Technische Universität München

TUM School of Computation, Information and Technology

# Homoscedasticity and Feedback Loops in Graphical Models

**Jun Wu**

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

**Doktors der Naturwissenschaften (Dr. rer. nat.)**

genehmigten Dissertation.

**Vorsitz:**

Prof. Dr. Matthias Scherer

**Prüfer der Dissertation:**

1. Prof. Mathias Drton, Ph.D.
2. Prof. Dr. Thomas Kahle,
   Otto-von-Guericke-Universität Magdeburg
3. Prof. Dr. Seyed Jalal Etesami

Die Dissertation wurde am 28.06.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 28.08.2023 angenommen.

# Abstract

Graphical models provide a tractable framework to encode stochastic dependencies between a set of random variables. In particular, they are able to capture noisy functional relations between the variables. In this framework, directed cycles in a graph imply the existence of feedback loops in the studied system, while bidirected edges are often used to represent dependencies induced by latent confounders. The focus of this thesis are graphical models specified via linear structural equations and the challenges that result from allowing the graph to contain directed cycles or bidirected edges.

We start the discussion by considering models given by bow-free mixed graphs that allow for feedback loops as well as certain types of latent confounding. We prove that the models in this class are of expected dimension, and we provide a sufficient condition for distributional equivalence between different graphs. For structure learning with observational data, we propose a greedy search scheme with model scores defined using maximum likelihood estimates (MLE).

Next, we consider the problem of computing MLEs in settings where one has both observational and interventional data. Following earlier work that only considered observational data, we develop a block-coordinate descent scheme that is applicable to models that may feature feedback loops. We lay out specific conditions on a graph or interventions, under which the resulting algorithm involves update steps with simple explicit solutions.

An important issue in graphical modeling is to clarify whether the graphical structure underlying a model is identifiable on the basis of the available data. The existing theory on this topic provides equivalence results as well as analyses of different conditions that ensure unique identifiability. We contribute to this literature by studying the class of graphical models with homoscedastic error variances in a generalization that allows for feedback loops. We give a definition of generic identifiability (distinguishability) based on a geometric perspective, and we develop general graphical criteria that, under mild conditions, are able to certify when the models given by two graphs are distinguishable. We also report on an extensive computational study that shows that nearly all simple directed graphs with at most 6 nodes are generically identifiable under the assumption of homoscedastic error variances.

Finally, we propose a class of models that weakens full homoscedasticity of error variances to a setting of partial homoscedasticity. In this setting, error variances are restricted to be equal only within the blocks of a partition of the considered random variables. For the class of directed acyclic graphs, we obtain a full characterization of model equivalence under such partial homoscedasticity.

# Zusammenfassung

Graphische Modelle sind nützliche Werkzeuge, um stochastische Abhängigkeiten zwischen einer Menge von Zufallsvariablen abzubilden. Insbesondere können sie verrauschte funktionale Beziehungen zwischen den Variablen erfassen. In diesem Rahmen deuten gerichtete Zyklen in einem Graphen auf das Vorhandensein von Rückkopplungsschleifen im untersuchten System hin, während bigerichtete Kanten oft dazu verwendet werden, Abhängigkeiten zu repräsentieren, die durch verborgene Störfaktoren verursacht werden. In dieser Arbeit betrachten wir lineare Strukturgleichungsmodelle, die mit Graphen assoziiert sind. Die betrachteten Graphen enhalten Zyklen oder bigerichtete Kanten.

Wir beginnen die Diskussion, indem wir Modelle betrachten, die durch bogenfreie gemischte Graphen gegeben sind und Rückkopplungsschleifen sowie bestimmte Arten verborgener Störfaktoren zulassen. Wir beweisen, dass die Modelle dieser Klasse von erwarteter Dimension sind und wir entwickeln eine hinreichende Bedingung für die verteilungstheoretische Äquivalenz verschiedener Graphen. Für das Strukturlernen mit beobachteten Daten untersuchen wir ein Greedy-Suchverfahren das Informationskriterien optimiert, welche über Maximum-Likelihood-Schätzung berechnet werden.

Als Nächstes betrachten wir das Problem der Berechnung von Maximum-Likelihood-Schätzern in Situationen, in denen sowohl Daten aus einer Beobachtungsstudie als auch Daten aus Interventionsexperimenten vorhanden sind. Aufbauend auf früheren Arbeiten, die nur Beobachtungsstudien betrachten, entwickeln wir ein Blockkoordinaten-Abstiegsverfahren, das auf Modelle anwendbar ist, die Rückkopplungsschleifen enthalten können. Wir legen spezifische Bedingungen für einen Graphen oder Interventionen fest, unter denen die Einzelschritte des resultierenden Algorithmus einfache explizite Lösungen erlauben.

Eine wichtige Fragestellung in der graphischen Modellierung besteht darin, zu klären, ob die graphische Struktur, die einem Modell zugrunde liegt, anhand der verfügbaren Daten identifizierbar ist. Die bestehende Theorie zu diesem Thema liefert Äquivalenzergebnisse sowie Analysen verschiedener Bedingungen, die eine eindeutige Identifizierbarkeit sicherstellen. Wir tragen zu dieser Literatur bei, indem wir die Klasse der graphischen Modelle mit homoskedastischen Fehlervarianzen in einer Verallgemeinerung untersuchen, die Rückkopplungsschleifen ermöglicht. Wir geben eine Definition der generischen Identifizierbarkeit (Unterscheidbarkeit) basierend auf einer geometrischen Perspektive und entwickeln allgemeine graphische Kriterien, die unter milden Bedingungen zertifizieren können, wann die Modelle, die durch zwei Graphen gegeben sind, unterscheidbar sind. Wir präsentieren eine umfangreiche Rechenstudie, die zeigt, dass nahezu alle einfachen gerichteten Graphen mit höchstens 6 Knoten unter der Annahme homoskedastischer Fehlervarianzen generisch identifizierbar sind.

Schließlich schlagen wir eine Klasse von Modellen vor, welche die vollständige Homoskedastizität der Fehlervarianzen auf ein Setup der partiellen Homoskedastizität abschwächt. In diesem Setup ist die Gleichheit der Fehlervarianzen nur innerhalb der Blöcke einer Partition der betrachteten Zufallsvariablen angenommen. Für die Klasse der gerichteten azyklischen Graphen erhalten wir eine vollständige Charakterisierung der Modelläquivalenz unter dieser partiellen Homoskedastizität.

# Acknowledgement

First and foremost, I am really grateful to my supervisor Mathias Drton for providing me with the opportunity to work on this research. I would like to thank him for his vision, suggestions and feedback, as well as his guidance and patience during the whole project, which was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 883818).

Secondly I would like to thank Benjamin Hollering for the insightful discussions and inspiring collaboration on the topic of algebraic matroids method.

Moreover, I would like to express my gratitude to my colleagues and the guests in the Statistics Research Group, particularly Carlos Améndola, a board game enthusiast. I am thankful for the scientific discussions, various after-work activities and exchange of thoughts on a wide range of topics. Without you my experience at TUM could not have been so enjoyable. I would also like to extend special thanks to Andrea Grant, the secretary in our group, on whose assistance I could always count in administrative matters. I am also very grateful for the help Stephan Haug has provided me in IT and teaching matters.

I would also like to thank the examination committee, which includes the chair Matthias Scherer and the examiners Thomas Kahle and Seyed Jalal Etesami, for taking the time to organize the examination and evaluate the thesis.

Finally, I am grateful to my family for their consistent support, not only during these years, but all throughout my academic journey.

