbol{M}]_{i,j}$ as $m(X_i, X_j)$. A variable that maximizes m_i is chosen as an exogenous variable. The advantage of using Equations (4.1) and (4.2) is that the difference between mutual information can be computed efficiently with only one-dimensional entropies when comparing the independencies of the residuals and predictor.

4.1.2 ANM: Nonlinear Functional Relations

Model Definition

ANM (Hoyer et al., 2008a; Peters et al., 2014) is a special case of SEM, which consists of a non-linear function and additive noise term. It is defined as a collection of P equations:

$$X_j = f_j(\mathrm{PA}_j) + N_j, \quad j = 1, ..., P$$

where f_j is a twice differentiable function, and N_j is a mutually independent noise term. When f_j is a linear function, and N_j follows non-Gaussian distribution, ANM reduces

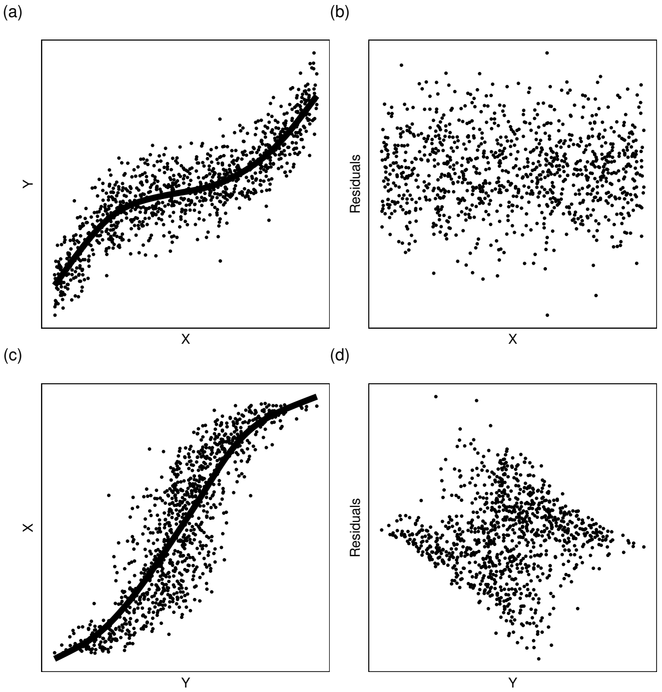


Fig. 4.2 Role of non-linearity in ANM. (a) Scatter plot and fitted linear regression model on true causal direction ($X \to Y$). (b) Predictor and residuals on the true causal direction. (c) Scatter plot and fitted linear regression model on anti-causal direction ($Y \to X$). (d) Predictor and residuals on the anti-causal direction. As with LiNGAM, the predictor and residuals are independent if regressed in the true causal direction but not in the anti-causal direction.

to LiNGAM. Assuming additive noise, the causal structure is identifiable from the joint distribution if

- f_j are linear and N_j are non-Gaussian noises (Shimizu et al., 2006).
- f_j are non-linear (Hoyer et al., 2008a; Peters et al., 2014).
- f_j are linear and N_j are Gaussian with equal variances (Peters and Bühlmann, 2014).

The principle of identifying ANM is depicted in Figure 4.2. As with LiNGAM, comparing Figures 4.2 (b) and (d) show that the predictor and regression residuals are independent only when regressed in the true direction. ANM leverages non-Gaussianity and non-linearity to determine the causal direction.

Estimation

For non-linear cases, the underlying graph is estimated using regression with a subsequent independence test (RESIT) algorithm (Mooij et al., 2009; Peters et al., 2014). RESIT recursively performs regression and independence tests to estimate the topological order of variables, starting from the sink. A generalized additive model (Hastie and Tibshirani, 2017) or Gaussian process regression (Williams and Rasmussen, 1995) is used for regression. In addition, regression by minimizing the dependency between the predictor and residuals is proposed, which does not assume a specific probability distribution (Mooij et al., 2009). RESIT can consistently estimate the causal structure if the data follow the assumption and the perfect independence test is provided. However, unlike that for DirectLiNGAM, the computational cost for the independence test may be high when the sample size or number of variables is large.

4.1.3 LSNM: Conditional Variance Modulated by Predictors

Although ANM can represent a wide range of data-generating processes, it fails to determine the correct causal structure under a typical violation of the model: heteroscedasticity. An example of heteroscedastic data is depicted in Figure 4.3. The true causal direction is $X \to Y$. Figure 4.3 (a) shows a non-linear relation between X and Y and heteroscedasticity, where the value of X modulates the conditional variance of Y. Given the relationship, regressing on the true causal direction is insufficient: the predictor and residuals are not independent; hence, the causal direction cannot be determined. Conditional variance between the variables must be considered to identify the causal direction.

Model Definition

In LSNM or *heteroscedastic noise models* (HNM) (Immer et al., 2023; Kikuchi, 2023; Strobl and Lasko, 2023), the scale of the noise term is modulated by other variables. LSNM is defined as a collection of P equations of the following form:

$$X_j = f_j(\mathrm{PA}_j) + s_j(\mathrm{PA}_j)N_j, \quad j = 1, \ldots, P$$

where f_j and scaling function $s_j > 0$ can be non-linear. This type of heteroscedasticity $s_j(\mathrm{PA}_j)N_j$ is called multiplicative heteroscedasticity (Harvey, 1976). If s_j is a constant function, LSNM reduces to ANM. LSNM is identifiable in linear, nonlinear, and multivariate settings (Immer et al., 2023; Strobl and Lasko, 2023; Xu et al., 2022).

Estimation

Generalized root causal inference (GRCI) (Strobl and Lasko, 2023), an estimation method for LSNM, leverages a similar procedure as RESIT. The core difference is that, after performing a regression in which the effect of the variance remains, GRCI *regresses out* the effect by performing an additional regression on the conditional variance (more precisely, the conditional absolute deviation) given the predictor. To constrain the search space, GRCI first estimates the skeleton of $\mathscr{G}$ using the PC-stable algorithm (Colombo et al., 2014), recursively finds the sink, and obtains the causal order of the variables. Before evaluating the independence between the residual and predictor, GRCI partials out the effect of PA_j by transforming the regression residuals using their estimated conditional mean absolute deviation (MAD). GRCI then obtains a unique DAG estimation using the PC-stable algorithm with conditioning sets restricted to preceding variables, followed by an orientation of directed edges according to the causal order.

The principle of the identification of LSNM is shown in Figure 4.3. Figures 4.3 (b) and (c) show that the remaining effect of the variance induced by the predictor is regressed using the estimated conditional variance.

Other estimation methods involve heteroscedastic noise causal models (HEC) (Xu et al., 2022), which assume that N_j is a standard Gaussian variable and the distributions of X_j have compact support. Causal autoregressive flow (CAREFL) (Khemakhem et al., 2021) models the scaling functions by $e^{s(X)}$, where s represents autoregressive transformations. However, HEM and CAREFL are only applicable when $P = 2$.

Although GRCI considerably improves the estimation accuracy of recovering DAG under HNM, the numerical results indicate substantial performance degradation under a linear multivariate setting. We suspect that estimating the conditional MAD of regression residuals with splines is too flexible, making identifying causal direction based on the independence test unstable in a linear multivariate setting.

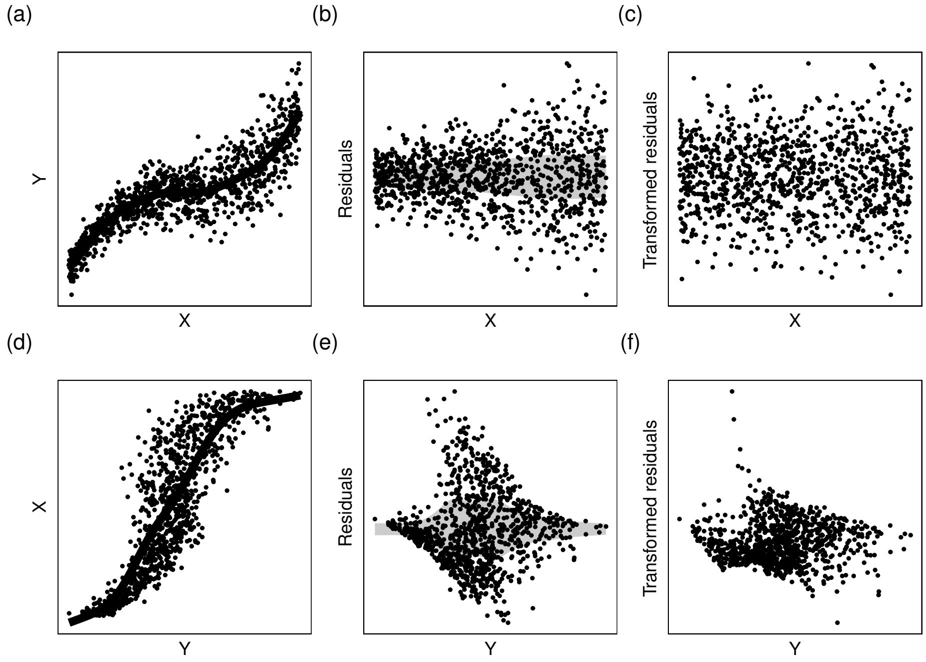


Fig. 4.3 Role of the location-scale noise in LSNM. (a) Scatter plot and fitted linear regression model on true causal direction ($X \to Y$). (b) Predictor and the residuals on the true causal direction. The gray area denotes the conditional standard deviation of the residuals given X. (c) Predictor and transformed residuals on the true causal direction. The predictor and the transformed residuals are independent. (d) Scatter plot and fitted linear regression model on anti-causal direction ($Y \to X$). (e) Predictor and residuals on the anti-causal direction. (f) Predictor and the transformed residuals on the anti-causal direction. By regressing out the effect of the variance, GRCI estimates the causal direction by leveraging the independence test.

4.1.4 TiMINo: Time-Series with Nonlinear Relations

Previously introduced LiNGAM, ANM, and LSNM assume that the data is independently and identically distributed (i.i.d.), indicating no temporal dependencies in the data. This subsection introduces an FCM that represents the time structure.

Model Definition

Let $X^t = \{X_1^t, X_2^t, ..., X_P^t\}$ be a set of P time-series. Assuming that the corresponding full time graph is acyclic, a *time-series model with independent noise* (TiMINo) (Peters et al., 2013) with maximum lag L is defined by a collection of P equations:

$$X_j^t = f_j\left(\mathrm{PA}_j^t, ..., \mathrm{PA}_j^{t-L}, N_j^t\right), \quad j = 1, ..., P \tag{4.3}$$

where $\mathrm{PA}_j^t \subseteq X^t \backslash X_j^t$ denote a set of instantaneous parents of X_j^t at present time t and $\mathrm{PA}_j^{t-\tau} \subseteq X^{t-\tau}(\tau > 1)$ denote a set of lagged parents, which is a set of variables with a direct connection from the previous timestep $X_i^{t-\tau}$ to X_j^t. N_j^t is a noise term that is jointly independent over j and t and for each j, and identically distributed in t. If $L = 0$, TiMINo reduces to ANM. TiMINo has been shown to be identifiable in the following cases (Peters et al., 2013):

1. Equation (4.3) consists of IFMOC. In this case, the corresponding summary time graph can contain cycles.
2. PA_j^t contains at least one $X_j^{t-\tau}$ (There is temporal dependency on each X_j^t), and the joint distribution is faithful w.r.t the underlying full time graph. In this case, the corresponding summary time graph must be acyclic. Thus, even when the functions and noises are not IFMOC (e.g., linear function and Gaussian noise), we can recover the causal structure using temporal dependencies.

From the latter, we can see that even when the functions and noises are not IFMOC (e.g. linear function and Gaussian noise), we can recover the causal structure using temporal dependencies.

Estimation

Estimation of TiMINo is done similarly to RESIT (Mooij et al., 2009; Peters et al., 2014), recursively searching a sink by performing regression and measuring independence between predictors and regression residuals. If TiMINo cannot find a sink that shows independence, it stops the iteration and returns the causal order identified so far.

