Title
Constraint-Based Learning of Interventional Markov Equivalence Classes on High-Dimensional Data

Author
Wang, Hao

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Constraint-Based Learning of Interventional Markov Equivalence Classes on High-Dimensional Data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Hao Wang

2022

ABSTRACT OF THE DISSERTATION

Constraint-Based Learning of Interventional Markov Equivalence Classes on High-Dimensional Data

by

Hao Wang
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2022
Professor Qing Zhou, Chair

Directed Acyclic Graphs (DAGs) are a powerful tool to model the network of dependencies among variables. They provide a basis for causal discovery, and have been widely used in many fields, especially biology. Unfortunately, structure learning is quite non-trivial for DAG. One major difficulty is that some DAGs are unidentifiable with observational data only, and undirected edges cannot be resolved to directed edges.

The opportunity to apply interventions motivates interest in the smaller interventional Markov equivalence class. In this dissertation, we discuss how to modify the classic PC algorithm for causal discovery so that it can be used safely on interventional data. We introduce invariance relations on conditional distributions with different intervention targets that provide a powerful rule for edge orientation. There are several advantages of this rule: first, it does not require the Gaussian distribution assumption, instead a general structural equation model (SEM) of DAG is sufficient; second, it works for both (structural) intervention and soft intervention. Finally, we can merge some data blocks with different interventions for edge orientation.

A new constraint-based method is proposed to recover the interventional essential graph from the CPDAG (or called as observational essential graph) based on the invariance rule. We also establish consistency guarantees for both an interventional PC and an edge orientation algorithm under a sparse high-dimensional setting. Such high-dimensional consistency results are rarely seen in this area. It is also worthwhile to emphasize that the constraints on the family of interventions throughout this dissertation are mild. Finally, simulations are used to show the effectiveness of our method.

The dissertation of Hao Wang is approved.

Arash A. Amini

Hongquan Xu

Mark S. Handcock

Qing Zhou, Committee Chair

University of California, Los Angeles

2022

To my parents.

I am feeling your love every moment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Professor Qing Zhou, for his continuous support and guidance throughout my Ph.D. studies. Thanks so much for giving me this opportunity to start a wonderful journey. And it is impossible for me to reach the destination without his help. What I have learned from him is beyond knowledge, also kindness, patience and enthusiasm. I would be more diligent and do better if I had the second chance.

I am also very grateful to have Professors Mark S. Handcock, Hongquan Xu and Arash A. Amini in my committee for their comments and suggestions. I can still remember the first year of my graduate study in their classes.

VITA

2012–2016 B.S. in Mathematics, Tsinghua Univeristy, Beijing, China

2016–present Ph.D. student in Statistics, UCLA, Los Angeles, USA

CHAPTER 1

Introduction

Directed acyclic graphs (DAGs) are commonly discussed in causal inference, for example [SGS00] or [Pea00], as the parents of one vertex in the graph could be recognized as 'causes' and the arrows naturally represent the 'causal relations'. Learning the structure of the DAG from data is a core problem in this field, but it's quite challenging. One difficulty is the number of DAGs grows superexponentially as the number of nodes increases; see [Rob77]. Another difficulty is that, for multivariate Gaussian distribution, some DAGs have exactly the same behaviors in probability, which means they are unidentifiable under general setting; see the criterion for Markov equivalence class (MEC) in [VP90] or [AMP97].

A number of approaches for structure learning has been developed in the past few years, and which can be classified as constraint-based or score-based. PC algorithm is a well-known representative of the constraint-based methods. The popular PC algorithm proposed by [SG91] can estimate the completed partially directed acyclic graph (CPDAG) from observational data by considering the conditional independence sets implied from data; and [KB07] establish theoretical guarantees for the PC algorithm under the sparse and high-dimensional setting. With the high-dimensional consistency results, also its scalable algorithm computation in real world, the PC algorithm and its variants are widely implemented to sparse high-dimensional datasets. The advantage of the PC algorithm is straightforward and intuitive: sparsity allows the PC algorithm to finish structure learning with limited number of conditional independence tests, even the number of variables grows fastly with the data sample size.

The second approach of structure learning is score-based, and in which some score function has be constructed and the method would find the estimation of DAG by optimizing the score function over the possible space of DAGs or CPDAGs. The Greedy Equivalence Search (GES) ([CM02]) conducts greedy search to optimize the score function with forward and backward phases. GES is a popular algorithm over the score-based methods, and [CM02] also proves its consistency under the classic setting (fixed number of variables) with the BIC score. Recently, [NHM18] proves the consistency results of GES also its variant under sparse high-dimensional setting. Another theoretical result around the score-based method is [GB13] proves the the high-dimensional consistency of DAG structure learning with ℓ_0-penalized maximum likelihood estimation, however, there is no algorithm to implement this method due to the ℓ_0 penalty. The main difficulty around score-based methods is the quick growth of search space size as the number of vertices increases, which challenging both computation in practice and theoretical proof. Considering this, some optimization methods are also introduced into this field ([BEd08], [YAZ20]).

As mentioned in the first paragraph, observational data has its limitation in identifiability. It could be a large pain point in causal inference if some of interests is still left as undirected edges in the estimated CPDAG. Intervention suggested by [Pea00] and [SGS00] can somehow overcome this kind of limitation of observational data. With the help of intervention, some undirected edges in CPDAG become identifiable. [HB12] introduce a series of concepts to interventional case, such as the interventional Markov equivalence class and $\mathcal{I}$-essential graph. DAGs among the same (interventional) Markov equivalence class are unidentifiable. And we can learn the power of intervention by comparing the difference between observational MEC and the corresponding interventional MEC. The interventional MEC should be smaller than MEC if intervention can provide useful information on edge orientation. In other words, the interventional MEC is a more precise partition over the space of DAGs.

In this case, graph structure learning will target to the interventional MEC, and the ideal output of structure learning algorithm with interventional data is the $\mathcal{I}$-essential graph $\mathcal{G}_{\mathcal{I}}$

instead of the CPDAG (or called as essential graph). Even the true DAG $\mathcal{G}$ can be recovered, if the intervention family is well-designed such that the intervention MEC only contains one DAG, i.e. $\mathcal{I}$-essential graph equals to the true underlying DAG, $\mathcal{G}_{\mathcal{I}} = \mathcal{G}$.

About the learning of interventional MEC, one approach is to extend the existing score-based methods to interventional dataset. For example, [HB12] shows that the Greedy Equivalence Search (GES) can also be applied to interventional data, and then build the Greedy Interventional Equivalence Search (GIES). They provides some simulation results to evaluate the GIES, but lack of theoretical results. And [HB15] introduces Gaussian likelihood framework for interventional data and derives the classic version consistency of the BIC criterion for estimating the interventional MEC. Different to these score-based methods, [HG08] designs a constraint-based method to recover $\mathcal{I}$-essential graph. They find the MEC from observational data first, and then orient undirected edges via intervention experiments by testing the invariance between pre- and post-intervention distributions.

In this dissertation, we focus on constraint-based method to approach the $\mathcal{I}$-essential graph. There are three main contributions of this work. First, this dissertation discusses how to extend the traditional PC algorithm to interventional dataset: we conduct conditional independence test for every interventional data block and determine the existence of edge based on all these test results. Second, we introduce the invariance relations on conditional distributions with different intervention targets, which provides a powerful rule for edge orientation. And furthermore, we can merge some data blocks to enhance the power of tests based on the invariance rule. Finally, in this work, we induces the high-dimensional consistency results for both skeleton recovery and edge orientation. Thus we can guarantee the consistency of the whole framework, interventional PC algorithm plus edge orientation, with mild assumptions.

One motivation for this work is the advantage of constraint-based method. There are two aspects of this advantage. The score optimization is more like a black-box, and constraint-based method can provide some information during the learning process. Considering the

expensive experimental data, it is ideal to combine the structure learning method and intervention design together, which could get some help from constraint-based method. Secondly, it is highly non-trivial to extend existing theoretical results of score-based methods to the interventional case, especially under high-dimensional setting. To the best of our knowledge, this work provides rarely seen consistency results for structure learning with interventional data.

The most related existing work is from [HG08], as both their method and our work rely on two stages for structure learning: recover skeleton and then do edge orientation, but our focuses are quite different. They start from the estimated CPDAG and mainly focus on the framework of active learning, lack of details of edge orientation and no consistency results. For edge orientation, they only discussed single node intervention and the corresponding simple intervention family. It is worthwhile to emphasize that our work makes mild constraints on the intervention, and our method can applied to those interventions on multiple nodes, also complicated intervention family $\mathcal{I}$.

The rest of the dissertation is organized as follows. In Chapter 2, we introduce some background knowledge of DAG with notations used throughout this dissertation, and the motivation of interventions. A brief introduction of PC algorithm can be found in Section 2.5. In Chapter 3, we discuss the difficulty of skeleton learning with interventional data, and then extend the original PC algorithm to interventional case. Chapter 4 focuses on the edge orientation and contains the main work of this dissertation. In Section 4.1, we describe the neighborhood behavior of the reversible edges in the essential graph, which provides the intuition to create edge orientation rule also several useful lemma. In Section 4.2, we start from the simple case to introduce how to conduct edge orientation with the invariance rule on conditional distributions, and then extend our work to more general case in Section 4.3. In Chapter 5, some assumptions are introduced to make the consistency guarantees possible in the high-dimensional setting. Simulations are posted in Chapter 6 to evaluate the methods introduced in this dissertation. Appendix of proofs for consistency results can be found in

Chapter 7. Finally, a brief summary and discussion about future work is in Chapter 8.

CHAPTER 2

Preliminaries

2.1 Graphs

A graph $\mathcal{G}$ can be represented by (V, E), where $V = [p] \coloneqq \{1, 2, \ldots, p\}$ is the vertex set and $E \subset V \times V$ is the edge set consisting of some ordered pairs of vertices. In this work, the vertex set V is identified with a set of random variables $X_1, \ldots, X_p$. If $(i, j) \in E$ and $(j, i) \notin E$, the edge between X_i and X_j is directed, represented by an arrowhead as $X_i \to X_j$; if $(i, j) \in E$ also $(j, i) \in E$, the edge between X_i and X_j is undirected, denoted as $X_i - X_j$. A graph is (un)directed if all edges of the graph are (un)directed, while a partially directed graph contains both directed and undirected edges.

We use $X_i \leftrightarrow X_j$ to represent the edge between X_i and X_j, no matter it is directed (in any direction) or undirected, and we say X_i and X_j are adjacent. Given a graph $\mathcal{G} = (V, E)$, suppose there is a sequence $\{k_0, k_1, \ldots, k_m\}$ such that $k_0 = i$ and $k_m = j$: if for every $l = 0, \ldots, m-1$, we have $X_{k_l} \leftrightarrow X_{k_{l+1}}$, then we say $X_{k_0}, \ldots, X_{k_m}$ form a trail between X_i and X_j of length m; if for every $l = 0, \ldots, m-1$, we have $X_{k_l} \to X_{k_{l+1}}$ or $X_{k_l} - X_{k_{l+1}}$, then we say $X_{k_0}, \ldots, X_{k_m}$ form a path between X_i and X_j of length m. A path is directed if there exists at least one directed edge. A cycle in $\mathcal{G}$ is a path that starts and ends at the same vertex, i.e. $X_i, \ldots, X_k$ where $X_i = X_k$, and a graph is acyclic if there is no cycle. We call a graph as a directed acyclic graph (DAG) if it is directed and acyclic. We call a graph a partially directed acyclic graph (PDAG) if it is acyclic and contains both directed and undirected edges.

Given a subset $A \subset V$, we say the subgraph $\mathcal{G}_A = (A, E_A)$, with $E_A = E \cap (A \times A)$, is an induced graph by A. A graph $\mathcal{G}_1 = (V_1, E_1)$ is larger than a graph $\mathcal{G}_2 = (V_2, E_2)$, i.e. $\mathcal{G}_2 \subset \mathcal{G}_1$, if $V_2 \subset V_1$ and $E_2 \subset E_1$. The graph theoretic union is defined as $\cup \mathcal{G}_i = (\cup V_i, \cup E_i)$.

Let $\mathcal{G}$ be a DAG. If $X_i \to X_s \leftarrow X_j$, and X_i and X_j are not adjacent in DAG $\mathcal{G}$, then we say the ordered triple of vertices, (X_i, X_s, X_j), forms a v-structure in graph. Vertex X_s is called as a collider on a trail if there exists $X_i \to X_s \leftarrow X_j$ for some vertices X_i and X_j on the trail. Notice that it doesn't require X_i and X_j are non-adjacent when the collider is defined. The collider in a v-structure is sometimes called unshielded collider. If there exists a directed path $X_i, \ldots, X_j$, then we say X_i is an ancestor of X_j and X_j is a descendant of X_i. We use Ancestor(j) to denote X_j's ancestors, and for convenience, let $j \in$ Ancestor(j). A trail $X_{k_0} \leftrightarrow \cdots \leftrightarrow X_{k_m}$ is active given a vertex set $\mathcal{S} \subset V$ if (1) X_{k_l} or one of its descendants are in $\mathcal{S}$ whenever there is a collider X_{k_l} in $X_{k_{l-1}} \to X_{k_l} \leftarrow X_{k_{l+1}}$; (2) no other node along the trail is in $\mathcal{S}$; otherwise, we say the trail is inactive or blocked by $\mathcal{S}$. Given $\mathcal{X}, \mathcal{Y}, \mathcal{S}$ three sets of nodes in $\mathcal{G}$, we say $\mathcal{X}$ and $\mathcal{Y}$ are $d-$separated by $\mathcal{S}$ if there is no active trail between any node $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ given $\mathcal{S}$. An ordering π of the vertices $X_1, X_2, \ldots, X_p$ is a topological ordering relative to $\mathcal{G}$ if whenever we have $X_i \to X_j \in E$, then $i \prec j$ in the ordering π. Given a graph $\mathcal{G} = (V, E)$ and a vertex $s \in V$, the neighborhood of vertex s, denoted by ne(s), is the set of all vertices adjacent to s, i.e. ne$(s) = \{a \in V | (a, s) \in E \text{ or } (s, a) \in E\}$.

For a partial directed acyclic graph (PDAG) $\mathcal{G}$, the acyclicity constraint implies the PDAG can be decomposed into several disjoint chain components $\{\mathcal{K}_i\}_{i=1,\ldots,m}$; see section (2.2.3) in [KFB09]: let $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_m$ be a disjoint partition of the vertex set V such that (1) the induced graph $\mathcal{G}_{\mathcal{K}_i}$ contains no directed edges and (2) for any pair $X \in \mathcal{K}_i$ and $Y \in \mathcal{K}_j$ with $i < j$, an edge between X and Y can only be directed as $X \to Y$. And a PDAG is also called a chain graph.

An undirected graph is chordal if any loop $X_{i_1} - X_{i_1} - \cdots - X_{i_k} - X_{i_1}$ for $k \geq 4$ has a chord that is an edge connecting X_i and X_j for two nonconsecutive nodes X_i, X_j. The skeleton of a graph $\mathcal{G} = (V, E)$ is its underlying undirected graph, i.e. $\mathcal{G}_{ske} = (V, E_{ske})$ with

$E_{ske} = \{(i,j)|(i,j) \in E \text{ or } (j,i) \in E\}$. A graph is said to be chordal if its skeleton is chordal.

2.2 Structural Equation Model (SEM)

As mentioned in Section 1, causal relations among random variables can be represented by arrows in DAG $\mathcal{G}$. Here we use the well-known structural equation model (SEM) to interpret the causal effects contained in DAGs. Suppose there is a true causal DAG $\mathcal{G} = (V, E)$ with vertex set $[p] = \{1, 2, \ldots, p\}$. Upon its graph structure, a Gaussian DAG model can be represented as a linear structural equation model,

$$X_j = \sum_{k \in \text{pa}(j)} \beta_{kj} X_k + \varepsilon_j, \quad j = 1, 2, \ldots, p, \tag{2.1}$$

where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are independent and $\varepsilon_j \sim \mathcal{N}(0, \omega_j^2)$. Here $pa(j)$ represents the parent node set of j, and $\beta_{kj} \neq 0$ only if $(k, j) \in E$. The coefficient β_{kj} represents the causal effect of X_k on X_j. The SEM (2.1) defines a joint Gaussian distribution for

$$X = (X_1, X_2, \ldots, X_p) \sim \mathcal{N}_p(0, \mathbf{\Sigma}), \tag{2.2}$$

such that its probability density $f(\cdot)$ factorises,

$$f(x) = \prod_{j=1}^{p} f(x_j | x_{pa(j)}). \tag{2.3}$$

Consider an $n \times p$ data matrix $\mathbf{X}$ with i.i.d. rows generated from the SEM (2.1),

$$\mathbf{X} = \mathbf{XB} + \mathbf{E}, \tag{2.4}$$

where $\mathbf{B} = (\beta_{kj})_{p \times p}$ is the coefficient matrix and $\mathbf{E}$ represents the noise matrix whose rows are i.i.d. from $\mathcal{N}_p(0, \mathbf{\Omega})$ with $\mathbf{\Omega} = \text{diag}(\omega_1^2, \omega_2^2, \ldots, \omega_p^2)$. Equation (2.4) can be treated as a

matrix expression of (2.1), and leads to

$$\mathbf{\Sigma} = (\mathbf{I} - \mathbf{B})^{-T}\mathbf{\Omega}(\mathbf{I} - \mathbf{B})^{-1},$$

an identity that expresses the covariance matrix $\mathbf{\Sigma}$ in terms of SEM parameters in (2.4).

2.3 Markov Equivalence Class (MEC)

Given a DAG $\mathcal{G}$ and a density $f(\cdot)$, we say the distribution $f(\cdot)$ is faithful to the graph $\mathcal{G}$ if for every triple of disjoint sets $\mathcal{X}, \mathcal{Y}, \mathcal{S} \subset V$,

$$X_{\mathcal{X}} \perp X_{\mathcal{Y}} | X_{\mathcal{S}} \Longleftrightarrow \mathcal{S} \text{ } d\text{-separates } \mathcal{X} \text{ and } \mathcal{Y} \text{ in } \mathcal{G}.$$

Let $\mathcal{D}(\mathcal{G})$ represent the set of all d-separation relations in $\mathcal{G}$. Then we say $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if $\mathcal{D}(\mathcal{G}_1) = \mathcal{D}(\mathcal{G}_2)$. Use notation $\sim$ to denote the Markov equivalence, i.e. $\mathcal{G}_1 \sim \mathcal{G}_2$ if $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent.

Definition 1. *Given a DAG $\mathcal{G}$, the Markov equivalence class (MEC) of $\mathcal{G}$, denoted by $[\mathcal{G}]$, is the set of DAGs that are equivalent to $\mathcal{G}$, that is, $[\mathcal{G}] = \{\tilde{\mathcal{G}} : \tilde{\mathcal{G}} \sim \mathcal{G}\}$.*

From graph structure perspective, more practically, two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same skeleton and the same v-structures.

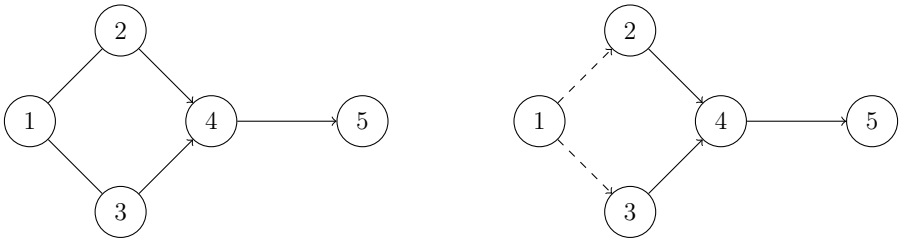
Markov equivalence sets a huge challenge for causal graph structure learning. DAGs in the same equivalence class imply exactly the same set of condition independence statements, which means in general it is impossible to distinguish the structures of equivalent DAGs from observational data only. For Gaussian linear SEM, all DAGs within the Markov equivalence class will have the same likelihood function such that they are non-identifiable. The essential graph, also known as CPDAG, helps us to understand the limitation of observational data in graph structure learning:

Definition 2. *Given a DAG $\mathcal{G}$, its essential graph (or CPDAG) is $\mathcal{G}_{ess} = \cup_{\tilde{\mathcal{G}} \in [\mathcal{G}]} \tilde{\mathcal{G}}$.*

To represent an equivalence class, there are two types of edges defined in a DAG $\mathcal{G}$: (1) a directed edge $i \to j$ is compelled in $\mathcal{G}$ if for every DAG $\tilde{\mathcal{G}}$ equivalent to $\mathcal{G}$, the edge $i \to j$ exists in $\tilde{\mathcal{G}}$; (2) if an edge is not compelled in $\mathcal{G}$, then it is reversible. Now we can give another definition of the completed PDAG.

Definition 3. *The CPDAG of an equivalence class is the PDAG consisting of a directed edge for every compelled edge in the equivalence class, and an undirected edge for every reversible edge in the equivalence class.*

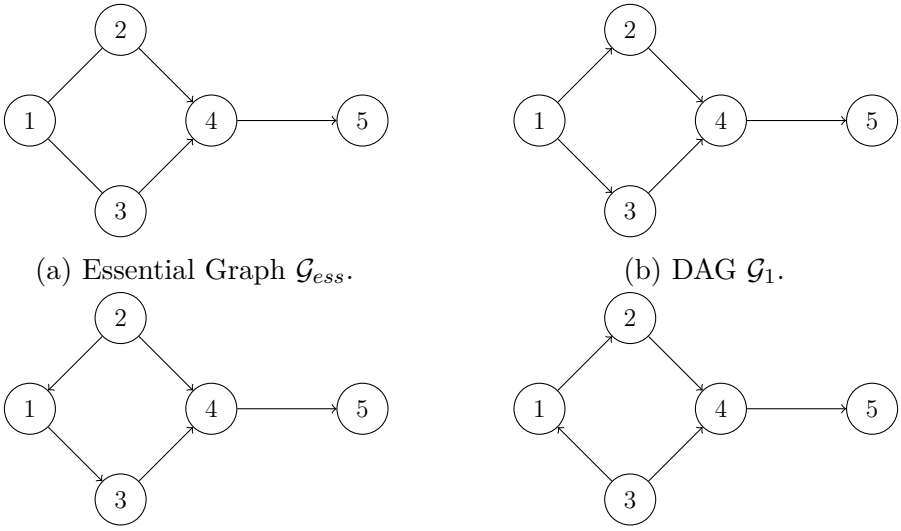
For example, in Figure 2.1, to keep the existing v-structure $2 \to 4 \leftarrow 3$, both $2 \to 4$ and $3 \to 4$ are irreversible; then the direction $5 \to 4$ is not allowed, as avoiding to induce new v-structure. In other words, edges $2 \to 4, 3 \to 4, 4 \to 5$ are compelled, as it will change the v-structure in DAG $\mathcal{G}$ if the direction of any one of these edges has been reversed. Meanwhile, edges $1 \to 2$ and $1 \to 3$ are reversible, since both $2 \to 1 \to 3$ and $2 \leftarrow 1 \leftarrow 3$ exist in the Markov equivalence class $[\mathcal{G}]$. Notice that it does not mean that there exists $2 \to 1 \leftarrow 3$ when we say 'both $1 \to 2$ and $1 \to 3$ are reversible'.



(a) Essential Graph $\mathcal{G}_{ess}$. (b) DAG $\mathcal{G}$.

Figure 2.1: An example of the essential graph and compelled/reversible edges.

Here we also list all DAGs according to the essential graph (a) in Figure 2.1, as an instance for the Markov equivalence class. There are three DAGs represented by the essential graph in Figure 2.2. Consider those DAGs having the same skeleton, i.e. $1-2, 1-3, 2-4, 3-4, 4-5$, the total number of such DAGs are $2^5 - 2 = 30$.



(a) Essential Graph $\mathcal{G}_{ess}$.

(b) DAG $\mathcal{G}_1$.

(c) DAG $\mathcal{G}_2$.

(d) DAG $\mathcal{G}_3$.

Figure 2.2: DAGs in the Markov equivalence class.

2.4 Experimental Intervention

Some edges would remain undirected in the essential graph, which obscures the causal interpretation of the edges. That is the motivation to introduce experimental interventions. In an experiment, the process of intervention forces one or several nodes to take values from an external distribution $f_{int}(\cdot)$, independent of the original joint distribution. A typical example of intervention, for instance, could be gene knockout or knockdown experiments in biology.

Let $I \subset [p]$ denote the intervention target, i.e. the set of intervened nodes. Let $X^I = (X_1^I, X_2^I, \ldots, X_p^I)$ be the random vector X under intervention I, whose distribution is given by a modified SEM (2.1):

$$X_j^I = \begin{cases} U_j, & \text{if } j \in I, \\ \sum_{k \in pa(j)} \beta_{kj} X_k^I + \varepsilon_j, & \text{if } j \notin I, \end{cases} \tag{2.5}$$

where U_j is a random variable that defines the distribution of X_j when it is under intervention. Notice that the equation (2.6) requires some independence assumptions on the intervention

variables U_I: (1) for any $j \in I$, U_j is independent of $\{\varepsilon_k, k \notin I\}$; (2) $U_i, i \in I$ are mutually independent. The intervention considered here is classified as stochastic intervention; cf. [KHN04]. Then (2.3) can be modified as,

$$f^I(x) = \prod_{j \in [p] \setminus I} f(x_j \mid x_{pa(j)}) \prod_{j \in I} f_{U_j}(x_j), \tag{2.6}$$

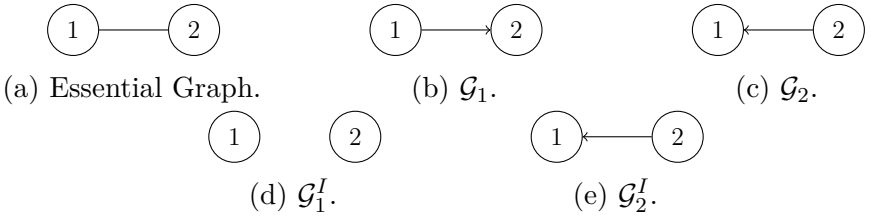
where $f^I(\cdot)$ is the joint density under the intervention target I.

Meanwhile, assuming the intervention variables are Gaussian with mean 0, i.e. $U_i \sim \mathcal{N}(0, \tau_i^2)$ for any $i \in I$, the random vector X^I preserves normality,

$$X^I \sim \mathcal{N}_p(0, \mathbf{\Sigma}^I), \tag{2.7}$$

where $\mathbf{\Sigma}^I$ is the modified covariance matrix under intervention target I. Equation (2.5) can be rewritten into an interventional matrix expression similar to (2.4), and furthermore the interventional covariance matrix $\mathbf{\Sigma}^I$ can be found; for details see [HB15].

A modified graph is also induced after intervention. Intervention on X_I will effectively remove all arrows pointing to nodes X_I, and we denote the modified graph by $\mathcal{G}^I$. For instance, consider a DAG with two nodes in Figure 2.3. Since $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent, it is impossible to distinguish these two without the help of intervention. Suppose we intervene on node 2, i.e. $I = \{2\}$, then Figure 2.3(d,e) shows that the arrow $1 \to 2$ will be cut off in $\mathcal{G}_1^I$, but $2 \to 1$ is not changed in graph $\mathcal{G}_2^I$. This difference between two graphs $\mathcal{G}_1^I$ and $\mathcal{G}_2^I$ after intervention enlightens the development of learning methods using interventional data.