# Contents

**Abstract** iii

**Zusammenfassung** v

**Acknowledgement** vii

**List of Figures** xi

**List of Tables** xiii

**Notation and Acronyms** xv

**1 Introduction** 1

**2 Preliminaries** 3
2.1 Backgrounds . . . 3
2.2 Brief literature review . . . 6
2.3 Organization of the Thesis . . . 7

**3 Structure Learning for Simple Mixed Cyclic Graphs** 8
3.1 Models defined by simple mixed graphs . . . 8
3.2 Sufficient Conditions for Distributional Equivalence . . . 10
3.2.1 Useful Lemmas . . . 11
3.2.2 Constructing Covariance Matrices . . . 13
3.3 Greedy Search . . . 14
3.4 Numerical Experiments . . . 15
3.4.1 Simulation Studies . . . 16
3.4.2 Protein Expression Data . . . 17
3.5 Discussion . . . 19

**4 BCD Algorithm for Interventional Data in Directed Graphical Models** 21
4.1 Optimization Problems for Gaussian Errors . . . 22
4.2 Maximum likelihood degrees, a concrete example . . . 25
4.3 Block-coordinate Descent . . . 27
4.3.1 Arbitrary directed graphs, special interventions . . . 27
4.3.2 Special directed graphs under arbitrary interventions . . . 32
4.3.3 Properties . . . 33
4.4 Simulation Studies . . . 37
4.5 Discussion . . . 39
4.5.1 Mixed graphs . . . 39
4.5.2 Stability . . . 39

**5 Identifiability of Linear SEMs using Algebraic Matroids** **45**
5.1 Matroid Approach: Preliminaries . . . . . 45
5.2 Jacobian . . . . . 48
5.3 Graphical Conditions for Distinguishing SEMs with Matroids . . . . . 51
5.4 Computational Study . . . . . 62
5.4.1 Summary . . . . . 62
5.4.2 p = 2 . . . . . 64
5.4.3 p = 3 . . . . . 65
5.4.4 p = 4 . . . . . 69
5.4.5 p = 5 . . . . . 70
5.4.6 p = 6 . . . . . 70
5.5 Long Matrices in Intermediate Steps . . . . . 72

**6 Partial Homoscedasticity in Causal Discovery with Linear Models** **74**
6.1 Setup . . . . . 74
6.2 Equal Variance Constraints and Model Characterization . . . . . 75
6.2.1 Equal Variance Constraints . . . . . 75
6.2.2 Characterization of the Models . . . . . 79
6.3 Equivalence Classes and CPDAG . . . . . 80
6.4 Greedy Search and Simulation Studies . . . . . 86
6.4.1 Greedy Search Scheme . . . . . 86
6.4.2 Simulation Studies . . . . . 87
6.5 Discussion . . . . . 87

**7 Conclusions** **91**

**Bibliography** **93**

# List of Figures


2.1 A directed graph and the corresponding mixed graph with unobserved $X_2$. . . . . 4
2.2 Two simple mixed graphs. . . . . 6

3.1 Commutative diagram illustrating two ways of obtaining a matrix $\Sigma \in M_{G_1} \cap M_{G_2}$. One is the parametrization $\phi_{G_1}$, while the other is a composition of maps (including the map $H$ defined via Lemma 3.9), that we denote $\Psi$. . . . . 14
3.2 Estimated graphs corresponding to the case of minimum number of edges (10 edges, dataset 1) and maximum number of edges (13 edges, dataset 6). . . . . 18
3.3 Edges appearing in at least 9 (a) and 13 (b) skeletons from the estimated graphs. . . . . 18
3.4 Curves of scores versus the time (seconds) in 300 random restarts greedy search for dataset 1 and dataset 6. . . . . 19

4.1 Original graph, and manipulated graph with intervention target $I = \{2\}$. 22
4.2 Graph of a linear SEM, with ML degree 23 for observational data. . . 25
4.3 Examples of functions from the 4 cases in Lemma 4.5. Two roots in case (ii) and one root in case(iii). In case (i) and (iv) the function has no roots. . . . . 35
4.4 A 2-cycle with parameters. . . . . 36
4.5 Original graph with bidirected edge, and manipulated graph with intervention target $I = \{2\}$. . . . . 40

5.1 A directed cyclic graph, with edge weights and equal error variances. . 46
5.2 The diamond graph $G$ used in Example 5.12. . . . . 50
5.3 Two 4-cycle of different directions, the key edges for selecting column set are highlighted in red. . . . . 56
5.4 The two forbidden subgraph structures for transitive triangle (shielded collider) free graphs. . . . . 57
5.5 Two graphs with the same out-degree sequence but which have no transitive triangles (shielded colliders). . . . . 59
5.6 This displays the subgraph relating $i, j$, and $l$ in the proof of Theorem 5.30. Since $L$ is parentally closed and $j \in L$ and $l \in \mathrm{pa}(j) \cap \mathrm{ne}(i)$, it must be that $l \in L$. . . . . 61
5.7 Example: necessity of checking unions of minimal parentally closed sets. 62
5.8 Two graphs with the same Jacobian matroid. . . . . 68
5.9 Two graphs with the same Jacobian matroid, p=4. . . . . 70
5.10 Two graphs with the same Jacobian matroid, $p$=5. . . . . 70

6.1 Under the constraint $\omega_{11} = \omega_{22}$, $G_1$ and $G_2$ generate different models. 75
6.2 The three subcases when there exists a node $k \in \mathrm{de}(i) \cap A_i$. . . . . 78
6.3 The active path $q$. . . . . 79
6.4 The four orientation rules. . . . . 83
6.5 $i_3 \to i_4$ from R1. . . . . 85
6.6 $i_3 \to i_4$ from R2. . . . . 85
6.7 A DAG and the corresponding CPDAG, under a fixed partition. . . . 85
6.8 Box-plots of SHD by groups of $p$ and $n$, sparse graphs 2 blocks. . . . . 88
6.9 Box-plots of SHD by groups of $p$ and $n$, dense graphs, 2 blocks. . . . . 88
6.10 Box-plots of SHD by groups of $p$ and $n$, sparse graphs, $\lceil p/3 \rceil + 1$ blocks. 89
6.11 Box-plots of SHD by groups of $p$ and $n$, dense graphs, $\lceil p/3 \rceil + 1$ blocks. 89

# List of Tables


3.1 Statistics on p=5. . . . . 16
3.2 Difference between the dimensions of the true graph and of the estimated graph. . . . . 17
3.3 Statistics on the Number of Edges. . . . . 17

4.1 Results on directed graphs . . . . . 38

5.1 Number of pairs of 4-node simple directed graphs, with same number of edges, that **cannot** be distinguished by Theorem 5.21 (outdegree sequence method) or Theorem 5.30 (outdegree sequence and parentally closed sets method). . . . . 70
5.2 Number of pairs of 5-node simple directed graphs, with same number of edges, that **cannot** be distinguished by Theorem 5.21 (outdegree sequence method) or Theorem 5.30 (outdegree sequence and parentally closed sets method). . . . . 71
5.3 Number of pairs of 6-node simple directed graphs, with same number of edges, that **cannot** be distinguished by Theorem 5.21 (outdegree sequence method) or Theorem 5.30 (outdegree sequence and parentally closed sets method). . . . . 71

# Notation and Acronyms

| | |
|---|---|
| $B$ | Set of bidirected edges. |
| $D$ | Set of directed edges. |
| $G$ | A graph. |
| $G_{\overline{I}}$ | Manipulated graph of $G$ with intervention target $I$. |
| $I$ | Identity matrix. |
| $I_k$ | Intervention target set. |
| $J$ | Jacobian matrix of a parameterization. |
| $K$ | Precision matrix of $X$. |
| $M_G$ | Model given by graph $G$. |
| $M_{G,\Pi}$ | Model given by graph $G$ under partition $\Pi$. |
| $O$ | Zero matrix. |
| $S$ | Sample covariance matrix of $X$. |
| $U[a, b]$ | Uniform distribution on $[a, b]$. |
| $V$ | Set of nodes. |
| $X$ | A random vector. |
| $\Lambda$ | Adjacency matrix with edge weights. |
| $\Omega$ | Covariance matrix of the random error $\varepsilon$. |
| $\Pi$ | A fixed partition of nodes. |
| $\Sigma$ | Covariance matrix of $X$. |
| $\mathrm{an}(i)$ | Set of ancestors of $i$, excluding $i$. |
| $\mathrm{ch}(i)$ | Set of children of $i$. |
| $\mathrm{de}(i)$ | Set of descendants of $i$, including $i$. |
| $\ell(\cdot)$ | Log-likelihood function. |
| $\mathbb{R}^D$ | Set of real matrices with support in $D$. |
| $\mathbf{X}$ | A data matrix of size $p \times n$. |
| $\omega$ | Variance vector of the random error $\varepsilon$. |
| $\mathcal{C}(i, G)$ | Strongly connected component in $G$ containing $i$. |
| $\mathcal{I}$ | Collection of intervention target sets. |
| $\mathcal{M}$ | A matroid. |
| $\mathcal{P}(A)$ | Powerset of the set $A$. |
| $\mathcal{R}$ | Standardization map of a covariance matrice $\Sigma$. |
| $PD$ | Cone of positive definite symmetric matrices. |
| $PD(B)$ | Subcone of $PD$ with support over $B$. |
| $\mathrm{ne}(i)$ | Set of neighbors of $i$. |
| $\omega_{ij}$ | The $i, j$ entry in the covariance matrix of the random error $\varepsilon$. |
| $\mathrm{pa}(i)$ | Set of parents of $i$. |
| $\perp_d$ | $d$-separation. |
| $\phi_G$ | Covariance parametrization. |