4.2 Structure Learning with Continuous Optimization

The estimation methods for FCMs introduced in previous sections involve combinatorial optimization, in which the search grows exponentially with the number of variables. This section introduces *continuous optimization based structure learning methods*, a framework for score-based learning of CBNs/SEMs to perform estimations with a continuous optimization problem. For instance, given a dataset $\mathbf{X} = [\mathbf{x}_1|...|\mathbf{x}_P] \in \mathbb{R}^{N \times P}$ of P variables and N observations, many of the existing score-based structure learning methods seek to obtain a graph $\mathscr{G}$ that minimizes a certain score function $Q(\mathscr{G};\mathbf{X}) : \mathbb{G} \to \mathbb{R}$ (e.g., Akaike information criterion (AIC) (Akaike, 1974)) over the set of DAGs $\mathbb{G}$:

$$\min_{\mathscr{G}} Q(\mathscr{G};\mathbf{X}) \quad s.t. \quad \mathscr{G} \in \text{DAGs},$$

where the optimization is formulated as a combinatorial optimization problem. Conversely, continuous optimization-based methods convert the problem to a continuous optimization problem:

$$\min_{W} F(W;\mathbf{X}) \quad s.t. \quad h(W) = 0, \tag{4.4}$$

where W is a $P \times P$ (weighted) adjacency matrix corresponding to $\mathscr{G}$, and $h(W) = 0$ is an algebraic constraint that equals to zero if and only if W corresponds to DAG. A common score function for linear SEM is the squared loss function $F(W;\mathbf{X}) = \|\mathbf{X} - \mathbf{X}W\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm.

We start with the NOTEARS (Zheng et al., 2018) that gives an algebraic characterization of DAGs and formulates the structure learning problem as a continuous optimization problem. Then, we introduce NOTEARS-MLP (Zheng et al., 2020), which is an extension to the non-linear case using MLP, and finally explain the NTS-NOTEARS (Sun et al., 2021) that leverages CNN to exploit non-linear functional relations and temporal dependencies.

4.2.1 NOTEARS: Linear Relations

To formulate the algebraic constraint $h(W) = 0$, which is also referred to as the *acyclicity constraint*, Zheng et al. (2018) proposed using the trace of a matrix exponential, and provided the following theorem:

Theorem 4.5. *Let a matrix $W \in \mathbb{R}^{P \times P}$ be a weighted adjacency matrix corresponding to a graph $\mathscr{G}$. $\mathscr{G}$ is a DAG if and only if:*

$$h(W) = \operatorname{tr}\left(e^{W \circ W}\right) - P = 0, \tag{4.5}$$

where $\circ$ is the elementwise product, and e^A is the matrix exponential of A.

For simplicity, consider a binary adjacency matrix $B = \{0,1\}^{P\times P}$. A trace of an adjacency matrix $\mathrm{tr}(B^k) = \sum_{j=1}^{P}[B^k]_{j,j}$ shows the number of closed walks of length k. Therefore, B has no cycles if and only if $\sum_{k=1}^{\infty}\sum_{j=1}^{P}[B^k]_{j,j} = 0$. Using the matrix exponential $e^B = \sum_{k=0}^{\infty}\frac{B^k}{k!}$, we get

$$\begin{aligned}\mathrm{tr}\left(e^B\right) &= \sum_{k=0}^{\infty}\sum_{j=1}^{P}\frac{[B^k]_{j,j}}{k!}\\ &= \frac{\mathrm{tr}\,(I)}{0!} + \sum_{k=1}^{\infty}\sum_{j=1}^{P}\frac{[B^k]_{j,j}}{k!}\\ &= P + \sum_{k=1}^{\infty}\sum_{j=1}^{P}\frac{[B^k]_{j,j}}{k!}.\end{aligned} \tag{4.6}$$

From Equation (4.6), we can see that B is acyclic if and only if

$$\sum_{k=1}^{\infty}\sum_{j=1}^{P}\frac{[B^k]_{j,j}}{k!} = \mathrm{tr}\left(e^B\right) - P = 0. \tag{4.7}$$

As Equation (4.7) is valid for nonnegative matrix, it can be extended to the weighted adjacency matrix using the Hadamard product as in Equation (4.5) (Zheng et al., 2018). Moreover, Equation (4.5) satisfies the following desirable properties regarding optimization (Zheng et al., 2018):

1. The values of h quantify the "DAG-ness" of the graph ($h(W) > h(W^{'})$ indicates that W has more cycles than $W^{'}$ or the cycles in W are more weighted than in $W^{'}$);

2. h is smooth, and its derivatives can be computed easily ($\nabla h(W) = \left(e^{W\circ W}\right)^T \circ 2W$).

Using Equation (4.5) as a constraint, Zheng et al. (2018) formulated a structure-learning problem as a continuous optimization problem and proposed an estimation algorithm for linear data (NOTEARS). Let $\mathbf{X} = [\mathbf{x}_1|...|\mathbf{x}_P] \in \mathbb{R}^{N\times P}$ be a dataset comprising N independent and identically distributed (i.i.d.) observations. NOTEARS solves the following constrained optimization problem:

$$\min_{W}\frac{1}{2N}\|\mathbf{X}-\mathbf{X}W\|_F^2 + \lambda\|W\|_1 \quad \text{subject to} \quad h(W) = 0,$$

where $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_1 = \|\mathrm{vec}(\cdot)\|_1$ is the vector L^1-norm and λ denotes the penalty factor. The constrained optimization problem is converted into an unconstrained optimization problem using the augmented Lagrangian method, followed by optimization using L-BFGS (Byrd et al., 1995).

NOTEARS and its extension generally leverage squared loss, and it has been seen that this is equivalent to assuming standard Gaussian noise; the estimation is hindered when the assumption does not hold (Cai et al., 2021). The violation of the assumption can easily occur by scaling the variables (Kaiser and Sipos, 2021; Reisach et al., 2021), often referred to as *scale sensitivity*.

4.2.2 NOTEARS-MLP: Nonlinear Relations

NOTEARS was later extended to non-linear functional relationships using MLP, named NOTEARS-MLP (Zheng et al., 2020). Consider an MLP that consists of h hidden layers with m_l hidden units in each layer and an activation function σ, given by

$$\mathrm{MLP}(\mathbf{X}; A_j^{(1)}, ..., A_j^{(h+1)}) = A^{(h+1)}\sigma(\cdots A^{(2)}\sigma(A^{(1)}(\mathbf{X})),$$

where $A^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$ is a connectivity matrix with $m_0 = P, m_{h+1} = 1$. The sigmoid function is often used for the activation functions σ. Let $\theta_j = \{A_j^{(1)}, ..., A_j^{(h+1)}\}$ be parameters of the j-th MLP to predict $\mathrm{E}\left[X_j|\mathrm{PA}_j\right]$. Thus, P MLPs are fitted in total. Estimation of NOTEARS-MLP is formulated by the following optimization problem:

$$\min_{\theta} \frac{1}{N} \sum_{j=1}^{P} \ell\left(\mathbf{x}_j, \mathrm{MLP}(\mathbf{X}; \theta_j)\right) + \lambda \|A_j^{(1)}\|_1 \quad \text{s.t.} \quad h(W(\theta)) = 0,$$

where $\theta = (\theta_1, ..., \theta_P)$, and ℓ denotes the loss function, which the squared loss is typically used. A weighted adjacency for $W(\theta)$ for NOTEARS-MLP is calculated based on the first layer of each MLP by

$$[W(\theta)]_{kj} = \|[A_j^{(1)}]_{\cdot,k}\|_2,$$

where $[W(\theta)]_{k,j} = 0$ if MLP$(\mathbf{X}, \theta_k)$ is independent of X_j.

4.2.3 NTS-NOTEARS: Nonlinear Relations with Temporal Dependencies

NOTEARS and NOTEARS-MLP assume i.i.d. data, that is, there is no time structure. NOTEARS for nonparametric temporal DAGs (NTS-NOTEARS) (Sun et al., 2021) effectively captures temporal dependency using CNN.

Recall the notation for time-series introduced on TiMINo (4.3), where $X^t = \{X_1^t, X_2^t, ..., X_P^t\}$ is a set of P time-series. NTS-NOTEARS exploits the temporal dependencies by fitting CNNs to each X_j^t, where the first layer of each CNN is a convolutional layer with kernel size S, a stride of 1, and no padding; the parameters are expressed as ϕ_j, which is a set of weight matrices of shape $P \times (L+1)$. The corresponding weight of ϕ_j with respect to the target variable of the instantaneous step is set to zero to avoid estimating X_j^t using its own value. The remaining layers are fully connected layers with parameter ψ_j, which is a set of weight matrices. Therefore, parameters of NTS-NOTEARS is given by $\theta = (\theta_1, ..., \theta_P)$, where $\theta_j = (\phi_j, \psi_j)$.

Given N i.i.d. observations $\mathbf{X}^t = \{\mathbf{x}^{t,(n)}\}_{n=1}^N$, NTS-NOTEARS estimates using the following optimization problem:

$$\min_\theta F(\mathbf{X}^t, \theta) \quad \text{subject to} \quad h(W(\theta)) = \operatorname{tr}\left(e^{W(\theta)\circ W(\theta)}\right) - P = 0,$$

where

$$F(\mathbf{X}^t, \theta) = \frac{1}{N-L} \sum_{n=L+1}^{N} \sum_{j=1}^{P} \ell\left(\mathrm{x}_j^{t:t-L,(n)}, \mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j)\right)$$
$$+ \sum_{\tau=0}^{L} \lambda_1^{L-\tau} \|\phi_j^{(L-\tau)}\|_1 + \frac{1}{2}\lambda_2 \|\theta_j\|_2^2.$$

Here, $\lambda_1 = (\lambda^0, ..., \lambda^L)$ and λ_2 are regularization parameters. $\phi_j^{(L)}$ denotes a collection of the last column of the S convolutional kernels, corresponding to the weights of the instantaneous step. ℓ denotes the loss function, which squared loss is used for NTS-NOTEARS. As NTS-NOTEARS models temporal dependencies, the weighted adjacency matrix for the instantaneous step and lagged effects can be obtained:

$$[W^\tau(\theta)]_{i,j} = \|\text{i-th element across all } \phi_j(\tau)\|_2,$$

where $[W^\tau(\theta)]_{i,j}$ represents the connection strength between $X_i^{t-\tau}$ to X_j^t. Therefore, constraining acyclicity is only necessary for the instantaneous step ($\tau = 0$) and $W^0(\theta)$ is sufficient for the weighted adjacency matrix of NTS-NOTEARS (Sun et al., 2021).

4.3 Estimation of LSNM with Continuous Optimization

This section describes the contribution of Article I (Differentiable Causal Discovery under Heteroscedastic Noise), which focuses on developing a continuous optimization-based estimation method for LSNM (Section 4.1.3):

$$X_j = f_j(\mathrm{PA}_j) + s_j(\mathrm{PA}_j)N_j. \quad j = 1, ..., P$$

The major aspects that differentiate our estimation method from existing methods are as follows:

1. We model the conditional expectation of each variable and conditional variance to model the induced variance by the predictors.
2. We leverage approximation using log probability during optimization without assuming a specific probability distribution for the noise term, as opposed to the existing continuous optimization-based methods that implicitly assume standard Gaussian noise.

Here, we describe the proposed method given in Article I. Let p_j denote the probability density function of $\tilde{N}_j := s_j(\mathrm{PA}_j)N_j = X_j - f_j(\mathrm{PA}_j)$ and $(\sigma_j)^2$ be the conditional variance of $\tilde{N}_j$ given PA_j. Then, we can write the probability density function of $\tilde{N}_j$ standardized to unit variance by

$$\tilde{p}_j\left(\frac{\tilde{N}_j}{\sigma_j}\right) = \sigma_j p_j\left(\frac{\tilde{N}_j}{\sigma_j}\right). \tag{4.8}$$

Using Equation (4.8), given N i.i.d. observations $\mathbf{X} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$, the log-likelihood of LSNM is defined as follows:

$$\begin{aligned}\log \mathscr{L}(\mathbf{X}) &= \frac{1}{N}\log\prod_{n=1}^{N}\prod_{j=1}^{P} p_j\left(\frac{\tilde{N}_j^{(n)}}{\sigma_j^{(n)}}\right)\\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{P}\log\frac{1}{\sigma_j^{(n)}}\tilde{p}_j\left(\frac{\tilde{N}_j^{(n)}}{\sigma_j^{(n)}}\right)\\ &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{P}\log\sigma_j^{(n)} + \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{P}\log\tilde{p}_j\left(\frac{\mathrm{x}_j^{(n)} - f_j(\mathrm{PA}_j^{(n)})}{\sigma_j^{(n)}}\right),\end{aligned} \tag{4.9}$$

where $(\sigma_j^{(n)})^2$ represents the conditional variance of the *n-th* sample before standardization. Using the same approach of NOTEARS-MLP (Section 4.2.2), we adopt MLP with parameter set θ_j^A to model $f_j(\mathrm{PA}_j^{(n)})$, and MLP with parameter set θ_j^C to model $\sigma_j^{(n)}$ for each j, resulting in $2P$ MLPs. Therefore, unlike NOTEARS-MLP, we obtain weighted adjacency matrices, representing the connection strength of conditional expectations and variances. Following

Equation (4.5), an acyclicity constraint for our approach is given by

$$h(W(\theta^A), W(\theta^C)) = \operatorname{tr}\left(e^{W(\theta^A)\circ W(\theta^A)+W(\theta^C)\circ W(\theta^C)}\right) - P = 0, \tag{4.10}$$

where $\theta^A = (\theta_1^A, ..., \theta_P^A)$ and $\theta^C = (\theta_1^C, ..., \theta_P^C)$.