(a) Essential Graph. (b) $\mathcal{G}_1$. (c) $\mathcal{G}_2$.

(d) $\mathcal{G}_1^I$. (e) $\mathcal{G}_2^I$.

Figure 2.3: An example of the Markov equivalence class and the effect of intervention.

It may be convenient to treat intervention as an additional variable in the DAG $\mathcal{G}$. Defining F_j as intervention on X_j, the parents of X_j is augmented to $pa(j) \cup \{F_j\}$ in the augmented graph with the following conditional distribution:

$$f(x_j \mid x_{pa(j)}, F_j) = \begin{cases} f(x_j \mid x_{pa(j)}), & F_j = idle; \\ f_{U_j}(x_j), & F_j = do(X_j = U_j). \end{cases} \tag{2.8}$$

Here, $do(X_j = U_j)$ is the *do*-operator [Pea00], denoting the intervention that forces X_j be U_j; *idle* means no intervention is applied on X_j. To understand the effect from intervention target I, sometimes it is better to investigate the augmented graph with auxiliary variables $\{F_m\}_{m \in I}$.

In general, experimental data collection is not limited to only one intervention target, and often there will be a family of intervention targets, $\mathcal{I} = \{I_1, I_2, \dots I_B\}$. Here B is the number of intervention targets in the family $\mathcal{I}$, i.e. $B = |\mathcal{I}|$. Recall the definition of $\mathcal{D}(\mathcal{G})$. For any intervention graph $\mathcal{G}^I$, let $\mathcal{D}(\mathcal{G}^I)$ represent the set of all d-separation statements in $\mathcal{G}^I$. Then the interventional Markov equivalence can be defined by considering the d-separation sets for the family of targets $\mathcal{I}$.

Definition 4. *Given a family of intervention targets $\mathcal{I}$, we say two Markov equivalent DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are $\mathcal{I}$-Markov equivalent, denoted by $\mathcal{G}_1 \sim_I \mathcal{G}_2$, if $\mathcal{D}(\mathcal{G}_1^I) = \mathcal{D}(\mathcal{G}_2^I)$ for all $I \in \mathcal{I}$.*

There are other equivalent ways to define Markov equivalence, also $\mathcal{I}$-Markov equivalence, see [HB12] Section 2.2 for more details. By Definition 4, interventional Markov equivalence requires that intervention graphs are always observational Markov equivalent over the family of targets, and from the graph structure perspective, if $\mathcal{G}_1 \sim_{\mathcal{I}} \mathcal{G}_2$ then $\mathcal{G}_1^I$ and $\mathcal{G}_2^I$ have the same skeleton and the same v-structures for all $I \in \mathcal{I}$. Extend Definition 1 to interventional case:

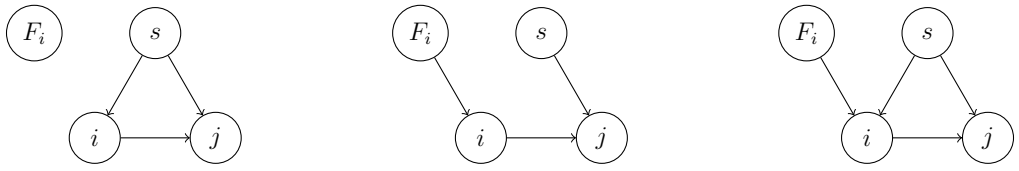
Definition 5. *Given a DAG $\mathcal{G}$, and a family of intervention targets $\mathcal{I}$, the $\mathcal{I}$-Markov equivalence class (MEC) of $\mathcal{G}$ is the set of graphs $[\mathcal{G}]_{\mathcal{I}} = \{\tilde{\mathcal{G}} : \tilde{\mathcal{G}} \sim_{\mathcal{I}} \mathcal{G}\}$.*

Up to now, the interventions mentioned in this section will make the intervened variable independent of its original causes. In other words, the interventions will destroy the arrows pointing to the intervened node, i.e. changing the structure of the original DAG. That's why this kind of intervention is referred as the structural intervention. Some other names are like surgical, ideal, or independent intervention. Sometimes the intervention is not that strong, or not ideal in the experiment. There is another weaker form of the intervention called as soft intervention.

The soft intervention will not affect the structure of DAG, which instead changes the conditional distributions among intervened node and its parents, i.e. the parameters of $f(x_j|x_{pa(j)})$. Someone refers this kind of intervention as the parametric intervention, and other names are: partial, conditional or dependent intervention. Similar to (2.8), the conditional distribution for soft intervention can be defined as:

$$f(x_j \mid x_{pa(j)}, F_j) = \begin{cases} f(x_j \mid x_{pa(j)}), & F_j = idle; \\ \tilde{f}(x_j \mid x_{pa(j)}), & F_j = do, \end{cases} \tag{2.9}$$

where $f(x_j \mid x_{pa(j)}) \neq \tilde{f}(x_j \mid x_{pa(j)})$. As the soft intervention does not change the structure, the augmented graph for post-intervention is also different; see Figure 2.4 as an example. [Ebe07] has a good summary of the different kinds of interventions.



(a) no intervention (b) intervention (c) soft intervention

Figure 2.4: Graph difference between structural and soft intervention while $F_i = do$.

The main focus of this dissertation will be put on the structural intervention, and we always refer to the structural intervention when we say intervention. Some concepts defined for structural intervention may not be extended to the soft intervention, for example, $\mathcal{I}$-Markov

equivalence. In Definition 4, $\mathcal{I}$-Markov equivalence is defined based on the d-separation sets, however soft intervention cannot provide extra information to change the d-separation relations as it will not affect the graph structure.

It is seen that the design of intervention target family $\mathcal{I}$ plays an important role in the definition of $\mathcal{I}$-Markov equivalence. We do not want to impose restrictive constraints on $\mathcal{I}$ in this work. Our methods apply to any type of intervention target family, $\mathcal{I} = \{I_1, I_2, \ldots, I_B\}$. A general $n \times p$ data matrix $\mathbf{X}$ generated under $\mathcal{I}$ consists of a number of data blocks $\mathbf{X}^i$ with n_i rows and p columns for $i = 1, 2, \ldots, B$. Each row within the same block $\mathbf{X}^i$ is drawn i.i.d. from $\mathcal{N}(0, \mathbf{\Sigma}^i)$, but data rows from different blocks are not identically distributed. Here after $\mathbf{\Sigma}^i$ corresponds to $\mathbf{\Sigma}^{I_i}$ in (2.7) to simplify the notation. An observational data block if exists could be treated as $I = \{\emptyset\}$ for notation consistency. To make it easier to understand, the general setting of data under a family of intervention targets $\mathcal{I} = \{I_1, I_2, \ldots I_B\}$ throughout this dissertation can be represented as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \\ \vdots \\ \mathbf{X}^B \end{pmatrix} \sim \begin{pmatrix} \mathcal{N}(0, \mathbf{\Sigma}^1) \\ \mathcal{N}(0, \mathbf{\Sigma}^2) \\ \vdots \\ \mathcal{N}(0, \mathbf{\Sigma}^B) \end{pmatrix} \quad \text{with } \sum_{i=1}^{B} n_i = n.$$

The goal of this work is to learn the MEC $[\mathcal{G}]_{\mathcal{I}}$ from the data $\mathbf{X}$.

2.5 Skeleton Learning on Observational Data

To recover the true causal DAG $\mathcal{G}$, we first find its skeleton $\mathcal{G}_{ske}$. Given the vertex set V, a complete undirected graph can be constructed, with an undirected edge between every pair of vertices. A general strategy of skeleton recovery is to eliminate some edges from the complete graph, leading to E_{ske} usually much smaller than the edge set of the complete

graph if $\mathcal{G}$ is sparse.

Any edge in skeleton can be detected with conditional independence constraints, as under faithfulness assumption,

$$\text{there is an edge between nodes } X_i \text{ and } X_j \iff X_i \not\perp X_j \mid X_{\mathcal{S}} \text{ for any } \mathcal{S} \subset V\backslash\{i,j\}; \quad (2.10)$$

see [SGS00] for details. The rule in (2.10) can be applied to skeleton recovery naively, by checking conditional independencies given all $\mathcal{S} \subset V\backslash\{i,j\}$. However, it faces the challenges of computational effectiveness and the reliability of high order conditional independence test, especially for high-dimensional case.

The PC algorithm [SG91] develops a better approach to test all these conditional independence relations effectively. In PC algorithm, different to (2.10), we start the conditional independence test from empty conditioning set, i.e. $|\mathcal{S}| = 0$, and increase $|\mathcal{S}|$ gradually. Another difference is PC algorithm only considers the subset of neighborhood as the set of variables conditioned, i.e. $\mathcal{S} \subset \text{ne}(i)\backslash\{j\}$.

Algorithm 1 shows the population level PC algorithm. Suppose we have oracle information on conditional independencies, i.e. line 11 in Algorithm 1 guaranteed, then the output of PC algorithm is the true skeleton of the DAG $\mathcal{G}$; see [KB07]. We summarize one useful result in Lemma 1 below. Lemma 1 can help us bound the number of tests required during skeleton recovery with PC algorithm.

Lemma 1. *Given a DAG $\mathcal{G}$ with faithfulness distribution. The population level PC algorithm constructs the true skeleton of the $\mathcal{G}$. And the maximal reached value of l: $m^* \in \{s-1, s\}$, here $s = \max_{i\in 1,\ldots,p} |ne(i)|$.*

With the help of the separation sets recorded in Algorithm 1, the second part of the PC algorithm extends the skeleton to CPDAG through several criteria, which is also well-known as the Meek's rule, see Algorithm 2.

Algorithm 1 The PC algorithm (skeleton estimation)

INPUT: vertex Set V, conditional independence information
OUTPUT: estimated skeleton $\mathcal{G}_{ske}$, seperation set S
from the complete undirected graph $\tilde{\mathcal{G}}$
$l = -1$; $\mathcal{G} = \tilde{\mathcal{G}}$
repeat
  $l = l + 1$
  repeat
    select an ordered pair of nodes i, j that are adjacent in $\mathcal{G}$ such that $|\text{ne}(i)\backslash\{j\}| \geq l$
    repeat
      choose (new) $\mathcal{S} \subset \text{ne}(i)\backslash\{j\}$ with $|\mathcal{S}| = l$.
      if i and j are conditionally independent given $\mathcal{S}$ **then**
        delete edge i, j
        denote this new graph by $\mathcal{G}$
        save $\mathcal{S}$ in $S(i, j)$ and $S(j, i)$
      end if
    until edge i, j is deleted or all $\mathcal{S} \subset \text{ne}(i)\backslash\{j\}$ with $|\mathcal{S}| = l$ have been chosen
  until all ordered pairs of adjacent variables i and j such that $|\text{ne}(i)\backslash\{j\}| \geq l$ and $\mathcal{S} \subset \text{ne}(i)\backslash\{j\}$ with $|\mathcal{S}| = l$ have been tested for conditional independence
until for each ordered pair of adjacent nodes i, j: $|\text{ne}(i)\backslash\{j\}| < l$

Algorithm 2 extending the skeleton to the essential graph (or called as CPDAG)

INPUT: skeleton $\mathcal{G}_{ske}$, separation sets S
OUTPUT: essential Graph $\mathcal{G}_{ess}$
for all pairs of nonadjacent variables i, j with common neighbour k **do**
if $k \notin S(i, j)$ **then**
  replace $i - k - j$ in $\mathcal{G}_{ske}$ by $i \to k \leftarrow j$
end if
orient more edges through the following rules:
R1 orient $j - k$ into $j \to k$ whenever there is an arrow $i \to j$ such that i and k are nonadjacent
R2 orient $i - j$ into $i \to j$ whenever there is a chain $i \to k \to j$
R3 orient $i - j$ into $i \to j$ whenever there are two chains $i - k \to j$ and $i - l \to j$ such that k and l are nonadjacent
R3 orient $i - j$ into $i \to j$ whenever there are two chains $i - k \to l$ and $k \to l \to j$ such that k and l are nonadjacent

CHAPTER 3

Learning skeleton

3.1 Difficulty with Interventional Data

The main difficulty is from interventional data, and such influence performs quite complicatedly. Suppose there is an edge $i \to j$ in a DAG $\mathcal{G}$. Then the arrow will disappear in the corresponding interventional graph $\mathcal{G}^I$ if node j is intervened, i.e. $j \in I$. To test the existence of edge between node i and j, either i or j under intervention could make the conclusion unreliable, since the direction remains unknown in skeleton recovery step.

Definition 6. *If there exists an intervention design $I \in \mathcal{I}$ such that $I \cap \{i, j\} = \emptyset$, we say the correlation between node i and j is accessible under $\mathcal{I}$.*

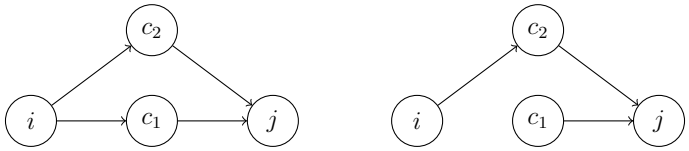
Intervention I will not change the existence of edge $i - j$ in skeleton recovery if $I \cap \{i, j\} = \emptyset$, but the estimated CPDAG may be influenced as Meek's rule relies on the separating set. In PC algorithm, we increase the separating set size gradually to find the minimal $\mathcal{S}^*$ such that:

$$X_i \perp X_j | X_{\mathcal{S}} \iff \mathcal{S} \text{ d-separate } i \text{ and } j \iff \text{all trails between } i \text{ and } j \text{ are blocked by } \mathcal{S}.$$

Thus for any $k \in \mathcal{S}^*$, there are three possible cases $\cdots \to k \to \cdots$, $\cdots \leftarrow k \leftarrow \cdots$ and $\cdots \leftarrow k \to \cdots$. Notice it is possible that a separating set contains the collider of a v-structure, however the minimal separating set $\mathcal{S}^*$ will not include any collider.

Suppose $\mathcal{S}_{ij}$ is a minimal separating set for node i and j: if node k in $i \leftarrow k \to j$ is

intervened, the minimal separating set $\mathcal{S}_{ij}$ is still valid (nothing happens in this case); if node k in $i \to k \leftarrow j$ is intervened, the minimal separating set $\mathcal{S}_{ij}$ is still valid (the trail is already blocked by collider); if node k in an open trail $i \to k \to j$ is intervened, the minimal separating set is $\mathcal{S}_{ij} \backslash \{k\}$ (intervention cuts off the trail). Thus the last case may result in a smaller separating set, and finally the vertex k will be recognized as a collider and incorrect $v-$structure $i \to k \leftarrow j$ could be added into CPDAG.



(a) True DAG. (b) Intervene on c_1.

Figure 3.1: The intervention blocks an active trail $i \to c_1 \to j$ in $\mathcal{G}$.

For example, in Figure 3.1, the intervention on c_1 blocks the active trail $i \to c_1 \to j$ in $\mathcal{G}$ such that there is only one active trail left in the interventional graph $\mathcal{G}^{\{c_1\}}$. In this simple case, they are leading to different separating sets $\mathcal{S}_{ij}$, $\{c_1, c_2\}$ for (a) and $\{c_2\}$ for (b). Suppose we have both observational and interventional data, i.e. $\mathcal{I} = \{\emptyset, \{c_1\}\}$, and the true skeleton can still be recovered by the power of observational data. While applying Meek's rule, with the separating set $\mathcal{S}_{ij} = \{c_2\}$, $i - c_1 - j$ will be identified as a v-structure $i \to c_1 \leftarrow j$ since the common neighbor $c_1 \notin \mathcal{S}$. Thus we may recognize $\{c_1\}$ as a collider incorrectly by Meek's rule, if we don't make a calibration on the separating set. And we can choose the largest minimal separating set for calibration to avoid this issue, if there are more than one minimal separating sets corresponding to different interventional data blocks.

Meanwhile, the change of structure from intervention actually makes nodes in each data block has its own corresponding partial correlation, which means we cannot treat interventional data as a whole to calculate the partial correlation. Intuitively, such mixture data could be regarded as a whole drawn from an average distribution, but which is helpless to establish the theory.

3.2 Skeleton Learning on Interventional Data

Under Gaussian assumption, the conditional independence test (2.10) is equivalent to testing the partial correlation $\rho_{i,j|\mathcal{S}}$. Generally, there are three methods to calculate the partial correlation: linear regression, inverse of covariance matrix and incursive method. Here we choose linear regression: to test $X_i \not\perp X_j \mid X_{\mathcal{S}}$, regress X_j on X_i with $X_{\mathcal{S}}$ and the coefficient of X_i determines the conditional independence. More details on the relation between partial neighborhood regression coefficient and conditional independence will be discussed in the latter section; see (5.5).

As discussed in the previous section, the interventional setting over data blocks could affect the coefficient of regression. Even those nodes not involved in the regression can still make a huge impact, which means it is hard to merge some blocks and do fewer tests. Consider this, we do regression within each block given specific intervention. For the re-labeled family of intervention targets $\mathcal{I}_{(i,j)} = \{I_1, I_2, \ldots, I_{B_{(i,j)}}\}$ with $B_{(i,j)} = |\mathcal{I}_{(i,j)}|$, where $\mathcal{I}_{(i,j)}$ is the set of intervention targets that are informative with respect to edge $i - j$, i.e. $\mathcal{I}_{(i,j)} = \{I \in \mathcal{I} \mid I \cap \{i, j\} = \emptyset\}$, we use $\mathcal{H}^k_{ij|\mathcal{S}}$ to represent $X_j \sim X_i + X_{\mathcal{S}}$ on data block $\mathbf{X}_{I_k}$ and the corresponding underlying linear model,

$$\mathcal{H}^k_{ij|\mathcal{S}} : X_j = \beta^k_{ij|\mathcal{S}} X_i + \beta^k_{\mathcal{S}j|\mathcal{S}} X_{\mathcal{S}} + \varepsilon^k_{ij|\mathcal{S}} \text{ with } \varepsilon^k_{ij|\mathcal{S}} \sim \mathcal{N}(0, (\sigma^k_{ij|\mathcal{S}})^2), \tag{3.1}$$

for $k = 1, 2, \ldots, B_{(i,j)}$. The reason we define a new notation $\mathcal{I}_{(i,j)}$ here instead of using $\mathcal{I}$ directly is that from Definiton 6 some intervention blocks may be infeasible to conduct PC algorithm, actually $\mathcal{I}_{(i,j)} \subset \mathcal{I}$ and $B_{(i,j)} \leq |\mathcal{I}|$.

To determine a single coefficient of linear regression nonzero or not, it is a classic question

that we can use the $t-$statistic,

$$T^k_{ij|\mathcal{S}} = \frac{\hat{\beta}^k_{ij|\mathcal{S}}}{s^k_{ij|\mathcal{S}}\sqrt{\left((\mathbf{X}^k_{ij|\mathcal{S}})^T\mathbf{X}^k_{ij|\mathcal{S}}\right)^{-1}_{ii}}}: \text{ if } |T^k_{ij|\mathcal{S}}| \geq \alpha_n \text{ we accept } \beta^k_{ij|\mathcal{S}} \neq 0, \tag{3.2}$$

where $s^k_{ij|\mathcal{S}}$ is an unbiased estimator of $\sigma^k_{ij|\mathcal{S}}$, a well-known result from linear regression,

$$s^k_{ij|\mathcal{S}} = SSR^k_{ij|\mathcal{S}}/(n^k_{ij|\mathcal{S}} - p^k_{ij|\mathcal{S}}), \tag{3.3}$$

here α_n is the critical value of test and $\mathbf{X}^k_{ij|\mathcal{S}}$ is the design matrix. In equation (3.3), $SSR^k_{ij|\mathcal{S}}$ is the Residual Sum of Squares defined from linear regression, and p is the number of predictors.

Lemma 2. *If nodes i and j are blocked by $\mathcal{S}$ in $\mathcal{G}$ and $I \cap \{i, j\} = \emptyset$, then i and j are also blocked by $\mathcal{S}$ in $\mathcal{G}^I$.*

Algorithm 1 shows the main part of PC algorithm given by [KB07], in which PC algorithm repeatedly selects the subset $\mathcal{S}$ from the neighbor of node i, i.e. $\mathcal{S} \subset \text{Adj}(\mathcal{G}, i)\backslash\{j\}$, and remove edge $i - j$ from the estimated skeleton once the conditional independence test shows $X_i \perp X_j \mid X_\mathcal{S}$. Under interventional data setting, as discussed in Section 3.1, it's challenging to conduct the conditional independence test over interventional blocks. So based on Lemma 2 and $t-$statistic (3.2), we can combine these tests to determine the conditional independence between i and j given $\mathcal{S}$, that is,

$$\text{If } |T^k_{ij|\mathcal{S}}| \leq \alpha_n \text{ for all } k = 1, \ldots, B_{(i,j)} \text{ then we accept } H_0 : \rho_{ij|\mathcal{S}} = 0; \tag{3.4}$$

otherwise, reject H_0. Now we replace the line 11 in Algorithm 1 with (3.4) also (3.2) to conduct the conditional independence test with finite samples generated by intervention.

CHAPTER 4

From Skeleton to $\mathcal{I}$-Essential Graph

4.1 Interventional Essential Graph and Reversible Edges

Definition 7 ([AMP97], Definition 3.3)**.** *Let $\mathcal{G}$ be a graph. An arrow $i \to j \in \mathcal{G}$ is strongly protected $\in \mathcal{G}$ if $i \to j$ occurs in at least one of the following four configurations as an induced subgraph of $\mathcal{G}$:*

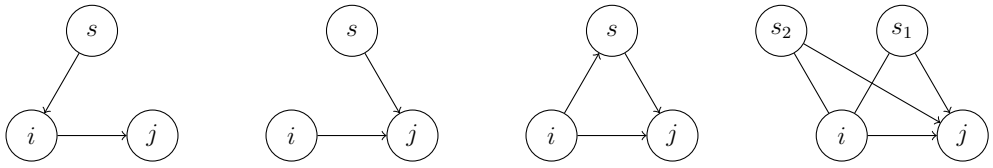


Figure 4.1: Four configurations of strongly protected arrow $i \to j$.

The configurations of strongly protected arrow in Definition 7 guarantee that the direction of this kind of arrow is irreversible. In configuration (a), it would induce new v-structure $s \to i \leftarrow j$ if the direction of arrow $i \to j$ has been reversed; in (b), it would eliminate the existing v-structure $i \to j \leftarrow j$; in (c), it would lead to a loop; in (d), to avoid new v-structure $s_1 \to i \leftarrow s_2$, there must exist $i \to s_1$ or $i \to s_2$, then the reversal of $i \to j$ would create a loop. The strongly protected arrow shows the status of compelled edges of DAG $\mathcal{G}$. And we have the following lemma based on this.

Lemma 3 ([AMP97], Theorem 4.1)**.** *$\mathcal{G}_{ess}$ is essential graph for some DAG $\mathcal{G}$ if and only if $\mathcal{G}_{ess}$ satisfies the following four conditions.*

(1) $\mathcal{G}_{ess}$ *is a chain graph.*

(2) Every chain component of $\mathcal{G}_{ess}$ *is chordal.*

(3) The configuration $a \to b - c$ *does not occur as an induced subgraph of* $\mathcal{G}_{ess}$*.*

(4) Every arrow $a \to b \in \mathcal{G}_{ess}$ *is strongly protect in* $\mathcal{G}_{ess}$*.*

As mentioned in its original paper [AMP97], Lemma 3 depicts the characterization of essential graph precisely. Since an essential graph $\mathcal{G}$ is a chain graph, let $\mathcal{K}_1, \ldots, \mathcal{K}_m$ be a disjoint partition of $\mathcal{G}$. Then in essential graph $\mathcal{G}$, every reversible edge must belong to some chain component $\mathcal{K}_l$ and if $|\mathcal{K}_l| \geq 3$, the reversible edge might be an edge of some triangle. For reversible edge $i - j \in \mathcal{K}_l$, if there exists a directed edge $k \to i$, then $k \to j$, otherwise the configuration $k \to i - j$ occurs; and also $k \to s$ for any other $s \in \mathcal{K}_l$, since $\mathcal{K}_l$ is a chordal and there always exists a trail from i to s.

For a reversible edge $i - j$ in the essential graph $\mathcal{G}_{ess}$, the neighborhood of vertex i or j can be characterized with the help of Lemma 3. Without loss of generality, focus on j's neighborhood $\text{ne}(j)$. If there exists $s \in \text{ne}(j)$ with direction $s \to j$, s must be a common parent of $i - j$ in $\mathcal{G}_{ess}$, otherwise the configuration $k \to i - j$ occurs, shown as Figure 4.2(b). A common parent s of $i - j$ must belong to the upstream chain component, i.e. if $i - j \in \mathcal{K}_l$ and $s \in \mathcal{K}_t$ then $t < l$.

If $s \in \text{ne}(j)$ belongs to the downstream chain component, it can be a common child of $i - j$ or child of j only, shown as Figure 4.2(c, e). Intervention on vertices from downstream cannot affect $i - j$, in this case they are not our interests.

The most complicated case is that $s \in \text{ne}(j)$ is in the same chain component with $i - j$. Since $s \in \text{ne}(j)$, s must be connected with j through an undirected edge. If s is also connected with i, s is a common linked node forming a triangle (i, j, s), respective to Figure 4.2(a); and define $\text{lk}(ij)$ for the set of common linked nodes, see definition 8(b). If s is not connected to i, s can be a single linked node within $\text{ne}(j)$, see Figure 4.2(d), which is the simple case. Another case is there exists a trail between s and i even though no direct edge

connecting them. Because the trail connecting s and i belongs to one chain component, the loop $i - j - \cdots - i$ consists of two or several triangles, see Figure 4.2(f) as an example.

To formalize case (f), we say a vertex s is a neighboring triangle node of j if (1) $s - j \in \mathcal{G}_{ess}$ (2) there exists an undirected trail from i to s of length ≥ 2 and not passing j. Based on this definition, Figure 4.2(f) actually shows a specific case in which the trail contained starting from i to s has length 2 exactly.