| | |
|---|---|
| $\pi$ | A block in a partition. |
| $\psi_G$ | Precision parametrization. |
| $\rho(\Lambda)$ | Spectral radius of the matrix $\Lambda$. |
| $\sigma_{ij}$ | The $i, j$ entry in the covariance matrix of $X$. |
| $\tau$ | A trek. |
| $\deg(i)$ | Degree of node $i$. |
| $\mathrm{diag}(\cdot)$ | Vector-to-matrix diag operator. |
| $\dim(M)$ | Dimension of algebraic model $M$. |
| $\varepsilon$ | A random error vector. |
| $i \leftrightarrow j$ or $\{i, j\}$) | Bidirected edge between $i$ and $j$. |
| $i \to j$ or $(i, j)$ | Directed edge from $i$ to $j$. |
| $n$ | Sample size of one dataset. |
| $p$ | Size of the set of nodes. |
| $s$ | Inverse of the common variance. |
| BCD | Blockwise Coordinate Descent algorithm. |
| CPDAG | Completed Partially Directed Acyclic Graph. |
| DAG | Directed Acyclic Graph. |
| GES | Greedy Equivalence Search algorithm. |
| PC | Peter-Clark algorithm. |
| RICF | Residual Iterative Conditional Fitting algorithm. |
| SEM | Structural Equation Model. |
| SHD | Structural Hamming Distance. |

# Chapter 1

# Introduction

A graphical model is a powerful framework for describing the statistical dependencies between random variables. It can be used to model multivariate distributions in various application areas. Directed acyclic graphical (DAG) models encode functional relations between variables through a directed graph. The edge directions have natural connection to causality, allowing them to represent causal models.

For the problem of linear Gaussian structural equation models (SEMs) with the associated DAGs, Markov equivalence provides an elegant characterization of equivalent classes through conditional independence information [Spirtes et al., 2000, Pearl, 2009]. This characterization yields model identifiability results and also aids structure learning tasks. However, in certain applications like gene expression network, there may exist feedback loops in the graph structure [Sachs et al., 2005]. This limitation motivates the study of graphical models with feedback loops. The problem is notoriously challenging, especially when latent variables are present [Evans, 2020]. In this type of models, meaningful and concise identifiability results necessitate special assumptions. An important condition is the presence of homoscedastic errors, i.e. all random errors have equal variances [Peters and Bühlmann, 2014, Chen et al., 2019].

In this thesis, we present some advances in structure learning and parameter estimation of linear SEMs, when the graphical models may contain cycles. Additionally, we discuss the structural identifiability properties under the assumption of (partially) homoscedastic errors.

A bow-free mixed graph is simple, meaning that there is at most one (bi)directed between every pair of node [Améndola et al., 2020]. This type of model restricts the presence of common unobserved parent variable (coufounders) for two observed variables only in the absence of direct causal effects. We demonstrate that the simple cyclic models have expected dimension, and analogous sufficient conditions for distribution equivalence still hold when compared bow-free acyclic models that have been studied in Nowzohour et al. [2017].

Residual Iterative Conditional Fitting (RICF) is an algorithm used to compute the maximum likelihood estimation (MLE) of parameters in bow-free acyclic models with observational data [Drton and Richardson, 2004]. Blockwise coordinate descent (BCD) is an extension of RICF that also works for general mixed cyclic graphs [Drton et al., 2019b]. Building upon these algorithm, we develop a BCD-type algorithm that can compute the MLE of parameters in general directed cyclic models with both observational and interventional data.

Structural identifiability provides the theoretic foundation for well-defined graphical models. The assumption of equal error variances, or homoscedastic errors, is restrictive but also really strong in distinguishing DAG models. The class of DAGs is identifiable, indicating every DAG has a unique model and the size of every equivalence class is 1 [Peters and Bühlmann, 2014, Chen et al., 2019]. We employ algebraic matroids [Hollering and Sullivant, 2021] to investigate the structural identifiability of directed cyclic graphs with homoscedastic errors. Distinguishing criteria are developed, and symbolic computation checks are performed. Furthermore, we extend the homoscedastic errors assumption to "groupwise homoscedastic errors" in DAG models, where variables (nodes) are partitioned into several blocks, with each block corresponding to a common error variance value. In this case, we derive an analogue of classic Markov equivalence class theory and completed partially directed acyclic graph (CPDAG) construction procedure.

# Chapter 2

# Preliminaries

## 2.1 Backgrounds

The main model we focus on in this thesis is the **linear stuctural equation model**, which takes the following form. Let $\varepsilon = (\varepsilon_i : i \in V)$ be a vector of random errors and $X = (X_i : i \in V)$ be a random vector satisfying the structural equation system:

$$X = \Lambda^T X + \varepsilon, \tag{2.1}$$

in which $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{V \times V}$ with the unknown coefficient $\lambda_{ij}$ for $(i \neq j)$ being the direct effect of $X_j$ on $X_i$. The error vector $\varepsilon$ is assumed to have postive definite covariance matrix $\Omega = (\omega_{ij}) \in \mathbb{R}^{V \times V}$. When $I - \Lambda$ is invertible, the equation system has a unique solution $X = (I - \Lambda)^{-T}\varepsilon$; here $I$ is the identity matrix. The covariance matrix of this solution $X$ is

$$\mathrm{Var}[X] = \Sigma = (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1}. \tag{2.2}$$

The linear structural equation model can be naturally represented by a graph. For independent errors $\varepsilon$ (i.e., $\Omega$ is diagonal), the linear SEM is associated to a directed graph $G = (V, D)$, where $V$ is the set of nodes and $D \subseteq V \times V$ is the edge set. Every node in $V$ represents a random variable. Elements in $D$ are ordered pairs $(i, j)$, $i \neq j$, also denoted by $i \to j$, encoding the causal relationships between random variables. If some errors are allowed to be dependent, the correlated node pairs are connected by bidirected edges. The resulting graph is a mixed graph $G' = (V, D, B)$ that can be interpreted as representing direct effects as well as latent confounding (some latent variables have effects on the variables corresponding to the node pair joined by a bidirected edge). As above the directed edge set $D$ contains ordered pairs, and elements in $B$ are unordered pairs $\{i, j\}$ representing bidirected edges.

**Example 2.1.** *The linear SEM encoded by the graph in (a) is*

$$\begin{aligned} X_1 &= \lambda_{21}X_2 + \varepsilon_1, \\ X_2 &= \varepsilon_2, \\ X_3 &= \lambda_{13}X_1 + \lambda_{23}X_2 + \varepsilon_3, \\ X_4 &= \lambda_{34}X_3 + \varepsilon_4. \end{aligned}$$

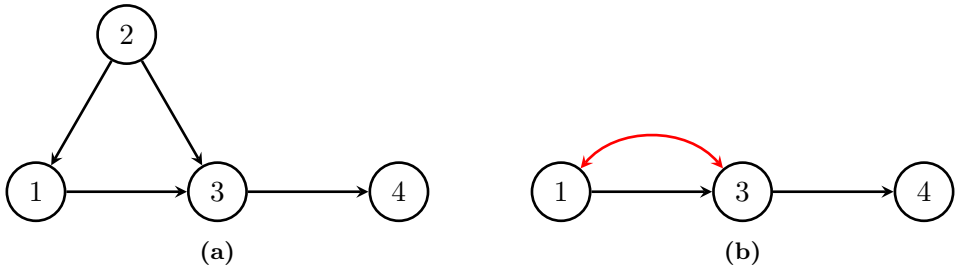



**Figure 2.1:** A directed graph and the corresponding mixed graph with unobserved $X_2$.

*But when $X_2$ is unobserved, we can define $\widetilde{\varepsilon}_1 = \lambda_{21}X_2 + \varepsilon_1$ and $\widetilde{\varepsilon}_3 = \lambda_{23}X_2 + \varepsilon_3$. The new equation system becomes*

$$\begin{aligned} X_1 &= \widetilde{\varepsilon}_1, \\ X_3 &= \lambda_{13}X_1 + \widetilde{\varepsilon}_3, \\ X_4 &= \lambda_{34}X_3 + \varepsilon_4. \end{aligned}$$

*Notice that the new errors are not independent anymore:* $\mathrm{Cov}[\tilde{\varepsilon}_1, \tilde{\varepsilon}_3] = \lambda_{21}\lambda_{23}\mathrm{Var}[\varepsilon_2]$.