Many existing continuous optimization-based methods leverage square loss for the loss function, assuming that the noise terms follow standard Gaussian noise (Cai et al., 2021; Reisach et al., 2021). However, the noise distributions are unknown in practice and exhibit non-Gaussianity. Therefore, we use the estimation method of approximating log-probability $\log \tilde{p}_j$, in which we choose the approximation function from the two candidates according to whether the variable is super-Gaussian or sub-Gaussian (Hyvärinen and Oja, 1998; Hyvärinen et al., 2001). During optimization, $\log \tilde{p}_j$ is determined as follows:

$$\log \tilde{p}_j(z) = \begin{cases} -2\log\cosh(z) & \text{if } \gamma_j > 0, \\ -\left(z^2/2 - \log\cosh(z)\right) & \text{else} \end{cases}, \tag{4.11}$$

where scaler γ_j is calculated by

$$\gamma_j = \mathrm{E}\left[-\tanh(z)z + (1 - \tanh(z)^2)\right]. \tag{4.12}$$

γ_j is positive if z is super-Gaussian and negative if z is sub-Gaussian. We expect that the approximated $\log \tilde{p}_j$ improves the estimation performance compared to the squared loss function when the noise distribution is not Gaussian. Although the approximation assumes non-Gaussian noise distribution and symmetric probability density functions, the results of the numerical experiments indicate that a relatively high estimation accuracy can be obtained even when the noise terms follow a Gaussian or Gumbel distribution.

Using the (negative of the) log-likelihood (4.9) and the acyclicity constraint (4.10), the optimization problem for our method is given by

$$\min_{\theta^A, \theta^C} F(\mathbf{X}^t, \theta^A, \theta^C) \quad \text{subject to} \quad h(W(\theta^A), W(\theta^C)) = 0, \tag{4.13}$$

where

$$F(\mathbf{X}^t, \theta^A, \theta^C) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{P} \log \mathrm{MLP}(\mathbf{x}^{(n)}; \theta_j^C) - \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{P} \log \tilde{p}_j \left(\frac{\mathrm{x}_j^{(n)} - \mathrm{MLP}(\mathbf{x}^{(n)}; \theta_j^A)}{\mathrm{MLP}(\mathbf{x}^{(n)}; \theta_j^C)} \right)$$
$$+ \sum_{j=1}^{P} \left(\lambda_1 \|A_j^{(1)}\|_1 + \lambda_2 \|C_j^{(1)}\|_1 \right).$$

λ_1 and λ_2 are regularization parameters, and $A_j^{(1)}$ and $C_j^{(1)}$ denote the weight matrices of the first layer of the MLP for predicting conditional expectation and variance, respectively.

Following NOTEARS, the constrained optimization problem (4.13) is converted to an unconstrained optimization problem using the augmented Lagrangian method. After optimization using L-BFGS (Byrd et al., 1995), we estimate the weighted adjacency matrix as $\tilde{W}(\theta^A, \theta^C) = 1/2(W(\theta^A) + W(\theta^C))$ and round off the small values to zero with a small threshold $w > 0$ to remove redundant edges and remaining cycles in the graph (Zhou, 2009).

4.4 TS-LSNM: An Extension to Time-Series

Article II (Structure Learning for Group of Variables with Nonlinear Time-Series Data with Location-Scale Noise) generalizes the work on Article I to be capable of capturing temporal dependencies and provide a novel approach for learning causal structure among groups of variables with a continuous optimization-based approach. We first propose a new FCM that extends LSNM (Section 4.1.3) to time series, which is a special case of TiMINo (Section 4.1.4) and show that the proposed model is structurally identifiable under some assumptions. Then, we provide a corresponding estimation method leveraging the same approach as NTS-NOTEARS (Section 4.2.3).

4.4.1 Model Definition

We first define a novel FCM: *time-series location-scale noise model* (TS-LSNM).

Definition 4.6 (Time-Series Location-Scale Noise Model). *Let $X^t = \{X_1^t, X_2^t, ..., X_P^t\}$ be a set of P time series, and assume that the corresponding full time graph is acyclic. A* time-series location-scale noise model *(TS-LSNM) with maximum lag $L \geq 0$ is defined by a collection of P equations:*

$$X_j^t = f_j\left(\mathrm{PA}_j^t, ..., \mathrm{PA}_j^{t-L}\right) + s_j\left(\mathrm{PA}_j^t, ..., \mathrm{PA}_j^{t-L}\right) N_j^t, \quad j = 1, ..., P \tag{4.14}$$

where $\mathrm{PA}_j^t \subseteq X^t \backslash X_j^t$ *denote a set of instantaneous parents of* X_j^t *at time* t *and* $\mathrm{PA}_j^{t-\tau} \subseteq X^{t-\tau}$ $(\tau > 0)$ *denote a set of lagged parents, which is a set of variables with a direct connection from the previous time step to* X_j^t *(e.g.* $X_i^{t-\tau}$ *to* X_j^t*).* N_j^t *is a noise term that is mutually independent over* j *and* t*, and identically distributed in* t*.* f_j *and scaling function* $s_j > 0$ *are at least twice differentiable functions. Additionally, we assume causal stationarity (Runge, 2018) and causal minimality (Peters et al., 2011; Spirtes et al., 2000); we also assume that the joint distribution of* X^t *satisfies the causal Markov property with respect to the underlying graph.*

TS-LSNM is a special case of TiMINo (Section 4.1.4), in particular, if no temporal dependency is assumed $(L = 0)$, TS-LSNM is reduced to LSNM (Section 4.1.3).

4.4.2 Structural Identifiability

The structural identifiability of TS-LSNM comes from the identifiability results of LSNM and TiMINo. Note that there are some exceptions, that TS-LSNM, as well as LSNM, is not identifiable in some pathological cases (e.g., linear Gaussian with constant function s_j and no temporal dependency) (Immer et al., 2023; Strobl and Lasko, 2023).

Corollary 4.7 (Structural identifiability of TS-LSNM)**.** *TS-LSNM given in Definition 4.6 is structurally identifiable from the joint distribution.*

Proof. As LSNM belongs to an IFMOC (Definition 4.4), it is structurally identifiable under the assumption of causal minimality, no cycles and no latent confounders (Strobl and Lasko, 2023). Because we can recover the underlying graph of TiMINo if the data-generating functions come from the IFMOC (Peters et al., 2013), TS-LSNM is also structurally identifiable from the joint distribution. □

The main advantage of handling various characteristics of the data is that we can use as much available information as possible. For instance, for the linear Gaussian data we may capture the heteroscedasticity to determine the causal direction, and even for the linear Gaussian data without heteroscedasticity, we might able to exploit the temporal dependencies to infer causal structure.

4.4.3 Estimation Algorithm

In this subsection, we briefly describe the estimation method for TS-LSNM. Set $\tilde{\mathrm{PA}}_j = \cup_{\tau=0}^{L} \mathrm{PA}_j^{t-\tau}$ and $\tilde{N}_j^t := s_j(\tilde{\mathrm{PA}}_j) N_j^t$. Given N observations $\mathbf{X}^t = \{\mathbf{x}^{t,(n)}\}_{n=1}^{N}$, log-likelihood of

TS-LSNM (4.14) is calculated as follows:

$$\log \mathscr{L}(\mathbf{X}^t) = -\frac{1}{N-L}\sum_{n=L+1}^{N}\sum_{j=1}^{P}\log\sigma_j^{t,(n)} + \frac{1}{N-L}\sum_{n=L+1}^{N}\sum_{j=1}^{P}\log\tilde{p}_j\left(\frac{\mathrm{x}_j^{t,(n)} - f_j(\tilde{\mathrm{PA}}_j^{(n)})}{\sigma_j^{t,(n)}}\right), \tag{4.15}$$

where $\tilde{p}_j$ denotes the probability density functions of noise $\tilde{N}_j^t$ standardized to unit variance and $(\sigma_j^{t,(n)})^2$ represents the conditional variance of the *n-th* sample before standardization.

We adopt the same approach as NTS-NOTEARS (Section 4.2.3), which leverages CNNs to capture temporal dependencies and model each variable's conditional expectation and variance separately for heteroscedasticity, as presented in the previous section. Therefore, we create two CNNs to estimate $f_j(\tilde{\mathrm{PA}}_j^{(n)})$ and $\sigma_j^{t,(n)}$ for each target variable, resulting in $2P$ CNNs. The estimations $\hat{f}_j(\tilde{\mathrm{PA}}_j^{(n)})$ and $\hat{\sigma}_j^{t,(n)}$ are given by CNNs with parameters $\theta_j^A = (\phi_j^A, \psi_j^A)$ and $\theta_j^C = (\phi_j^C, \psi_j^C)$, respectively:

$$\begin{aligned}
\hat{f}_j(\tilde{\mathrm{PA}}_j^{(n)}) &= \mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^A), \\
\hat{\sigma}_j^{t,(n)} &= \mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^C),
\end{aligned}$$

where ϕ and ψ denote the parameter of the convolutional layer and fully connected layers, respectively. Figure 4.4 illustrates how CNNs capture time structure and heteroscedasticity and obtain $f_j(\tilde{\mathrm{PA}}_j^{(n)})$ and $\sigma_j^{t,(n)}$. When $L = 0$, $\mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^A)$ and $\mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^C)$ with kernel size S can be represented by $\mathrm{MLP}(\mathbf{x}^{(n)}; \theta_j^A)$ and $\mathrm{MLP}(\mathbf{x}^{(n)}; \theta_j^C)$ with S nodes on the first hidden layer, respectively.

Following NTS-NOTEARS, the weighted adjacency matrix for TS-LSNM is calculated using the weights of the convolutional layers of CNNs. We calculate two weighted adjacency matrices $W^\tau(\theta^A)$ and $W^\tau(\theta^C)$ for each time lag τ, representing the overall connection strengths with respect to the conditional expectations and variances, respectively. To obtain a weighted adjacency matrix $W^\tau(\theta^A, \theta^C)$ that represents the connection strengths of the conditional expectation and variance, we calculate $W^\tau(\theta^A, \theta^C) = W^\tau(\theta^A) + W^\tau(\theta^C)$, where element i, j of $W^\tau(\theta^A, \theta^C)$ indicates the overall connection strength from $X_i^{t-\tau}$ to X_j^t. $W^0(\theta^A, \theta^C)$ represents the dependency structure of the current time step t.