Definition 8. *For a reversible edge $i - j$ in the essential graph $\mathcal{G}_{ess}$, define:*

- *(a) the common parent set of $i - j$ as*

$$cp(ij) = \{s \in V \mid s \to i, s \to j \text{ in } \mathcal{G}_{ess}\};$$

- *(b) the common linked set of $i - j$ as*

$$lk(ij) = \{s \in V \mid s - i, s - j \text{ in } \mathcal{G}_{ess}\};$$

- *(c) the blocker set of $i - j$ as*

$$\mathcal{L}_{ij} = cp(ij) \cup lk(ij).$$

Proposition 4. *Given a reversible edge $i - j$ in the essential graph $\mathcal{G}_{ess}$, a neighborhood node $s \in ne(j)$ must belong to one of following configurations:*

- *(a) s is a common linked node, i.e. $s \in lk(ij)$;*
- *(b) s is a common parent node, i.e. $s \in cp(ij)$;*
- *(c) s is a common child node, i.e. $s \in \{s \in V \mid i \to s, j \to s \text{ in } \mathcal{G}_{ess}\}$;*

- *(d) s is a single linked node, i.e.*

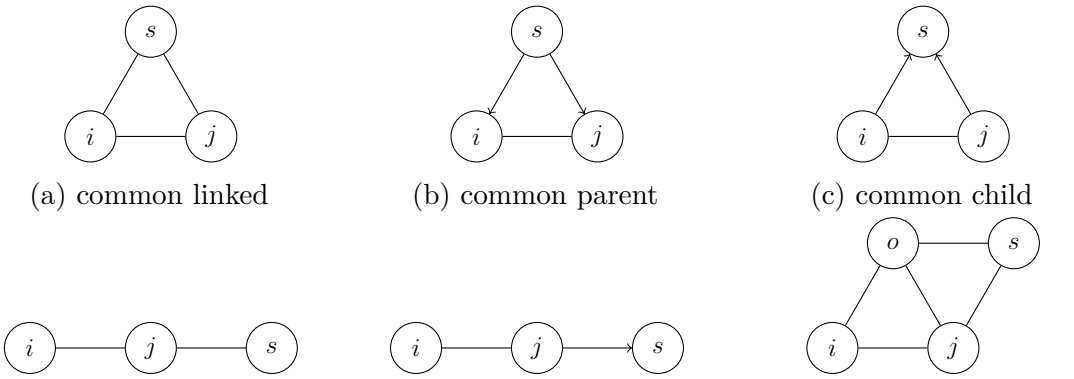
$$s \in \{s \in V \mid j - s \text{ and } i - j - s \text{ is the only undirected trail between } i \text{ and } s \text{ in } \mathcal{G}_{ess}\};$$

- *(e) s is a single child node, i.e.*

$$s \in \{s \in V \mid j \to s \text{ and there is no } i \to s \text{ in } \mathcal{G}_{ess}\};$$

- *(f) s is a neighboring triangle node, i.e.*

$$s \in \{s \in V \mid j - s \text{ and there exists an undirected trail from } i \text{ to } s \text{ of length } \geq 2 \text{ not passing } j \text{ in } \mathcal{G}_{ess}\}.$$



(a) common linked (b) common parent (c) common child

(d) single linked (e) single child (f) neighboring triangle

Figure 4.2: Neighborhood behavior of vertex j.

Figure 4.2 lists all the possible cases of $\text{ne}(j)$ by enumeration, where $\text{lk}(ij)$ and $\text{cp}(ij)$ are our interests also their union $\mathcal{L}_{ij}$; see Definition 8. Notice that trail $i-o-s-j$ in Figure 4.2(f) also goes through vertex o, which is a part of $\text{lk}(ij)$. Consider the true underlying DAG $\mathcal{G}$, all trails connecting i and j can be classifed to three cases: (1) the direct connection $i \leftrightarrow j$ itself; (2) the trail is inactive given empty observed set; (3) the trail is active given empty observed set. If the trail belongs to case (3), it must pass through $\mathcal{L}_{ij}$. That is the motivation

of methods around $\mathcal{L}_{ij}$ introduced in next section, also the reason why other nodes can be ignored in the discussion.

Lemma 5. *Let $\mathcal{G}_{ess}$ be the essential graph of a DAG $\mathcal{G}$ and $i - j$ be a reversible edge of $\mathcal{G}_{ess}$. Then, any trail connecting i and j of length ≥ 2 is blocked by empty set if the trail does not pass through $\mathcal{L}_{ij}$.*

Lemma 6. *Let $\mathcal{G}_{ess}$ be the essential graph of a DAG $\mathcal{G}$ and $i - j$ be a reversible edge of $\mathcal{G}_{ess}$. Then, any trail connecting i and j of length ≥ 2 is blocked by $\mathcal{L}_{ij}$ if the trail does not pass through $\mathcal{L}_{ij}$ and the direction of edge $i - j$ is $i \to j$.*

Lemma 5 and 6 shows the good property on reversible edge $i - j$ in the essential graph, and $\mathcal{L}_{ij}$ plays an important role in this discussion. All trails not passing through $\mathcal{L}_{ij}$ are blocked by empty set, and will not be activated by $\mathcal{L}_{ij}$ if the direction is $i \to j$. In other words, only those trails that passes through $\mathcal{L}_{ij}$ should be added into consideration when determining the d-separation relations in DAG $\mathcal{G}$.

Lemma 7. *Let $\mathcal{G}_{ess}$ be the essential graph of a DAG $\mathcal{G}$ and $i - j$ be a reversible edge of $\mathcal{G}_{ess}$. Then any neighboring triangle node s defined in Proposition 4 must be a common child of $o - j$ if the direction of edge $i - j$ is $i \to j$. Here o is the closest node to s on the undirected trail $i - \cdots - o - s$.*

Lemma 7 ensures the neighboring triangle node s is not the parent of node j in DAG $\mathcal{G}$ if the direction is $i \to j$. And s is in the downstream of $i - j$, thus the intervention on s cannot affect the reversible edge $i - j$.

For instance, an essential graph $\mathcal{G} = \mathcal{K}_1 \cup \mathcal{K}_2 \cup \mathcal{K}_3 \cup \mathcal{K}_4 = \{1, 2, 3\} \cup \{4, 5, 6\} \cup \{7\} \cup \{8\}$ in Figure 4.3. And for any reversible edge $i - j$ we can define $\mathcal{L}_{ij}$, still in Figure 4.3,

$$\mathcal{L}_{12} = \mathcal{L}_{13} = \emptyset, \quad \mathcal{L}_{45} = \{2, 3, 6\}, \quad \mathcal{L}_{46} = \{2, 3, 5\}, \quad \mathcal{L}_{56} = \{2, 3, 4\},$$

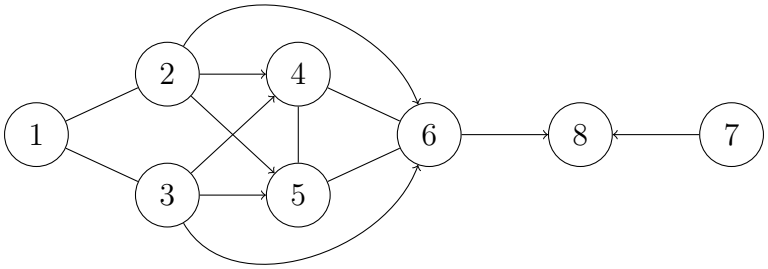


Figure 4.3: An example of the essential graph.

and all reversible edges could be classified into two classes: (1) $|\mathcal{L}_{ij}| = 0$ (2) $|\mathcal{L}_{ij}| \neq 0$. If the reversible edge with $|\mathcal{L}_{ij}| = 0$, in this case, it can be in the root chain component $\mathcal{K}_1$ under some chain component partition $\mathcal{K}_1, \ldots, \mathcal{K}_m$. To describe the $\mathcal{I}$-Markov equivalence class, we give the definition of $\mathcal{I}$-essential graph $\mathcal{G}_\mathcal{I}$ at first:

Definition 9 ([HB12], Definition 11)**.** *The $\mathcal{I}$-essential graph $\mathcal{G}_\mathcal{I}$ associated with $\mathcal{G}$ and $\mathcal{I}$ is the graph*

$$\mathcal{G}_\mathcal{I} = \cup_{\mathcal{G}' \in [\mathcal{G}]_\mathcal{I}} \mathcal{G}',$$

that is, $\mathcal{G}_\mathcal{I}$ is the smallest graph larger than every $\mathcal{G}' \in [\mathcal{G}]_\mathcal{I}$.

Recall $[\mathcal{G}]_\mathcal{I}$ represents the $\mathcal{I}$-Markov equivalence class of graph $\mathcal{G}$ and the union is graph theoretic union. Under this definition, $\mathcal{G}_\mathcal{I}$ is a subset of $\mathcal{G}_{ess}$, and its difference shows the power of the family of targets $\mathcal{I}$. [HB12] extend the characterization of essential graph in Lemma 3 to the interventional case; see Lemma 8 below about $\mathcal{I}$-essential graph.

Lemma 8 ([HB12], Theorem 18)**.** *$\mathcal{G}_\mathcal{I}$ is $\mathcal{I}$-essential graph for some DAG D if and only if*

(1) $\mathcal{G}_\mathcal{I}$ is a chain graph;

(2) Every chain component of $\mathcal{G}_\mathcal{I}$ is chordal;

(3) No induced subgraph of the form $a \to b - c$;

(4) Every arrow $a \to b \in G_\mathcal{I}$ is $\mathcal{I}$-strongly protected in $\mathcal{G}_\mathcal{I}$;

(5) $\mathcal{G}_{\mathcal{I}}$ has no line $a - b$ for which there exists some $I \in \mathcal{I}$ such that $|I \cap \{a, b\}| = 1$.

An arrow $i \to j$ is $\mathcal{I}$-strongly protected if there exists $I \in \mathcal{I}$ such that $|\{i, j\} \cap I| = 1$ or the arrow satisfies one of four configurations mentioned in Definition 7. The difference between the definition of 'strongly protected' and '$\mathcal{I}$-strongly protected' is from the power of intervention. The intervention can guarantee new identifiable edge orientation. In other words, some new irreversible arrows would be brought into $\mathcal{G}_{\mathcal{I}}$ by the intervention, besides those strongly protected arrows. Hence $\mathcal{I}$-strongly protected arrows can be understood as the combination of irreversible edges brought from intervention and the irreversible edges given by graph structure itself.

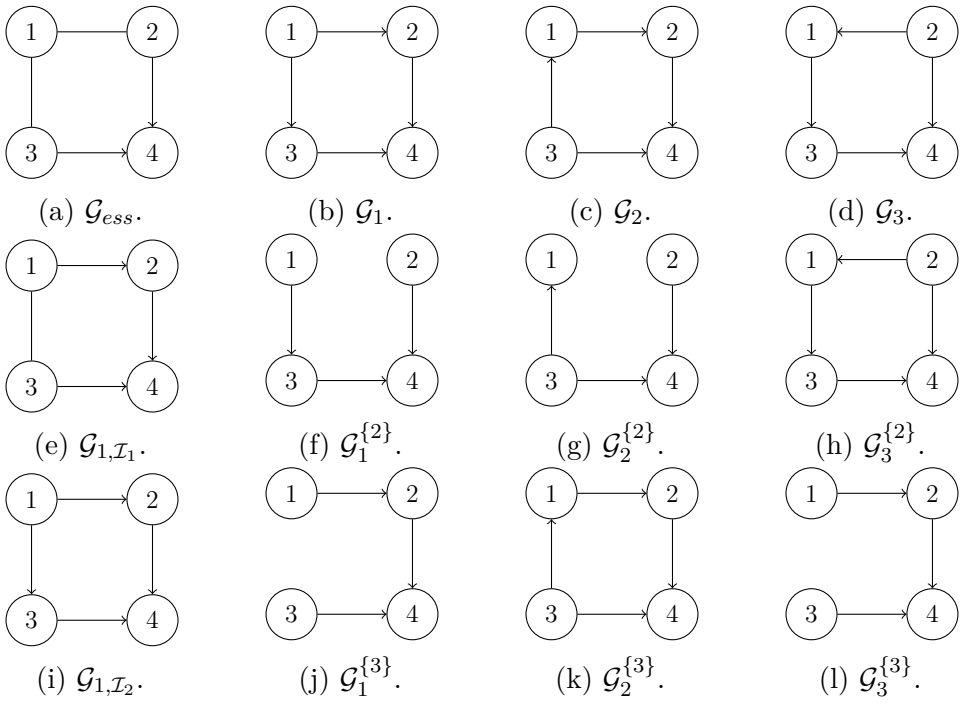
Lemma 8 motivates us to approach the $\mathcal{I}$-essential graph sequentially. Given essential graph $\mathcal{G}_{ess}$ and the intervention family $\mathcal{I}$, define edge set

$$E_{\mathcal{I}} = \{(i, j); i < j, i - j \in \mathcal{G}_{ess} | \exists I \in \mathcal{I} \text{ such that } |I \cap \{i, j\}| = 1\}, \tag{4.1}$$

which contains all the undirected edges can be determined with interventional data. In practice, our algorithm can infer the edge direction for $E_{\mathcal{I}}$ one by one, and apply Meek's rule to get the $\mathcal{I}$-essential graph $\mathcal{G}_{\mathcal{I}}$ after all edges in $E_{\mathcal{I}}$ are done.

The first row of Figure 4.4 shows three DAGs that are observationally Markov equivalent, as they, $\mathcal{G}_1, \mathcal{G}_2$ and $\mathcal{G}_3$, have the same skeleton and a single v-structure $2 \to 4 \leftarrow 3$. Now suppose $\mathcal{G}_1$ is the true DAG, which is targeted to recover from interventional data. For family of targets $\mathcal{I}_1 = \{2\}$, edge $1 - 2$ can be oriented, i.e. $E_{\mathcal{I}_1} = \{(1, 2)\}$ defined in (4.1). With the help of $\mathcal{I}_1$, there is a partition over three DAGs: $[\mathcal{G}_1]_{\mathcal{I}_1} = \{\mathcal{G}_1, \mathcal{G}_2\}$ and $[\mathcal{G}_2]_{\mathcal{I}_1} = \{\mathcal{G}_3\}$, also $\mathcal{G}_{1,\mathcal{I}_1}$ in Figure 4.4(e) is the $\mathcal{I}$-essential graph corresponding to the interventional Markov equivalence class $[\mathcal{G}_2]_{\mathcal{I}_1}$.

The true DAG $\mathcal{G}$ is indistinguishable with single intervention on $\{2\}$. To enhance the power of intervention, if $\{3\}$ is also intervened, now $[\mathcal{G}_1]_{\mathcal{I}_1}$ can be partitioned more precisely; check the difference between $\mathcal{G}_1^{\{3\}}$ and $\mathcal{G}_2^{\{3\}}$ in Figure 4.4. Let $\mathcal{I}_2 = \{\{2\}, \{3\}\}$, then $[\mathcal{G}_1]_{\mathcal{I}_2} =$



(a) $\mathcal{G}_{ess}$. (b) $\mathcal{G}_1$. (c) $\mathcal{G}_2$. (d) $\mathcal{G}_3$.

(e) $\mathcal{G}_{1,\mathcal{I}_1}$. (f) $\mathcal{G}_1^{\{2\}}$. (g) $\mathcal{G}_2^{\{2\}}$. (h) $\mathcal{G}_3^{\{2\}}$.

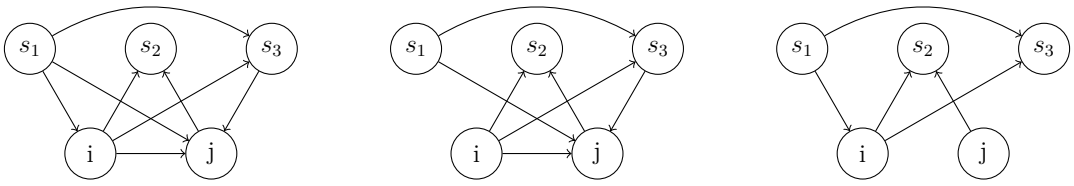
(i) $\mathcal{G}_{1,\mathcal{I}_2}$. (j) $\mathcal{G}_1^{\{3\}}$. (k) $\mathcal{G}_2^{\{3\}}$. (l) $\mathcal{G}_3^{\{3\}}$.

Figure 4.4: An example of the edge orientation with intervention.

$\{\mathcal{G}_1\}$. In other words, we can find the true DAG $\mathcal{G}_1$ with the family of targets $\mathcal{I}_2$, i.e. $\mathcal{G}_{1,\mathcal{I}_1} = \mathcal{G}_1$. The edge set $E_{\mathcal{I}_2} = \{(1,2),(1,3)\}$ and can be oriented sequentially in practice.

4.2 Edge Orientation with Experimental Data

To determine the direction of reversible edges, one preferred way is to test the difference from intervention. Figure 4.5 gives an example to show this difference. And the motivation



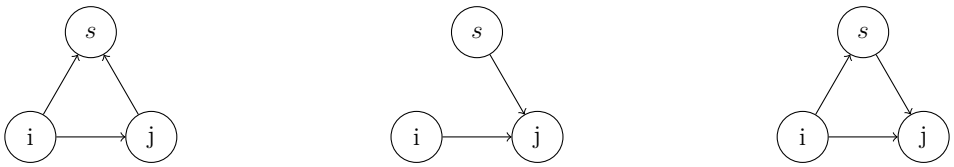
(a) True DAG. (b) Intervene on i. (c) Intervene on j.

Figure 4.5: An example of the difference from intervention.

of our algorithm comes from one observation: if the true direction is $i \to j$, the intervention

of node j will damage the local neighborhood of node j, but the intervention of node i will not.

Assume $i \to j$ and i is under intervention, consider the common linked node, i.e. (a) in Figure 4.2, there are three possible cases in the true underlying DAG: case (a) and (c) still hold the graph structure in Figure 4.6, which implies the conditional relations and correlation coefficients are invariant under intervention. Since the intervention cut the trail $i \leftarrow s \to j$ in case (b), still seeking for some invariant quantity under intervention, conditioning on node s is a good choice, as conditioning can be regarded as a blocking operation.



(a) s is common child. (b) s is common parent. (c) $s : i \to s \to j$.

Figure 4.6: Three cases of common linked node in $i \to j$ with intervened i.

We assume the joint distribution for X is defined by a set of SEMs with a DAG $\mathcal{G}$, as in Section 2.2.

Theorem 9. *Consider a reversible edge $i - j$ in an essential graph $\mathcal{G}_{ess}$. Suppose the underlying DAG is $\mathcal{G}$. If the edge direction is $i \to j$ in $\mathcal{G}$, then the conditional distribution $[X_j \mid X_i, \{X_s, s \in \mathcal{L}_{ij}\}]$ is invariant after intervention on X_i.*

From Theorem 9, to orient the direction of reversible edge, it suffices to check whether the conditional distribution $[X_j \mid X_i, \{X_s, s \in \mathcal{L}_{ij}\}]$ has changed under intervention. It is hard to compare conditional distributions directly, so corresponding partial regression coefficients can be a good substitute, as the invariant conditional distribution ensures the invariant regression.

If X_i is under intervention in some intervention target,

$$\mathcal{H}_{\text{obs},j\sim i} : X_j \sim X_{i,\text{obs}} + X_{\mathcal{L}_{ij}}, \quad \mathcal{H}_{\text{int},j\sim i} : X_j \sim X_{i,\text{int}} + X_{\mathcal{L}_{ij}},$$

the first regression $\mathcal{H}_{\text{obs},j\sim i}$ uses observational data block, while the second one using interventional block with $I = \{i\}$. To represent the underlying linear equation, we always use subscript 0 in all notations related to $\mathcal{H}_{\text{obs},j\sim i}$ and 1 for $\mathcal{H}_{\text{int},j\sim i}$, and we use $\mathcal{L}$ to replace $\mathcal{L}_{ij}$ for simplicity,

$$X_j = \beta_{0,ij} X_{i,\text{obs}} + \beta_{0,\mathcal{L}j} X_{\mathcal{L}} + \varepsilon_{0,ij}, \quad X_j = \beta_{1,ij} X_{i,\text{int}} + \beta_{1,\mathcal{L}j} X_{\mathcal{L}} + \varepsilon_{1,ij},$$

with the invariant conditional distributions guaranteed by Theorem 9. Then the null hypothesis,

$$H_0 : X_i \to X_j \Longrightarrow H_0 : \beta_{0,ij} = \beta_{1,ij} \text{ and } \text{var}(\varepsilon_{0,ij}) = \text{var}(\varepsilon_{1,ij}), \tag{4.2}$$

now we transfer the orientation problem to a two-sample test about the partial coefficients of neighborhood linear regression.

To test the null hypothesis (4.2), the test statistics can be used is,

$$T_{i\to j} = \frac{\hat{\beta}_{0,ij} - \hat{\beta}_{1,ij}}{s_{p,ij}\sqrt{\left(\mathbf{X}_{0,ij}^T \mathbf{X}_{0,ij}\right)^{-1}_{ii} + \left(\mathbf{X}_{1,ij}^T \mathbf{X}_{1,ij}\right)^{-1}_{ii}}} \sim t_{n_{1,ij}+n_{2,ij}-2p_{ij}}, \tag{4.3}$$

in (4.3),

$$s_{p,ij} = \sqrt{\frac{(n_{0,ij} - p_{ij}) s_{0,ij}^2 + (n_{1,ij} - p_{ij}) s_{1,ij}^2}{n_{0,ij} + n_{1,ij} - 2p_{ij}}}, \quad p_{ij} = |\mathcal{L}| + 1,$$

and Section 7.3 gives all details about the test statistics derivation. Then the direction of edge can be determined through,

$$\textbf{if } |T_{i\to j}| > \alpha \textbf{ so } H_0 \text{ is rejected}, \tag{4.4}$$

here α is the critical value of test. If H_0 is rejected, the alternative hypothesis leads to a

different conditional distribution under intervention, based on above discussion also Theorem 9, we can orient $X_i - X_j$ as $X_i \leftarrow X_j$ in this case.

Remark 1. *It's not enough to use common parents only as regressors in the neighborhood linear regression, as possibly there may be some common parents in DAG $\mathcal{G}$ hidden as common linked nodes in the essential graph $\mathcal{G}_{ess}$, and the ignorance of such hidden common parents is unacceptable. Notice that it is crucial to condition on all the common parents.*

Remark 2. *It is worthwhile to mention that, if j is under intervention with $i \to j$, the partial regression coefficient $\beta^{\mathcal{G}_j}_{ij|\mathcal{L}_{ij}} = 0$ may not hold, even no edge between i and j exists under intervention. Suppose s is a hidden common child of $i - j$, coefficient $\beta^{\mathcal{G}_j}_{ij|s}$ is nonzero because conditioning on s activates the trial $i \to s \leftarrow j$.*

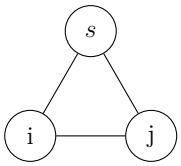
4.3 Extend to A General Intervention Target

In previous section, we introduce the intuition of our method under single intervention, i.e. only one node of $\{i, j\}$ is intervened. Move to the general case, as we have no constraint on the intervention target, some intervention target will be 'bad' for our algorithm. Obviously, the data block with intervention target I_k that both X_i and X_j are under intervention, that is $|I_k \cap \{i, j\}| = 2$, cannot provide any information about the reversible edge $i - j$.

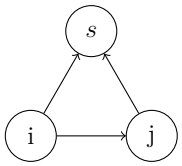
Meanwhile, the node set $[p]$ can be separated to three parts by $i - j$ corresponding to one typological sort: the upstream of $i - j$, the midstream between $i - j$ and the downstream of $i - j$. Intervention on the nodes of downstream could be safe in the most cases, since it cannot influence the joint density of $[X_i, X_j, X_{\mathcal{L}}]$ at all. The one special case is the common child hidden in $\mathcal{L}$, but which is avoidable in practice. Compared to downstream, intervention on upstream will not change the structure of induced subgraph, and our interest, the conditional density $[X_j \mid X_i, X_{\mathcal{L}}]$ also keeps invariant.

Figure 4.7 shows a different case: if intervention target I_k contains any node from $lk(ij)$, i.e. $|I_k \cap lk(ij)| \neq 0$, this intervention can endanger the coefficient comparison in the regres-

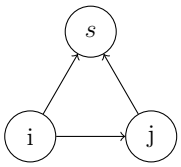
sion $X_j \sim X_i + X_s$ while the $s \in lk(ij)$ is actually the common child node in the true DAG. And those common child nodes hidden in the essential graph prevents us merging such kind of blocks, since the $[X_j \mid X_i, X_s]$ has changed over block 1 and 2.



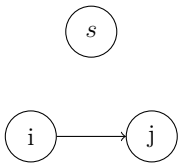
(a) Essential Graph.



(b) True DAG.



(c) Block 1 with $I = \{i\}$.



(d) Block 2 with $I = \{i, s\}$.

Figure 4.7: Example about common child with intervention.

To solve this difficulty, use $lk(ij)$ to classify and merge different intervention targets. Recall our algorithm, to judge one undirected edge $i - j$, we need to run regressions $X_j \sim X_i + X_{cp(ij)} + X_{lk(ij)}$ over two paired data design matrices. Start from the most natural case, in which we can construct a pair of two design matrices that have no intersection with $lk(ij)$. Here define two index sets,

$$\begin{aligned} S_{obs} &= \{k \in [B] \mid I_k \cap \{i, j\} = \emptyset, I_k \cap lk(ij) = \emptyset\}, \\ S_{int} &= \{k \in [B] \mid I_k \cap \{i, j\} = \{i\}, I_k \cap lk(ij) = \emptyset\}. \end{aligned} \tag{4.5}$$

If $S_{obs} \neq \emptyset$ and $S_{int} \neq \emptyset$, we can merge all data blocks indexed in S_{obs} to construct the observational design matrix, that is $\mathbf{X}_{0,ij}$ in (4.3), similarly construct $\mathbf{X}_{1,ij}$ with S_{int}. Actually, data block with intervention target $I_k \in S_{obs}, |I_k| \neq 0$ can be regarded as 'purely observational data', and which implies that such kind of data blocks can be merged with the observational block $I_k = \emptyset$ if there is. That's important to include observational data into this merge step as generally data block with $I_k = \emptyset$ has the largest sample size.

Due to the flexibility of intervention family $\mathcal{I}$, one or both of (S_{obs}, S_{int}) defined in (4.5)

could be empty. In this case, if we still plan to do edge orientation, for any $\mathcal{L}_{sub} \subset lk(ij)$,

$$\begin{aligned} S_{obs} &= \{k \in [B] \mid I_k \cap \{i,j\} = \emptyset, I_k \cap lk(ij) = \mathcal{L}_{sub}\}, \\ S_{int} &= \{k \in [B] \mid I_k \cap \{i,j\} = \{i\}, I_k \cap lk(ij) = \mathcal{L}_{sub}\}, \end{aligned} \tag{4.6}$$

such that $S_{obs} \neq \emptyset$ and $S_{int} \neq \emptyset$, our algorithm still works on this dataset. The intuition of (4.6) is that some intervention targets can ensure the same local environment around undirected edge $i - j$ such that it is still possible to approach the difference from intervened node i only. Notice that $\mathcal{L}_{sub}$ can be empty, and which implies (4.5) is a special case of (4.6).

Theorem 10. *For a reversible edge $i - j$ and corresponding $\mathcal{L}_{ij}$ in an essential graph $\mathcal{G}_{ess}$. Suppose the underlying DAG is $\mathcal{G}$. If the edge direction is $i \to j$ in $\mathcal{G}$. Given the family of interventional targets $\mathcal{I}$, for any $\mathcal{L}_{sub} \subset lk(ij)$, if we can find (S_{obs}, S_{int}) defined in (4.6), then the following conclusions hold:*

(a) $[X_j \mid X_i, \{X_s, s \in \mathcal{L}_{ij}\}]$ *is invariant over interventions* $\{I_k, k \in S_{obs}\}$*, which we denote as* $[X_j \mid X_i, \{X_s, s \in \mathcal{L}_{ij}\}]_{obs}$.

(b) $[X_j \mid X_i, \{X_s, s \in \mathcal{L}_{ij}\}]$ *is invariant over interventions* $\{I_k, k \in S_{int}\}$*, which we denote as* $[X_j \mid X_i, \{X_s, s \in \mathcal{L}_{ij}\}]_{int}$*;*

(c) $[X_j \mid X_i, \{X_s, s \in \mathcal{L}_{ij}\}]_{obs}$ *is the same as* $[X_j \mid X_i, \{X_s, s \in \mathcal{L}_{ij}\}]_{int}$.