In a mixed graph $G = (V, D, B)$, if $i \to j \in D$, we say that $i$ is a **parent** of $j$ and $j$ is a **child** of $i$. Introducing notation for the sets of parents or children, we denote this also by $i \in \mathrm{pa}(j)$ and $j \in \mathrm{ch}(i)$. Similarly, the notation $\mathrm{an}(i)$ denotes the set of **ancestors** of $i$, and $\mathrm{de}(i)$ denotes the set of **descendants** of $i$. For simplicity, we adopt the convention that $i \notin \mathrm{an}(i)$ and $i \in \mathrm{de}(i)$. If $i \leftrightarrow j \in B$ then $i(j)$ is a neighbor of $j(i)$: $i \in \mathrm{ne}(j)$ and $j \in \mathrm{ne}(i)$. Writing $\mathrm{ne}(i)$ for the set of all neighbors of a node $i$, it holds for directed graphs without bidirected edges that $\mathrm{ne}(i) = \mathrm{pa}(i) \cup \mathrm{ch}(i)$. When dealing with multiple graphs, we use subscripts to indicate the corresponding set in each graph. For example, $\mathrm{pa}_1(i)$ represents the parent set of node $i$ in $G_1$. In all those above cases if there exist an edge between nodes $i$ and $j$, we say that $i$ and $j$ are **adjacent**.

A **collider triple** in the mixed graph $G = (V, D, B)$ is a triple of nodes $(i, j, k)$ in which $i, j$ and $j, k$ are adjacent with $j$ being a head on both edges. In other words, the two edges form a path of the form $i \to j \leftarrow k$, $i \leftrightarrow j \leftarrow k$, $i \to j \leftrightarrow k$ or $i \leftrightarrow j \leftrightarrow k$. The middle node $j$ is a **collider**. When the nodes $i, k$ are not adjacent, we say that the collider is **unshielded**, otherwise it is **shielded**. The **skeleton** of $G$ is the undirected graph obtained by replacing all edges with undirected edges.

In a directed graph $G = (V, D)$, a **path** is an alternating sequence of nodes from $V$ and edges from $D$, such that each edge in the sequence is an edge between the nodes that precede and succeed it. A path can contain a node more than once. Given a fixed set $S \subseteq V$, two nodes $i, j \notin S$ are **d-connected** by $S$ if $G$ contains a path from $i$ to $j$ that has all colliders in $S$ and all non-colliders outside $S$. If it is not the case, we say that $i$ and $j$ are **d-separated** by $S$, which is also written as $i \perp_d j \mid S$. A **trek** is a path without collider triples, i.e., it can only take one of two forms:

$$i_l^L \leftarrow \cdots \leftarrow i_0^L \leftrightarrow i_0^R \to \cdots \to i_r^R,$$

which is possible only in mixed graphs with bidirected edges, or

$$i_l^L \leftarrow \cdots \leftarrow i_1^L \leftarrow i_0 \to i_1^R \to \cdots \to i_r^R,$$

where $L$ and $R$ superscripts correspond to left-hand side and right-hand side of the trek. In the second case $i_0$ is the **top node** and in both sides.

A directed graph is called **strongly connected** if for each node pair $(i, j)$, there exist a directed path from $i$ to $j$ and a directed path from $j$ to $i$. A **strongly connected component** of a directed graph $G$ is a subgraph that is strongly connected, and no additional edges or nodes can be added to the subgraph without breaking the strong connectedness. The strongly connected component containing $i$ in a directed graph $G$ is denoted by $\mathcal{C}(i, G)$. If there is no ambiguity with respect to the graph, it is also denoted by $\mathcal{C}(i)$.

For Gaussian errors, the random vector $X$ follows a multivariate Gaussian distribution, which is uniquely determined by the covariance matrix, once it has been centered. Every distribution in the model then corresponds to a covariance matrix, $\mathrm{Var}[X]$. We define the model as the set of covariance matrices can be generated from covariance map (2.2), and hence require the matrix $I - \Lambda$ to be invertible so that the linear SEM is well-defined. We list some notations and definitions below, which work for the general mixed graph setup.

Let $\mathbb{R}^D$ be the set of real $V \times V$ matrices $\Lambda = (\lambda_{ij})$ with support in $D$, i.e.,

$$\mathbb{R}^D := \{\Lambda \in V \times V : \ \Lambda_{ij} = 0 \text{ if } i \to j \notin D\}.$$

We also define $\mathbb{R}^D_{\mathrm{reg}}$ to be the subset of matrices $\Lambda \in \mathbb{R}^D$ for which $I - \Lambda$ is invertible. Let $PD$ be the cone of positive definite symmetric $V \times V$-matrices, and define $PD(B)$ to be the subcone with support over $B$, that is,

$$PD(B) = \left\{ \Omega = (\omega_{ij}) \in PD : \ \omega_{ij} = 0 \ \text{ if } \ i \neq j \text{ and } i \leftrightarrow j \notin B \right\}.$$

**Definition 2.2.** *The **linear Gaussian model** given by the mixed graph $G = (V, D, B)$ is the family of all multivariate normal distributions on $\mathbb{R}^V$ with covariance matrix in*

$$M_G = \left\{ (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1} : \ \Lambda \in \mathbb{R}^D_{\mathrm{reg}}, \, \Omega \in PD(B) \right\}.$$

*The **covariance parametrization** of the model is the map*

$$\begin{aligned} \phi_G : \mathbb{R}^D \times PD(B) &\mapsto PD, \\ (\Lambda, \Omega) &\mapsto (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}. \end{aligned} \tag{2.3}$$

A classical result known as the trek rule provides a combinatorial description of the coordinates of $\phi_G$ (see, e.g., Theorem 4.1 in the review of Drton [2018]).

**Theorem 2.3** (Trek rule)**.** *Let $G = (V, E)$ be a DAG, and let $\Lambda \in \mathbb{R}^E$ and $\boldsymbol{\omega} \in (0, \infty)^V$. For $i, j \in V$, let $\mathcal{T}(i, j)$ be the set of all treks between $i$ and $j$. For a trek $\tau$ with top node $i_0$, we define the trek monomial*

$$\tau(\Lambda, \boldsymbol{\omega}) = \boldsymbol{\omega}_{i_0} \prod_{k \to l \in \tau} \lambda_{kl}.$$

*Then the covariance between $X_i$ and $X_j$ equals the sum of all trek monomials for treks between $i$ and $j$, i.e.,*

$$\phi_G(\Lambda, \boldsymbol{\omega})_{ij} = \sum_{\tau \in \mathcal{T}(i,j)} \tau(\Lambda, \boldsymbol{\omega}), \quad i, j \in V.$$

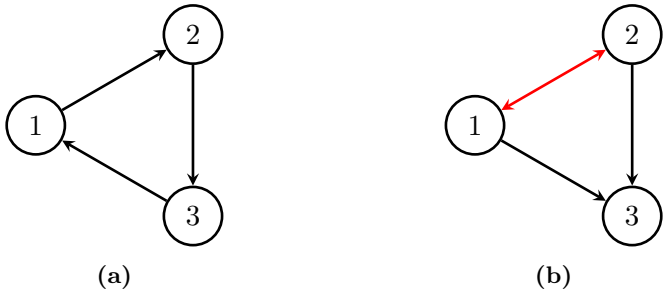



**Figure 2.2:** Two simple mixed graphs.

The trek rule gives an explicit description of the entries in the covariance matrix of the distribution. Another perspective is implicit: a conditional independence corresponds to an algebraic constraint on the covariance matrix, as stated in the following theorem (see [Drton, 2018, Section 10], [Richardson and Spirtes, 2002, Section 8]).