Finally, using the log-likelihood (4.15) with acyclicity constraint (4.5) and regularization terms with respect to the model weights $\theta^A = (\theta_1^A, ..., \theta_P^A)$ and $\theta^C = (\theta_1^C, ..., \theta_P^C)$, we obtain the following constrained optimization problem for TS-LSNM:

$$\min_{\theta^A, \theta^C} F(\mathbf{X}^t, \theta^A, \theta^C) \quad \text{subject to} \quad h(W^0(\theta^A, \theta^C)) = 0, \tag{4.16}$$

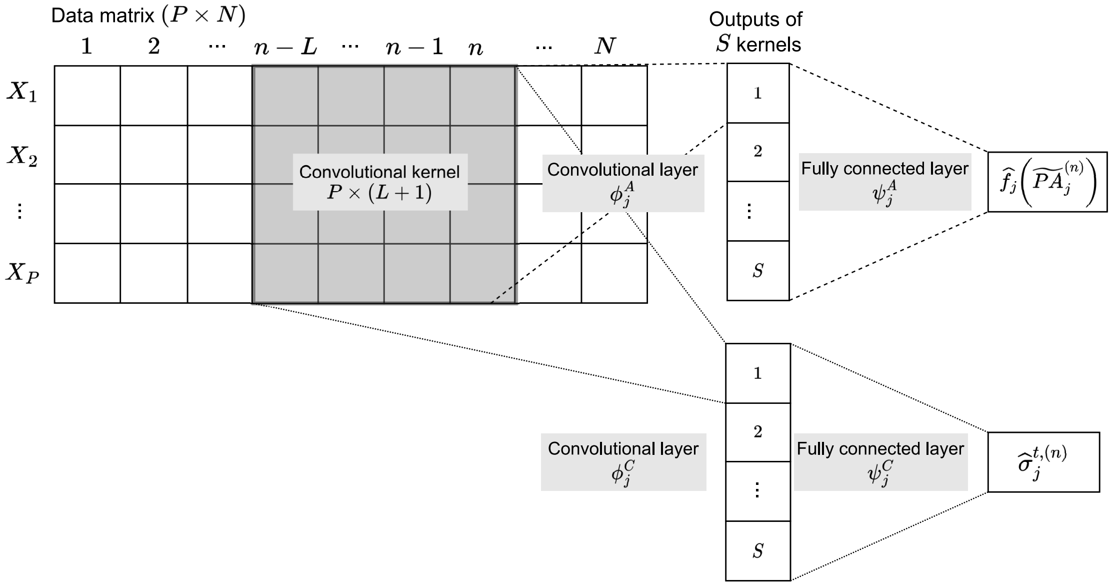


Fig. 4.4 Schematic of how CNNs are used to capture time structure and heteroscedasticity

where

$$
\begin{aligned}
F(\mathbf{X}^t, \theta^A, \theta^C) = & \frac{1}{N-L} \sum_{n=L+1}^{N} \sum_{j=1}^{P} \log \mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^C) \\
& - \frac{1}{N-L} \sum_{n=L+1}^{N} \sum_{j=1}^{P} \log \tilde{p}_j \left(\frac{\mathrm{x}_j^{t:t-L,(n)} - \mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^A)}{\mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^C)} \right) \\
& + \sum_{j=1}^{P} \left(\lambda_1 \|\phi_j^A\|_1 + \lambda_1 \|\phi_j^C\|_1 + \frac{1}{2}\lambda_2 \|\theta_j^A\|_2^2 + \frac{1}{2}\lambda_2 \|\theta_j^C\|_2^2 \right).
\end{aligned}
$$

Here, λ_1 and λ_2 are regularization parameters, and $\log \tilde{p}_j$ is calculated using the approximation of the log probability (4.11). Following previous works, the constrained optimization problem (4.16) is converted to an unconstrained optimization problem using the augmented Lagrangian method. L-BFGS is used for optimization, followed by a post-processing step that rounds off the small values of $W^\tau(\theta^A, \theta^C)$ to zero with a small threshold $w > 0$ for each τ.

4.5 Numerical Experiments

In this section, we report the results of numerical experiments on synthetic data conducted to assess the performance of TS-LSNM. We first present the results on data with no temporal dependencies in Section 4.5.1. Then, we show the results on time-series data in Section 4.5.2.

4.5.1 Linear/Nonlinear Data with no Temporal Dependencies

Here, we present the results on data with heteroscedasticity for linear cases and nonlinear cases, but with no temporal dependencies. Baseline algorithms selected for comparison are NOTEARS (Zheng et al., 2018) and DirectLiNGAM (Shimizu et al., 2011) for the linear cases, and NOTEARS-MLP (Zheng et al., 2020), RESIT (Peters et al., 2014) and GRCI (Strobl and Lasko, 2023) for the non-linear cases. We also included the performance of the empty graph as a naive baseline. DirectLiNGAM and RESIT are included to investigate the performance of the independence-based estimation algorithms, which do not assume heteroscedasticity. For each algorithm, we used the parameter set defined in the corresponding study and codes.

Setup

In the experiments, ground truth DAGs were generated from Erdös-Rényi model with $2P$ edges (ER2). For the linear cases, each variable was generated by

$$X_j = \sum_{i \in \mathrm{pa}(j)} [W]_{i,j} X_i + s_j(\mathrm{PA}_j) N_j,$$

where $\mathrm{pa}(j)$ denotes the index set of PA_j. For the non-linear cases, we used index models

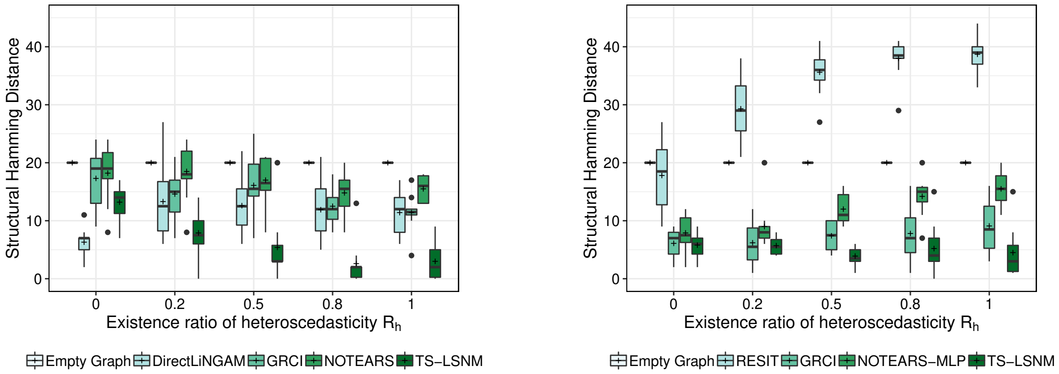
$$X_j^t = \tanh\left(f_j^{(1)}(\mathrm{PA}_j)\right) + \cos\left(f_j^{(2)}(\mathrm{PA}_j)\right) + \sin\left(f_j^{(3)}(\mathrm{PA}_j)\right) + s_j\left(\mathrm{PA}_j\right) N_j^t, \tag{4.17}$$

where

$$f_j^{(w)} = \sum_{i \in \mathrm{pa}(j)} [W^{(w)}]_{i,j} X_i. \quad w = 1,2,3$$

For the scaling function, we used $s_j(\mathrm{PA}_j) = \exp\left(\sum_{i \in \mathrm{pa}(j)} [C]_{i,j} X_i\right)$, and noise terms were generated from $N_j^t \sim \mathrm{U}(-1/\sqrt{3}, 1/\sqrt{3})$. Each non zero element of $P \times P$ weight matrix W and C were drawn randomly from $[W]_{i,j} \sim \pm\mathrm{U}(0.5, 2.0)$ and $[C]_{i,j} \sim \pm\mathrm{U}(0.4, 0.8)$, respectively. All variables were scaled to zero mean and unit variance. Each experiment ran 10 times, and structural Hamming distance (lower the better) was used as an evaluation metric.

We introduced an existence ratio of heteroscedasticity R_h, which controls how often heteroscedasticity occurs in the generated data. We first generated a binary matrix $B \in \{0,1\}^{P \times P}$ from ER2 graph, and then randomly selected each non-zero element of B with probability R_h and obtained a binary matrix $\tilde{B}$. For example, if $R_h = 0$ we get $\tilde{B}$ with all



(a) Linear case

(b) Nonlinear case

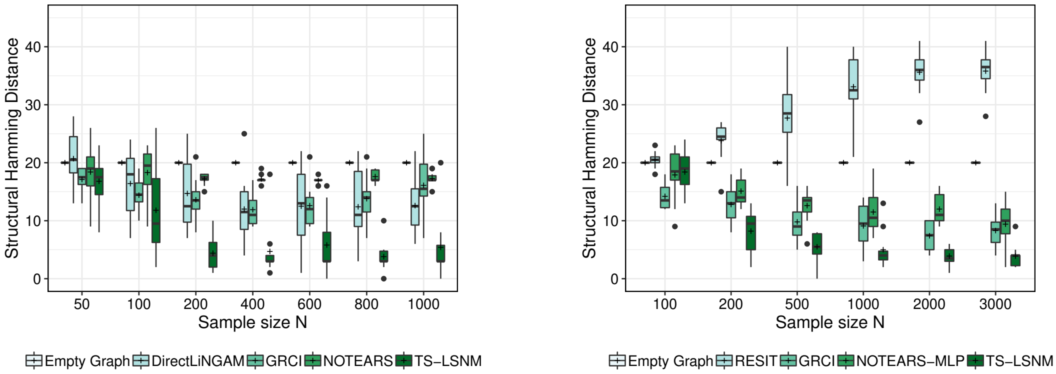
Fig. 4.5 Result on changing existence ratio of heteroscedasticity R_h

zeros, and if $R_h = 1.0$, we get $\tilde{B} = B$. Thereafter, we generated each element of W and C according to B and $\tilde{B}$, respectively.

We used for TS-LSNM the same parameter settings as for NOTEARS and NOTEARS-MLP, where $\lambda_1 = 0.01$, $\lambda_2 = 0.01$, $w = 0.3$, and sigmoid function was used for the activation function on the hidden layer. The maximum time lag was set to $L = 0$, which means that we did not estimate temporal dependencies in this experiment. MLP with no hidden layer on the linear cases and one hidden layer with 10 nodes on non-linear cases was used for modeling the conditional expectations, and MLP with one hidden layer with 10 nodes for both linear and non-linear cases for the conditional variances. The exponential activation function was applied to the output of conditional variances to ensure positivity. To improve convergence, we initialized the weights of TS-LSNM with the fitted result of NOTEARS on the linear cases and NOTEARS-MLP on the non-linear cases.

Results

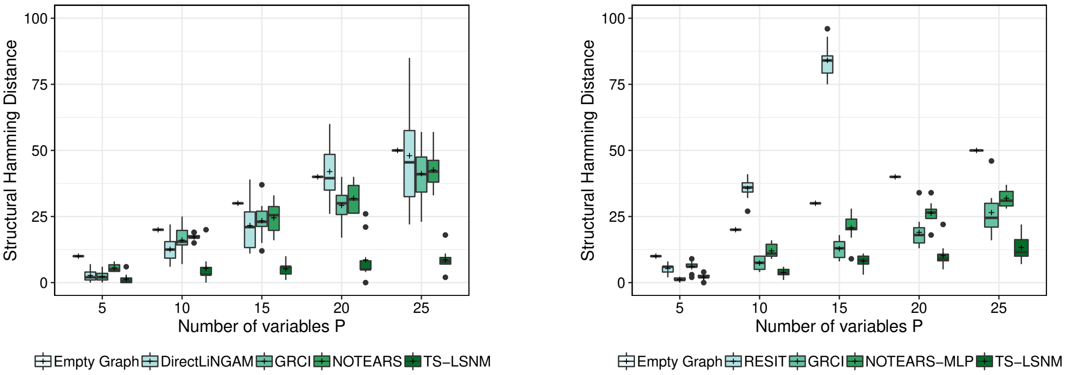
We first conducted an experiment with different existence ratios of heteroscedasticity R_h. We changed $R_h = \{0, 0.2, 0.5, 0.8, 1.0\}$ with number of variables $P = 10$, sample size $N = 1000$ for the linear case and $N = 2000$ for the non-linear case. The results are given in Figure 4.5. As R_h increases, TS-LSNM generally outperformed the others. TS-LSNM outperformed NOTEARS and NOTEARS-MLP under no heteroscedasticity ($R_h = 0$). This verifies the effect of using log-likelihood with variance estimation as an objective function, compared to the squared loss function, which is equivalent to the standard Gaussian noise assumption. GRCI matched TS-LSNM in the non-linear case, although its performance degraded in the linear case. DirectLiNGAM and RESIT did not perform well even in small R_h.



(a) Linear case

(b) Nonlinear case

Fig. 4.6 Result on changing sample size N



(a) Linear case

(b) Nonlinear case

Fig. 4.7 Result on changing number of variables P

Then, to evaluate the estimation performance under different sample sizes, we set $R_h = 0.5$, $P = 10$ and changed $N = \{50, 100, 200, 400, ..., 1000\}$ for the linear case and $N = \{50, 100, 200, 500, 1000, 2000, 3000\}$ for the non-linear case. As shown in Figure 4.6, TS-LSNM generally outperformed others in different sample sizes. In this setting, TS-LSNM needed at least $N = 200$ in the linear case and $N = 1000$ for the non-linear case, thanks to the initialization using fitted NOTEARS and NOTEARS-MLP. From Figure 4.6a, the performance of GRCI on the linear case gradually decreases from $N > 400$, which implies that there is a case that GRCI converges to the wrong solution with large sample size.

Finally, to investigate the change of estimation performance under different number of variables, we set $R_h = 0.5$ and changed $P = \{5, 10, 15, 20, 25\}$ with $N = 1000$ for the linear case and $N = 2000$ for the non-linear case. As shown in Figure 4.7, TS-LSNM considerably outperformed the others. In Figure 4.7b, GRCI outperformed TS-TS-LSNM only in the non-linear case with a small number of variables $P = 5$. TS-LSNM performed better on $p > 5$ compared to GRCI and the difference in performance got larger as P increased. This is likely to be the effect of the GRCI estimation procedure, which recursively finds sink nodes, where mistakes that occurred in the preceding iterations propagate through the whole procedure. TS-LSNM does not have that property, which is an advantage of continuous optimization-based methods.