Remark 3. *The proofs of both Theorems 9 and 10 are based on the graph structure, and no assumption on the joint distribution is involved in the invariance rule. The invariance rule does not need Gaussian assumption, just a general SEM according to a DAG. Therefore, we can apply this rule to other distributions, even discrete case.*

Remark 4. *It's preferred to compare two coefficients instead of testing zero or nonzero of marginal correlation ρ_{ij}. Both Theorems 9 and 10 can be extended to the soft interventional case, as the d-separation discussion among the proofs still works for the soft intervention.*

Due to high-dimensional setting and sparsity, the danger set $lk(ij)$ would be much smaller than the edge set $[p]$. Consider this, the merge step can be very useful such that both $\mathbf{X}_{0,ij}$ and $\mathbf{X}_{1,ij}$ have adequate data samples. In practice, since observational data is cheap and common, it is ideal to conduct the comparison based on (4.5). But still we are not preferred to set any constraint on the intervention family $\mathcal{I}$ in this dissertation, and also to ensure the integrity of the theoretical result, (4.6) is introduced.

In this case, given interventional target $\mathcal{I}$ and undirected edge $i - j$, the crucial step is to search any $\mathcal{L}_{sub} \subset lk(ij)$ satisfied (4.6) such that both S_{obs} and S_{int} are nonempty. It is possible that we can't establish any test, if there is no feasible pair of $(\mathcal{S}_{obs}, \mathcal{S}_{int})$ given specific interventional target $\mathcal{I}$. Meanwhile, sometimes there are more than one $\mathcal{L}_{sub} \subset lk(ij)$ can be found, and in this case multiple tests can be conducted for edge orientation. Number the pairs of $(\mathcal{S}^k_{obs}, \mathcal{S}^k_{int})$ with $k = 1, 2, \ldots, K$ where K is the total number of such pairs. And similar to (3.2), to handle multiple tests,

$$X_i \to X_j \Longrightarrow \beta^k_{0,ij} = \beta^k_{1,ij} \text{ and } \operatorname{var}(\varepsilon^k_{0,ij}) = \operatorname{var}(\varepsilon^k_{1,ij}) \text{ for each } (\mathcal{S}^k_{obs}, \mathcal{S}^k_{int}),$$

with $k = 1, 2, \ldots, K$. The number of such tests K should be quite limited, and for simplicity the rest of this dissertation will focus on the case having only one test, especially for the consistency proof part; check Section 5.3 for details. Now we can give Algorithm 3 to summarize the method introduced in this section.

The Algorithm 3 conducts only one side of the edge orientation, as it is always testing the invariance relations after intervention on X_i. For reversible edge $i - j$, obviously not only $X_j \sim X_i + X_{\mathcal{L}_{ij}}$, the reversal one $X_i \sim X_j + X_{\mathcal{L}_{ij}}$ can also be used for the edge orientation. At the population level, these two tests should be consistent, which means one always rejects H_0 while another not. Sometimes we don't have various enough intervention targets. In practice, the orientation can be decided from one side regression only, or remain as an undirected edge in result if two tests are inconsistent. And in this case, the orientation rule in lines 5-12 of

Algorithm 3 Recover the Interventional Essential Graph (Population)

INPUT: essential graph $\mathcal{G}_{ess}$, intervention target $\mathcal{I}$
OUTPUT: estimated $\mathcal{I}$-Essential graph $\hat{\mathcal{G}}_{\mathcal{I}}$
repeat
  select a pair of nodes (i, j) from edge set $E_{\mathcal{I}}$ defined in (4.1)
  if for (i, j), there exists a qualified pair $(\mathcal{S}_{obs}, \mathcal{S}_{int})$ satisfied rule (4.6) **then**
    run regression $X_j \sim X_{i,obs} + X_{\mathcal{L}_{ij}}$ on merged block indexed by $\mathcal{S}_{obs}$
    run regression $X_j \sim X_{i,int} + X_{\mathcal{L}_{ij}}$ on merged block indexed by $\mathcal{S}_{int}$
    if $\beta_{0,ij} = \beta_{1,ij}$ **then**
      orient $i - j$ as $i \to j$
    else
      orient $i - j$ as $j \to i$
    end if
  else
    continue.
  end if
until all undirected edges have been tested.

Algorithm 3 needs to be revised slightly to include the criteria for two tests.

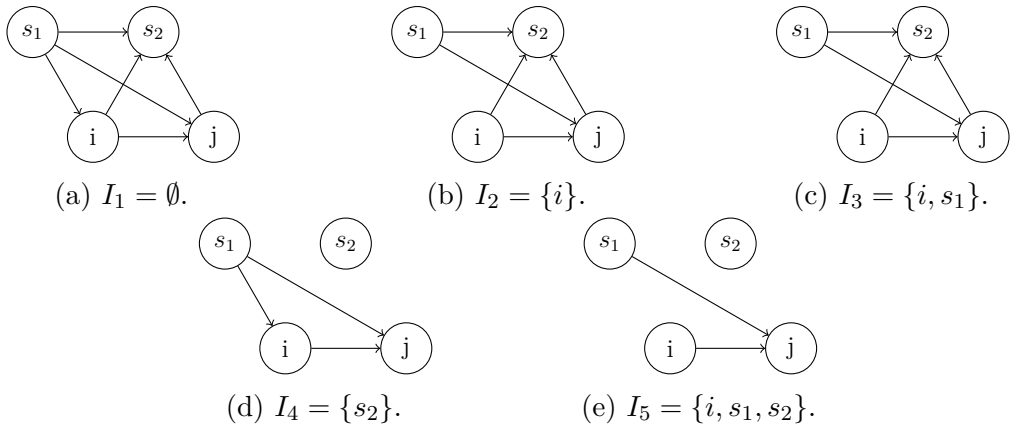
To make it easier for understanding, Figure 4.8 serves as an example on how to implement algorithm 3 in practice. There is one common parent and one common linked node in Figure 4.8, and the family of intervention targets is $\mathcal{I} = \{\emptyset, \{i\}, \{i, s_1\}, \{s_2\}, \{i, s_1, s_2\}\}$. Figure 4.9 shows all interventional graphs.

(a) Essential Graph. (b) True DAG.

Figure 4.8: Undirected edge $i - j$ with one common parent s_1 and one common linked node s_2.

Then there exists two possible pairs of $(\mathcal{S}_{obs}, \mathcal{S}_{int})$ considering $lk(ij) = \{s_2\}$. The first one is $(\mathcal{S}^{(1)}_{obs}, \mathcal{S}^{(1)}_{int}) = (\{1\}, \{2, 3\})$, and another is $(\mathcal{S}^{(2)}_{obs}, \mathcal{S}^{(2)}_{int}) = (\{4\}, \{5\})$. For $(\mathcal{S}^{(1)}_{obs}, \mathcal{S}^{(1)}_{int})$, we can merge data blocks $\mathbf{X}_2$ and $\mathbf{X}_3$, corresponding to $I_2 = \{i\}$ and $I_3 = \{i, s_1\}$, to prepare interventional data block $\mathbf{X}^{(1)}_{1,ij}$, implied by $\mathcal{S}^{(1)}_{int} = \{2, 3\}$; observational data block $\mathbf{X}^{(1)}_{0,ij}$ is just

$\mathbf{X}_1$ as $\mathcal{S}_{obs}^{(1)} = \{1\}$. Finally conduct the partial neighborhood regressions $X_j \sim X_i + X_{s_1} + X_{s_2}$ on both $\mathbf{X}_{0,ij}^{(1)}$ and $\mathbf{X}_{1,ij}^{(1)}$, and make the edge orientation based on the results from test (4.2).



(a) $I_1 = \emptyset$. (b) $I_2 = \{i\}$. (c) $I_3 = \{i, s_1\}$.

(d) $I_4 = \{s_2\}$. (e) $I_5 = \{i, s_1, s_2\}$.

Figure 4.9: The interventional graphs with $\mathcal{I}$.

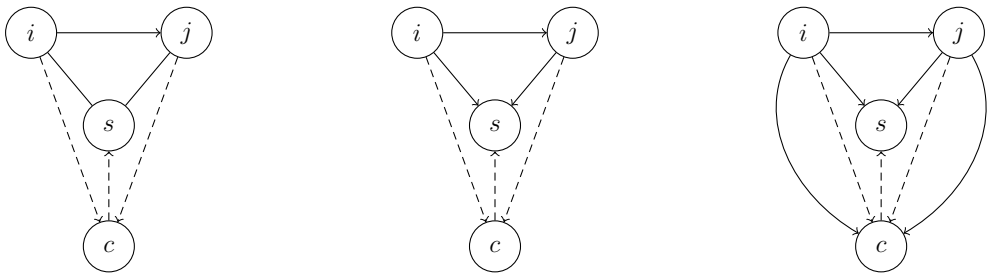
We can construct another test on $(\mathcal{S}_{obs}^{(2)}, \mathcal{S}_{int}^{(2)})$ separately. Observational data block $\mathbf{X}_{0,ij}^{(2)}$ is $\mathbf{X}_4$, and interventional data block $\mathbf{X}_{1,ij}^{(2)}$ is $\mathbf{X}_5$ given $(\mathcal{S}_{obs}^{(2)}, \mathcal{S}_{int}^{(2)}) = (\{4\}, \{5\})$. In practice, edge orientation can admit the consistent result only, and abandon this attempt to leave the edge remain undirected if two test results are inconsistent as expected.

4.4 Proof of Section 4

Proof of Lemma 5. It suffices to show that all the trails not passing through $\mathcal{L}_{ij}$ will have at least one collider. Based on Proposition 4, the trail must go though the node in case (c, d, e) if not passing via $\mathcal{L}_{ij}$. For the common child case (c), it is trivial as s is collider in $i \to s \leftarrow j$. For (d) and (e), there are three possible cases of node s: $i \to j \to s, i \leftarrow j \to s, i \leftarrow j \leftarrow s$. We use $i \to j \to s$ as an example to show why the trail connecting i and j passing through s cannot avoid collider. First, the trail connecting i and j must look like $i \to j \to s \to \cdots$ as collider s is not preferred; then it will create a loop if there is no collider in $i \to j \to s \to \cdots \leftrightarrow i$. So there must be a collider in the trail. Similar discussion for $i \leftarrow j \to s$ and $i \leftarrow j \leftarrow s$. □

Proof of Lemma 6. Due to Lemma 5, all the trails not passing through $\mathcal{L}_{ij}$ will have at least one collider, and the collider itself cannot be in $\mathcal{L}_{ij}$. Thus it suffices to show that there exists at least one collider, and any descendants of which are not included in $\mathcal{L}_{ij}$, , as d-separation requires all colliders meet the rule.

In this case, we can focus on the first collider in trail $j \cdots \to c \leftarrow \cdots i$, i.e. the nearest collider to j. Let dashed line represent the connection between two vertices with trail of length ≥ 1. In Figure 4.10, $s \in \mathcal{L}_{ij}$ is the descendant of collider c and the direction between i and j is $i \to j$.



(a) s is the descendant. (b) $i \to s$ and $j \to s$. (c) $i \to c$ and $j \to c$.

Figure 4.10: The descendant of collider in trail connecting i and j.

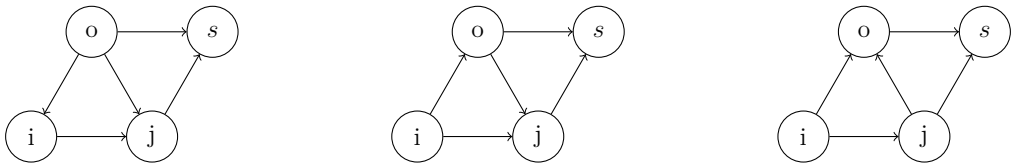
Considering the trail between j and c, the arrow must point away from j in this trail to avoid new v-structure $i \to j \leftarrow \cdots$. As c is the nearest collider to j, there is no more colliders in the trail $j \to \cdots \to c$, which means c is the descendant of j. Now the direction of s and j must be $j \to s$ to avoid cycle among s, j and c. Then $i \to s$ to avoid cycle among i, j and s.

Right now we show s must be a hidden common child in $\mathcal{L}_{ij}$. Suppose $c \to \cdots \to s_1 \to s$, it requires $s_1 \in \text{ne}(i) \cap \text{ne}(j)$ such that there is no v-structures $i \to s \leftarrow s_1$ and $j \to s \leftarrow s_1$. Then we can repeat the discussion of last paragraph on vertex s_1 to show that s_1 is also the common child of i and j. Finally, c is the common child of i and j in $\mathcal{G}$.

Vertices i, j and s belong to one chain component in $\mathcal{G}_{ess}$. Recall the definition of chain graph, it is impossible that $i \to c$ and $j \to c$ are visible in $\mathcal{G}_{ess}$. In other words, c is the

hidden common child of i and j, i.e. $c \in \mathcal{L}_{ij}$. It implies contradiction as we have assumed that the trail does not pass through $\mathcal{L}_{ij}$. □

Proof of Lemma 7. Recall the formulation of neighboring triangle node. If the length of undirected edge connecting i with s equals to 2, i.e. $i - o - s$, Figure 4.11 shows that s is the common child of o and j. Once $i \to j$, there must be $j \to s$ to avoid inducing new $v-$structure (i, j, s); set $o \to s$ to avoid a loop $o \to j \to s$ in case (a) and to avoid new $v-$structure (i, o, s) in case (b, c). Thus s is the common child of $o - s$ while $i \to j$ in all cases.

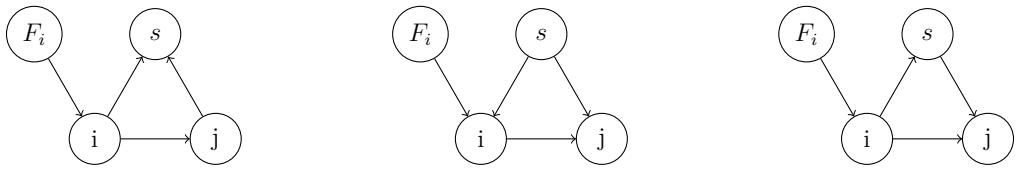


(a) $i \leftarrow o \to j$ (b) $i \to o \to j$ (c) $i \to o \leftarrow j$

Figure 4.11: Three cases of the neighboring triangle node s.

Furthermore, for the general case, suppose the undirected edge connecting i and s with length $k + 1 \geq 3$, the trail can be represented as $i - \cdots - o_k - o_{k+1} - s$. First, it is trivial to show $j \to s$ as avoiding new v-structure $i \to j \leftarrow s$. Then assume the statement is true for k, o_{k+1} is the common child of o_k and j. If s is not the common child of o_{k+1} and s, there must be an edge between o_k and s in the essential graph to eliminate possible new v-structure $o_k \to o_{k+1} \leftarrow s$. In this case, there exists an undirected trail $i - \cdots - o_k - s$ with length k. Contradict! Thus o_{k+1} is the common child of o_k and j, and the whole statement is true. □

Proof of Theorem 9. We regard intervention on X_i as an additional parent F_i for X_i, and denote this parent variable as node F_i in the graph. Based on Lemma 5, all trails with length ≥ 3 connecting F_i and j must go through s if the trail is active. Consider this, without loss of generality, we simplify the whole problem to three cases; see Figure 4.12.



(a) s is common child. (b) s is common parent. (c) $s : i \to s \to j$.

Figure 4.12: Introduce auxiliary node f to represent intervention on i.

It is easy to see that $\{i, s\}$ d-separates F_i and j in Fig 4.12(b); and both i and $\{i, s\}$ d-separate F_i and j in Fig 4.12(a)(c). Return to the general case, given DAG $\mathcal{G}$ augmented with additional node F_i, all trails between F_i and j are inactive, and they are blocked by $\{i\} \cup \mathcal{L}_{ij}$. Thus we have X_j independent of F_i given $X_i, \{X_s, s \in \mathcal{L}_{ij}\}$, which implies that the conditional distribution $[X_j | X_i, \{X_s, s \in \mathcal{L}_{ij}\}]$ is invariant after intervention on X_i. □

Proof of Theorem 10. Add additional parent variables $\{f_m\}_{m \in I}$ for $\{X_m\}_{m \in I}$ corresponding to the intervention target I. And meanwhile there is an augmented graph with nodes $\{f_m\}_{m \in I}$. Consider the trail $f_m \to \cdots \leftrightarrow s \leftrightarrow j$ in the underlying true DAG $\mathcal{G}$, s must be in the ne(j).

Then for any $s \in \text{ne}(j)$ but $s \notin \mathcal{L}_{ij}$, recall the discussion in Section 4.1, s can be one of the four cases: (1) single child of j; (2) common child of j; (3) single linked node of j; (4) neighbored triangle node of j. For (1) and (2), it is trivial that the trail will look like $f_m \to \cdots \leftrightarrow s \leftarrow j$, which is blocked by some v-structure in the trail, if the observed set $\mathcal{S} = \emptyset$. Consider the observed set $\mathcal{S} = \{i\} \cup \mathcal{L}_{ij}$, it is possible that the v-structure collider is included in $\mathcal{S}$ such that the trail is activated. Assume $s_0 \in \mathcal{S} = \{i\} \cup \mathcal{L}_{ij}$ is the collider in the trail, it is easy to find:

$$f_m \to \cdots \to s_0 \leftarrow s \leftarrow j \text{ is active given } \mathcal{S} \iff f_m \to \cdots \to s_0 \leftarrow j \text{ is active given } \mathcal{S}, \quad (4.7)$$

notice that if LHS exists, there must have a trail in the RHS of (4.7), since $s_0 \in \mathcal{S} \subset \text{ne}(j)$ and $s_0 \to j$ would lead to a loop. Actually (4.7) shows that we can simplify the discussion

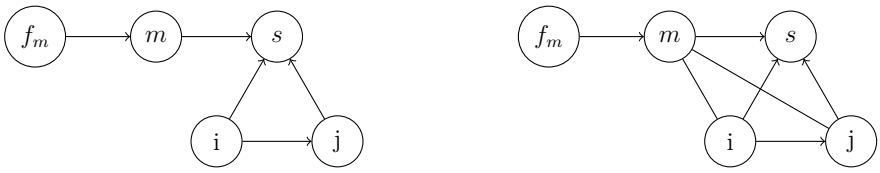
to the trails go through $\{i\} \cup \mathcal{L}_{ij}$, which is the topic in the next paragraph. For (3) and (4), rely on Lemma 7, the trail will still look like $f_m \to \cdots \leftrightarrow s \leftarrow j$, similarly to (1) and (2).

Next focus on the trails like $f_m \to \cdots \leftrightarrow s \leftrightarrow j$ with $s \in \{i\} \cup \mathcal{L}_{ij}$. If $s \in pa(j)$, then by local Markov property, $pa(j)$ blocks all trails from f_m to j. If $s \in \{i\} \cup \mathcal{L}_{ij} - pa(j)$, s must be a child of j in the true underlying DAG $\mathcal{G}$. In other words, only if $s \in \text{lk}(ij) \in \mathcal{L}_{ij}$ and actually s is a hidden common child, the trail could be active given the observed set $\mathcal{S} = \{i\} \cup \mathcal{L}_{ij}$.

Based on the discussion above, given the observed set $\mathcal{S} = \{i\} \cup \mathcal{L}_{ij}$, any trail like $f_m \to \cdots \leftrightarrow s \leftrightarrow j$ would be blocked by $\mathcal{S}$ unless s is a hidden common child node. Thus to determine f_m and j is d-separated or not given $\mathcal{S}$ in the augmented graph, it suffices to check the status of trail via hidden common child nodes.

Consider node $m \notin \mathcal{L}_{ij}$, assume $f_m \to m \to s \leftarrow j$ is active given observed s; see Figure 4.13(a). However it is illegal graph, as it contains two v-structures $m \to s \leftarrow j$ and $m \to s \leftarrow i$. Both will lead to the directed edge $j \to s$ in the essential graph. If we eliminate v-structures by connecting m to both i and j, then $m \in \mathcal{L}_{ij}$ conflicts the assumption; see Figure 4.13(b).

Actually m can only be a hidden common child to make the trail active, since observed m would block the trail like $f_m \to m \to j$. It is a stronger result but we cannot distinguish any hidden common child from the essential graph in practice, thus let's focus on $\text{lk}(ij)$.



(a) illegal graph. (b) no v-structure.

Figure 4.13: Example for non-hidden common child node m.

Finally, we already show there is no active trail between f_m and j given the separated

set $\mathcal{S} = \{i\} \cup \mathcal{L}_{ij}$ for any $m \notin \text{lk}(ij)$. In order words, if $i \to j$, then

$$X_j \perp F_m \mid \{X_i, X_s, s \in \mathcal{L}_{ij}\},$$

for any $m \notin \text{lk}(ij)$. Return to the definition of $\mathcal{S}_{obs}$,

$$X_j \perp F_m \mid \{X_i, X_s, s \in \mathcal{L}_{ij}, F_{\mathcal{L}_{sub}} = do(\mathcal{L}^*_{sub}), F_{\text{lk}(ij)-\mathcal{L}_{sub}} = idle\}, \tag{4.8}$$

for any $m \notin \text{lk}(ij)$ where $\mathcal{L}^*_{sub}$ denotes the intervened values.

Now (4.8) suffices to prove the result in (a). To prove (b),

$$X_j \perp F_i \mid \{X_i, X_s, s \in \mathcal{L}_{ij}, F_{\mathcal{L}_{sub}} = do(\mathcal{L}^*_{sub}), F_{\text{lk}(ij)-\mathcal{L}_{sub}} = idle\}, \tag{4.9}$$

actually (4.9) is a special case of (4.8), and no additional comment is required, as the trail like $f_i \to i \to j$ is blocked by $\{i\}$ trivially.

Base on the definition of (S_{obs}, S_{int}) defined in (4.6), we can combine (a) and (b) to get result in (c). □

CHAPTER 5

High-Dimensional Consistency

Let us revisit notations defined in Section 2 before moving to the consistency results. A Gaussian DAG model can be represented as a linear structural equation model,

$$X_j = \sum_{k \in \mathrm{pa}(j)} \beta_{kj} X_k + \varepsilon_j, \quad j = 1, 2, \ldots, p, \tag{5.1}$$

where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are independent and $\varepsilon_j \sim \mathcal{N}(0, \omega_j^2)$. Here $pa(j)$ represents the parent node set of j, and $\beta_{kj} \neq 0$ only if $(k, j) \in E$. The coefficient β_{kj} represents the causal effect of X_k on X_j. The SEM (5.1) defines a joint Gaussian distribution for,

$$X = (X_1, X_2, \ldots, X_p) \sim \mathcal{N}_p(0, \mathbf{\Sigma}), \tag{5.2}$$

Our methods apply to any type of intervention target family, $\mathcal{I} = \{I_1, I_2, \ldots, I_B\}$. A general $n \times p$ data matrix $\mathbf{X}$ generated under $\mathcal{I}$ consists of a number of data blocks $\mathbf{X}^i$ with n_i rows and p columns for $i = 1, 2, \ldots, B$. Each row within the same block $\mathbf{X}^i$ is drawn i.i.d. from $\mathcal{N}(0, \mathbf{\Sigma}^i)$, but data rows from different blocks are not identically distributed. Here $\mathbf{\Sigma}^i$ corresponds to $\mathbf{\Sigma}^{I_i}$ in (5.2) to simplify the notation. An observational data block if exists could be treated as $I = \{\emptyset\}$ for notation consistency. To make it easier to understand, the general setting of data under a family of intervention targets $\mathcal{I} = \{I_1, I_2, \ldots I_B\}$ is:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \\ \vdots \\ \mathbf{X}^B \end{pmatrix} \sim \begin{pmatrix} \mathcal{N}(0, \mathbf{\Sigma}^1) \\ \mathcal{N}(0, \mathbf{\Sigma}^2) \\ \vdots \\ \mathcal{N}(0, \mathbf{\Sigma}^B) \end{pmatrix} \quad \text{with } \sum_{i=1}^{B} n_i = n.$$

And now we can start the discussion on consistency results.

5.1 Assumptions

Some assumptions are required to establish the theoretical results under the sparse high-dimensional settings.

Assumption 1. *Dimension assumption:* $p = O(n^a)$ *with some non-negative constant* $a \geq 0$.

For $a \geq 1$, we have a high-dimensional setting of $p \gg n$.

Assumption 2. *Sparsity assumption:*

$$s = \max_{i \in 1,\ldots,p} |ne(X_i)| = O(n^{1-b}), \text{ with constant } 1/2 < b \leq 1.$$

Assumption 3. *There is a uniform upper bound of diagonal entries over all covariance matrices:*

$$\max_{i=1,\ldots,B} \left\{ \max_{j=1,\ldots,p} \left(\mathbf{\Sigma}^i\right)_{jj} \right\} \leq \bar{\sigma}^2. \tag{5.3}$$

The upper bound on diagonal entries in (5.3) can be used as an upper bound on $\text{var}(X_i)$. Suppose the interventional distribution is $\mathcal{N}(0, \tau^2)$ for some intervened node, it's easy to see assumption 3 also bounds the variation of intervened value.

Assumption 4. *There is a uniform lower bound of the minimal eigenvalue over all covariance matrices:*

$$\min_{i=1,\ldots,B} \left\{ \lambda_{min} \left(\mathbf{\Sigma}^i \right) \right\} \geq \sigma_*^2, \tag{5.4}$$

here $\lambda_{min}(\cdot)$ *represents the minimal eigenvalue.*

Assumption 4 provides a uniform lower bound for variance of noise in neighborhood linear regression, see details in proof chapter, and also implies a lower bound for marginal variance $\text{var}(X_i)$ and the variance of interventional variable τ.

Definition 10. *Strong Faithfulness with respect to the SEM (2.1):* $\inf_{i,j,S}\{|\rho_{i,j|S}| : \rho_{ij|S} \neq 0\} \geq c_n$ *for* $c_n = \Omega(n^{-d})$*, where the constant* $d \in (0, b/2)$*.*

Strong faithfulness assumption reveals the signal strength contained in a network. One intuition is the weak signal strength will cause the failure of structure recovery. That's why many researchers relies on this assumption, especially in discussion of consistency: PC algorithm [KB07], GES [NHM18]. [GB13] substitute strong faithfulness with beta-min condition, which requires a part of edges with enough signal strength. This dissertation does not require the original version of strong faithfulness, in this case it is listed as a definition not assumption for reference here.

Since we always use partial regression coefficients instead of partial correlation coefficients, it is worthwhile to mention that,

$$\beta_{ij|\mathcal{S}} = \rho_{ij|\mathcal{S}} \frac{\sigma_{ji|\mathcal{S}}}{\sigma_{ij|\mathcal{S}}} \Longrightarrow \rho_{ij|\mathcal{S}} = \beta_{ij|\mathcal{S}} \frac{\sigma_{ij|\mathcal{S}}}{\sigma_{ji|\mathcal{S}}}, \tag{5.5}$$

here $\sigma^2_{ij|\mathcal{S}}$ is the conditional variance,

$$\sigma^2_{ij\mathcal{S}} = \text{Var}(X_i \mid X_j, X_{\mathcal{S}}) \Longrightarrow \sigma^2_{ij|\mathcal{S}} \text{ is the variance of noise in } X_i \sim X_j + X_{\mathcal{S}},$$

see Section 5.1.3 in [Lau96], here $\beta_{ij|\mathcal{S}}$ represents the coefficient of X_j in regression $X_i \sim X_j + X_{\mathcal{S}}$. Then with Assumption 4 and 10,

$$\beta_{ij|\mathcal{S}} = \rho_{ij|\mathcal{S}} \frac{\sigma_{j i|\mathcal{S}}}{\sigma_{ij|\mathcal{S}}} \Longrightarrow |\beta_{ij|\mathcal{S}}| \geq \psi_n = c_n \sigma_* / \sigma, \tag{5.6}$$

therefore $\psi_n^{-1} = O(n^d)$.