**Theorem 2.4.** *Let $G = (V, E)$ be a DAG. Let $i, j$ be two distinct nodes, and let $S \subseteq V \setminus \{i, j\}$.*

*(i) If $X$ is a multivariate normal random vector with covariance matrix $\Sigma$, then the conditional independence $X_i \perp\!\!\!\perp X_j \mid X_S$ holds if and only if $\det(\Sigma_{iS,jS}) = 0$.*

*(ii) The conditional independence constraint $\det(\Sigma_{iS,jS}) = 0$ holds for all covariance matrices $\Sigma \in M_G$ if and only if the d-separation $i \perp_d j \mid S$ holds in $G$.*

*(iii) A matrix $\Sigma \in PD$ is in $M_G$ if and only if $\det(\Sigma_{iS,jS}) = 0$ for all triples $(i, j, S)$ with $i \perp_d j \mid S$ in $G$.*

Analogous to the famous Markov equivalence theory for DAGs, different mixed graphs can induce the same model. We use the general terminology "distributional equivalence", which means the covariance parametrization of two graphs give exact the same region. This makes the model well-defined without structural identifiability issue. Markov equivalence only requires the same conditional independence relations, which is usually weaker but the same as distribution equivalence for DAGs.

**Definition 2.5.** *Two mixed graphs $G_1$ and $G_2$ are distributionally equivalent if $M_{G_1} = M_{G_2}$.*

## 2.2 Brief literature review

The research on structural equation models dates back to Wright's path diagrams [Wright, 1921, 1934], and Haavelmo's simultaneous equations [Haavelmo, 1943]. Much more recently, structural equation models are summarized into a general causal modeling framework and related to formalization of causal effects of experimental interventions [Spirtes et al., 2000, Pearl, 2009].

A large amount of works focus on DAG models [Koller and Friedman, 2009, Lauritzen, 1996, Maathuis et al., 2019], while the topic of feedback loops is much less popular. The study of cyclic graphs started from conditional independence relations and $d$-separations [Spirtes, 1995, Richardson, 1996b,a], which is a natural extension

of Markov equivalence theory in DAGs. Recent progresses of this approach lie in the assumption of strong dependence in cycles and the new concept of $\sigma$-seperation. But the new separation rule is not applicable in Gaussian models; see, e.g., Bongers et al. [2021], Forré and Mooij [2018], Mooij and Claassen [2020]. Algebraic methods can be applied to graphical model analysis, for both DAGs and cyclic graphs. An overview of algebraic methods and issues in linear Gaussian SEMs is presented in Drton [2018]. More results can be found in Sullivant [2018].

Structure learning from (observational) data is a fundamental problem in the area of graphical models [Drton and Maathuis, 2017]. Most of structure learning algorithms fall into the two catagory: score-based methods and constraint-based methods. A score-based method assigns a score for each graph and search the graph with highest score, which usually need a parameterized form of SEM; see, e.g., van de Geer and Bühlmann [2013], Chickering [2002, 2003], Solus et al. [2021]. A constraint-based method exploits conditional independence relations for learning the structure and the needed independence tests can be either parametric or nonparametric; see, e.g., Rantanen et al. [2020], Hyttinen et al. [2014, 2012], Richardson [1996b], Forré and Mooij [2018]. Constraint-based methods (independent tests) are more flexible, while causal models with feedback loops or latent variables can generally not be characterized by conditional independence constraints alone [Drton et al., 2020, van Ommen and Mooij, 2017]. There are also some hybrid methods, which combine score-based and constraint-based methods by searching over a restricted space obtained from conditional independence relations. Score-based method for structure learning of bow-free acyclic graphs is proposed in Nowzohour et al. [2017]. Inevitably, there are issues about identifiability, model equivalence and model dimension for cyclic graphical models. The general equivalence properties and even determining the model dimension constitute problems that are not fully understood.

## 2.3 Organization of the Thesis

The reminder of this thesis is structured as follows. Chapter 3 shows that simple cyclic mixed models are of expected dimension. IT also generalizes the sufficient conditions for distributional equivalence that were given for bow-free acyclic mixed models in Nowzohour et al. [2017]. Chapter 4 discusses a block-coordinate descent scheme for computing the MLE in linear structural equation models when there are no hidden variables but one has access to a combination of multiple observational and interventional environments. The chapter exhibit formulas for block updates in tractable special cases. Chapter 5 is about structural identifibility of simple cyclic models under homoscedastic error assumption. Graphical criteria for distinguishing models are provided, and a computational study develops additional conjectures. Finally, Chapter 6 describes the DAG model equivalence results under groupwise homoscedastic errors.

# Chapter 3

# Structure Learning for Simple Mixed Cyclic Graphs

In structural equation modeling, mixed graphs are commonly used to represent dependencies induced by hidden confounders [Evans, 2019]. Here, a mixed graph is a graph that may feature two types of edges: directed edges and bidirected edges. The bidirected edges are included to indicate that errors in the structural equations may be correlated (due to hidden confounding), as we detailed in Chapter 2.

This chapter is based on a publication in the Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI) [Améndola et al., 2020]. This publication focuses on mixed graphs that are simple, i.e., they contain at most one edge between any node pair. This property has also been termed 'bow-free' in related literature.

The chapter presents a dimension theorem and a sufficient condition for distributional equivalence of simple mixed graphs. Furthermore, a structure learning method based on a greedy search with an extended Bayesian information criterion is proposed. The dimension theorem and the greedy search method were my primary contributions to the mentioned UAI paper.

## 3.1 Models defined by simple mixed graphs

Taking up the definitions from Chapter 2, the model is given by the covariance parameterization (Definition 2.2). The dimension of the model, i.e., the set of its covariance matrices, equals the maximal rank of the Jacobian of the parametrization map [Geiger et al., 2001]. This maximal rank can be at most the number of free parameters $|V| + |D| + |B|$. Tight equality with this parameter count is not always achieved for general mixed graphs, but as we show now simple mixed graphs are special in this respect. This is the case even when the mixed graph contains directed cycles.

Let $J_G$ be the Jacobian of the covariance parametrization $\phi_G$. The map $g : (\Lambda, \Sigma) \mapsto (I - \Lambda)^T \Sigma (I - \Lambda)$ on $\mathbb{R}^D_{\text{reg}} \times PD$ computes $\Omega$ from $(\Lambda, \Sigma)$. A positive definite matrix $\Sigma \in M_G$ if and only if $\exists (\Lambda, \Omega) \in \mathbb{R}^D_{\text{reg}} \times PD(B)$ s.t. $\Sigma = \phi_G(\Lambda, \Omega)$, which is again equivalent to $g(\Lambda, \Sigma) = \Omega$. We can rewrite the latter conditions equivalently as

$$
\begin{aligned}
&g_{ij}(\Lambda, \Sigma) = \left[(I - \Lambda)^T \Sigma (I - \Lambda)\right]_{ij} = 0, \\
&\forall \{i, j\} \in N := \{\{i, j\} : i, j \in V,\ i \neq j,\ \{i, j\} \notin B\}.
\end{aligned}
$$

Consider now the $N \times D$ Jacobian matrix $\mathbf{J}(\Lambda, \Sigma)$ whose entries are the partial derivatives

$$\mathbf{J}(\Lambda, \Sigma)_{\{i,j\},(k,l)} = \frac{\partial g_{ij}(\Lambda, \Sigma)}{\partial \lambda_{kl}} \tag{3.1}$$

with $\{i, j\} \in N$ and $k \to l \in D$. For $i \neq j$, $g_{ij}$ is multilinear in $\Lambda$ and

$$\frac{\partial g_{ij}(\Lambda, \Sigma)}{\partial \lambda_{kl}} = \begin{cases} -[(I - \Lambda)^T \Sigma]_{jk} & \text{if } l = i, \\ -[(I - \Lambda)^T \Sigma]_{ik} & \text{if } l = j, \\ 0 & \text{if } l \notin \{i, j\}. \end{cases} \tag{3.2}$$

The following lemma relates the rank of $J_G$ and the rank of $\mathbf{J}$.