4.5.2 Nonlinear Time-Series Data with Location-Scale Noise

In this section, we show the results of numerical experiments on non-linear time-series data with heteroscedasticity. We compared TS-LSNM with NTS-NOTEARS (Sun et al., 2021), which can capture non-linearity as well as temporal dependency, though does not model heteroscedasticity.

Setup

In this experiment, ground truth DAGs were generated from Erdös-Rényi model with $2P$ edges (ER2). For the time-lagged effects, following (Sun et al., 2021), we created a connection from $X_i^{t-\tau}$ to X_j^t with a probability of $1/P$, which indicates that on average, there was one connection from each $X_i^{t-\tau}$ to X_j^t. After generating the connections between the variables, similar to Equation (4.17), each variable was generated using the following function based on the index models:

$$X_j^t = \tanh\left(f_j^{(1)}(\tilde{\mathrm{PA}}_j)\right) + \cos\left(f_j^{(2)}(\tilde{\mathrm{PA}}_j)\right) + \sin\left(f_j^{(3)}(\tilde{\mathrm{PA}}_j)\right) + s_j\left(\tilde{\mathrm{PA}}_j\right) N_j^t, \tag{4.18}$$

where

$$f_j^{(w)} = \sum_{\tau=0}^{L} \sum_{i \in \mathrm{pa}^\tau(j)} [W_\tau^{(w)}]_{i,j} X_i^{t-\tau}. \quad w = 1,2,3$$

For the scaling function s_j, for each j, we randomly selected a strictly positive nonlinear function from a set

$$\{1/(1+\exp\left(g(\tilde{\mathrm{PA}}_j)\right)) + 0.5,\ \exp\left(g(\tilde{\mathrm{PA}}_j)\right),\ \tanh(g(\tilde{\mathrm{PA}}_j)) + 1.5\}, \tag{4.19}$$

where

$$g(\tilde{\mathrm{PA}}_j) = \sum_{\tau=0}^{L} \sum_{i \in \mathrm{pa}^\tau(j)} [C_\tau]_{i,j} X_i^{t-\tau}.$$

The connection weights $W_\tau^{(1)}, W_\tau^{(2)}, W_\tau^{(3)}$ were sampled from $\pm U(0.5, 2.0)$, and C_τ was sampled from $\pm U(0.4, 0.8)$.

The parameters for TS-LSNM and NTS-NOTEARS were determined by performing a grid search in the condition of $P = 20$, and $N_j^t \sim \mathrm{U}(-1/\sqrt{3}, 1/\sqrt{3})$ with parameter space $\lambda_1 = \lambda_2 \in \{0.05, 0.01, 0.005\}, w \in \{0.3, 0.2, 0.1\}$, resulting in $\lambda_1 = \lambda_2 = 0.005$ for both methods, $w = 0.1$ for NTS-NOTEARS and $w = 0.2$ for TS-LSNM. The number of hidden

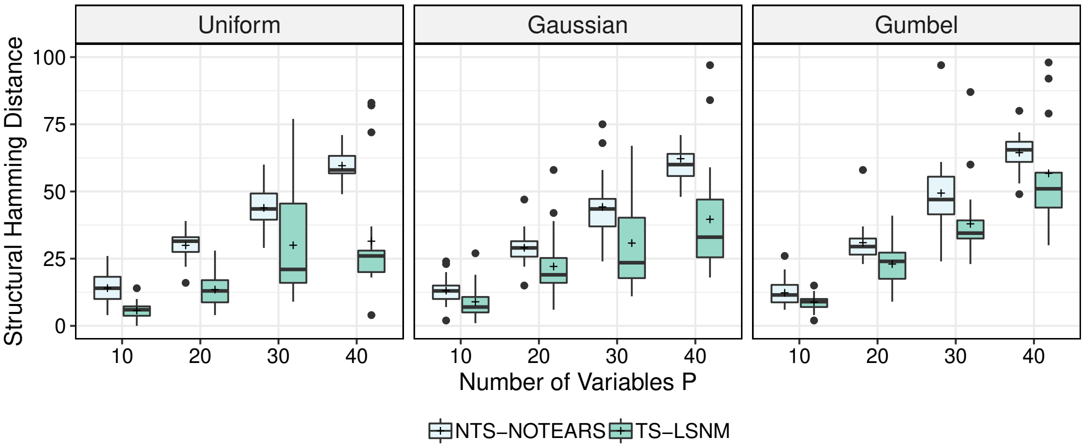


Fig. 4.8 Results on synthetic data using different number of variables P and different noise distributions (lower the better)

layers was set to 1, and the kernel size S was set to 10. The maximum length L of the temporal dependencies of the data and models was set to 1.

We conducted experiments on the combination of different numbers of variables $P = \{10, 20, 30, 40\}$ and noise distributions $N_j^t \sim \{\mathrm{U}(-1/\sqrt{3}, 1/\sqrt{3}), \mathcal{N}(0,1), \mathrm{Gumbel}(0, \sqrt{6}/\pi)\}$. We generated 2000 data points, scaled all the variables to zero-mean unit variance, and shuffled the column order. We used the structural Hamming distance as an evaluation metric, and each experiment was performed 20 times.

Results

The results are shown in Figure 4.8 (lower the better). We can see that TS-LSNM generally outperformed NTS-NOTEARS with respect to SHD, indicating the effectiveness of modeling heteroscedasticity. We can also see that the estimation performance of TS-LSNM decreased when the noise term follows the Gaussian distribution or the Gumbel distribution; those violate the symmetric assumption of the log-probability (Equation (4.11)), though we still obtained better results than NTS-NOTEARS. Both methods occasionally output large SHD, possibly caused by a convergence on a local minimum. In practice, one can fit the model several times and adopt a result that minimizes the objective function.

Chapter 5

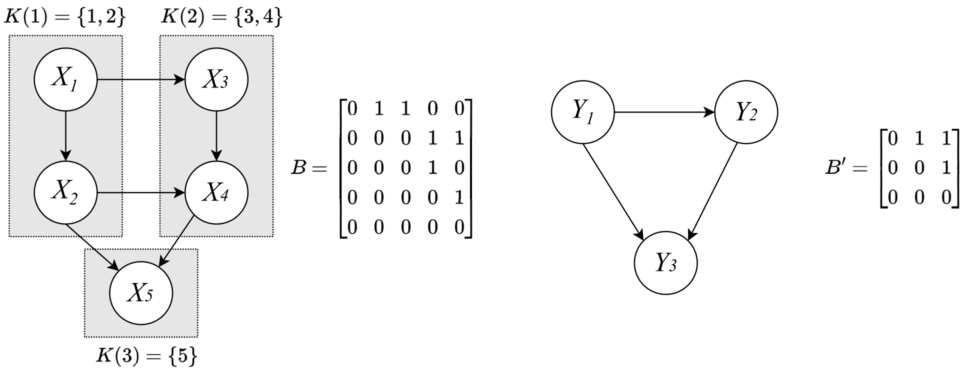
Causal Discovery for Groups of Variables on Data Beyond Linear Functional Relations

In this chapter, we describe part of the contributions of Article II (Structure Learning for Group of Variables with Nonlinear Time-Series Data with Location-Scale Noise), which proposes a novel method for learning causal structure among groups of variables not limited to data with linear functional relations. We first explain the problem definition in Section 5.1 and then introduce existing methods based on FCMs in Section 5.2. After that, we introduce the proposed method in Section 5.3 and the results of numerical experiments on synthetic data in Section 5.4.

5.1 Problem Definition

In this section, we introduce a problem definition of performing causal discovery when there exists a *group of variables*. For instance, the relationship between brain regions rather than the individual measurement positions in functional magnetic resonance imaging (fMRI) data is of interest to researchers in neuroscience (Smith et al., 2011). In manufacturing, multiple measurements obtained from the same machine show relatively strong correlations. Therefore, one can consider obtaining a graph representing causal structure among groups of variables, which is more comprehensive compared to a graph of individual variables.

A standard approach to infer such a graph is to aggregate variables by calculating, for example, the sum of the variables in the same group (Scheines and Spirtes, 2008). Another option is to select one variable per group (Marazopoulou et al., 2016). Although


$K(1) = \{1, 2\}$
$K(2) = \{3, 4\}$
X_1
X_3
X_2
X_4
X_5
$K(3) = \{5\}$
$B = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$
Y_1
Y_2
Y_3
$B' = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

(a) Variable DAG

(b) Group DAG

Fig. 5.1 Example of a variable DAG and a group DAG.

these approaches can reduce the computation time by reducing dimensionality, they reduce the performance of existing causal discovery methods due to changes in the conditional dependencies between variables (Scheines and Spirtes, 2008; Spirtes et al., 2000) or the cancellation of dependence (Wahl et al., 2023).

Here we first describe the problem definition using the terms based on (Parviainen and Kaski, 2017). Recall that a graph $\mathscr{G}$ representing the causal structure of P variables can be parameterized by the adjacency matrix $B \in \{0, 1\}^{P \times P}$, where $[B]_{i,j} = 1$ if and only if a direct connection from X_i to X_j exists. Assume that each variable belongs to one of $M (M \leq P)$ groups, and let $K = \{K(1), ..., K(M)\}$ be a set of index sets for each group. Groups $\{1, ..., M\}$ is an ordered set, where group l has no connection to group $k (< l)$. Let $Y = \{Y_1, ..., Y_M\}$ be a supervertex obtained by contracting variables X in the same group on $\mathscr{G}$. A graph on X is called a *variable graph* and a graph on Y is called a *group graph*; the corresponding adjacency matrices $B \in \{0, 1\}^{P \times P}$ and $B^{'} \in \{0, 1\}^{M \times M}$ are referred to as *variable adjacency matrices* and *group adjacency matrices*, respectively.

An example of a variable DAG with grouping K and the corresponding group DAG is shown in Figure 5.1. $B^{'}$ encapsulates the connections between the groups, where $[B^{'}]_{k,l} = 1$ if and only if $\exists [B]_{i \in K(k),\ j \in K(l)} = 1$. The group graph of $\mathscr{G}$ is further assumed to be DAG, where $\mathscr{G}$ is called *group-acyclic* given the grouping K.

The goal is to estimate B' from observational data, which we call the corresponding graph, *group DAG*. Many existing causal discovery methods perform estimation under $M = P$, e.g., the number of groups is equal to that of variables; we call this the corresponding graph, *variable DAG*.

5.2 Existing Methods

This section introduces existing works on causal discovery among the group of variables based on FCMs. There are several works we do not include the details: Causal autoregressive flow model (Khemakhem et al., 2021) leveraging normalizing flows are used to capture data non-linearity. The two group vector causal inference method (Wahl et al., 2023) can also capture non-linearity; however, they are limited to inference of a causal direction between two groups and cannot be applied to three or more groups. For the conditional-independence-based structure-learning methods, which is out of scope for this thesis, Parviainen and Kaski (2017) proposed an estimation method in which a directed acyclic graph is constructed over the individual variables to infer connections between groups.

5.2.1 GroupLiNGAM: Infer Groupings and Relations

Model Definition

GroupLiNGAM (Kawahara et al., 2010) is a generalization of LiNGAM used to describe causal structure among a group of variables. Let $l(j)$ be an index of a group to which a variable X_j belongs. Then, GroupLiNGAM is represented by a collection of P equations, as follows:

$$X_j = \sum_{l(i)<l(j), i\neq j} b_{ji} X_i + N_j, \quad j = 1, \dots, P \tag{5.1}$$

where noise terms N_j are generated from non-Gaussian distributions, are mutually independent over the groups, and need not be independent of each other in the same group. By setting a vector of variables in group l as $\boldsymbol{X}_l = X_{j\in K(l)}$ and a vector of corresponding noise terms as $\boldsymbol{N}_l = N_{j\in K(l)}$, Equation (5.1) can be represented by

$$\boldsymbol{X}_l = \sum_{k<l} \mathbf{B}_{l,k} \boldsymbol{X}_k + \boldsymbol{N}_l, \quad l = 1, \dots, M$$

where $\mathbf{B}_{l,k}$ is a matrix of connection strength from a vector of group k to group l. Each group of variables is generated from the variables in preceding groups.