5.2 High-Dimensional Consistency of Algorithm 1

To ensure the interventional PC algorithm can recover every edge in the graph skeleton, as discussed in Section 3.1, also Definition 6, there is one assumption about the intervention design:

Assumption 5. *For any (i, j), the correlation between node i and j is accessible under the family of targets $\mathcal{I}$.*

Even the edge between i and j is accessible under $\mathcal{I}$, it still cannot guarantee the success of edge detection. We hope at least one design $I \in \mathcal{I}$ could keep the signal strength for node i and j, that is, the strong faithfulness bound given in Assumption 10 still hold in graph $\mathcal{G}_{\{I\}}$ under intervention I. Let $\rho^I_{ij|S}$ represent the partial correlation with intervention design I, to estimate the whole skeleton, we give the following assumption on interventional target:

Assumption 6. *For any (i, j, S) with $\rho_{ij|S} \neq 0$, there exists an intervention target $I \in \mathcal{I}$ such that $|\rho^I_{ij|S}| \geq c_n = \Omega(n^{-d})$, where the constant $d \in (0, (b-q)/2)$.*

This assumption guarantees that strong faithfulness is preserved under intervention. It could be satisfied naturally by adding observational data, i.e. $\emptyset \in \mathcal{I}$, which often happens in the real life as observational data is always cheaper and much easier to collect compared to interventional data. Assumption 6 contains an interventional version of strong faithfulness assumption, which avoids some extreme intervention target. For example, if intervention

target $I = [p]$, i.e. all nodes intervened, in this case no information about the graph structure could still remain in the data. A similar assumption is proposed in [HB15], in which they impose faithfulness assumption for all intervention targets. Here, in Assumption 6, we don't require a uniform strong faithfulness assumption over all intervention targets, instead assuming for any $(i, j, \mathcal{S})$ there exists at least one feasible intervention target, which plays an important role in judgement of edge $i - j$.

To understand the constant value d, the smaller d means the stronger signal required for the graph structure learning. Assumption 6 shows that two constants can affect the range of d. For big b and small q, the weaker signal strength is acceptable, as upper bound $(b-q)/2$ has been lifted. It meets intuition that the sparser graph with less number of interventions can lead to an easier task for structure learning.

Consider the probability of error when testing correlation $\rho_{ij|\mathcal{S}}$ in (3.4):

$$P(E_{ij|\mathcal{S}}) = P(E^I_{ij|\mathcal{S}}) + P(E^{II}_{ij|\mathcal{S}}) = P\left(|T^k_{ij|\mathcal{S}}| \geq \alpha_n, \exists k \mid \rho_{ij|\mathcal{S}} = 0\right) + P\left(|T^k_{ij|\mathcal{S}}| \leq \alpha_n, \forall k \mid \rho_{ij|\mathcal{S}} \neq 0\right). \tag{5.7}$$

For the first term in (5.7),

$$\begin{aligned} P(E^I_{ij|\mathcal{S}}) &= P\left(|T^k_{ij|\mathcal{S}}| \geq \alpha_n, \exists k \mid \rho_{ij|\mathcal{S}} = 0\right) \\ &= P\left(|T^k_{ij|\mathcal{S}}| \geq \alpha_n, \exists k \mid \beta^k_{ij|\mathcal{S}} = 0, \forall k\right) \\ &= 1 - P\left(|T^k_{ij|\mathcal{S}}| \leq \alpha_n, \forall k \mid \beta^k_{ij|\mathcal{S}} = 0, \forall k\right) \\ &= 1 - \prod_{k=1}^{B_{(i,j)}} P\left(|T^k_{ij|\mathcal{S}}| \leq \alpha_n \mid \beta^k_{ij|\mathcal{S}} = 0\right) \\ &\leq 1 - (1-\Delta)^{B_{(i,j)}} \leq |\mathcal{I}|\Delta. \end{aligned} \tag{5.8}$$

Let Δ represent a uniform bound such that $P\left(|T^k_{ij|\mathcal{S}}| \geq \alpha_n \mid \beta^k_{ij|\mathcal{S}} = 0\right) \leq \Delta$ exists for any k. The bound derived in (5.8) motivates a constraint for the number of intervention

blocks:

Assumption 7. *The sample size of each intervention block, $n_k \gtrsim n^{1-q}$, for all $k = 1, \ldots, B$. This implies that $B \precsim n^q$ for $0 < q < b$, where B is the number of intervention blocks.*

For the second term in (5.7),

$$P(E^{II}_{ij|\mathcal{S}}) = P\left(|T^k_{ij|\mathcal{S}}| \leq \alpha_n, \forall k \mid \rho_{ij|\mathcal{S}} \neq 0\right) = P\left(|T^k_{ij|\mathcal{S}}| \leq \alpha_n, \forall k \mid \beta^k_{ij|\mathcal{S}} \neq 0, \exists k\right), \tag{5.9}$$

here Assumption 6 implies that there always exists k such that $\beta^k_{ij|\mathcal{S}} \neq 0$ once $\rho_{ij|\mathcal{S}} \neq 0$ is nonzero, which induces the second equality in (5.9) as the conditioning set is the same. For every $(i, j, \mathcal{S})$, the intervention design ensures correlation satisfying strong faithfulness is crucial in discussion. Here we use the superscript small o to mark all the notations related this intervention design and corresponding block. For example, $T^o_{ij|\mathcal{S}}$ is the t-statistic in this data block, and similarly for other notations, then the Type II error,

$$P(E^{II}_{ij|\mathcal{S}}) \leq P\left(|T^o_{ij|\mathcal{S}}| \leq \alpha_n \mid \beta^o_{ij|\mathcal{S}} \neq 0\right), \tag{5.10}$$

and the bound of (5.10) requires $0 < d < (b - q)/2$ in Assumption 10, which is stricter than the common assumption $0 < d < b/2$ in some exsiting results.

We can conclude the consistency result of algorithm 1 by bounding the errors of all tests with assumptions introduced in the last and this section. It is worth to mention that the consistency result keeps valid in both fixed p and high dimensional setting.

Theorem 11. *Under assumption 1-7, let $\hat{\mathcal{G}}_{ske}$ be the output of algorithm 1. Then there exists a sequence $\alpha_n \to \infty$ such that,*

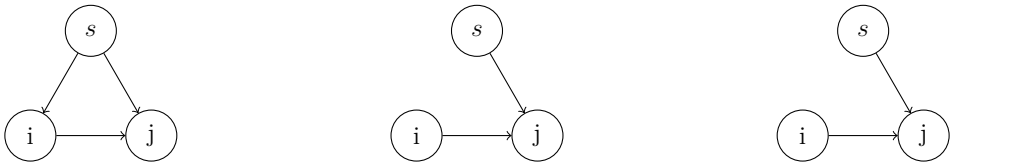
$$\lim_{n\to\infty} P(\hat{\mathcal{G}}_{ske} = \mathcal{G}_{ske}) = 1,$$

$\mathcal{G}_{ske}$ is the skeleton of graph $\mathcal{G}$.

Remark 5. *The correctness of Algorithm 2 totally depends on the skeleton and separation sets [Mee95], which means the high-dimensional consistency of Algorithm 1 also guarantees the CPDAG recovery, since all randomness is included in Theorem 11.*

5.3 High-Dimensional Consistency of Algorithm 3

In Section 4.3, the node set $[p]$ can be separated to three parts by $i - j$ corresponding to one typological sort: the upstream of $i - j$, the midstream between $i - j$ and the downstream of $i - j$. Intervention on the nodes of downstream could be safe in the most cases, since it cannot influence the joint density of $[X_i, X_j, X_{\mathcal{L}}]$ at all. The one special case is the common child hidden in $\mathcal{L}$, but which is avoidable in practice. Compared to downstream, intervention on upstream will not change the structure of induced subgraph, and our interest, the conditional density $[X_j|X_i, X_{\mathcal{L}}]$ also keeps invariant. But the joint density may be impacted hugely by this kind of intervention, for instance, we choose the simplest upstream node, a common parent:



(a) Observational Graph. (b) Block 1, $I_1 = \{i\}$. (c) Block 2, $I_2 = \{i, s\}$.

Figure 5.1: Common parent node in $i \to j$ with different intervention targets.

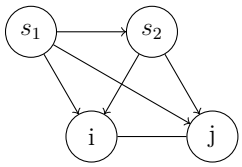
In Figure 5.1, we do regress $X_j \sim X_i + X_s$. If (a) is true, the conditional distribution $[X_j \mid X_i, X_s]$ will be invariant in (b) and (c), which implies the regression coefficients and variance of regression noise would keep exactly the same over interventional blocks $\{i\}, \{i, s\}$, thus it is reasonable to merge these two data blocks together; see Theorem 10. The potential danger difficulty in theoretical analysis of the least-squares estimator $\widehat{\beta}$ is about the intervention on node X_s, suppose $X_s \sim \mathcal{N}(0, \tau_s^2)$, the joint density of $[X_i, X_j, X_s]$ would be different given different τ_s. The most straightforward impact is that we cannot treat the rows in

blocks $I_1 = \{i\}$ and $I_2 = \{i, s\}$ as identical data samples. Let $\mathcal{N}(0, \Sigma^{\{i\}})$ represent the interventional distribution in block I_1 and $\mathcal{N}(0, \Sigma^{\{i,s\}})$ for I_2, generally $\Sigma^{\{i\}}$ and $\Sigma^{\{i,s\}}$ are different covariance matrices.

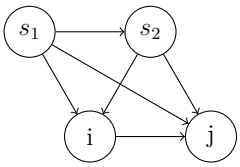
From another perspective, if we set τ_s as an extreme large value, the whole local system would be more variant than before. It brings a challenge in the coefficient test, since the power of test is related to the variance of estimated coefficients. The variance in distribution of $\hat{\beta}_{j|i,s}$ will change due to different τ_s.

To eliminate the impact from upstream intervention nodes, one way is to set the interventional distribution $\mathcal{N}(0, \tau_s^2)$ with $\tau_s^s = \text{var}(X_s)$, i.e. the intervention should keep the variance magnitude on intervened node s. In practice, someone can estimate $\text{var}(X_s)$ from purely observational data and then conduct the experiments. Then consider the theoretical part, now two covariance matrices $\Sigma^{\{i\}}$ and $\Sigma^{\{i,s\}}$ are the same.

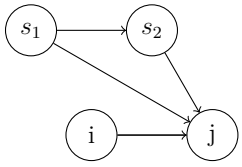
Figure 5.2 shows another different case about the effect while intervening on the node from $cp(ij)$. Compared to single intervention on $\{i\}$, in block 2 with $\{i, s_2\}$, the extra intervention on s_2 will affect the covariance matrix of joint density $[X_j, X_i, X_{s_1}, X_{s_2}]$, as the $cov(j, s_1)$ changed. This impact cannot be corrected by adjusting the interventional distribution on X_{s_1} or X_{s_2}, which introduced in last paragraph has its own limit.



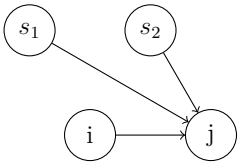
(a) Essential Graph.



(b) True DAG.



(c) Block 1 with $I = \{i\}$.



(d) Block 2 with $I = \{i, s_2\}$.

Figure 5.2: Example about path cut off by intervention.

However in Figure 5.2 we can still merge block 1 with block 2 as the conditional distri-

bution $[X_j \mid X_i, X_{s_1}, X_{s_2}]$ keeps invariant. To ensure the power of test in (4.3) is still at a good stage, mild assumption is set on the covariance matrix of each block, since intuitively any single block with extreme value can harm the whole test on the merged bigger block; see assumption (4). Furthermore, check Lemma 12 about details on how to bound $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}$, a key part of our test statistic, with mixture data.

Figure 5.1 and Figure 5.2 focus on the upstream of $i - j$, i.e. the interaction between common parent node with intervention; the midstream of $i - j$ has the similar behavior, as the path is $i \to s \to j$ and s is the parent of j. To solve this difficulty, also considering rows of our design matrix in regression are independent but not identical, Lemma 12 is introduced to handle such mixture data.

Lemma 12. *Suppose random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ has C submatrices and each submatrix $\mathbf{X}_i$ is drawn from the $\mathbf{\Sigma}^i$-Gaussian ensemble with $n_i \geq d$,*

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \\ \vdots \\ \mathbf{X}^C \end{pmatrix} \sim \begin{pmatrix} \mathcal{N}(0, \mathbf{\Sigma}^1) \\ \mathcal{N}(0, \mathbf{\Sigma}^2) \\ \vdots \\ \mathcal{N}(0, \mathbf{\Sigma}^C) \end{pmatrix} \quad \textit{with } \sum_{i=1}^{C} n_i = n,$$

then for all $1 > \delta > 0$,

$$P\left(\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}} \geq \frac{2}{\sigma_* \delta \sqrt{n}}\right) \leq Ce^{-n_*(1-\delta)^2/2} \quad \textit{for } j = 1, \ldots, d, \tag{5.11}$$

here $\sigma_* = \min_{i=1,\ldots,C}\{\gamma_{min}(\sqrt{\mathbf{\Sigma}^i})\}$ and $n_* = \min_{i=1,\ldots,C}\{n_i\}$.

Remark 6. *A relatively less number of submatrices can provide tighter probability bound as the number C shown in (5.11). This could be satisfied as we already discussed how to ensure the data sample identical by choosing good intervention variables in practice, for example, set the interventional distribution as $\mathcal{N}(0, \tau_s^2)$ with $\tau_s^s = var(X_s)$. A constant number C is preferred which doesn't vary as n increases. But it is actually not the pain point for the*

proof of consistency, as $\exp(-n_*(1-\delta)^2/2)$ *in* (5.11) *will dominate the bound such that the probability decays to zero quickly. Notice that* C *cannot equal to* n*. If each row of the data matrix has its own intervention such that* $C = n$ *and then* $n_* = 1$*, the probability bound in* (5.11) *is meaningless.*

Remark 7. *This lemma requires a uniform lower bound about minimal eigenvalues over all covariance matrices, which implies a constraint on the magnitude of variance for those interventional distributions in Assumption 4.*

Algorithm 3 can orient reversible edge in the essential graph $\mathcal{G}_{ess}$ only if there exists a qualified pair $(\mathcal{S}_{obs}, \mathcal{S}_{int})$ satisfied rule (4.6) under the family of intervention targets $\mathcal{I}$. Define edge set:

$$\tilde{E}_{\mathcal{I}} = \{(i,j); i < j, i - j \in \mathcal{G}_{ess} \mid \exists(\mathcal{S}_{obs}, \mathcal{S}_{int}) \text{ satisfied rule (4.6)}\}, \tag{5.12}$$

here $\tilde{E}_{\mathcal{I}}$ contains all edges feasible to orient by Algorithm 3. Compare $\tilde{E}_{\mathcal{I}}$ with $E_{\mathcal{I}}$ defined in (4.1), generally $\tilde{E}_{\mathcal{I}} \subseteq E_{\mathcal{I}}$ without any constraint on $\mathcal{I}$.

Assumption 8. *Given the family of targets* $\mathcal{I}$*,* $\tilde{E}_{\mathcal{I}} = E_{\mathcal{I}}$ *in the essential graph* $\mathcal{G}_{ess}$*.*

Finally, for consistency result, we focus on single test, i.e. assuming that we can always establish the test based on (4.5). To prove the consistency, let ζ_{ij} represent the fraction of data used in regression $X_j \sim X_i + X_{\mathcal{L}_{ij}}$ where $\mathcal{L}_{ij}$ is determined by the neighborhood of $i-j$ in the essential graph $\mathcal{G}_{ess}$ then,

$$\zeta_{ij} n = \zeta_{0,ij} n + \zeta_{1,ij} n = n_{0,ij} + n_{1,ij}.$$

Ideally ζ_{ij} can equal to one.

Assumption 9. *Consider essential graph $\mathcal{G}_{ess}$, given the family of targets $\mathcal{I}$,*

$$\zeta_{0,ij} \wedge \zeta_{1,ij} \geq \zeta \Longrightarrow \frac{\sqrt{\zeta_{0,ij}\zeta_{1,ij}}}{\sqrt{\zeta_{0,ij}} + \sqrt{\zeta_{1,ij}}} \geq \sqrt{\zeta^2}/(\sqrt{1} + \sqrt{1}) = \zeta/2, \tag{5.13}$$

for any $(i, j) \in E_{\mathcal{I}}$ and set $\zeta \gtrsim n^{-q}$, where the constant $q \in (0, 1)$.

Remark 8. *Equation (5.13) reveals two perspectives about the sample size of two design matrices: (1) Both $\zeta_{0,ij}$ and $\zeta_{1,ij}$ should be large enough. (2) It's better to have the closer $\zeta_{0,ij}$ and $\zeta_{1,ij}$ leading to balanced sample sizes if the sum ζ_{ij} is constant as the optimal split is to maximize the ratio $\sqrt{\zeta_{0,ij}(\zeta_{ij} - \zeta_{0,ij})}/\sqrt{\zeta_{0,ij}}\sqrt{(\zeta_{ij} - \zeta_{0,ij})}$.*

Assumption 9 requires constant q to control the magnitude of the sample size. We use the consistent notation with Assumption 7, as both of them are focused on the sample size. There is no real constraint on the range of constant q given in the Assumption 9, but actually this constant should be considered together with d in Assumption 10.

Assumption 10. *(Bounds on the gaps of partial correlations between pre- and post-intervention) For each $(\mathcal{S}_{obs}, \mathcal{S}_{int})$ defined in 4.6, let $\rho_{0,ij|\mathcal{L}}$ represent the partial correlation corresponding to $\mathcal{S}_{obs}$ and use $\rho_{1,ij|\mathcal{L}}$ for $\mathcal{S}_{int}$,*

$$\inf_{i,j,\mathcal{L}} \{\big||\rho_{0,ij|\mathcal{L}}| - |\rho_{1,ij|\mathcal{L}}|\big|\} \geq c_n, \tag{5.14}$$

for $c_n = \Omega(n^{-d})$, where the constant $d \in (0, (1 - q)/2)$.

In (4.2), it is shown that the partial regression coefficients would be invariant with $i \to j$ when regressing X_j on X_i, i.e. $\beta_{0,ij} = \beta_{1,ij}$. Consider the null hypothesis is not true, if $j \to i$, the magnitude of $|\beta_{0,ij} - \beta_{1,ij}|$ is crucial as it represents the signal strength of this intervention. If there is no hidden common child added into the regression, all trials connecting i and j will be blocked by $\mathcal{L}$ such that $\beta_{1,ij}$ equals to 0. In this case, the strong faithfulness assumption will guarantee a trivial bound for the absolute value of difference, i.e. Assumption 10.

(a) Observational block. (b) Intervention $\{i\}$.

Figure 5.3: Intervene on node i while $j \to i$.

However, if any hidden common child is added into the regression, there is one active trail between i and j since the common child is conditioned, which means now $\beta_{1,ij}$ is nonzero. In this case, the bound of signal strength is non-trivial. Figure 5.3 shows this issue for instance, in which the coefficient $\beta_{1,ij}$ of regression $X_j \sim X_i + X_s$ on interventional block (b) is nonzero.

To bound the signal, without loss of generality, assume both $\beta_{0,ij}$ and $\beta_{1,ij}$ are positive:

$$|\beta_{0,ij} - \beta_{1,ij}| = \left|\rho_{0,ij|s}\frac{\sigma_{0,ji|s}}{\sigma_{0,ij|s}} - \rho_{1,ij|s}\frac{\sigma_{1,ji|s}}{\sigma_{1,ij|s}}\right| \geq \left|\rho_{0,ij|s}\frac{\sigma_*}{\sigma} - \rho_{1,ij|s}\frac{\sigma}{\sigma_*}\right|, \tag{5.15}$$

furthermore as $\rho_{ij} = 0$ in the interventional graph,

$$\rho_{1,ij|s} = \frac{\rho_{ij} - \rho_{is}\rho_{sj}}{\sqrt{1-\rho_{is}^2}\sqrt{1-\rho_{sj}^2}} = -\rho_{0,is|j}\rho_{0,sj|i},$$

then (5.15),

$$|\beta_{0,ij} - \beta_{1,ij}| \geq \left|\left|\rho_{0,ij|s}\right|\frac{\sigma_*}{\sigma} - \left|\rho_{0,is|j}\rho_{0,sj|i}\right|\frac{\sigma}{\sigma_*}\right|, \tag{5.16}$$

(5.16) is the simplest case with three nodes.

The general intuition of Assumption 10 is the direct cause of the $j \to i$ in the graph should be stronger than other causes via trails like $i \to s \leftarrow j$. And (5.16) shows that this actually works for the simplest case, if we assume $\rho_{0,ij|s}, \rho_{0,is|j}$ and $\rho_{0,sj|i}$ have the same magnitude, as all of them are smaller than 1. However, it is not trivial to give explicit expression of the

partial correlations for the general case. We still choose to list this Assumption 10 as (5.14). Another potential gap is the role of σ and σ_* in (5.15), and we suppose it can be ignored for asymptotic results.

Now we can give our main result for this section.

Theorem 13. *Under assumptions 1-4 and 8-10, let $\hat{\mathcal{G}}_{\mathcal{I}}$ be the output of algorithm 3. Then there exists a sequence $\alpha_n \to \infty$ such that,*

$$\lim_{n\to\infty} P(\hat{\mathcal{G}}_{\mathcal{I}} = \mathcal{G}_{\mathcal{I}}) = 1,$$

where $\mathcal{G}_{\mathcal{I}}$ is the interventional essential graph.

Actually in proof of Theorem 13, we set $\alpha_n = O(n^{1/2-d-q/2})$ to show the consistency results. And for proof, it is hard to give a precise estimation of the number of reversible edges. Since we start from CPDAG, one acceptable upper bound is $ps/2$, which is the worst case that all edges are reversible. Then we need to do at most $ps/2 \times 2 = ps$ times of regressions. In practice, the actual number of tests will be much smaller than ps.

CHAPTER 6

Simulation Results

6.1 Single Edge Orientation Simulation

Algorithm 3 introduced in previous chapter could be simplified to a single edge orientation solution, i.e. focus on only one reversible edge instead of the recovery of the whole interventional essential graph. To show the power of Algorithm 3 also the efficiency of test statistics in (4.3), this simulation would prepare 1000 undirected edges for edge orientation, then the simulation can calculate the error or accuracy of the test based on the results of 1000 runnings.

For data generation, the first step is to build the adjacency matrix $\mathbf{A}$ of the graph. To ensure the randomness of the graph structure, here the entry of $\mathbf{A}$ is drawn from the Bernoulli distribution $Bernoulli(s)$ and all upper triangle including diagonal entries are set to zero to meet the DAG requirements. Then, corresponding to generated adjacency matrix $\mathbf{A}$, we can prepare the coefficient matrix $\mathbf{B}$, and finally collect data samples using SEM introduced in (2.1).

Here are some parameters:

- $p = 20, s = 4/p = 0.2$;
- $\beta_{ij} \sim Unif\left((-0.8, -0.1) \cup (0.1, 0.8)\right)$;
- the sample size set is $n \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$;
- the noise distribution is $\epsilon_i \sim \mathcal{N}(0, 0.1)$ in SEM (2.1);

- the intervention distribution is $\tau_i \sim \mathcal{N}(0, 0.1)$ if some node is intervened.

Once we generate a graph, immediately we can find its CPDAG or essential graph, only some reversible edges are left to be oriented. It is hard to give a precise estimation of the number of reversible edges in one CPDAG, but simulation shows that this number is relatively small due to the graph sparsity. Generally, we can only get several reversible edges from one randomly generated graph. Considering this, in this simulation we are using many randomly generated DAGs with different graph structures to ensure there are 1000 undirected edges in the simulation. It is preferred to have different graph structures as it can show the edge orientation rule works under various circumstances.

The last setting is the intervention family $\mathcal{I}$, here we include the single interventions on all nodes also with purely observational data:

$$\mathcal{I} = \{\emptyset, \{1\}, \{2\} \ldots, \{20\}\},$$

this intervention family guarantees that any reversible edge $i - j$ can be oriented with our algorithm. And given $i - j$ with true direction $i \to j$, there are two directions of tests for edge orientation, and one shows the Type I error of edge orientation while another is referred to the Type II error.

Figure 6.1 meets our expectations about the test statistics in (4.3). The choice of α can control Type I error, and Type II error will decay quickly as sample size increases.

As the intervention family $\mathcal{I}$ given in this section includes single intervention target for each node, we can conduct two edge orientation tests for every undirected edge. For example, consider the undirected edge $i - j$, let us orient edge by checking the invariance relation between pre- and post-intervention on node i first, and then implement on node j. If both results are consistent and correct according to the true DAG structure, we will count this edge orientation as a success. Based on this, we calculate the ratio of successes over $N = 1000$ runs, and referred as accuracy in Figure 6.1.

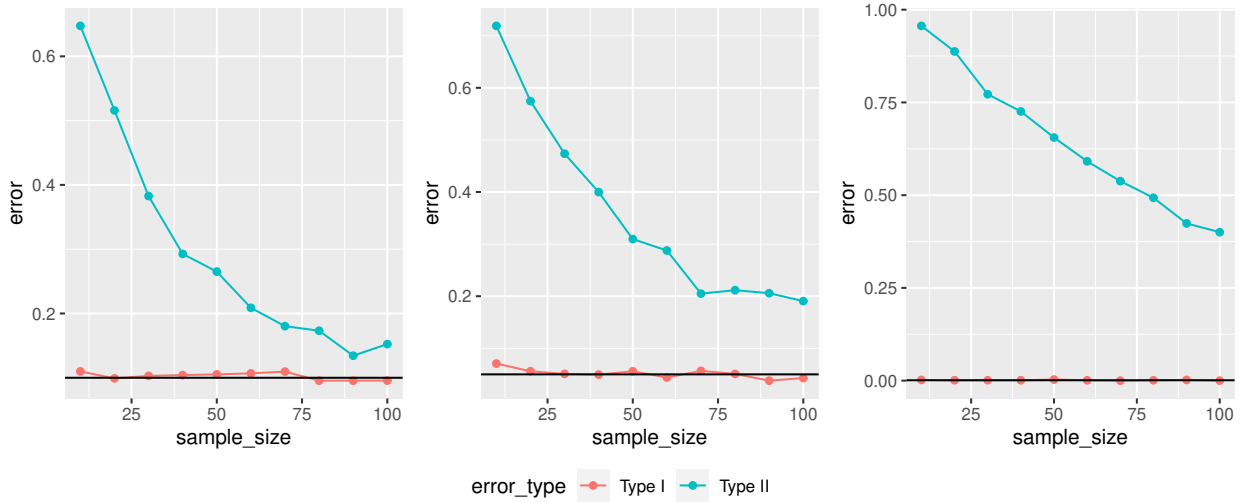


Figure 6.1: Type I and type II error with different $\alpha \in \{0.1, 0.05, 0.001\}$.