**Lemma 3.1.** *For $\Lambda \in \mathbb{R}^D_{\text{reg}}$, $\Omega \in PD(B)$, let $\Sigma = \phi_G(\Lambda, \Omega)$. Then the rank of the Jacobian $J_G(\Lambda, \Omega)$ is equal to*

$$\text{rank}\left(\mathbf{J}\left(\Lambda, \Sigma\right)\right) + |B| + |V|.$$

*Proof.* On $\mathbb{R}^D_{\text{reg}} \times PD(B)$, define the map

$$h : (\Lambda, \Omega) \mapsto (\Lambda, \phi_G(\Lambda, \Omega)).$$

Composing with $g$ we have

$$(g \circ h)(\Lambda, \Omega) = \Omega. \tag{3.3}$$

Differentiating this equation with respect to the free entries (i.e., nonzero) in $\Lambda$ and $\Omega$ gives

$$\frac{\partial}{\partial \Lambda} g(\Lambda, \Sigma)\Big|_{\Sigma=\phi_G(\Lambda,\Omega)} + \frac{\partial}{\partial \Sigma} g(\Lambda, \Sigma)\Big|_{\Sigma=\phi_G(\Lambda,\Omega)} \frac{\partial}{\partial \Lambda} \phi_G(\Lambda, \Omega) \;=\; O, \tag{3.4}$$

$$\frac{\partial}{\partial \Sigma} g(\Lambda, \Sigma)\Big|_{\Sigma=\phi_G(\Lambda,\Omega)} \frac{\partial}{\partial \Omega} \phi_G(\Lambda, \Omega) \;=\; \begin{pmatrix} O \\ I_{|B|+|V|} \end{pmatrix}, \tag{3.5}$$

where the rows are indexed by unordered pairs $\{i, j\}$. In the partitioning of the rows, the pairs in $N$ are listed first.

The entry-wise equation (3.2) can be written as

$$\frac{\partial}{\partial \Lambda} g(\Lambda, \Sigma) = \begin{pmatrix} \mathbf{J}(\Lambda, \Sigma) \\ O \end{pmatrix}$$

with the same ordering of rows.

From (3.4) and (3.5) we can obtain that

$$\frac{\partial}{\partial \Sigma} g(\Lambda, \Sigma)|_{\Sigma=\phi_G(\Lambda,\Omega)} \cdot J_G(\Lambda, \Omega) = \begin{pmatrix} -\mathbf{J}(\Lambda, \Sigma)|_{\Sigma=\phi(\Lambda,\Omega)} & O \\ O & I_{|B|+|V|} \end{pmatrix}, \tag{3.6}$$

with rows and columns partitioned as $(N, B \cup V)$ and $(D, B \cup V)$, respectively. We restrict $g$ by fixing $\Lambda \in \mathbb{R}^D_{\text{reg}}$ gives the bijection $\Sigma \mapsto (I-\Lambda)^T \Sigma (I-\Lambda)$, consequently the matrix $\frac{\partial g}{\partial \Sigma}$ is invertible. And hence the rank of $J_G$ equals the rank of the partitioned matrix on the right-hand side, which is $\text{rank}\left(\mathbf{J}\left(\Lambda, \Sigma\right)\right) + |B| + |V|$. □

The maximal rank of $J_G$ can be achieved at $(\Lambda, \Omega) = (O, I)$. Indeed, we have the following lemma which links the Jacobian rank and the graph property.

**Lemma 3.2.** *The Jacobian $J_G(O, I)$ has full column rank $|V| + |B| + |D|$ if and only if the mixed graph $G$ is simple.*

*Proof.* It suffices to show that the matrix $\mathbf{J}(O, I)$ defined in (3.2) has full column rank $|D|$ if and only if $G$ is simple.

First we suppose that $G$ is simple, in this case $k \to l \in D$ implies $k \leftrightarrow l \in N$. When $(\Lambda, \Omega) = (O, I)$ we have $\Sigma = \phi_G(O, I) = I$ and $(I - \Lambda)^T \Sigma = I$. From (3.2) we know that the only nonzero entry in column indexed by $(k, l)$ is

$$\mathbf{J}(O, I)_{\{k,l\},(k,l)} = -1.$$

Rearranging the row indices such that all $\{k, l\}$ pairs of $k \to l \in D$ first, the matrix form becomes

$$\mathbf{J}(O, I) = \begin{pmatrix} -I_D \\ O \end{pmatrix},$$

which obviously has rank $|D|$.

If $G$ is not simple, we can assume that $k$ and $l$ are connected by two edges. Without loss of generality, we further assume that $k \to l \in D$. If $k \leftrightarrow l \in B$, then $\{k, l\} \notin N$ and the column of $\mathbf{J}(O, I)$ indexed by $(k, l)$ is zero (the only nonzero entry $\{k, l\}, (k, l)$ is missing), and $\operatorname{rank}(\mathbf{J}(O, I)) < |D|$. Otherwise $k \leftrightarrow l \notin B$ and $l \to k \in D$, the two columns of $\mathbf{J}(O, I)$ indexed by $(k, l)$ and $(l, k)$ are identical and we also have $\operatorname{rank}(\mathbf{J}(O, I)) < |D|$. □

We arrive at the main result of this section, which clarifies that models given by simple mixed graphs are of expected dimension.

**Theorem 3.3.** *If the graph $G$ is simple, then*

$$\dim(M_G) = |V| + |D| + |B|.$$

*Proof.* Directly from Lemma 3.2. □

## 3.2 Sufficient Conditions for Distributional Equivalence

In this section, we show that the sufficient condition for distributional equivalence from Nowzohour et al. [2017] admits an extension to our setting of possibly cyclic graphs. To this end, we define $\overline{M_G}$ to be the closure of $M_G$ (in Euclidean topology). Two mixed graphs $G_1$ and $G_2$ are then distributionally equivalent up to closure if $\overline{M_{G_1}} = \overline{M_{G_2}}$.

**Theorem 3.4.** *Let $G_1$ and $G_2$ be two simple mixed graphs with same skeleton and collider triples. Then $G_1$ and $G_2$ are distributionally equivalent up to closure.*

Note that the likelihood functions of two models that are equal up to closure have the same supremum.

While our proof of Theorem 3.4 (developed in §3.2.1-3.2.2) concludes equality up to closure, we do not know any examples where the models are not exactly equal. The condition in Theorem 3.4 is also far from being necessary, e.g., it does not include the well-known characterization of Markov equivalence for directed acyclic graphs (DAGs) as a special case. However, the theorem is useful to assert equivalence in our simulations (Section 3.4.1). We are also not aware of better (tractable) conditions in the literature. Indeed, distributional equivalence for cyclic mixed graphs is a subtle problem as the following example shows.

**Example 3.5.** *Let $G_1$ and $G_2$ be the two simple mixed graphs displayed in Figure 2.2(a) and (b), respectively. By Theorem 3.3, both $M_{G_1}$ and $M_{G_2}$ are full-dimensional (i.e., 6-dimensional) subsets of the cone of positive definite $3\times3$ matrices. Graph $G_2$ is acyclic, and $M_{G_2}$ is easily seen to be equal to $PD$. However, as observed in Drton et al. [2019a], the set $M_{G_1}$ is a strict subset of $M_{G_2} = PD$.*

### 3.2.1 Useful Lemmas

Let $G_1 = (V, D_1, B_1)$ and $G_2 = (V, D_2, B_2)$ be two mixed graphs. Let $(\Lambda_1, \Omega_1) \in \mathbb{R}^{D_1}_{\text{reg}} \times PD(B_1)$ be parameters for $G_1$. The essence of the proof of Theorem 3.4 is a strategy to find parameters $(\Lambda_2, \Omega_2) \in \mathbb{R}^{D_2}_{\text{reg}} \times PD(B_2)$ such that $\phi_{G_2}(\Lambda_2, \Omega_2) = \phi_{G_1}(\Lambda_1, \Omega_1)$. The key steps of the construction are a reduction to correlation matrices and an edge-relabeling considered in the acyclic case by Nowzohour et al. [2017]. However, the cyclic case brings about new subtleties in this approach.

Let $\mathcal{R} : PD \to PD$ be the standardization map that takes covariance matrices to correlation matrices via $\mathcal{R}(\Sigma)_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$.