Estimation

The corresponding estimation algorithm, also termed GroupLiNGAM, seeks to find the causal structure among groups and the grouping of the variables. Given that a group of variables $\boldsymbol{X}_l$ is exogenous if and only if $\boldsymbol{X}_l$ is independent of the residuals obtained by regressing $\boldsymbol{X}_s = X_{j\notin K(l)}$ on $\boldsymbol{X}_l$ (Kawahara et al., 2010), GroupLiNGAM recursively finds an exogenous group until the full causal order of the group is inferred. The search space for finding the grouping of the variables and the exogenous group grows exponentially in the number of variables, hence the algorithm is infeasible for a large number of variables.

5.2.2 DirectGroupLiNGAM: Efficient Estimation Under Known Groupings

DirectGroupLiNGAM (Entner and Hoyer, 2012) is an extension of DirectLiNGAM for estimating the causal structure among groups of variables, assuming that the grouping of the variables is known in prior. The estimation algorithm is similar to DirectLiNGAM, where regression and identification of an exogenous group are recursively performed based on the observation of the following lemma (Entner and Hoyer, 2012):

Lemma 5.1. *Suppose that a data is generated from GroupLiNGAM:*

$$\boldsymbol{X}_l = \sum_{k<l} \mathbf{B}_{l,k}\boldsymbol{X}_k + \boldsymbol{N}_l. \quad l = 1,...,M$$

Let $\boldsymbol{R}_l^k := \boldsymbol{X}_l - \mathbf{C}\boldsymbol{X}_k$ *be the regression residuals when regressing* $\boldsymbol{X}_l$ *on* $\boldsymbol{X}_k$ *using ordinary least squares. Then, group* k *is exogenous if and only if* $\boldsymbol{X}_k \perp\!\!\!\perp \boldsymbol{R}_l^k$ *satisfies for all* $l \neq k$.

The main difference in the estimation algorithm lies in how an exogenous group is identified; three approaches are provided (Entner and Hoyer, 2012). The typical approach termed DirectGroupLiNGAM combines p-values obtained by the independence test under the null hypothesis of $\boldsymbol{X}_k \perp\!\!\!\perp \boldsymbol{R}_l^k$ with the Fisher's method (Fisher, 1970) and selects a group that likely minimizes the p-values.

5.3 Causal Discovery on Groups of Variables with Continuous Optimization

In this section, we give a summary of the second major contribution of Article II on developing an estimation for groups of variables that can be used with the estimation of TS-LSNM and other existing continuous optimization-based methods. This is done by defining an algebraic constraint that captures the causal structure among groups of variables.

Recall that a variable graph with P variables represented by a binary adjacency matrix $B \in \{0,1\}^{P\times P}$ with grouping $K = \{K(1),...,K(M)\}$ of M groups can be converted to a group graph represented by a group adjacency matrix $B^{'} \in \{0,1\}^{M\times M}$ (Section 5.1). We extend B and $B^{'}$ to the weighted adjacency matrix. Suppose we have a weighted adjacency matrix $W \in \mathbb{R}^{P\times P}$ representing the connection strength between individual variables. The corresponding weighted group adjacency matrix $W^{'} \in \mathbb{R}^{M\times M} \geq 0$ can be calculated as follows:

$$[W^{'}]_{k,l} = \begin{cases} 0 & \text{if } k = l, \\ \sum_{i\in K(k)} \sum_{j\in K(l)} [W \circ W]_{i,j} & \text{else} \end{cases}. \tag{5.2}$$

Here, $[W^{'}]_{k,l}$ denotes the total amount of squared connection strengths from the variables in groups k to that in group l. We do not claim that this calculation is optimal; for example, we can use the absolute values of W to calculate $W^{'}$. The diagonal elements of $W^{'}$ are set to zero to enable connections between variables within the same group.

By substituting $W^{'}$ into the algebraic constraint h (4.5), we obtain a constraint for the group DAGs, which we call the *group DAG constraint*:

$$h(W^{'}) = \mathrm{tr}(e^{W^{'}\circ W^{'}}) - M = 0. \tag{5.3}$$

The group DAG constraint is satisfied if and only if the corresponding variable graph is group-acyclic.

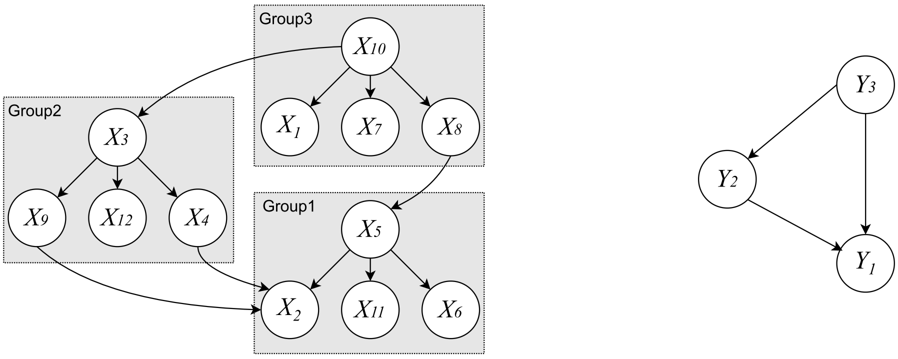
Corollary 5.2 (Acyclicity of group DAGs)**.** *A variable graph $\mathscr{G}$ with grouping K represented by a weighted adjacency matrix W is group-acyclic if and only if the constraint in Equation (5.3) is satisfied.*

Proof. From Equation (5.2), we can see that $[W^{'}]_{k,l} > 0$ if and only if $\exists[W]_{i\in K(k),j\in K(l)} \neq 0$; thus, $W^{'}$ represents the weighted adjacency matrix of the group graph of $\mathscr{G}$. From Theorem 4.5, $\mathrm{tr}(e^{W^{'}\circ W^{'}}) - M = 0$ is satisfied if and only if the group graph of $\mathscr{G}$ is acyclic, which implies that $\mathscr{G}$ is group-acyclic. □

By replacing the group DAG constraint with the algebraic constraint (4.5), we can estimate the structure among the groups of variables using methods adopting the existing algebraic DAG constraint.

5.4 Numerical Experiments

This section reiterates the results of the numerical experiments conducted to assess the effect of the group DAG constraint using synthetic data, given in Section 4 of Article II. Compared to the numerical experiment in Section 4.5, the main difference was how to generate a true graph.



(a) Variable DAG

(b) Group DAG

Fig. 5.2 Simulated variable DAG and group DAG (Figure 2 of Article II)

For each variable, we randomly assigned group $l \in \{1,...,M\}$ such that the groups had an equal number of variables. Then, we randomly selected a parent variable for each group and generated an intragroup DAG with a tree structure having a depth of 1. This operation simulated the observation that variables in the same group had similar values. Subsequently, starting from the first group $k = 1$, we assigned a connection from the variables in subsequent groups l $(l > k)$ to the variables in group k with a probability of 0.1, where at least one connection from group l to group k was established. As a result, we obtained an adjacency matrix $B \in \{0,1\}^{P \times P}$ representing a group-acyclic graph. An example of a simulated variable DAG and the corresponding group DAG for $P = 12$ and $M = 3$ are shown in Figure 5.2.

5.4.1 Nonlinear Data

First, we report the results of experiments on nonlinear data with heteroscedasticity but no temporal dependencies to examine the effect of the group DAG constraint applied to the existing estimation algorithm.

Setup

Based on each randomly generated graph, we generated data from Equation (4.18) with $L = 0$ and $s_j = 1$. We compared the following four methods: NOTEARS-MLP (Zheng et al., 2020), NOTEARS-MLP using randomly selected variables for each group (NOTEARS-MLP-SEL), NOTEARS-MLP using the average value of the variables for each group (NOTEARS-MLP-AVE), and NOTEARS-MLP with a group DAG constraint (NOTEARS-MLP-ACY). All four methods had the same parameter settings of NOTEARS-MLP given in the original paper,

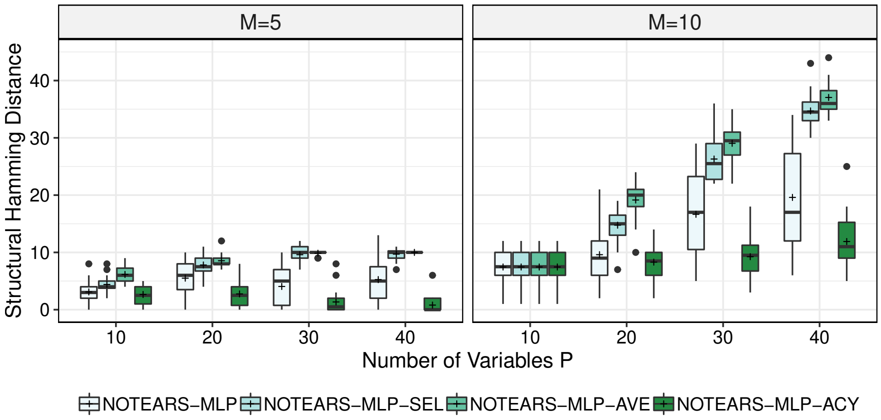


Fig. 5.3 Results on synthetic data using different number of variables P, number of groups M (Figure 3 of Article II)

i.e., $\lambda_1 = \lambda_2 = 0.01$, and $w = 0.3$, and the MLPs consisted of a single hidden layer with 10 nodes.

Results

The results for different numbers of variables $P = \{10, 20, 30, 40\}$ and number of groups $M = \{5, 10\}$ are presented in Figure 5.3. As shown, NOTEARS-MLP-ACY outperformed the other methods, indicating the effectiveness of the group DAG constraint. Interestingly, NOTEARS-MLP-AVE and NOTEARS-MLP-SEL exhibited worse performance than NOTEARS-MLP, indicating that aggregating the information of the groups leads to inferior results. The cases of $P = 10$ and $M = 10$ corresponded to the estimation of the variable DAGs; thus, the four methods exhibited identical results.

5.4.2 Nonlinear Time-Series Data with Location-Scale Noise

Next, we describe the results on nonlinear time-series data with location-scale noise to assess the performance of TS-LSNM and TS-LSNM with the group DAG constraint.

Setup

Nonlinear time-series data with location-scale noise was generated using Equation (4.18), with the maximum length L of the temporal dependencies set to 1.

We compared three methods: NTS-NOTEARS, TS-LSNM, and TS-LSNM with group DAG constraint (TS-LSNM-ACY). The parameters for each method were determined by performing a grid search in the condition of $M = 10$ with other settings as the same as Section 4.8. As a result, $\lambda_1 = \lambda_2 = 0.01$ for all methods, $w = 0.2$ for NTS-NOTEARS and TS-LSNM-ACY, and $w = 0.3$ for TS-LSNM were chosen.

As NTS-NOTEARS and TS-LSNM do not necessarily return a group DAG, we recursively remove edges with the smallest absolute value from the estimated group adjacency matrix until we obtain a group-acyclic graph, which is analogous to the postprocessing in (Ng et al., 2020).

Results

The results are shown in Figure 5.4 (lower the better). TS-LSNM-ACY generally exhibited the best performance, followed by TS-LSNM, indicating the effectiveness of group DAG constraint and capturing the heteroscedasticity. Because of the DAG constraint, TS-LSNM-ACY achieved a relatively low SHD even if the number of variables increased.

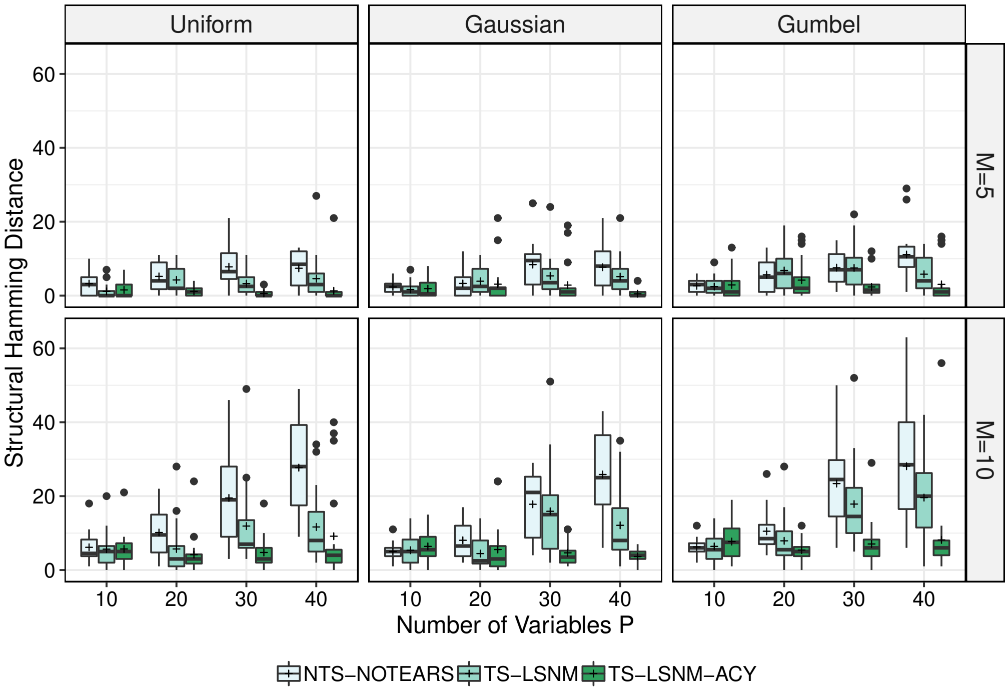


Fig. 5.4 Results on synthetic data using different number of variables P, number of groups M, and different noise distributions (Figure 4 of Article II)

Chapter 6

Application to Real-world Manufacturing Data

Finally, we summarize the results obtained from numerical experiments on real-world data collected from a ceramic substrate manufacturing process, described in Article Ⅱ.