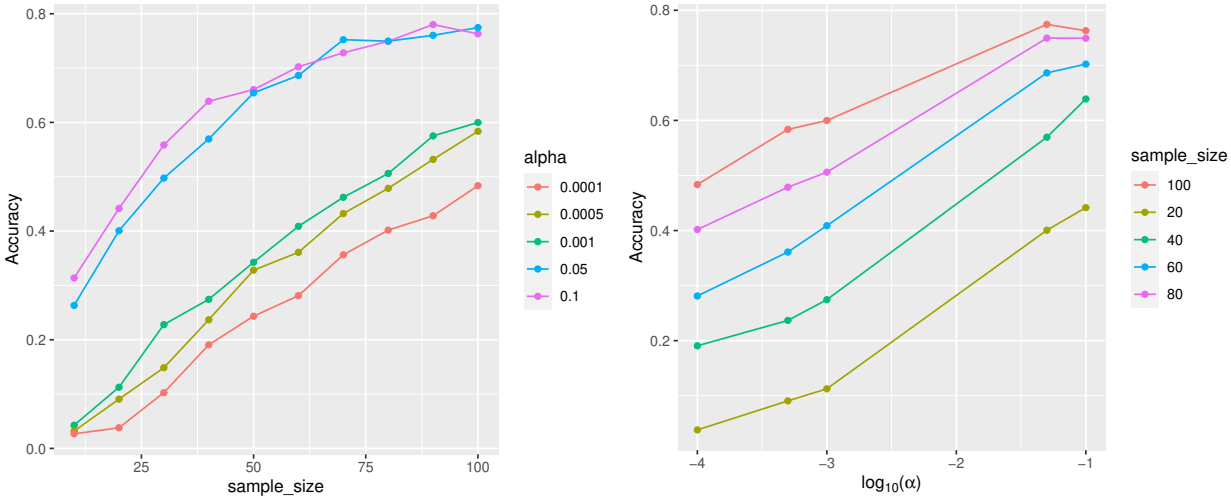


Figure 6.2: Accuracy performance when two tests are combined.

6.2 Interventional Essential Graph Recovery

We discuss how to modify the classic PC algorithm in section 3 and suggest Algorithm 1 to recover the skeleton from interventional data. Then Algorithm 3 makes it possible to find the interventional essential graph. Combine these algorithms together,

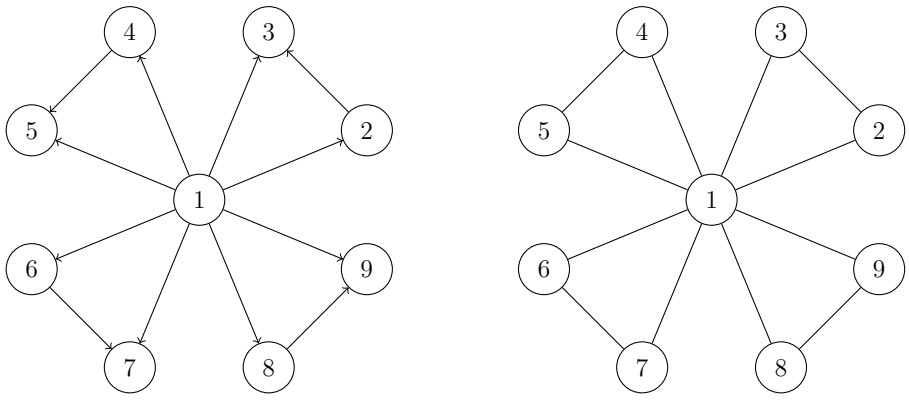
$$\text{Fullly-Connected Graph} \xrightarrow{\text{int-PC}} \text{Skeleton} \xrightarrow{\text{Meek's rule}} \text{CPDAG} \xrightarrow{\text{EO}} \mathcal{I}\text{-Essential Graph},$$

here we use int-PC to represent our intervention PC algorithm, and EO for edge orientation step.

To build the theory, multiple tests are required to ensure the test reliable. However, in practice, the limit of multiple tests is the sample size of each intervention data block would be quite limited, even though we prove the consistency of the int-PC algorithm. To increase the sample size, in this simulation, int-PC algorithm will remove any data blocks that either node i or node j is under intervention while determining the existence of edge $i - j$, and merge all other blocks.

It is tough to generate data for this simulation, as the edge orientation requires there exists intervention target feasible for the undirected edge. In our practice, the number of undirected edges remained in the CPDAG is quite limited if we create the graph structure randomly. Consider this, in this section, we design three DAG structures and corresponding intervention families to evaluate the performance of algorithms.

Table 6.1 shows the configurations of simulations in this section. This design meets two expectations to show the power of edge orientation: there are enough undirected edges in the CPDAG; and the intervention family can help us orient them as many as possible. Actually, among our configurations, all the edges in the CPDAGs are undirected. And intervention makes the $\mathcal{I}$-essential graph is the same as the true DAG.

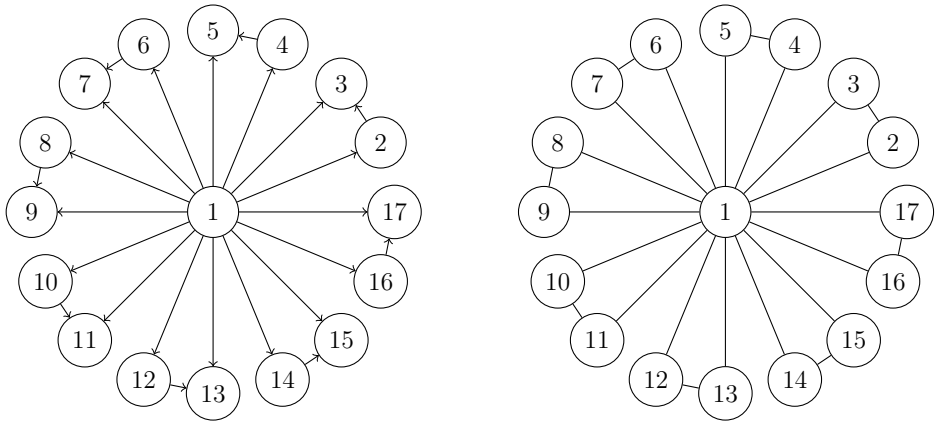


(a) True DAG. (b) Essential Graph.

Figure 6.3: The DAG used for data generation with $p = 9$.

	p	num of blocks	total sample size	graph structure	intervention family
simulation 1	9	6	6*50	Figure 6.3	{{1}, {2}, {3, 4}, {6, 8}, {5, 7, 9}, ∅}
simulation 2	17	10	10*50	Figure 6.4	{{1}, {2}, {3}, {4, 6}, {5, 7}, {8, 10}, {9, 11}, {12, 14, 16}, {13, 15, 17}, ∅}
simulation 3	49	26	26*20	Figure 6.5	{{1}, {2}, {3}, {4, 6}, {5, 7}, {8, 10}, {9, 11}, {12, 14, 16}, {13, 15, 17}, {18}, {19}, {20, 22}, {21, 23}, {24, 26}, {25, 27}, {28, 30, 32}, {29, 31, 33}, {34}, {35}, {36, 38}, {37, 39}, {40, 42}, {41, 43}, {44, 46, 48}, {45, 47, 49}, ∅}

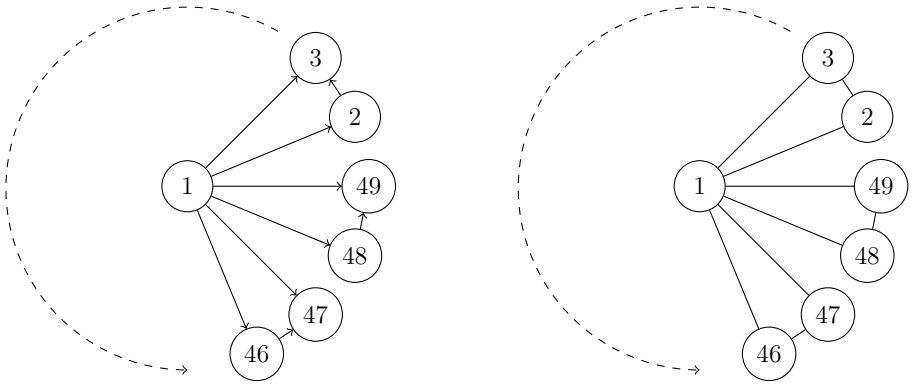
Table 6.1: Configurations of Structure Recovery Simulations.



(a) True DAG. (b) Essential Graph.

Figure 6.4: The DAG used for data generation with $p = 17$.

Here are two tuning parameters, one α_{pc} is used in int-PC algorithm and another α_{eo}



(a) True DAG. (b) Essential Graph.

Figure 6.5: The DAG used for data generation with $p = 49$.

works for edge orientation. In this simulation, we choose the int-PC parameters from $\alpha_{pc} \in \{0.1, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ and set edge orientation parameter as $\alpha_{eo} = 0.05$. Notice that the score-based method GES we used for comparison has no tuning parameter.

Performance metrics are calculated by comparing the estimated $\hat{\mathcal{G}}_{\mathcal{I}}$ with the true $\mathcal{I}$-essential graph $\mathcal{G}_{\mathcal{I}}$. P is the number of edges in the estimated graph. TP is the number of true positive edges, corresponding to consistent edges between the estimated $\mathcal{I}$-essential graph and the true $\mathcal{I}$-essential graph. FP counts the number of edges in the estimated graph but not in the true skeleton. Inversely, M counts the number of edges in the true $\mathcal{I}$-essential graph but not in the skeleton of estimated graph. R is the number of reversed edges, and it includes two kinds of edges: (1) the directions of edge are inconsistent between true and estimated graph; or (2) the edge remains undirected in the estimated (true) graph, but has its direction in the true (estimated) graph. Notice undirected edge will be counted twice into these metrics.

Then structural Hamming distance (SHD) and Jaccard index (JI) can be defined to evaluate the overall performance of graph structure learning:

$$\text{SHD} = \text{R} + \text{FP} + \text{M}, \quad \text{JI} = \text{TP}/(s + \text{P} - \text{TP}),$$

where s the number of edges in the true $\mathcal{I}$-essential graph. A lower SHD or a higher JI indicates better performance of the estimation results. TRP and FPR are defined as

$$\text{TPR} = \text{TP}/s, \quad \text{FPR} = (\text{R} + \text{FP})/\text{P}.$$

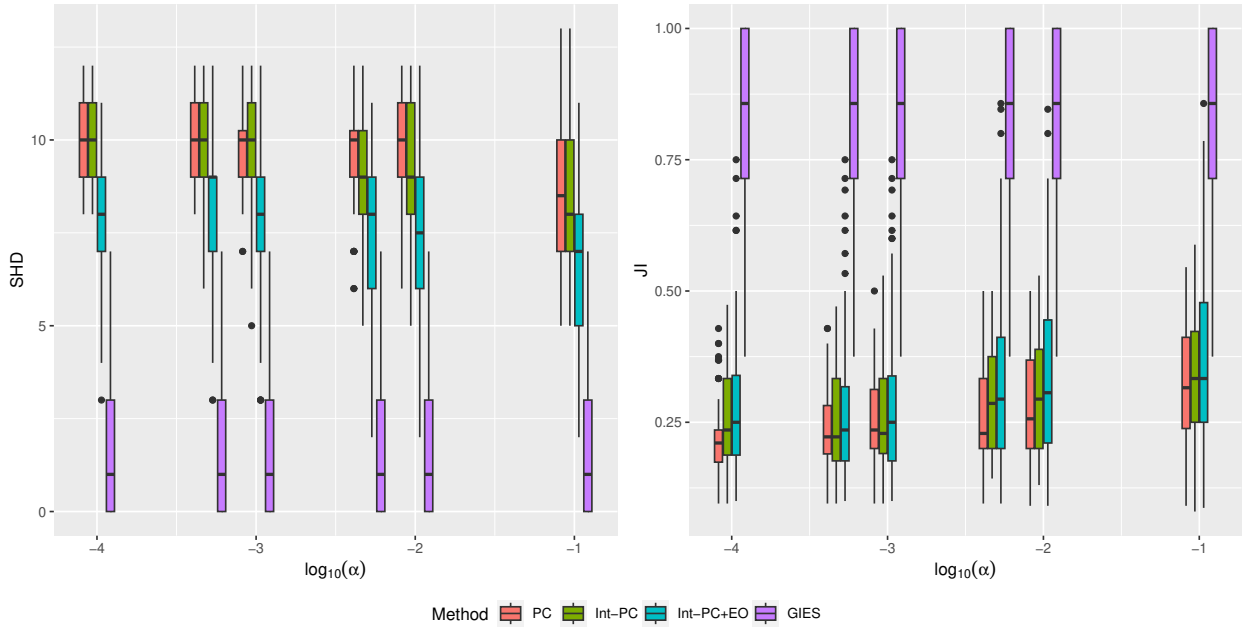


Figure 6.6: The comparison of methods with $p = 9$.

From Figure 6.6, 6.7 and 6.8, edge orientation works well over these three plots. The gap between the green box (int-PC) and blue box (int-PC+EO) shows the power of the edge orientation. However, GIES shows better performance with $p = 9$ and $p = 17$. The main reason is that GIES as a score-based method can over-perform PC algorithm in the skeleton and CPDAG recovery step when the number of variables are limited. As we mentioned in Section 1, the score-based method may meet difficulty to handle the huge search space when we increase the number of variables.

In Figure 6.8, we use the DAG with $p = 49$, and the SHD of GIES is much lower than int-PC plus EO, as Table 6.2 shows GIES has more false positives. Compared to GIES, constraint-based methods are more conservative at this time when introducing new edges

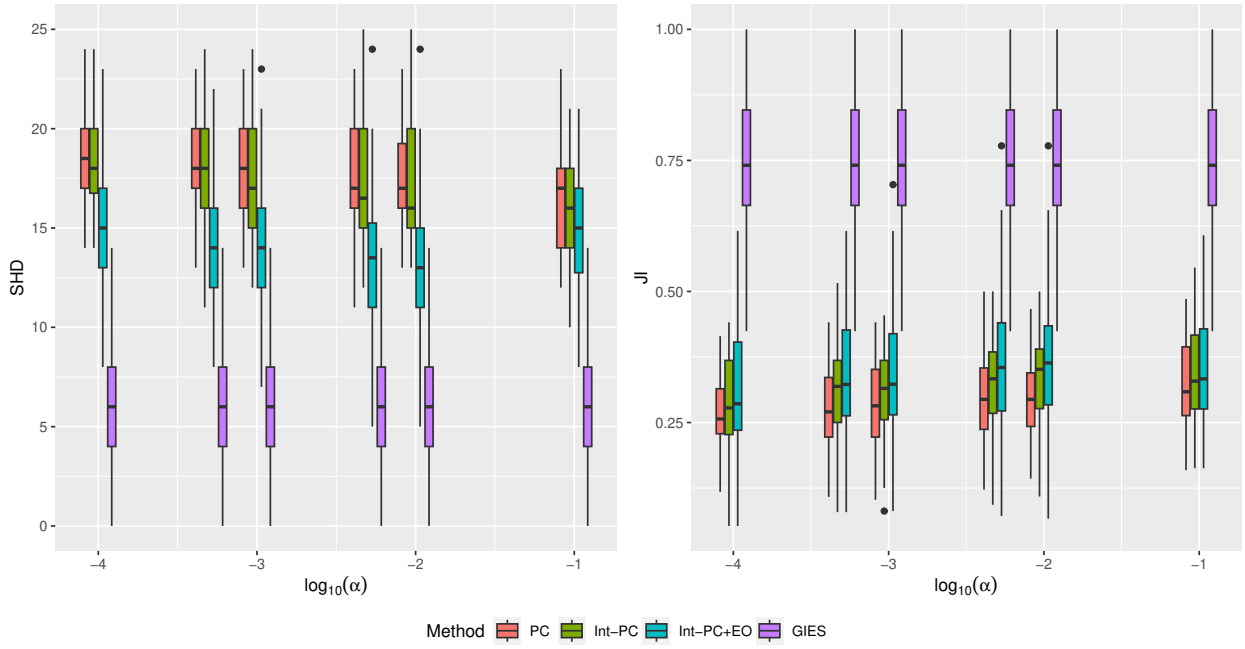


Figure 6.7: The comparison of methods with $p = 17$.

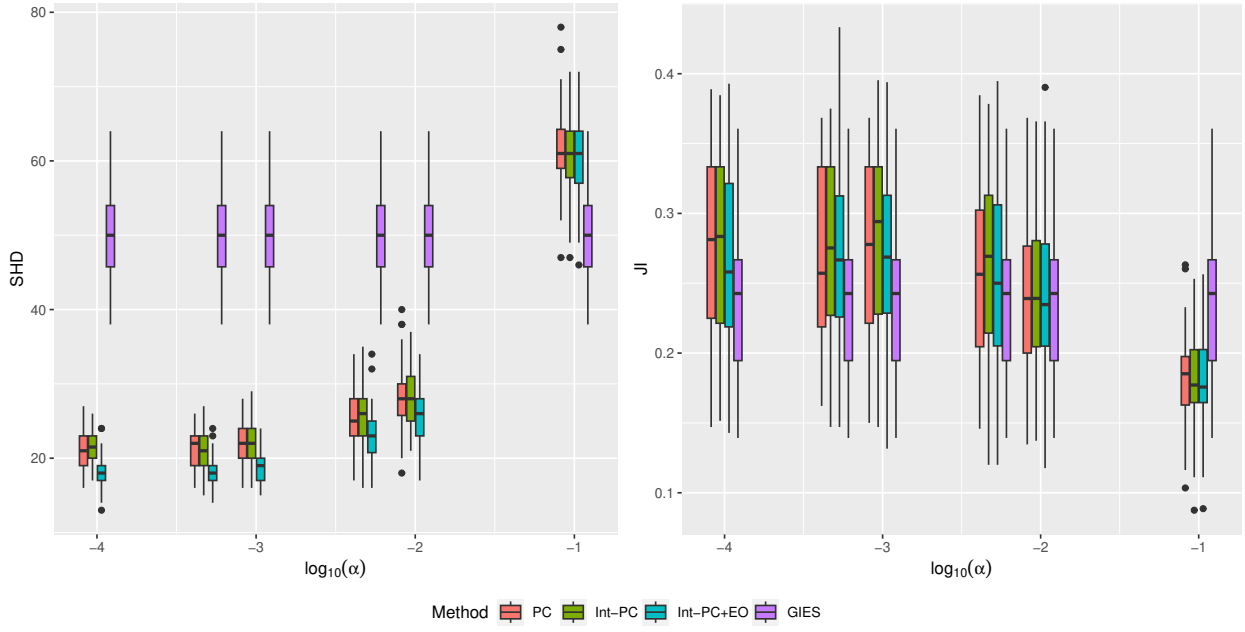


Figure 6.8: The comparison of methods with $p = 49$.

into the estimated results. Notice that the main trend and performance of boxes in these Figures are dominated by skeleton recovery, instead of the edge orientation. If we cannot

get the high quality estimation of skeleton, there is no space for the edge orientation.

	method	P	TP	R	FP	M	TPR	FPR	SHD	JI
$p = 9$	PC	12.48	5.31	7.03	0.14	2.46	0.44	0.58	9.63	0.28
	Int-PC.	12.89	5.78	7.06	0.05	2.10	0.48	0.56	9.21	0.30
	Int-PC + EO	10.65	5.58	5.02	0.05	2.10	0.47	0.48	7.17	0.34
	GIES	12.30	11.04	0.63	0.63	0.33	0.92	0.10	1.59	0.84
$p = 17$	PC	23.94	11.02	12.59	0.33	4.53	0.46	0.54	17.45	0.30
	Int-PC.	25.23	12.32	12.72	0.19	4.13	0.51	0.51	17.04	0.36
	Int-PC + EO	21.10	11.98	8.94	0.18	4.13	0.50	0.43	13.25	0.37
	GIES	25.63	20.95	1.76	2.92	1.29	0.87	0.18	5.97	0.74
$p = 49$	PC	21.00	9.68	10.39	0.93	9.73	0.40	0.54	21.05	0.27
	Int-PC.	21.37	10.00	10.55	0.82	9.75	0.42	0.53	21.12	0.28
	Int-PC + EO	17.15	8.72	7.68	0.75	9.75	0.36	0.49	18.18	0.27
	GIES	63.72	16.78	4.48	42.47	2.74	0.70	0.73	49.69	0.24

Table 6.2: Numerical results of structure learning with PC, Int-PC, Int-PC+EO and GIES.

6.3 Implement Edge Orientation on GES

Figure 6.6 and 6.7 shows that the score-based learning has some advantages when the number of variables is relatively small. Also observational data is generally cheaper than the experimental data in the real world practice. Based on this, since our edge orientation method can be used independently of the PC algorithm. In this section, we implement the GES algorithm on observational data first, and then implement our EO on the estimated CPDAG.

The graph structure and parameter settings are the same as the simulation 2 in Table 6.1. The only difference is: we increase the sample size of observational block for GES to learn the CPDAG. The size of each interventional block is 100 and observational block is 1000.

From both Figure 6.9 and Table 6.3, the edge orientation finishes its job quite well. Edge orientation decreases the SHD and increases the JI, both of which indicate the better

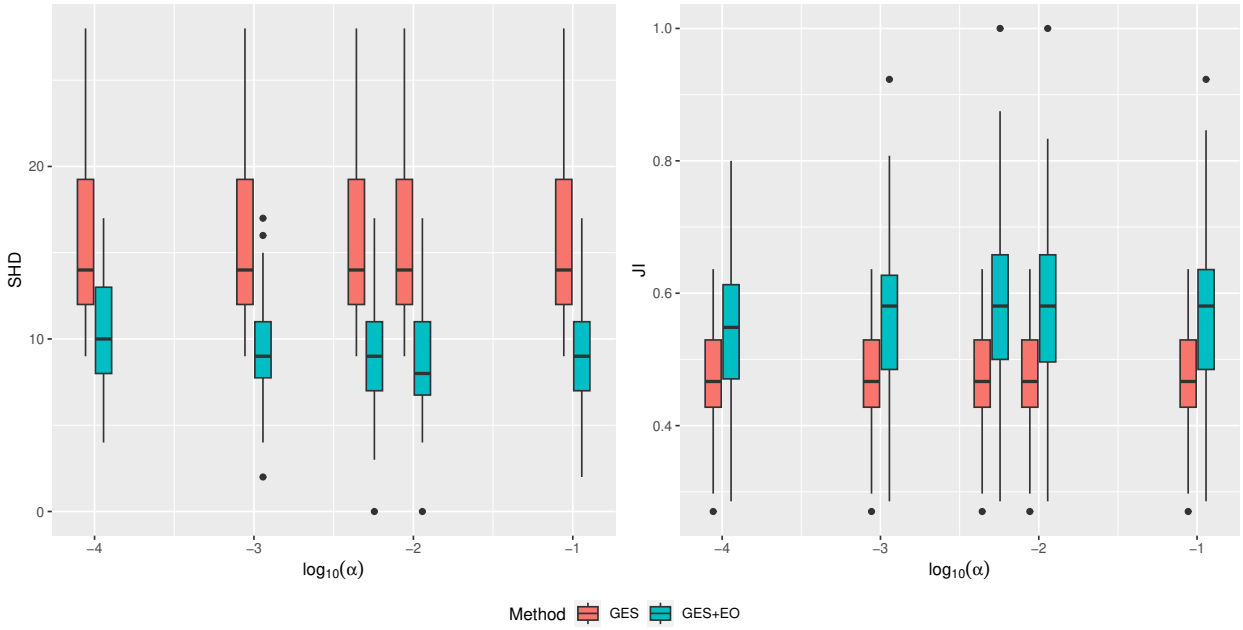


Figure 6.9: Implement edge orientation on GES output.

	method	P	TP	R	FP	M	TPR	FPR	SHD	JI
$p = 17$	GES	31.52	17.63	11.92	1.97	1.93	0.73	0.43	15.82	0.47
	GES + EO	24.41	17.54	5.12	1.75	1.93	0.73	0.28	8.80	0.58

Table 6.3: Numerical results of structure learning with GES and GES+EO.

performance of GES+EO.

CHAPTER 7

Proofs of Consistency

7.1 Some Ancillary Results

The neighborhood linear regression is widely used in this dissertation, the first part of this section will show some results about this. Consider regression,

$$X_j \sim X_i + X_{\mathcal{S}} \Longrightarrow X_j = \beta_{ij|\mathcal{S}} X_i + \beta_{\mathcal{S}i|\mathcal{S}} X_{\mathcal{S}} + \varepsilon_{ij|\mathcal{S}} \text{ with } \varepsilon_{ij|\mathcal{S}} \sim \mathcal{N}(0, (\sigma_{ij|\mathcal{S}})^2), \tag{7.1}$$

notice that $\{i\} \cup \{j\} \cup \mathcal{S} \subset [p]$. In this dissertation, multiple regressions are used to treat different interventional blocks. Here we use superscript k to distinguish different regressions:

$$X_j^k \sim X_i^k + X_{\mathcal{S}}^k \Longrightarrow X_j^k = \beta_{ij|\mathcal{S}}^k X_i^k + \beta_{\mathcal{S}i|\mathcal{S}}^k X_{\mathcal{S}}^k + \varepsilon_{ij|\mathcal{S}}^k \text{ with } \varepsilon_{ij|\mathcal{S}}^k \sim \mathcal{N}(0, (\sigma_{ij|\mathcal{S}}^k)^2). \tag{7.2}$$

For any $p \times p$ matrix $\mathbf{A}$ and a set $\mathcal{K} \subset [p]$, we define the new $|\mathcal{K}| \times |\mathcal{K}|$ matrix $\mathbf{A}_{\mathcal{K}}$ by extracting rows and columns from matrix $\mathbf{A}$ corresponding to $\mathcal{K}$. Someone prefers to call $\mathbf{A}_{\mathcal{K}}$ as the principal submatrix of $\mathbf{A}$. Suppose each row of the whole data matrix is drawn from multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{\Sigma})$, then the design matrix of regression in (7.1) with $1 + |\mathcal{S}|$ columns has the following properties.

Lemma 14. *Each row of the design matrix of regression in* (7.1) *also follows multivariate Gaussian distribution, and its covariance matrix is* $\mathbf{\Sigma}_{\{i\}\cup\mathcal{S}}$.

Proof of Lemma 14. Properties of multivariate Gaussian distribution from probability the-

ory. □

Lemma 15. *The covariance matrix* $\mathbf{\Sigma}_{\{i\}\cup\mathcal{S}}$ *satisfies*

$$\lambda_{min}\left(\mathbf{\Sigma}_{\{i\}\cup\mathcal{S}}\right) \geq \lambda_{min}\left(\mathbf{\Sigma}\right),$$

furthermore, back to the context of this dissertation, over all interventional targets in $\mathcal{I}$*,*

$$\lambda_{min}\left(\mathbf{\Sigma}^k_{\{i\}\cup\mathcal{S}}\right) \geq \lambda_{min}\left(\mathbf{\Sigma}^k\right) \geq \sigma_*^2,$$

with the help of assumption 4.