**Lemma 3.6.** *Let $G = (V, D, B)$ be simple and $\Sigma \in PD$. Then*

$$\Sigma \in M_G \text{ if and only if } \mathcal{R}(\Sigma) \in M_G.$$

*Proof.* We show one direction as the converse can be verified similarly. If $\Sigma \in M_G$ then

$$\Sigma = \phi_G(\Lambda, \Omega) = (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1}.$$

for some matrices $\Lambda \in \mathbb{R}^D_{\text{reg}}, \Omega \in PD(B)$. Setting $\Delta$ diagonal with entries $\Delta_{ii} = \Sigma_{ii}^{-\frac{1}{2}}$ it holds that

$$\begin{aligned}\mathcal{R}(\Sigma) &= \Delta\Sigma\Delta \\ &= (\Delta^{-1} - \Delta^{-1}\Lambda)^{-T}\Omega(\Delta^{-1} - \Delta^{-1}\Lambda)^{-1} \\ &= \phi_G(\tilde{\Lambda}, \tilde{\Omega})\end{aligned}$$

with $\tilde{\Lambda} = \Delta^{-1}\Lambda\Delta \in \mathbb{R}^D_{\text{reg}}$, $\tilde{\Omega} = \Delta\Omega\Delta \in PD(B)$. □

Throughout the rest of this section, let $G_1 = (V, D_1, B_1)$ and $G_2 = (V, D_2, B_2)$ be two mixed graphs. If the graphs have the same skeleton, then there is a natural way to copy the edge labels from one graph to the other. To describe the procedure, we

decompose an error covariance matrix, $\Omega$, into its diagonal and off-diagonal parts, denoted $\Omega^d$ and $\Omega^{od}$, respectively. So, $\Omega = \Omega^d + \Omega^{od}$.

**Definition 3.7.** *Let $G_1$ and $G_2$ be simple mixed graphs with the same skeleton. Given a choice $(\Lambda_1, \Omega_1) \in \mathbb{R}^{D_1}_{\text{reg}} \times PD(B_1)$, the* induced edge labeling on $G_2$ *is the pair of matrices $(\Lambda_2, \Omega_2^{od})$ obtained as*

$$(\Lambda_2)_{ij} = \begin{cases} (\Lambda_1)_{ij} & \text{if } i \to j \in G_1,\ i \to j \in G_2, \\ (\Lambda_1)_{ji} & \text{if } i \leftarrow j \in G_1,\ i \to j \in G_2, \\ (\Omega_1)_{ij} & \text{if } i \leftrightarrow j \in G_1,\ i \to j \in G_2, \\ 0 & \text{if } i \to j \notin G_2, \end{cases}$$

$$(\Omega_2^{od})_{ij} = \begin{cases} (\Lambda_1)_{ij} & \text{if } i \to j \in G_1, i \leftrightarrow j \in G_2, \\ (\Lambda_1)_{ji} & \text{if } i \leftarrow j \in G_1, i \leftrightarrow j \in G_2, \\ (\Omega_1)_{ij} & \text{if } i \leftrightarrow j \in G_1, i \leftrightarrow j \in G_2, \\ 0 & \text{if } i \leftrightarrow j \notin G_2 \text{ or } i = j. \end{cases}$$

For the construction from Definition 3.7, it holds that $\Lambda_2 \in \mathbb{R}^{D_2}_{\text{reg}}$. Moreover, $\Omega_2^{od}$ can be turned into a matrix in $PD(B_2)$ by addition of a diagonal matrix.

**Lemma 3.8.** *Let $G_1$ and $G_2$ be simple mixed graphs with same skeleton and collider triples, and let $(\Lambda_i, \Omega_i) \in \mathbb{R}^{D_i} \times PD(B_i)$ for $i = 1, 2$. If $(\Lambda_2, \Omega_2^{od})$ equals the edge labeling induced by $(\Lambda_1, \Omega_1)$ then*

$$det(I - \Lambda_1) = det(I - \Lambda_2).$$

*In particular, if $\Lambda_1 \in \mathbb{R}^{D_1}_{reg}$ then $\Lambda_2 \in \mathbb{R}^{D_2}_{reg}$.*

*Proof.* The determinants depend on the values of cycle products [Drton et al., 2019b, Lemma 1]. Let $S_V$ be the group of permutations of the nodes in $V$. For $\sigma \in S_V$, let $V(\sigma)$ be the set of nodes contained in a nontrivial cycle of $\sigma$. Then

$$\det(I - \Lambda) = \sum_{\sigma \in S_V(G)} (-1)^{\text{sgn}(\sigma)} \prod_{i \in V(\sigma)} \Lambda_{\sigma(i),i} \tag{3.7}$$

where $S_V(G)$ is the subset of permutations such that $i = \sigma(i)$ or $i \to \sigma(i) \in D$ for all $i \in V$. We remark that even though the lemma in Drton et al. [2019b] is stated for $\Lambda \in \mathbb{R}^D_{\text{reg}}$, the proof relies on Laplace expansion of the determinant which holds even if $I - \Lambda$ is not invertible.

Now, since collider triples are preserved, an edge that is part of a directed cycle of $G_1$ cannot be bidirected in $G_2$. Furthermore, if $G_1$ contains a cycle which has a directed edge that is reversed in $G_2$, then the cycle must be chordless in $G_1$ (that is, every node in the cycle can have only one child in the cycle) and must be fully reversed in $G_2$. Since the labels agree, the cycle products in (3.7) remain unchanged and therefore $\det(I - \Lambda_1) = \det(I - \Lambda_2)$. □

### 3.2.2 Constructing Covariance Matrices

The key to completing the proof of Theorem 3.4 is to show that a correlation matrix obtained from a generic choice of parameters $(\Lambda_1, \Omega_1) \in \mathbb{R}^{D_1}_{\text{reg}} \times PD(B_1)$ also belongs to $M_{G_2}$. Let $\circ$ denote the Hadamard (entrywise) product of matrices, and define $\mathcal{H} : \mathbb{R}^{D}_{\text{reg}} \to \mathbb{R}^{V \times V}$ by

$$\mathcal{H}(\Lambda) := (I - \Lambda)^{-1} \circ (I - \Lambda)^{-1}.$$

We denote the spectral radius of a matrix $\Lambda$ by $\rho(\Lambda)$.

**Lemma 3.9.** *Let $G_1, G_2$ be simple mixed graphs with same skeleton and collider triples. Let $(\Lambda_1, \Omega_1) \in \mathbb{R}^{D_1}_{reg} \times PD(B_1)$ such that $\Sigma = \phi_{G_1}(\Lambda_1, \Omega_1) \in M_{G_1}$ is a correlation matrix and consider the induced edge labeling $(\Lambda_2, \Omega_2^{od})$. If*

*(i) $\rho(\Lambda_j) < 1$ for $j = 1, 2$, and*

*(ii) $\det(\mathcal{H}(\Lambda_2)) \neq 0$,*

*then there exists a unique diagonal matrix $\Omega_2^d$ such that with $\Omega_2 = \Omega_2^d + \Omega_2^{od}$ it holds that $(\Lambda_2, \Omega_2) \in \mathbb{R}^{D_2}_{reg} \times PD(B_2)$ and $\Sigma = \phi_{G_2}(\Lambda_2, \Omega_2) \in M_{G_2}$.*

*Proof.* By Lemma 3.8, we have indeed that $\Lambda_2 \in \mathbb{R}^{D_2}_{\text{reg}}$. We need to construct $\Omega_2^d$ such that

$$\phi_{G_2}(\Lambda_2, \Omega_2) = (I - \Lambda_2)^{-T} \Omega_2^d (I - \Lambda_2)^{-1} + (I - \Lambda_2)^{-T} \Omega_2^{od} (I - \Lambda_2)^{-1} = \Sigma.$$

Since $\Sigma_{ii} = 1$, this requires for all $i \in V$ that

$$((I - \Lambda_2)^{-T} \Omega_2^d (I - \Lambda_2)^{-1})_{ii} = 1 - ((I - \Lambda_2)^{-T} \Omega_2^{od} (I - \Lambda_2)^{-1})_{ii}.$$

Solving for the diagonal of $\Omega_2^d$ is equivalent (see [Horn and Johnson, 1991, Lemma 5.1.3] ) to the linear system $Ax = b$ where

$$A = \mathcal{H}(\Lambda_2) = (I - \Lambda_2)^{-1} \circ (I - \Lambda_2)^{-1}$$

and the coordinates of the vector $b$ are

$$b_i = 1 - ((I - \Lambda_2)^{-T} \Omega_2^{od} (I - \Lambda_2)^{-1})_{ii}.$$

By hypothesis, $\det(\mathcal{H}(\Lambda_2)) \neq 0$ and the system has a unique solution. It thus remains to show that $\phi_{G_2}(\Lambda_2, \Omega_2)$ also matches $\Sigma$ in all off-diagonal entries.