6.1 Ceramic Substrate Manufacturing Process

A ceramic substrate is a ceramic structure with high porosity cell walls forming a honeycomb used to purify gases emitted from engines of automobiles (Yamada et al., 2002). For example, they convert harmful nitrogen oxide and carbon monoxide to less harmful nitrogen and carbon dioxide. Here, we are interested in a kneading process of manufacturing ceramic substrates. The process uses two kneaders (upper and middle) to mix the ingredients of the ceramic, each of which was cooled using a separate water-cooled chiller. The kneaded ingredients were cut to the same length and baked. The goal was to identify the cause of cutting torque, which is measured as an alternative characteristic of the viscosity of the ceramic and is closely related to crack failure during baking. A schematic of the kneading process and collected measurements are depicted in Figure 6.1.

The temperature, electricity (voltage and frequency), and pressure were measured at several positions of the kneaders and chillers. We identified 19 variables and 2000 data points after removing obvious outliers for fitting each model. We assigned groups to each variable according to domain knowledge; the details of these groups are presented in Table 6.1. Scatter plots of the data are depicted in Figure 6.2. From Figure 6.2, we can see that the data more or less exhibit nonlinearity as well as heteroscedasticity.

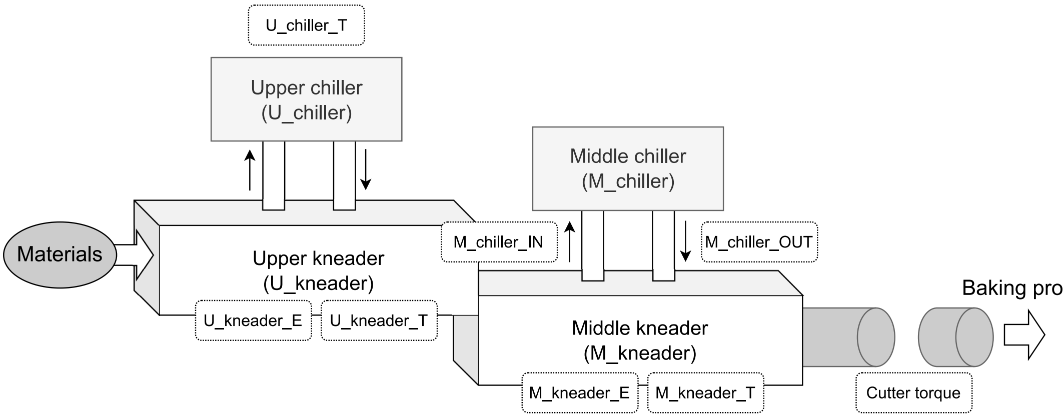


Fig. 6.1 Schematic of a kneading process in ceramic substrate manufacturing, and corresponding measurements. The rounded rectangles denote the group name specified in Table 6.1.

Table 6.1 Assigned groups for the ceramic manufacturing process data (Table 1 of Article II)

Group ID	Name	Description	# of variables
1	U_chiller_T	Upper chiller temperature	1
2	U_kneader_T	Upper kneader temperature	3
3	U_kneader_E	Upper kneader electricity	3
4	M_chiller_IN	Water entering middle chiller	3
5	M_chiller_OUT	Water exiting middle chiller	2
6	M_kneader_T	Middle kneader temperature	3
7	M_kneader_E	Middle kneader electricity	3
8	Cutter torque	Cutting torque	1

We compared group DAGs obtained from two methods that can model non-linear relations as well as time structure:

- TS-LSNM-ACY (Proposed): TS-LSNM with group acyclicity constraint
- NTS-NOTEARS (Sun et al., 2021)

We incorporated prior knowledge that the cutter torque is a sink by restricting the corresponding kernel weights to zero (Sun et al., 2021).

We used the same parameter settings obtained from the grid search performed in Section 5.4.1, where $\lambda_1 = \lambda_2 = 0.01, w = 0.2$ were used for both methods. In the numerical experiments on synthetic data, we assumed that the true time lags L were known in advance. However, in a real-world scenario, we must select an appropriate L value from the data. We fitted each model with a large time-lag value of $L = 5$ and estimated the weighted adjacency

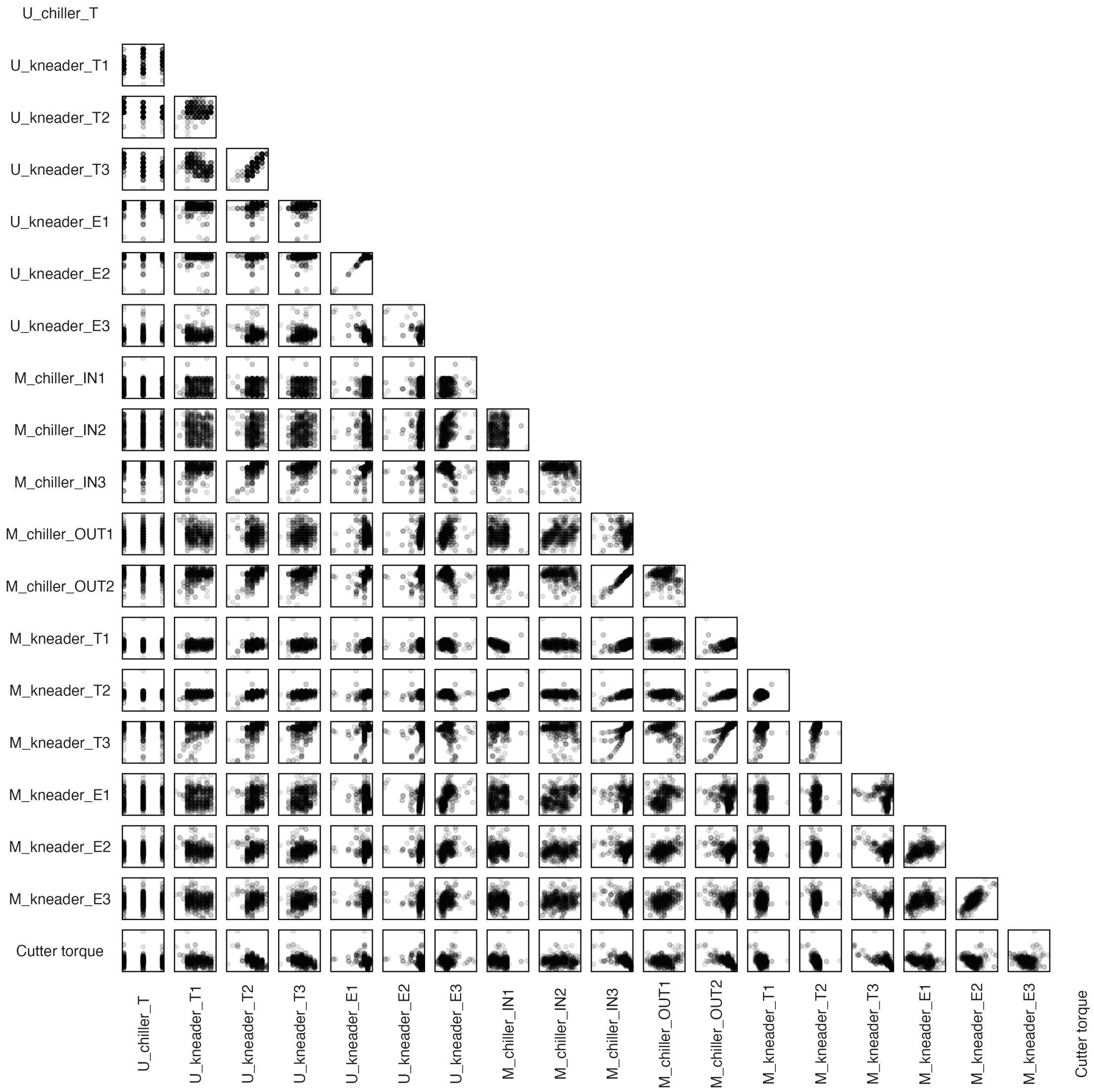


Fig. 6.2 Scatter plots of the ceramic substrate manufacturing process data. Data are anonymized by standardizing each variable to zero mean and unit variance, and the variable names are set to the group name of Table 6.1 with sequential numbers if multiple columns exist in a group.

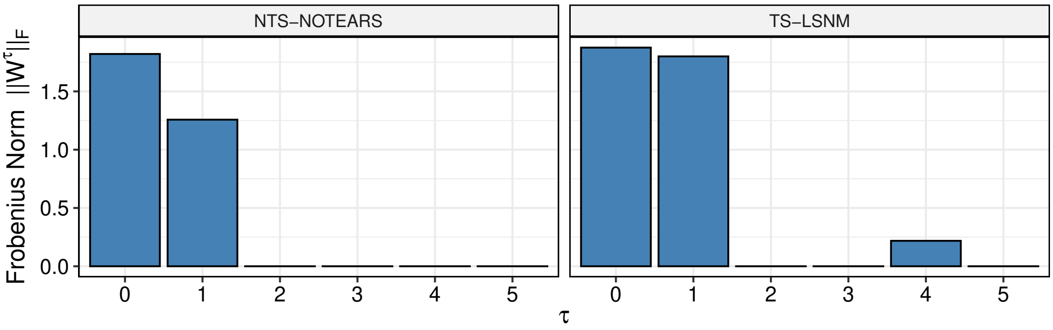
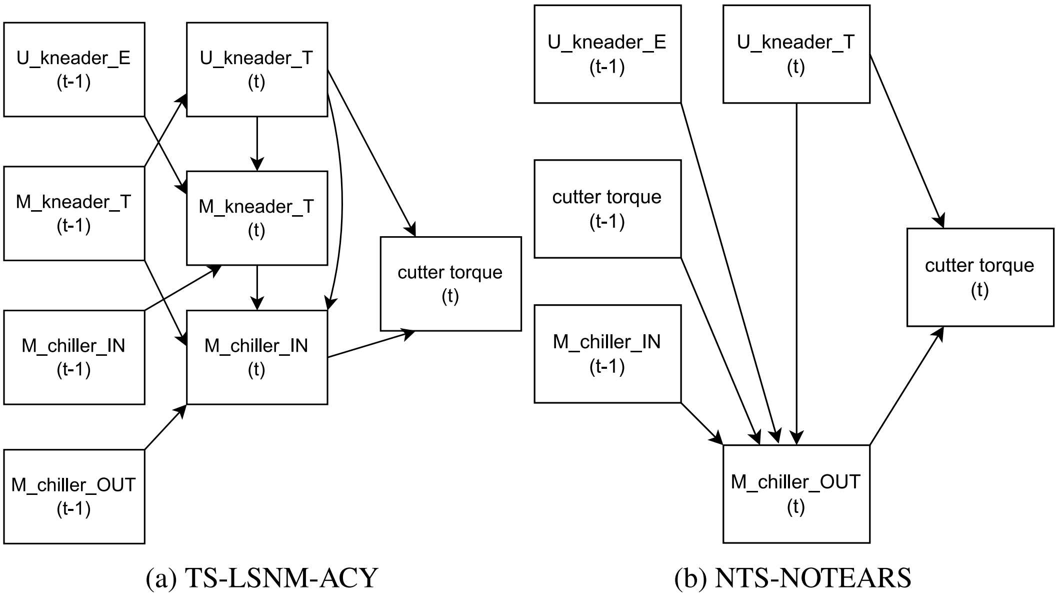


Fig. 6.3 Frobenius norm ($\|\tilde{W}^{\tau}(\theta^A, \theta^C)\|_F$) of the estimated weighted adjacency matrix on each time lag τ (Figure 10 of Article II)

matrix $\tilde{W}^{\tau}(\theta^A, \theta^C)$ $(\tau = 0, ..., L)$. We then calculated the Frobenius norm of the estimated weighted adjacency matrix for each time lag τ.