Proof of Lemma 15. A well-known result in linear algebra shows that,

$$\lambda_{min}(\mathbf{A}) \leq \lambda_l(\mathbf{A}_{\mathcal{K}}) \leq \lambda_{max}(\mathbf{A}) \quad l = 1, \ldots, |\mathcal{K}|, \tag{7.3}$$

so all eigenvalues of the the principal submatrix are located within the $[\lambda_{min}, \lambda_{max}]$. And it is suffice to show Lemma 14 from the properties of multivariate Gaussian distribution. □

Lemma 16. *Consider neighborhood linear regression defined in (7.2), there is a uniform lower bound of noise variance* $\sigma^k_{ij|\mathcal{S}}$ *for any* $(i, j, \mathcal{S})$ *over the whole intervention family* $\mathcal{I}$*,*

$$\min_{k=1,\ldots,B}\left(\min_{(i,j,\mathcal{S})}\sigma_{ij|\mathcal{S}}\right) \geq \sigma_*,$$

here σ_* *is defined in assumption 4.*

Proof of Lemma 16. When regressing X_j onto the variables $X_i + X_{\mathcal{S}}$, the Least-Square population regression coefficient vector is,

$$\tilde{\beta} = \underset{\beta\in\mathbb{R}^{1+|\mathcal{S}|}}{\operatorname{argmin}}\,\mathbb{E}(X_j - \mathbf{X}_{\{i\}\cup\mathcal{S}}\beta)^2 = \underset{\beta\in\mathbb{R}^{2+|\mathcal{S}|}:\beta_{jj}=0}{\operatorname{argmin}}\,\mathbb{E}(X_j - \mathbf{X}_{\{i\}\cup\{j\}\cup\mathcal{S}}\beta)^2,$$

using β_{kj} to represent the partial regression coefficient corresponding to X_k. Then the corresponding population residual standard variance can be defined as,

$$\sigma_{ij|\mathcal{S}} = [\mathbb{E}(X_j - \mathbf{X}_{\{i\}\cup\{j\}\cup\mathcal{S}}\tilde{\beta})^2]^{1/2}.$$

Since for $\tilde{\beta} \in \mathbb{R}^{2+|\mathcal{S}|}$ with $\tilde{\beta}_{jj} = 0$, define $\hat{\beta}$ such that $\hat{\beta}_{jj} = 1$ and $\hat{\beta}_{kj} = -\tilde{\beta}_{kj}$ if $k \neq j$,

$$\begin{aligned}&\mathbb{E}(X_j - \mathbf{X}_{\{i\}\cup\{j\}\cup\mathcal{S}}\tilde{\beta})^2 = \mathbb{E}(\mathbf{X}_{\{i\}\cup\{j\}\cup\mathcal{S}}\hat{\beta})^2 \\ =&\hat{\beta}^T\mathbb{E}\left(\mathbf{X}^T_{\{i\}\cup\{j\}\cup\mathcal{S}}\mathbf{X}_{\{i\}\cup\{j\}\cup\mathcal{S}}\right)\hat{\beta} \geq \lambda^2_{min}\left(\mathbf{\Sigma}_{\{i\}\cup\{j\}\cup\mathcal{S}}\right)\|\hat{\beta}\|^2 \geq \sigma^2_*, \end{aligned} \tag{7.4}$$

and extend (7.4) to the general case,

$$\min_{k=1,\ldots,B}\left(\min_{(i,j,\mathcal{S})}\sigma_{ij|\mathcal{S}}\right) \geq \sigma_*.$$

. □

Lemma 17. *Consider neighborhood linear regression defined in (7.1), there is a uniform upper bound of noise variance $\sigma^k_{ij|\mathcal{S}}$ for any $(i, j, \mathcal{S})$ over the whole intervention family $\mathcal{I}$,*

$$\max_{k=1,\ldots,B}\left(\max_{(i,j,\mathcal{S})}\sigma_{ij|\mathcal{S}}\right) \leq \bar{\sigma},$$

here $\bar{\sigma}$ is defined in assumption 3.

Proof of Lemma 17. Define $S(i, \mathcal{S}) = \{i\} \cup \mathcal{S}$, obviously,

$$S(i, \mathcal{S}_1) \subset S(i, \mathcal{S}_2) \Longrightarrow \sigma_{ij|\mathcal{S}_1} \geq \sigma_{ij|\mathcal{S}_2},$$

since any β satisfied for $S(i, \mathcal{S}_1)$ is also feasible for $S(i, \mathcal{S}_2)$. Then,

$$\varnothing \subset S(i, \mathcal{S}) \subset \mathrm{ne}(j) \Longrightarrow \sigma_{ij|\mathcal{S}} \leq (\mathbb{E}X^2_j)^{1/2} \leq \bar{\sigma},$$

here $X_j \sim \mathcal{N}(0, \sigma_j^2)$ or $\mathcal{N}(0, \tau_j^2)$, the marginal distribution of single node X_j. □

Lemma 18. *Suppose t_r follows a t distribution with r degree of freedom. Then for any $\delta > 0$,*

$$P(|t_r| \geq \delta) \leq 2e^{-\delta^2/4} + e^{-r/16}.$$

Proof of Lemma 18. Use the property of t distribution, for any constant $1 > c > 0$,

$$\begin{aligned}
&P(|t_r| \geq \delta) = P(\frac{|Z|}{\sqrt{\chi_r^2/r}} \geq \delta) \\
=&P(\frac{|Z|}{\sqrt{\chi_r^2/r}} \geq \delta, \chi_r^2/r \geq c) + P(\frac{|Z|}{\sqrt{\chi_r^2/r}} \geq \delta, \chi_r^2/r \leq c) \\
=&P(\frac{|Z|}{\sqrt{\chi_r^2/r}} \geq \delta, \chi_r^2/r \geq c) + P(\frac{|Z|}{\sqrt{\chi_r^2/r}} \geq \delta \mid \chi_r^2/r \leq c) P(\chi_r^2/r \leq c) \\
\leq&P(|Z| \geq c^{1/2}\delta) + P(\chi_r^2 \leq cr),
\end{aligned}$$

for the first term, use the classic normal tail bound,

$$P(|Z| \geq c^{1/2}\delta) \leq 2e^{-c\delta^2/2},$$

then from Lemma 1 in [LM00]:

$$P(\chi_r^2 - r \leq -2\sqrt{r}\sqrt{x}) \leq e^{-x} \Longrightarrow P(\frac{\chi_r^2}{r} \leq 1 - 2\sqrt{\frac{x}{r}}) \leq e^{-x}. \tag{7.5}$$

Modified for our problem,

$$P(\chi_r^2 \leq cr) \leq e^{-r(1-c)^2/4},$$

set $c = 1/2$ to get the final result. □

Here some basic knowledge about random matrix would be helpful.

Definition 11. *($\mathbf{\Sigma}$-Gaussian ensemble) A random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is drawn from the $\mathbf{\Sigma}$-Gaussian ensemble if its each row x_i^T is drawn i.i.d from a multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{\Sigma})$.*

Someone may prefer to call the associated sample covariance $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ as Wishart matrix. The eigenvalues of $\hat{\mathbf{\Sigma}}$,

$$\gamma_j(\hat{\mathbf{\Sigma}}) = \left(\sigma_j(\mathbf{X})/\sqrt{n}\right)^2 \quad \text{for } j = 1, \ldots, d,$$

here $\sigma_j(\mathbf{X})$ represents the j-th singular value of $\mathbf{X}$. And

$$\gamma_{\max}(\hat{\mathbf{\Sigma}}) \geq \hat{\mathbf{\Sigma}}_{jj} \quad \text{for } j = 1, \ldots, d,$$

i.e. the largest eigenvalue could be used as an upper bound for diagonal entries.

Lemma 19. *Suppose random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is drawn from the $\mathbf{\Sigma}$-Gaussian ensemble with $d \leq n^{1-b}$ here b is defined in assumption 2, then for all $1 > \delta > 0$,*

$$P\left(\sqrt{\hat{\mathbf{\Sigma}}^{-1}_{jj}} \geq \frac{2}{\sigma_*\delta}\right) \leq e^{-n(1-\delta)^2/2} \quad \textit{for } j = 1, \ldots, d,$$

with sufficient large n, here $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ as Wishart matrix and $\sigma_* = \gamma_{min}(\sqrt{\mathbf{\Sigma}})$.

Proof of Lemma 19. First from Theorem 6.1 in [Wai19], we know: for $n \geq d$,

$$P\left(\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{n}} \leq \gamma_{\min}(\sqrt{\mathbf{\Sigma}})(1-\delta) - \sqrt{\frac{\text{trace}(\mathbf{\Sigma})}{n}}\right) \leq e^{-n\delta^2/2},$$

here $\sigma_{\min}(\mathbf{X})$ is the minimum singular value. Then,

$$P\left(\frac{1}{\sigma_{\min}(\mathbf{X})/\sqrt{n}} \geq \frac{1}{\gamma_{\min}(\sqrt{\mathbf{\Sigma}})(1-\delta) - \sqrt{\text{trace}(\mathbf{\Sigma})/n}}\right) \leq e^{-n\delta^2/2},$$

with diagonal entry upper bound τ^2, $\sqrt{\text{trace}(\mathbf{\Sigma})/n} \leq \sqrt{d\tau^2/n} \leq n^{-b/2}\tau$,

$$P\left(\frac{1}{\sigma_{\min}(\mathbf{X})/\sqrt{n}} \geq \frac{1}{\sigma_*(1-\delta) - n^{-b/2}\tau}\right) \leq e^{-n\delta^2/2}.$$

Replace δ as $1-\delta$ for convenience,

$$P\left(\frac{1}{\sigma_{\min}(\mathbf{X})/\sqrt{n}} \geq \frac{1}{\sigma_*\delta - n^{-b/2}\tau}\right) \leq e^{-n(1-\delta)^2/2},$$

for n large enough such that $n^{-b/2}\tau \leq \sigma_*\delta/2$,

$$P\left(\frac{1}{\sigma_{\min}(\mathbf{X})/\sqrt{n}} \geq \frac{2}{\sigma_*\delta}\right) \leq e^{-n(1-\delta)^2/2} \Longrightarrow P\left(\frac{1}{\sqrt{\gamma_{\min}(\hat{\mathbf{\Sigma}})}} \geq \frac{2}{\sigma_*\delta}\right) \leq e^{-n(1-\delta)^2/2}, \quad (7.6)$$

finally as the Wishart matrix $\hat{\mathbf{\Sigma}} = \mathbf{X}^t\mathbf{X}/n$,

$$P\left(\sqrt{\gamma_{\max}(\hat{\mathbf{\Sigma}}^{-1})} \geq \frac{2}{\sigma_*\delta}\right) \leq e^{-n(1-\delta)^2/2} \Longrightarrow P\left(\sqrt{\hat{\mathbf{\Sigma}}^{-1}_{jj}} \geq \frac{2}{\sigma_*\delta}\right) \leq e^{-n(1-\delta)^2/2},$$

for any $j = 1, \ldots, d$. □

To deal with multiple blocks in Algorithm 2, here we give another version of Lemma 19, which may be loose but still helpful to bound the mixed Wishart matrix.

Lemma 20. *Suppose random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ has C submatrices and each submatrix $\mathbf{X}_i$ is drawn from the $\mathbf{\Sigma}^i$-Gaussian ensemble with $d \leq n_i^{1-b}$ here b is defined in assumption 2,*

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_C \end{pmatrix} \sim \begin{pmatrix} \mathcal{N}(0, \mathbf{\Sigma}^1) \\ \mathcal{N}(0, \mathbf{\Sigma}^2) \\ \vdots \\ \mathcal{N}(0, \mathbf{\Sigma}^C) \end{pmatrix} \quad \textit{with } \sum_{i=1}^{C} n_i = n,$$

then for all $1 > \delta > 0$,

$$P\left(\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}} \geq \frac{2}{\sigma_*\delta\sqrt{n}}\right) \leq Ce^{-n_*(1-\delta)^2/2} \quad \textit{for } j = 1, \ldots, d,$$

here $\sigma_* = \min_{i=1,\ldots,C}\{\gamma_{min}(\sqrt{\mathbf{\Sigma}^i})\}$ *and* $n_* = \min_{i=1,\ldots,C}\{n_i\}$.

Proof of Lemma 20. From (7.6), we have for any submatrix $\mathbf{X}_i$,

$$P\left(\frac{1}{\sigma_{\min}(\mathbf{X}_i)/\sqrt{n_i}} \geq \frac{2}{\sigma_*\delta}\right) \leq e^{-n_i(1-\delta)^2/2} \Longrightarrow P\left(\gamma_{\min}(\mathbf{X}_i^T\mathbf{X}_i) \leq \frac{1}{4}\sigma_*^2\delta^2 n_i\right) \leq e^{-n_i(1-\delta)^2/2},$$

by Weyl's inequality,

$$\gamma_{\min}\left(\sum_{i=1}^{C}\mathbf{X}_i^T\mathbf{X}_i\right) \geq \sum_{i=1}^{C}\gamma_{\min}\left(\mathbf{X}_i^T\mathbf{X}_i\right),$$

then,

$$\begin{aligned}
&P\left(\gamma_{\min}\left(\sum_{i=1}^{C}\mathbf{X}_i^T\mathbf{X}_i\right) \leq \frac{1}{4}\sigma_*^2\delta^2 n\right) \leq P\left(\sum_{i=1}^{C}\gamma_{\min}\left(\mathbf{X}_i^T\mathbf{X}_i\right) \leq \frac{1}{4}\sigma_*^2\delta^2 n\right) \\
&\leq P\left(\bigcup_{i=1}^{C}\left\{\gamma_{\min}\left(\mathbf{X}_i^T\mathbf{X}_i\right) \leq \frac{1}{4}\sigma_*^2\delta^2 n_i\right\}\right) \leq \sum_{i=1}^{C}P\left(\gamma_{\min}\left(\mathbf{X}_i^T\mathbf{X}_i\right) \leq \frac{1}{4}\sigma_*^2\delta^2 n_i\right) \leq Ce^{-n_*(1-\delta)^2/2},
\end{aligned}$$

here $n_* = n_1 \wedge n_2 \wedge \cdots \wedge n_C$. With the property of Wishart matrix,

$$P\left(\gamma_{\min}\left(\mathbf{X}^T\mathbf{X}\right) \leq \frac{1}{4}\sigma_*^2\delta^2 n\right) \leq Ce^{-n_*(1-\delta)^2/2} \Longrightarrow P\left(\sqrt{\gamma_{\max}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)} \geq \frac{2}{\sigma_*\delta\sqrt{n}}\right) \leq Ce^{-n_*(1-\delta)^2/2},$$

finally we find the bound for diagonal entries,

$$P\left(\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}} \geq \frac{2}{\sigma_*\delta\sqrt{n}}\right) \leq Ce^{-n_*(1-\delta)^2/2} \quad \text{for } j = 1, \ldots, d.$$

□

7.2 High-Dimensional Consistency of Algorithm 1

In Section 5.2, superscript o is introduced in (5.10) to label the interventional block which meets Assumption 10. For notation simplicity, we discard superscript o in this section's discussion. But let us keep in mind that we are always working on the block satisfying Assumption 10.

Define event $\mathcal{E}_{ij|\mathcal{S}}$,

$$\mathcal{E}_{ij|\mathcal{S}} = \left\{ \frac{|\beta_{ij|\mathcal{S}}|}{s_{ij|\mathcal{S}} \sqrt{\left(\left(\mathbf{X}_{ij|\mathcal{S}} \right)^T \mathbf{X}_{ij|\mathcal{S}} \right)^{-1}_{jj}}} \geq \frac{\psi_n \sigma_*}{4\sqrt{2}\bar{\sigma}} n^{1/2-q/2} \right\},$$

here ψ_n satisfies $\min_{i,j,\mathcal{S}:\beta_{ij|\mathcal{S}} \neq 0} |\beta_{ij|\mathcal{S}}| \geq \psi_n = O(n^{-d})$ in (5.6), thus we can rewrite,

$$\mathcal{E}_{ij|\mathcal{S}} = \left\{ \frac{|\beta_{ij|\mathcal{S}}|}{s_{ij|\mathcal{S}} \sqrt{\left(\left(\mathbf{X}_{ij|\mathcal{S}} \right)^T \mathbf{X}_{ij|\mathcal{S}} \right)^{-1}_{jj}}} \geq \kappa_n \right\}, \tag{7.7}$$

with

$$\kappa_n = \frac{\sigma_*}{4\sqrt{2}\bar{\sigma}} n^{1/2-q/2-d} = O(n^{1/2-q/2-d}).$$

To prove the results in this section, we will use Lemma 19 and 20. Recall Assumption 4, as the covariance matrix of interventional graph $\mathbf{\Sigma}^i$ is PSD for $i = 1, \ldots, B$,

$$\min_{i=1,\ldots,B} \left\{ \lambda_{min} \left(\sqrt{\mathbf{\Sigma}^i} \right) \right\} \geq \sigma_*, \tag{7.8}$$

we can use σ_* in Lemma 19 and 20. It is somehow abuse of notation when σ_* occurs in this dissertation, but which is always referred to the uniform lower bound of eigenvalues.

Lemma 21. *For any i, j and $\mathcal{S}$,*

$$\frac{|\beta_{ij|\mathcal{S}}|}{s_{ij|\mathcal{S}}\sqrt{\left(\left(\mathbf{X}_{ij|\mathcal{S}}\right)^T\mathbf{X}_{ij|\mathcal{S}}\right)^{-1}_{jj}}} \geq \frac{\sigma_*}{4\sqrt{2}\bar{\sigma}}n^{1/2-q/2-d},$$

with the probability at least $1-\exp(-(n^{1-q}-s-1)/8)-\exp(-n^{1-q}/8)$.

Proof of Lemma 21. For any i, j and $\mathcal{S}$,

$$P\left(\sqrt{n_{ij}}s_{ij|\mathcal{S}}\sqrt{\left(\left(\mathbf{X}_{ij|\mathcal{S}}\right)^T\mathbf{X}_{ij|\mathcal{S}}\right)^{-1}_{jj}} \geq NM\right) \leq P\left(s_{ij|\mathcal{S}} \geq N\right) + P\left(\sqrt{n_{ij}\left(\left(\mathbf{X}_{ij|\mathcal{S}}\right)^T\mathbf{X}_{ij|\mathcal{S}}\right)^{-1}_{jj}} \geq M\right).$$

For the first term,

$$\frac{(n_{ij}-p_{ij|\mathcal{S}})(s_{ij|\mathcal{S}})^2}{\sigma^2_{ij|\mathcal{S}}} \sim \chi^2_{n_{ij}-p_{ij|\mathcal{S}}},$$

then with assumption 3,

$$P\left(\left(s_{ij|\mathcal{S}}\right)^2 \geq N^2\right) = P\left(\chi^2_{n_{ij}-p_{ij|\mathcal{S}}}/(n_{ij}-p_{ij|\mathcal{S}}) \geq N^2/\sigma^2_{ij|\mathcal{S}}\right) \leq P\left(\chi^2_{n_{ij}-p_{ij|\mathcal{S}}}/(n_{ij}-p_{ij|\mathcal{S}}) \geq N^2/\bar{\sigma}^2\right).$$

Now recall the χ^2_n concentration,

$$P(\chi^2_n/n \geq 1+t) \leq e^{-nt/8} \text{ for } t \geq 1, \tag{7.9}$$

thus set $N=\sqrt{2}\bar{\sigma}$ in (7.9),

$$P\left(\left(s_{ij|\mathcal{S}}\right)^2 \geq 2\bar{\sigma}^2\right) \leq P\left(\chi^2_{n_{ij}-p_{ij|\mathcal{S}}}/(n_{ij}-p_{ij|\mathcal{S}}) \geq 2\right) \leq e^{-(n_{ij}-p_{ij|\mathcal{S}})/8} \leq e^{-(n^{1-q}-s-1)/8},$$

notice here $p_{ij|\mathcal{S}} = |\mathcal{S}|+1 \leq s+1$ from Lemma 1.

For the second part, recall the definition of Wishart matrix,

$$\sqrt{\left(\hat{\mathbf{\Sigma}}_{ij|\mathcal{S}}\right)^{-1}_{jj}} = \sqrt{n_{ij}\left(\left(\mathbf{X}_{ij|\mathcal{S}}\right)^T \mathbf{X}_{ij|\mathcal{S}}\right)^{-1}_{jj}},$$

set $M = 4/\sigma_*$,

$$P\left(\sqrt{n_{ij}\left(\left(\mathbf{X}_{ij|\mathcal{S}}\right)^T \mathbf{X}_{ij|\mathcal{S}}\right)^{-1}_{jj}} \geq 4/\sigma_*\right) = P\left(\sqrt{\left(\hat{\mathbf{\Sigma}}_{ij|\mathcal{S}}\right)^{-1}_{jj}} \geq 4/\sigma_*\right) \leq e^{-n_{ij}/8} \leq e^{-n^{1-q}/8}, \tag{7.10}$$

the inequality in (7.10) could be guaranteed by Lemma 19. Finally combine two parts,

$$P\left(\sqrt{n_{ij}} s_{ij|\mathcal{S}} \sqrt{\left(\left(\mathbf{X}_{ij|\mathcal{S}}\right)^T \mathbf{X}_{ij|\mathcal{S}}\right)^{-1}_{jj}} \geq 4\sqrt{2}\bar{\sigma}/\sigma_*\right) \leq e^{-(n^{1-q}-s-1)/8} + e^{-n^{1-q}/8},$$

which suffices to finish this proof. □

Proof of Theorem 11. For any i, j and $\mathcal{S}$,

$$P(E_{ij|\mathcal{S}}) = P(E^I_{ij|\mathcal{S}} \cup E^{II}_{ij|\mathcal{S}}) = P(E^I_{ij|\mathcal{S}}) + P(E^{II}_{ij|\mathcal{S}}),$$

from the discussion in main context, for the first term, Type I error,

$$P(E^I_{ij|\mathcal{S}}) \leq \sum_{k=1}^{B_{(i,j)}} P\left(|T^k_{ij|\mathcal{S}}| \geq \alpha_n \mid \beta^k_{ij|\mathcal{S}} = 0\right) \leq B \cdot P\left(|t_{n^{1-q}-s}| \geq \alpha_n\right), \tag{7.11}$$

the last inequality in (7.11) is guaranteed by Assumption 7 and the property of t-statistics.

For the Type II error, for notation simplicity,

$$\mathcal{V} = \frac{|\beta_{ij|\mathcal{S}}|}{s_{ij|\mathcal{S}}\sqrt{\left(\left(\mathbf{X}_{ij|\mathcal{S}}\right)^T \mathbf{X}_{ij|\mathcal{S}}\right)^{-1}_{jj}}}, \tag{7.12}$$

introduce $\mathcal{V}$ to rewrite the t-test error and event defined in (7.7),

$$\mathcal{E}_{ij|\mathcal{S}} = \{\mathcal{V} \geq \kappa_n\},$$

and,

$$P(E^{II}_{ij|\mathcal{S}}) \leq P\left(\left|T_{ij|\mathcal{S}}\right| \leq \alpha_n \mid \beta_{ij|\mathcal{S}} \neq 0\right) \leq P\left(\left|T_{ij|\mathcal{S}} - \mathcal{V}\right| \geq \mathcal{V} - \alpha_n\right) = P\left(\left|t_{n_{ij}-p_{ij|\mathcal{S}}}\right| \geq \mathcal{V} - \alpha_n\right).$$

Consider event $\mathcal{E}_{ij|\mathcal{S}}$ defined in (7.7),

$$P(E_{ij|\mathcal{S}}) = P(E^{I}_{ij|\mathcal{S}}) + P(E^{II}_{ij|\mathcal{S}}) \leq P(E^{I}_{ij|\mathcal{S}}) + P(E^{II}_{ij|\mathcal{S}} \mid \mathcal{E}_{ij|\mathcal{S}}) + P(\mathcal{E}^{c}_{ij|\mathcal{S}}),$$

if we set $\alpha_n = \kappa_n/2 = O(n^{1/2-q/2-d})$ and based on the definition of event $\mathcal{E}_{ij|\mathcal{S}}$,

$$P(E^{I}_{ij|\mathcal{S}}) \leq B \cdot P(|t_{n^{1-q}-s}| \geq \kappa_n/2), \quad P(E^{II}_{ij|\mathcal{S}} \mid \mathcal{E}_{ij|\mathcal{S}}) \leq P(|t_{n^{1-q}-s}| \geq \kappa_n/2), \tag{7.13}$$

then finally apply the tail bound of t distribution in Lemma 18 to (7.13),

$$P(E^{I}_{ij|\mathcal{S}}) + P(E^{II}_{ij|\mathcal{S}} \mid \mathcal{E}_{ij|\mathcal{S}}) \leq (B+1)\left(\exp(-\kappa_n^2/16) + \exp(-(n^{1-q}-s)/16)\right). \tag{7.14}$$

Combine results in (7.14) and Lemma 21,

$$\begin{aligned} P(E_{ij|\mathcal{S}}) \leq & n^q \exp(-n^{1-q-2d}/16) + n^q \exp(-(n^{1-q}-s)/16) \\ & + \exp(-(n^{1-q}-s-1)/16) + \exp(-n^{1-q}/8), \end{aligned} \tag{7.15}$$

the RHS of (7.15) is dominated by $O(-n^{1-q-2d})$. The number of tests required for skeleton

recovery in PC algorithm can be bounded with Lemma 1, therefore,

$$\begin{aligned} P(\text{Error}) &\leq p^{s+2} \sup_{i,j,\mathcal{S}} P(E_{ij|\mathcal{S}}) \\ &\precsim \exp(q\log(n) + (s+2)\log(p) - n^{1-q-2d}) \\ &= \exp(q\log(n) + a(n^{1-b}+2)\log(n) - n^{1-q-2d}) \to 0, \end{aligned}$$

as $q < b$ and $d < (b-q)/2$ defined in Assumption 6 and 7. □

7.3 Derive the Test Statistics

In edge orientation step, for each reversible edge, there are two regressions:

$$\begin{aligned} \mathcal{H}_{\text{obs},j\sim i} : X_j &= \beta_{0,ij} X_{i,\text{obs}} + \beta_{0,\mathcal{L}j} X_{\mathcal{L}} + \varepsilon_{0,ij}, \quad \varepsilon_{0,ij} \sim N(0, \sigma^2_{0,ij}), \\ \mathcal{H}_{\text{int},j\sim i} : X_j &= \beta_{1,ij} X_{i,\text{int}} + \beta_{1,\mathcal{L}j} X_{\mathcal{L}} + \varepsilon_{1,ij}, \quad \varepsilon_{1,ij} \sim N(0, \sigma^2_{1,ij}), \end{aligned}$$

with the linear regression coefficients,

$$\hat{\beta}_{0,ij} \sim N(\beta_{0,ij}, \sigma^2_{0,ij}(\mathbf{X}^T_{0,ij}\mathbf{X}_{0,ij})^{-1}_{ii}), \quad \hat{\beta}_{1,ij} \sim N(\beta_{1,ij}, \sigma^2_{1,ij}(\mathbf{X}^T_{1,ij}\mathbf{X}_{1,ij})^{-1}_{ii}).$$

Thus,

$$\hat{\beta}_{0,ij} - \hat{\beta}_{1,ij} \sim N(\beta_{0,ij} - \beta_{1,ij}, \sigma^2_{0,ij}(\mathbf{X}^T_{0,ij}\mathbf{X}_{0,ij})^{-1}_{ii} + \sigma^2_{1,ij}(\mathbf{X}^T_{1,ij}\mathbf{X}_{1,ij})^{-1}_{ii}),$$

meanwhile,

$$\frac{(n_{m,ij} - p_{ij})s^2_{m,ij}}{\sigma^2_{m,ij}} \sim \chi^2_{n_{m,ij}-p_{ij}} \text{ and here } s^2_{m,ij} = \frac{SSR_{m,ij}}{(n_{m,ij} - p_{ij})} \text{ for } m = 0, 1,$$

if under H_0, which means $\beta_{0,ij} = \beta_{1,ij}$ and $\sigma_{0,ij} = \sigma_{1,ij} = \sigma_{ij}$ then,

$$\frac{\hat{\beta}_{0,ij} - \hat{\beta}_{1,ij}}{\sigma_{ij}\sqrt{(\mathbf{X}_{0,ij}^T\mathbf{X}_{0,ij})_{ii}^{-1} + (\mathbf{X}_{1,ij}^T\mathbf{X}_{1,ij})_{ii}^{-1}}} \sim N(0,1),$$

and,

$$\{(n_{0,ij} - p_{ij})s_{0,ij}^2 + (n_{1,ij} - p_{ij})s_{1,ij}^2\}/\sigma_{ij}^2 \sim \chi^2_{n_{0,ij}+n_{1,ij}-2p_{ij}}.$$

Finally,

$$T = \frac{\hat{\beta}_{0,ij} - \hat{\beta}_{1,ij}}{s_{p,ij}\sqrt{(\mathbf{X}_{0,ij}^T\mathbf{X}_{0,ij})_{ii}^{-1} + (\mathbf{X}_{1,ij}^T\mathbf{X}_{1,ij})_{ii}^{-1}}} \sim t_{n_{0,ij}+n_{1,ij}-2p_{ij}},$$

here

$$s_{p,ij} = \sqrt{\frac{(n_{0,ij} - p_{ij})s_{0,ij}^2 + (n_{1,ij} - p_{ij})s_{1,ij}^2}{n_{0,ij} + n_{1,ij} - 2p_{ij}}}.$$

7.4 High-Dimensional Consistency of Algorithm 3

First we define $\mathcal{E}^1_{i\sim j}$,

$$\mathcal{E}^1_{i\to j} = \{s_{p,ij} \le \sqrt{2}\bar{\sigma}\}, \tag{7.16}$$

and $\mathcal{E}^2_{i\sim j}$,

$$\mathcal{E}^2_{i\to j} = \left\{\sqrt{\left(\mathbf{X}_{0,ij}^T\mathbf{X}_{0,ij}\right)_{ii}^{-1} + \left(\mathbf{X}_{1,ij}^T\mathbf{X}_{1,ij}\right)_{ii}^{-1}} \le \frac{1}{\sigma_*\sqrt{\zeta n}}\right\}. \tag{7.17}$$

Lemma 22. *For any i, j and constant $N \geq 2\bar{\sigma}^2$,*

$$s_{p,ij} \leq \sqrt{N},$$

with probability at least $1 - \exp(-(\zeta n - 2s)(N/\bar{\sigma}^2 - 1)/8)$.

Proof of Lemma 22. From the independence of data,

$$(n_{0,ij} - p_{ij})(s_{0,ij})^2/(\sigma_{0,ij})^2 + (n_{1,ij} - p_{ij})(s_{1,ij})^2/(\sigma_{1,ij})^2 \sim \chi^2_{n_{0,ij}+n_{1,ij}-2p_{ij}},$$

then,

$$\begin{aligned}
P(s_{p,ij} \geq \sqrt{N}) &= P(\sqrt{\frac{(n_{0,ij} - p_{ij})(s_{0,ij})^2 + (n_{1,ij} - p_{ij})(s_{1,ij})^2}{n_{0,ij} + n_{1,ij} - 2p_{ij}}} \geq \sqrt{N}) \\
&= P(\sqrt{\frac{(n_{0,ij} - p_{ij})(s_{0,ij})^2 + (n_{1,ij} - p_{ij})(s_{1,ij})^2}{n_{0,ij} + n_{1,ij} - 2p_{ij}}}/\bar{\sigma} \geq \sqrt{N}/\bar{\sigma}) \\
&\leq P(\sqrt{\frac{(n_{0,ij} - p_{ij})(s_{0,ij})^2/(\sigma_{0,ij})^2 + (n_{1,ij} - p_{ij})(s_{1,ij})^2/(\sigma_{1,ij})^2}{n_{0,ij} + n_{1,ij} - 2p_{ij}}} \geq \sqrt{N}/\bar{\sigma}) \\
&= P(\frac{\chi^2_{n_{0,ij}+n_{1,ij}-2p_{ij}}}{n_{0,ij} + n_{1,ij} - 2p_{ij}} \geq N/\bar{\sigma}^2),
\end{aligned}$$

since $\bar{\sigma} \geq \sigma_{m,ij}$ for $m = 1, 2$ and any i, j from Lemma 17. Now recall the χ^2_n concentration,

$$P(\chi^2_n/n \geq 1 + t) \leq e^{-nt/8} \text{ for } t \geq 1, \tag{7.18}$$

thus for any constant $N \geq 2\bar{\sigma}^2$,

$$P(s_{p,ij} \geq \sqrt{N}) \leq \exp(-(\zeta_{ij} n - 2p_{ij})(N/\bar{\sigma}^2 - 1)/8) \leq \exp(-(\zeta n - 2s)(N/\bar{\sigma}^2 - 1)/8),$$

so finally,

$$P(s_{p,ij} \leq \sqrt{N}) \geq 1 - \exp(-(\zeta n - 2s)(N/\bar{\sigma}^2 - 1)/8),$$

here ζ in defined in Assumption 9. □

Proposition 23. *For any i, j,*

$$s_{p,ij} \leq \sqrt{2}\bar{\sigma},$$

with probability at least $1 - \exp(-(\zeta n - 2s)/8)$.

Proof of Proposition 23. Set $N = 2\bar{\sigma}^2$ in Lemma 22. □

Lemma 24. *For any i, j,*

$$\sqrt{\left(\mathbf{X}_{0,ij}^T \mathbf{X}_{0,ij}\right)_{ii}^{-1} + \left(\mathbf{X}_{1,ij}^T \mathbf{X}_{1,ij}\right)_{ii}^{-1}} \leq \frac{2}{\sigma_* \delta \sqrt{\zeta n}},$$

with probability at least

$$1 - 2C \exp(-\zeta n (1-\delta)^2/2),$$

any $1 > \delta > 0$.

Proof of Lemma 24. With inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \in \mathbb{R}_+$,

$$\begin{aligned} P\left(\sqrt{\left(\mathbf{X}_{0,ij}^T \mathbf{X}_{0,ij}\right)_{ii}^{-1} + \left(\mathbf{X}_{1,ij}^T \mathbf{X}_{1,ij}\right)_{ii}^{-1}} \geq \frac{2}{\sigma_* \delta} \frac{\sqrt{n_{0,ij}} + \sqrt{n_{1,ij}}}{\sqrt{n_{0,ij} n_{1,ij}}}\right) \\ \leq P\left(\sqrt{\left(\mathbf{X}_{0,ij}^T \mathbf{X}_{0,ij}\right)_{ii}^{-1}} \geq \frac{2}{\sigma_* \delta \sqrt{n_{0,ij}}}\right) + P\left(\sqrt{\left(\mathbf{X}_{1,ij}^T \mathbf{X}_{1,ij}\right)_{ii}^{-1}} \geq \frac{2}{\sigma_* \delta \sqrt{n_{1,ij}}}\right), \end{aligned}$$

thus with Lemma 12,

$$P\left(\sqrt{\left(\mathbf{X}_{0,ij}^T\mathbf{X}_{0,ij}\right)_{ii}^{-1}+\left(\mathbf{X}_{1,ij}^T\mathbf{X}_{1,ij}\right)_{ii}^{-1}}\leq\frac{2}{\sigma_*\delta}\frac{\sqrt{\zeta_{0,ij}}+\sqrt{\zeta_{1,ij}}}{\sqrt{n\zeta_{0,ij}\zeta_{1,ij}}}\right)\geq 1-2C\exp(-\zeta n(1-\delta)^2/2),$$

finally use assumption 9,

$$P\left(\sqrt{\left(\mathbf{X}_{0,ij}^T\mathbf{X}_{0,ij}\right)_{ii}^{-1}+\left(\mathbf{X}_{1,ij}^T\mathbf{X}_{1,ij}\right)_{ii}^{-1}}\leq\frac{2}{\sigma_*\delta\sqrt{\zeta n}}\right)\geq 1-2C\exp(-\zeta n(1-\delta)^2/2).$$

□

Proposition 25. *For any i, j,*

$$\sqrt{\left(\mathbf{X}_{0,ij}^T\mathbf{X}_{0,ij}\right)_{ii}^{-1}+\left(\mathbf{X}_{1,ij}^T\mathbf{X}_{1,ij}\right)_{ii}^{-1}}\leq\frac{1}{\sigma_*\sqrt{\zeta n}},$$

with probability at least $1-2C\exp(-\zeta n/8)$.

Proof of Proposition 25. Set $\delta=1/2$ in Lemma 24. □

Now we can move to the main result of this section. Recall the test statistics,

$$T_{i\to j}=\frac{\hat{\beta}_{0,ij}-\hat{\beta}_{1,ij}}{s_{p,ij}\cdot\sqrt{\left(\mathbf{X}_{0,ij}^T\mathbf{X}_{0,ij}\right)_{ii}^{-1}+\left(\mathbf{X}_{1,ij}^T\mathbf{X}_{1,ij}\right)_{ii}^{-1}}},$$

for the notation simplicity, here we introduce a new notation $\mathcal{W}$,

$$\mathcal{W}=\sqrt{\left(\mathbf{X}_{0,ij}^T\mathbf{X}_{0,ij}\right)_{ii}^{-1}+\left(\mathbf{X}_{1,ij}^T\mathbf{X}_{1,ij}\right)_{ii}^{-1}},$$

then the test statistics can be expressed as,

$$T_{i\to j}=\frac{\hat{\beta}_{0,ij}-\hat{\beta}_{1,ij}}{s_{p,ij}\cdot\mathcal{W}}.$$

Also the event sets defined in (7.16) and (7.17) are,

$$\mathcal{E}^1_{i\to j} = \{s_{p,ij} \le \sqrt{2}\sigma\}, \quad \mathcal{E}^2_{i\to j} = \left\{\mathcal{W} \le \frac{1}{\sigma_*\sqrt{\zeta n}}\right\}.$$

Proof of Theorem 13. The error of single test for edge orientation on $(i, j, \mathcal{L})$ is,

$$\begin{aligned} P(E_{i\to j}) =& P(E^I_{i\to j} \cup E^{II}_{i\to j}) \\ =& P(E^I_{i\to j}) + P(E^{II}_{i\to j}) \\ =& P(|T_{i\to j}| \ge \alpha_n \mid \beta_{0,ij} = \beta_{1,ij} \text{ and } \sigma_{0,ij} = \sigma_{1,ij}) \\ &+ P(|T_{i\to j}| \le \alpha_n \mid \beta_{0,ij} \ne \beta_{1,ij} \text{ and } \sigma_{0,ij} \ne \sigma_{1,ij}). \end{aligned} \tag{7.19}$$

The type I error of (7.19),

$$P(E^I_{i\to j}) \le P\left(|T_{i\to j}| \ge \alpha_n \mid \beta_{0,ij} = \beta_{1,ij} \text{ and } \sigma_{0,ij} = \sigma_{1,ij}\right) \le P\left(|t_{n_{0,ij}+n_{1,ij}-p_{ij}}| \ge \alpha_n\right), \tag{7.20}$$

and Type II error of (7.19),

$$\begin{aligned} P(E^{II}_{i\to j}) =& P(|T_{i\to j}| \le \alpha_n \mid \beta_{0,ij} \ne \beta_{1,ij} \text{ and } \sigma_{0,ij} \ne \sigma_{1,ij}) \\ =& P\left(\left|\frac{\hat\beta_{0,ij} - \hat\beta_{1,ij}}{s_{p,ij}\cdot\mathcal{W}}\right| \le \alpha_n\right) \\ =& P\left(\left|\frac{(\hat\beta_{0,ij} - \beta_{0,ij}) - (\hat\beta_{1,ij} - \beta_{1,ij})}{\mathcal{W}}\right| \ge \frac{|\beta_{1,i\sim j} - \beta_{2,i\sim j}|}{\mathcal{W}} - \alpha_n s_{p,ij}\right). \end{aligned}$$

One important trick here:

$$\frac{(\hat\beta_{0,ij} - \beta_{0,ij}) - (\hat\beta_{1,ij} - \beta_{1,ij})}{\sqrt{(\sigma_{0,ij})^2\left(\mathbf{X}^T_{0,ij}\mathbf{X}_{0,ij}\right)^{-1}_{ii} + (\sigma_{1,ij})^2\left(\mathbf{X}^T_{1,ij}\mathbf{X}_{1,ij}\right)^{-1}_{ii}}} \sim \mathcal{N}(0,1),$$

and $\sigma_{0,ij} \wedge \sigma_{0,ij} \leq \bar{\sigma}$ from Lemma 17,

$$\begin{aligned}
P(E^{II}_{i\to j}) =& P\left(\left|\frac{(\hat{\beta}_{0,ij}-\beta_{0,ij})-(\hat{\beta}_{1,ij}-\beta_{1,ij})}{\bar{\sigma}\mathcal{W}}\right| \geq \frac{|\beta_{1,i\sim j}-\beta_{2,i\sim j}|}{\bar{\sigma}\mathcal{W}} - \alpha_n s_{p,ij}/\bar{\sigma}\right) \\
\leq& P\left(\left|\frac{(\hat{\beta}_{0,ij}-\beta_{0,ij})-(\hat{\beta}_{1,ij}-\beta_{1,ij})}{\sqrt{(\sigma_{0,ij})^2\left(\mathbf{X}^T_{0,ij}\mathbf{X}_{0,ij}\right)^{-1}_{ii}+(\sigma_{1,ij})^2\left(\mathbf{X}^T_{1,ij}\mathbf{X}_{1,ij}\right)^{-1}_{ii}}}\right| \geq \frac{|\beta_{0,ij}-\beta_{1,ij}|}{\bar{\sigma}\mathcal{W}} - \alpha_n s_{p,ij}/\bar{\sigma}\right) \\
=& P\left(|Z| \geq \frac{|\beta_{0,ij}-\beta_{1,ij}|}{\bar{\sigma}\mathcal{W}} - \alpha_n s_{p,ij}/\bar{\sigma}\right), \qquad (7.21)
\end{aligned}$$

here Z represents the standard normal distribution $\mathcal{N}(0,1)$ in (7.21).

Based on the definitons of $\mathcal{E}^1_{i\to j}$ and $\mathcal{E}^2_{i\to j}$ in (7.4),

$$\begin{aligned}
P(E^{II}_{i\to j} \mid \mathcal{E}^1_{i\to j} \cap \mathcal{E}^2_{i\to j}) \leq& P\left(|Z| \geq \frac{|\beta_{0,ij}-\beta_{1,ij}|}{\bar{\sigma}\mathcal{W}} - \alpha_n s_{p,ij}/\bar{\sigma}\right) \\
\leq& P\left(|Z| \geq \frac{|\beta_{0,ij}-\beta_{1,ij}|\sigma_*\sqrt{\zeta}}{\bar{\sigma}}\sqrt{n} - \sqrt{2}\alpha_n\right),
\end{aligned}$$

the Type II error conditioning on $\mathcal{E}^1_{i\to j} \cap \mathcal{E}^2_{i\to j}$. Then from Assumption 10,

$$P(E^{II}_{i\to j} \mid \mathcal{E}^1_{i\to j} \cap \mathcal{E}^2_{i\to j}) \leq P(|Z| \geq \frac{\psi_n \sigma_* \sqrt{\zeta}}{\bar{\sigma}}\sqrt{n} - \sqrt{2}\alpha_n),$$

next the bound of the Type II error,

$$P(E^{II}_{i\to j}) \leq P(E^{II}_{i\to j} \mid \mathcal{E}^1_{i\to j} \cap \mathcal{E}^2_{i\to j}) + P(\{\mathcal{E}^1_{i\to j}\}^c) + P(\{\mathcal{E}^2_{i\to j}\}^c), \qquad (7.22)$$

the last two probability we have given bounds in Proposition 23 and 25,

$$P(\{\mathcal{E}^1_{i\to j}\}^c) \leq \exp(-(\zeta n - 2s)/8), \quad P(\{\mathcal{E}^2_{i\to j}\}^c) \leq 2C\exp(-\zeta n/8). \qquad (7.23)$$

Consider $\zeta \gtrsim O(n^{-q})$, then from (7.4) we have,

$$P(\{\mathcal{E}^1_{i\to j}\}^c) + P(\{\mathcal{E}^2_{i\to j}\}^c) \precsim \exp(-n^{1-q}). \tag{7.24}$$

For the Type I error in (7.20),

$$P(|t_{n_{0,ij}+n_{1,ij}-p_{ij}}| \geq \alpha_n) \leq \exp(-\alpha_n^2/4) + \frac{1}{2}\exp(-(\zeta n - s)/16)), \tag{7.25}$$

with the help of Lemma 18.

Then for the first term in (7.22), set $\alpha_n = \eta_n\sqrt{n}$ with ancillary factor η_n,

$$P\left(|Z| \geq \frac{\psi_n\sigma_*\sqrt{\zeta}}{\bar{\sigma}}\sqrt{n} - \sqrt{2}\alpha_n\right) = P\left(|Z| \geq (\frac{\psi_n\sigma_*\sqrt{\zeta}}{\bar{\sigma}} - \sqrt{2}\eta_n)\sqrt{n}\right),$$

so the constraint on η_n is,

$$\eta_n < \frac{\psi_n\sigma_*\sqrt{\zeta}}{\sqrt{2}\bar{\sigma}}.$$

We can set,

$$\eta_n = \frac{\psi_n\sigma_*\sqrt{\zeta}}{2\sqrt{2}\bar{\sigma}} = O(n^{-d-q/2}) \Longrightarrow \alpha_n = O(n^{1/2-d-q/2}), \tag{7.26}$$

then,

$$P\left(|Z| \geq \left(\frac{\psi_n\sigma_*\sqrt{\zeta}}{\bar{\sigma}} - \sqrt{2}\eta_n\right)\sqrt{n}\right) = P\left(|Z| \geq \frac{\psi_n\sigma_*\sqrt{\zeta}}{2\bar{\sigma}}\sqrt{n}\right) \leq 2\exp(-\kappa_n^2 n/2),$$

here $\kappa_n = \psi_n\sigma_*\sqrt{\zeta}/2\bar{\sigma} = O(n^{-d-q/2})$. So summarize the result of this part,

$$P(E^{II}_{i\to j} \mid \mathcal{E}^1_{i\to j} \cap \mathcal{E}^2_{i\to j}) \precsim \exp(-n^{1-2d-q}). \tag{7.27}$$

Take α set in (7.26) into (7.25),

$$P(|t_{n_{0,ij}+n_{1,ij}-p_{ij}}| \geq \alpha_n) \precsim \exp(-n^{1-2d-q}). \tag{7.28}$$

Finally combine (7.19), (7.20), (7.22), (7.24), (7.27) and (7.28), we can get,

$$\sup_{i,j} P(E_{i\to j}) \precsim \exp(-n^{1-2d-q}),$$

as there are at most $ps/2$ reversible edges in the essential graph,

$$\begin{aligned} P(\text{Error}) &\leq ps \cdot \sup_{i,j} P(E_{i\to j}) \precsim ps \cdot \exp(-n^{1-2d-q}) \\ &\precsim \exp(-n^{1-2d-q} + (a-b+1)\log(n)) \to 0, \end{aligned}$$

the result is alway true when $1-2d-q>0$. □

CHAPTER 8

Summary and Discussion

8.1 Main Contributions

In this dissertation, the main target is to build a constraint-based method for structure learning from interventional data. Intervention provides the motivation that we can approach more precise estimation of graph structure beyond the traditional observational method. Meanwhile, intervention also sets some challenges for our work. For the graph structure, intervention will remove some arrows from the DAG; and from probability perspective, the joint distribution needs to modify after intervention. In this case, this dissertation is always facing multiple graphs and distributions. One straightforward difficulty is: the sample is independent but not always identical, which means it is not that trivial to find help from the existing results in probability theory. And besides these, similar to other work in this area, the number of DAGs in the search space also the complexity of graph structure itself will increase superexponentially as the number of nodes increases.

To overcome these difficulties, this work introduces two stages for the graph structure learning. For the first stage, to recover the skeleton of graph, this dissertation discusses how to extend the original PC algorithm to the interventional case. And correspondingly, some conditions are given to guide the intervention family design such that the algorithm can find the true skeleton. And in Section 4.1, this work provides a throughout description of the neighborhood of the reversible edge in the essential graph. And based on the intuition, we show and prove the invariance relations between intervention targets; see Theorem 9 and

10. Then furthermore, one edge orientation method is introduced by testing the invariance relations. It is worthwhile to emphasize that there are mild assumptions on the intervention family we discuss this part of work. We are not working on the single intervention, instead our methods can handle very complicated intervention family. And another contribution for edge orientation is: inspired by Theorem 10, we design a rule to merge intervention blocks while conducting the tests.

Next, a major work of this dissertation is to show the consistency results of our structure learning method under the sparse high-dimensional settings. Such kind of results are rarely seen in this area. And some assumptions given in this part are also quite innovative, and I do believe it can provide some good learning for interventional method; see Assumption 6 and 10. For the theoretical part, we give clear and explicit formulas of the test statistics we used for structure learning for the Gaussian graph.

Finally, simulation results are provided to evaluate the performance of the structure learning methods built in this work. Even though we focus on the two stages of graph structure learning, we can still decouple the edge orientation with the skeleton recovery in practice. In Section 6.3, GES plus EO shows better performance than GES only. If we have good estimation of the skeleton, the edge direction can be implemented without the PC algorithm. That is a good advantage of this kind of constraint-based method.

8.2 Future Directions

One potential improvement is to extend the theoretical part in this dissertation from Gaussian to other probability distributions. Edge orientation on the discrete case is an attractive direction, as it can be widely applied to many datasets in the real world. Different to the Gaussian graph assumption, we can assume the variables in the graph follow binomial or multinomial distribution. As mentioned in Remark 3, Theorem 10 can be applied to the discrete case with no change required. The difficulty is: in this work, we test the coefficient of

regressions, the result of which infers the invariance of conditional distribution. For discrete case, it is not a good idea to implement linear regression. G-test (or equivalently likelihood ratio test) could be helpful to solve the discrete case, however many technical details are still waiting to confirm.

Besides the discrete case, another approach is to extend this work to sub-Gaussian distribution. Even though we derive the test statistics based on the Gaussian distribution, it may still work for sub-Gaussian variables, as the behavior of probability tail is so similar in sub-Gaussian. We have strong confidence that this method can be applied to the sub-Gaussian case in practice. But it is not easy to provide similar theoretical results. In Chapter 7, we derive a tail bound for t-distribution, and work on the Gaussian ensemble. Many existing results on Gaussian distribution cannot be extended to sub-Gaussian, especially those high-dimensional tail bound.

There are some other possible directions in this area. We mentioned in introduction chapter that [GB13] proves the the high-dimensional consistency of DAG structure learning with ℓ_0-penalized maximum likelihood estimation. It is also attractive to use maximum likelihood score on intervention case. Since we show in this dissertation that the behavior of conditional distributions can be quite different between the interventions, which will also affect the maximum likelihood intuitively. Another interesting direction is about the intervention design. For graph structure learning, intervention design is crucial and deterministic. As the experimental data could be expensive, it will provide huge benefit if we can design the optimal interventions containing the maximum amount of information such that the learning algorithm can approach better $\mathcal{I}$-essential graph; see [LKD18], [ZM22]. Recently, some researchers, for example [SWU20] and [CP22], focus on the graph structure learning with unknown or uncertain interventions, which is also an interesting and popular topic in this area.

Bibliography

[AMP97] Steen A Andersson, David Madigan, Michael D Perlman, et al. "A characterization of Markov equivalence classes for acyclic digraphs." *The Annals of Statistics*, **25**(2):505–541, 1997.

[BEd08] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data." *The Journal of Machine Learning Research*, **9**:485–516, 2008.

[CM02] David Maxwell Chickering and Christopher Meek. "Finding optimal Bayesian networks." In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 94–102. Morgan Kaufmann Publishers Inc., 2002.

[CP22] Federico Castelletti and Stefano Peluso. "Network structure learning under uncertain interventions." *Journal of the American Statistical Association*, pp. 1–12, 2022.

[Ebe07] Frederick Eberhardt. "Causation and intervention." *Unpublished doctoral dissertation, Carnegie Mellon University*, p. 93, 2007.

[GB13] Sara van de Geer and Peter Bühlmann. "ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs." *The Annals of Statistics*, **41**(2):536–567, 2013.

[HB12] Alain Hauser and Peter Bühlmann. "Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs." *The Journal of Machine Learning Research*, **13**(1):2409–2464, 2012.

[HB15] Alain Hauser and Peter Bühlmann. "Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**(1):291–318, 2015.

[HG08] Yang-Bo He and Zhi Geng. "Active learning of causal networks with intervention experiments and optimal designs." *Journal of Machine Learning Research*, **9**(Nov):2523–2547, 2008.

[KB07] Markus Kalisch and Peter Bühlmann. "Estimating high-dimensional directed acyclic graphs with the PC-algorithm." *The Journal of Machine Learning Research*, **8**:613–636, 2007.

[KFB09] Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[KHN04] Kevin B Korb, Lucas R Hope, Ann E Nicholson, and Karl Axnick. "Varieties of causal intervention." In *Pacific Rim International Conference on Artificial Intelligence*, pp. 322–331. Springer, 2004.

[Lau96] Steffen L Lauritzen. *Graphical models*. Oxford University Press, 1996.

[LKD18] Erik Lindgren, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. "Experimental design for cost-aware learning of causal graphs." *Advances in Neural Information Processing Systems*, **31**, 2018.

[LM00] Beatrice Laurent and Pascal Massart. "Adaptive estimation of a quadratic functional by model selection." *The Annals of Statistics*, pp. 1302–1338, 2000.

[Mee95] Christopher Meek. "Causal inference and causal explanation with background knowledge." *Uncertainty in Artificial Intelligence*, **11**:403–410, 1995.

[NHM18] Preetam Nandy, Alain Hauser, Marloes H Maathuis, et al. "High-dimensional consistency in score-based and hybrid structure learning." *The Annals of Statistics*, **46**(6A):3151–3183, 2018.

[Pea00] Judea Pearl. *Causality: Models, reasoning and inference*. Cambridge Univ Press, 2000.

[Rob77] Robert W Robinson. "Counting unlabeled acyclic digraphs." In *Combinatorial mathematics V*, pp. 28–43. Springer, 1977.

[SG91] Peter Spirtes and Clark Glymour. "An algorithm for fast recovery of sparse causal graphs." *Social Science Computer Review*, **9**(1):62–72, 1991.

[SGS00] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. The MIT Press, second edition, 2000.

[SWU20] Chandler Squires, Yuhao Wang, and Caroline Uhler. "Permutation-based causal structure learning with unknown intervention targets." In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR, 2020.

[VP90] Thomas Verma and Judea Pearl. "Causal networks: Semantics and expressiveness." In *Machine intelligence and pattern recognition*, volume 9, pp. 69–76. Elsevier, 1990.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[YAZ20] Qiaoling Ye, Arash Amini, and Qing Zhou. "Optimizing regularized cholesky score for order-based learning of bayesian networks." *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[ZM22] Michele Zemplenyi and Jeffrey W Miller. "Bayesian optimal experimental design for inferring causal structure." *Bayesian Analysis*, **1**(1):1–28, 2022.