The results for selecting L are presented in Figure 6.3, where plateaus are observed for $L > 1$ for both methods. Therefore, we selected $L = 1$ for both methods and fitted the model again with $L = 1$. An alternative approach for determining L is to determine the value of the objective function, although we must fit the model multiple times.



(a) TS-LSNM-ACY

(b) NTS-NOTEARS

Fig. 6.4 Estimated window causal graph of group DAGs for ceramic substrate manufacturing process data (Figure 5 of Article II).

6.2 Results

The obtained group DAGs are shown in Figure 6.4, where groups estimated not to be the ancestors of the cutter torque are omitted. Both methods succeeded in recovering the relationship between the temperature of the kneader (U_kneader_T) and the cutter torque. The results matched the domain knowledge that the material's viscosity may change with temperature. The result of TS-LSNM-ACY, which revealed an arrow from U_kneader_T to the cooling water flowing into the chiller (M_chiller_IN), was more consistent with the domain knowledge than the result of NTS-NOTEARS, which revealed a connection from U_kneader_T to the cooling water flowing out of the chiller (M_chiller_OUT) because M_chiller_OUT is expected to be controlled by the chiller. Moreover, the result of TS-LSNM-ACY revealed the correct physical phenomenon in which the chillers cool the kneaders (M_chiller_IN → M_kneader_T). In contrast, the result of NTS-NOTEARS revealed no connection between the chillers and the kneaders, indicating that chillers do not cool the kneaders, which disagrees with expectations from domain knowledge (this was possible if both chillers were broken and the cooling performance was lost, though this was not the case). Therefore, we conclude that TS-LSNM-ACY obtained better estimation results than NTS-NOTEARS.

Chapter 7

Conclusions

This thesis presented novel methods for learning causal relationships from observational data. We developed a causal discovery method capable of estimating the characteristics of data collected from a manufacturing process. Although non-linearity and temporal dependency were addressed in previous studies, many works did not consider heteroscedasticity, where the variance of a quantity is modulated by others, despite the variance being assessed in traditional quality control methods. Moreover, although groups of variables need to be handled carefully in causal discovery, methods for estimating causal structure among groups of variables beyond linear relationships remained unclear.

After describing the motivation and scope of this thesis in Chapter 1, necessary mathematical backgrounds and definitions of causal models were introduced in Chapters 2 and 3. Then, starting from the introduction of existing FCMs relevant to this thesis, the proposed method for estimating LSNM, and its extension to time series data is described in Chapter 4. The problem definition of causal discovery for groups of variables and existing methods, followed by the proposed method, which can be applied to data beyond linear relation, is explained in Chapter 5. Finally in Chapter 6, the proposed method is applied to real-world data, and the results indicate the strength of the proposed method.

The main contribution of this thesis is twofold: (i) In Articles I and II, we proposed an estimation method for LSNM and then extended the model to exploit time structure. (ii) In Article II, a novel approach is presented for learning causal structure among groups of variables for use with continuous optimization-based structure learning methods. These works provide a causal discovery method that simultaneously handles non-linearity, temporal dependency, and heteroscedasticity and estimates causal structure among individual and groups of variables.

Although Articles I and II contributed to developing a causal discovery method suitable for manufacturing data, many open problems remain, highlighting the limitations of the proposed methods. First, the proposed methods assume no latent confounders. Although the important quality measures of a product are likely to be collected, there is no guarantee that all relevant information is measured, including confounders. Moreover, although continuous optimization-based methods can handle latent confounders (Bhattacharya et al., 2021), they are limited to linear relations, and the extension remains a challenging problem.

Second, on the causal discovery for groups of variables, we assume that the (true) grouping of variables is known. As we assume an acyclic group graph, incorrect grouping that violates this assumption may cause the estimation method to output a completely different graph.

Finally, the proposed estimation method for TS-LSNM fits two CNNs for each variable; hence, the optimization becomes difficult to converge as the number of variables and sample size increase. One possible scenario is to leverage a more efficient form of acyclicity constraint and optimization method (Bello et al., 2022). In addition, it would be interesting to see the effect of using the skeleton of a DAG inferred with the existing constraint-based approaches (e.g., FCI algorithm) to limit the search space.

References

Peter Martey Addo, Christelle Manibialoa, and Florent McIsaac. Exploring nonlinearity on the CO2 emissions, economic production and energy use nexus: a causal discovery approach. *Energy Reports*, 7:6196–6204, 2021.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Saeed Albukhitan. Developing digital transformation strategy for manufacturing. *Procedia Computer Science*, 170:664–671, 2020.

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2021.

James E Brady and Theodore T Allen. Six Sigma literature: a review and agenda for future research. *Quality and Reliability Engineering International*, 22(3):335–367, 2006.

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *International Conference on Machine Learning*, pages 2357–2369. PMLR, 2022.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Ruichu Cai, Weilin Chen, Jie Qiao, and Zhifeng Hao. On the role of entropy-based loss for learning causal structures with continuous optimization. *arXiv preprint arXiv:2106.02835*, 2021.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Max Chickering. Statistically efficient greedy equivalence search. In *Conference on Uncertainty in Artificial Intelligence*, pages 241–249. PMLR, 2020.

Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

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

Doris Entner and Patrik O Hoyer. Estimating a causal order among groups of variables in linear models. In *Artificial Neural Networks and Machine Learning–ICANN 2012: 22nd International Conference on Artificial Neural Networks, Lausanne, Switzerland, September 11-14, 2012, Proceedings, Part II 22*, pages 84–91. Springer, 2012.

Ronald Aylmer Fisher. Design of experiments. *British Medical Journal*, 1(3923):554, 1936.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pages 1–12, 1943.

Andrew C. Harvey. Estimating regression models with multiplicative heteroscedasticity. *Econometrica: Journal of the Econometric Society*, pages 461–465, 1976.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

Patrik Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21:689–696, 2008a.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008b.

Aapo Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. *Advances in Neural Information Processing Systems*, 10: 273–279, 1998.

Aapo Hyvärinen and Erkki Oja. Independent component analysis by general nonlinear hebbian-like learning rules. *signal processing*, 64(3):301–313, 1998.

Aapo Hyvärinen and Stephen M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan): 111–152, jan 2013.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14316–14332. PMLR, 23–29 Jul 2023.

Marcus Kaiser and Maksim Sipos. Unsuitability of notears for causal graph discovery. *arXiv preprint arXiv:2104.05441*, 2021.

Shyam Kumar Karna, Rajeshwar Sahai, et al. An overview on Taguchi method. *International Journal of Engineering and Mathematical Sciences*, 1(1):1–7, 2012.

Yoshinobu Kawahara, Kenneth Bollen, Shohei Shimizu, and Takashi Washio. GroupLiNGAM: Linear non-Gaussian acyclic models for sets of variables. *arXiv preprint arXiv:1006.5041*, 2010.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 2017.

Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvärinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, pages 3520–3528. PMLR, 2021.

Genta Kikuchi. Differentiable causal discovery under heteroscedastic noise. In *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part I*, pages 284–295. Springer, 2023.

Roger E Kirk. Experimental design. *Sage Handbook of Quantitative Methods in Psychology*, pages 23–45, 2009.

Gülser Köksal, Inci Batmaz, and Murat Caner Testik. A review of data mining applications for quality improvement in manufacturing industry. *Expert systems with Applications*, 38 (10):13448–13467, 2011.

Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. pages 366–374, 2008.

Rinat Landman, Jukka Kortela, Qiang Sun, and S-L Jämsä-Jounela. Fault propagation analysis of oscillations in control loops using data-driven causality and plant connectivity. *Computers & Chemical Engineering*, 71:446–456, 2014.

Jing Li and Jianjun Shi. Knowledge discovery from observational data for process control using causal Bayesian networks. *IIE Transactions*, 39(6):681–690, 2007.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1): 56–67, 2020.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Sisi Ma and Roshan Tourani. Predictive and causal implications of using Shapley value for model interpretation. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, pages 23–38. PMLR, 2020.

Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*, 9(1):381–386, 2020.

Subramani Mani and Gregory F Cooper. Causal discovery from medical textual data. In *Proceedings of the AMIA Symposium*, page 542. American Medical Informatics Association, 2000.

Katerina Marazopoulou, Rumi Ghosh, Prasanth Lade, and David Jensen. Causal discovery for manufacturing domains. *arXiv preprint arXiv:1605.04056*, 2016.

Justin Matejka and George Fitzmaurice. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1290–1294, 2017.

meettechniek.info. Accuracy, precision & resolution, 2013. https://meettechniek.info/measurement/accuracy.html.

Douglas C Montgomery. *Introduction to Statistical Quality Control 8th ed.* Wiley, 2019.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 745–752, 2009.

Jerzy Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (in Polish). English translation by DM Dabrowska and TP Speed (1990). *Statistical Science*, 5:465–480, 1923.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. *Advances in Neural Information Processing Systems*, 33, 2020.

Pekka Parviainen and Samuel Kaski. Learning structures of Bayesian networks for variable groups. *International Journal of Approximate Reasoning*, 88:110–127, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. pages 589–598, 2011.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in Neural Information Processing Systems*, 26, 2013.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15: 2009–2053, 2014.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7): 075310, 2018.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. PMLR, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019.

Michael Rüßmann, Markus Lorenz, Philipp Gerbert, Manuela Waldner, Jan Justus, Pascal Engel, and Michael Harnisch. Industry 4.0: The future of productivity and growth in manufacturing industries. *Boston Consulting Group*, 9(1):54–89, 2015.

Richard Scheines and Peter Spirtes. Causal structure search: Philosophical foundations and future problems. In *Neural Information Processing Systems 2008 Workshop, Causality: Objectives and Assessment, Whistler, Canada*, 2008.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and opportunities with causal discovery algorithms: application to Alzheimer's pathophysiology. *ScientificRreports*, 10(1):2975, 2020.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for FMRI. *Neuroimage*, 54(2):875–891, 2011.

Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000.

Erik Stolterman and Anna Croon Fors. Information technology and the good life. *Information Systems Research: Relevant Theory and Informed Practice*, pages 687–692, 2004.

Eric V Strobl. A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, 8:33–56, 2019.

Eric V Strobl and Thomas A Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *Journal of Computational Science*, 72:102099, 2023.

Xiangyu Sun, Guiliang Liu, Pascal Poupart, and Oliver Schulte. NTS-NOTEARS: Learning nonparametric temporal DAGs with time-series data and prior knowledge. *arXiv e-prints*, pages arXiv–2109, 2021.

Michail Tsagris, Giorgos Borboudakis, Vincenzo Lagani, and Ioannis Tsamardinos. Constraint-based causal discovery with mixed data. *International Journal of Data Science and Analytics*, 6:19–30, 2018.

Saurabh Vaidya, Prashant Ambad, and Santosh Bhosle. Industry 4.0–a glimpse. *Procedia Manufacturing*, 20:233–238, 2018.

Matej Vuković and Stefan Thalmann. Causal discovery in manufacturing: A structured literature review. *Journal of Manufacturing and Materials Processing*, 6(1):10, 2022.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Vector causal inference between two groups of variables. 37(10):12305–12312, 2023.

Weschler Instruments. Meter accuracy explained, 2020. https://weschler.com/blog/meter-accuracy-explained/.

Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in Neural Information Processing Systems*, 8, 1995.

Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585, 1921.

Paul Wunderlich and Oliver Niggemann. Structure learning methods for Bayesian networks to reduce alarm floods by identifying the root cause. In *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–8. IEEE, 2017.

Sascha Xu, Alexander Marx, Osman Mian, and Jilles Vreeken. Causal inference with heteroscedastic noise models. *Proceedings of the AAAI Workshop on Information Theoretic Causal Inference and Discovery (ITCI'22)*, 2022.

Keiichi Yamada, Masakazu Tanaka, and Keiji Itou. Development of reinforced thin wall ceramic substrate. *DENSO TECHNICAL REVIEW*, 7(1), 2002.

K Zhang and A Hyvärinen. On the identifiability of the post-nonlinear causal model. pages 647–655, 2009.

Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Causality: Objectives and Assessment*, pages 157–164. PMLR, 2010.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9492–9503, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

Shuheng Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. *Advances in Neural Information Processing Systems*, 22, 2009.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.