In general, if $\phi(\Lambda, \Omega)$ is a correlation matrix over a mixed graph $G$ and $\rho(\Lambda) < 1$, by [Nowzohour et al., 2017, Theorem 4], the entries for $i \neq j$ are given by

$$\phi_G(\Lambda, \Omega)_{ij} = \sum_{\tau \in S^{ij}} \prod_{s \to t \in \tau} \Lambda_{ts} \prod_{s \leftrightarrow t \in \tau} \Omega_{st}, \tag{3.8}$$

where $S_G^{ij}$ is the set of simple treks from $i$ to $j$. By assumption, $\rho(\Lambda_1), \rho(\Lambda_2) < 1$, and we may apply the representation in (3.8) to $G_1$ and $G_2$. In general, $S_{G_1}^{ij} \neq S_{G_2}^{ij}$. However, the fact that the graphs have the same skeleton and share collider triples implies that when replacing $(\Lambda, \Omega)$ by $(\Lambda_j, \Omega_j)$, $j = 1, 2$, in (3.8), the induced edge labeling guarantees that the right hand sides of the expression are equal. Hence,

$$\phi_{G_1}(\Lambda_1, \Omega_1) = \Sigma = \phi_{G_2}(\Lambda_2, \Omega_2)$$

as was the claim. □

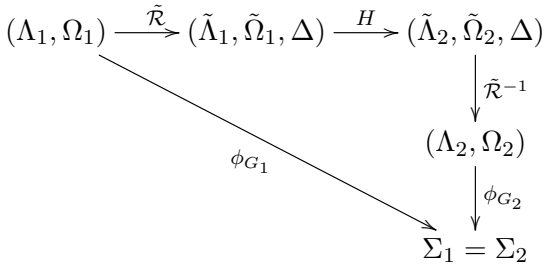



**Figure 3.1:** Commutative diagram illustrating two ways of obtaining a matrix $\Sigma \in M_{G_1} \cap M_{G_2}$. One is the parametrization $\phi_{G_1}$, while the other is a composition of maps (including the map $H$ defined via Lemma 3.9), that we denote $\Psi$.

With these preparations in place, we may complete the proof of the main result of this section.

*Proof of Theorem 3.4.* First, observe that the covariance parametrization

$$\phi_{G_1} : \mathbb{R}^{D_1} \times PD(B_1) \to M_{G_1} \subseteq PD$$

is a rational map. Next, consider the algebraic map

$$\Psi : \mathcal{U} \subset \mathbb{R}^{D_1} \times PD(B_1) \to M_{G_2} \subseteq PD$$

defined as follows. First, apply the standardization map on the parameter $(\Lambda_1, \Omega_1)$ to obtain $(\tilde{\Lambda}_1, \tilde{\Omega}_1, \Delta)$, as in the proof of Lemma 3.6. We denote this map by $\tilde{\mathcal{R}}$. As $(\tilde{\Lambda}_1, \tilde{\Omega}_1)$ define a correlation matrix we may obtain $(\tilde{\Lambda}_2, \tilde{\Omega}_2)$ from the procedure in Lemma 3.9, for representation of the same correlation matrix. Finally, destandardize $(\tilde{\Lambda}_2, \tilde{\Omega}_2)$ with the matrix $\Delta$ from the standardization map, and apply $\phi_{G_2}$.

Note that the map $\Psi$ is well-defined for input that satisfies the two conditions in Lemma 3.9. This domain includes an open subset $\mathcal{U} \subset \mathbb{R}^{D_1} \times PD(B_1)$. This subset is nonempty because $(0, I) \in \mathcal{U}$. The final application of $\phi_{G_2}$ to $(\Lambda_2, \Omega_2)$ gives a matrix in $M_{G_2}$, which by construction and Lemma 3.9 coincides with $\phi_{G_1}(\Lambda_1, \Omega_1)$. The diagram in Figure 3.1 illustrates the situation.

The map $\Psi$ is a composition of a rational map with algebraic maps that involve radicals (i.e., square roots in the standardization $\mathcal{R}$). Since $\Psi$ coincides with the rational map $\phi_{G_1}$ on the open set $\mathcal{U}$, they must be equal outside of an algebraic hypersurface (i.e., the zero set of a multivariate polynomial). This exceptional set has Lebesgue measure zero (see, e.g., the lemma in Okamoto [1973]). Covariance matrices in $M_{G_1}$ that are given by parameters $(\Lambda_1, \Omega_1)$ outside the exceptional set are also in $M_{G_2}$. We may conclude that $M_{G_1} \subseteq \overline{M_{G_2}}$ because the elements of the exceptional set are limits of sequences off the exceptional set. By symmetry, $\overline{M_{G_1}} = \overline{M_{G_2}}$ as claimed. □

## 3.3 Greedy Search

Since we know the exact dimension of models given by simple mixed graphs, we can apply model selection criteria that balance model complexity (dimension) and

model fit. We assign a score to each graph and want to find the graph with maximal score. Given the exponentially-growing large number of possible graphs, we follow prior work and consider a greedy search scheme, which starts from some random, or possibly also the empty graph and selects the DAG with highest score in the local neighborhood at each step. The procedure stops when local maximum score or a fixed maximum number of iterations is reached. To alleviate issues of local optima, we perform the greedy search multiple times starting from different initial graphs. The neighborhood of a graph $G$ is defined to be all simple mixed graphs that can be obtained from $G$ by one edge addition or one edge deletion, or one edge reversal [Nowzohour et al., 2017].

We denote the data matrix by $\mathbf{X} \in \mathbb{R}^{p\times n}$, in which each row represents one observation and is centered. Let $S = \mathbf{X}\mathbf{X}^T/n$ be the sample covariance matrix. The Gaussian log-likelihood function is

$$\ell(\Sigma; S) = -\frac{n}{2}\left[\log\det(2\pi\Sigma) + \operatorname{tr}(\Sigma^{-1}S)\right]. \tag{3.9}$$

The score of a graph $G$ takes the form

$$s(G) = \frac{1}{n}\left(\max_{\Sigma\in M_G} \ell(\Sigma; S) - \text{penalty}(p, k, n)\right), \tag{3.10}$$

where $p = |V|$ and $k = |D| + |B|$ is the number of edges. To compute the maximum log-likelihood in (3.10) we apply the block coordinate-descent algorithm from Drton et al. [2019b]. We will discuss an extension of the algorithm in next chapter. As for the penalty function, the standard Bayesian information criterion (BIC) takes $\text{penalty}(p, k, n) = \frac{1}{2}(p + k)\log n$, where $p + k$ is the model dimension. The extended Bayesian information criterion (eBIC) [Chen and Chen, 2008, Foygel and Drton, 2010] is induced from a prior distribution over graphs under which the number of edges is uniformly distributed. In our setup the penalty becomes

$$\text{penalty}(p, k, n) = \tfrac{1}{2}(p + k)\log(n) + \log(p^{2k}3^k), \tag{3.11}$$

where the last term reflects that there are $\binom{p(p+1)/2}{k}3^k \sim p^{2k}3^k$ simple mixed graphs with $k$ edges. This penalty tends to select sparser graph than than standard BIC penalty.

We remark that it would be desirable to additionally account for distributional equivalence and consider priors over equivalence classes of graphs. However, at this point, we do not have enough theoretical insights into distributional equivalence to make such an approach practical.

## 3.4 Numerical Experiments

Extending the work of Nowzohour et al. [2017] and Drton et al. [2019b], we implement the proposed greedy search scheme in R [R Core Team, 2020]. For illustration we apply the algorithm to simulated data and the well-known Sachs protein expression data [Sachs et al., 2005].

### 3.4.1 Simulation Studies

We consider graphs with $p \in \{5, 6\}$ nodes. For each setup, 100 simple mixed graphs are drawn uniformly at random by MCMC algorithm [Nowzohour et al., 2017]. The parameters $\Lambda_{ij}$'s and $\Omega_{ij}$'s are sampled uniformly from $[-0.9, -0.5] \cup [0.5, 0.9]$. The diagonal entries of $\Omega$ are set as the sum of the absolute values of the entries in the row in $\Omega$ plus an independent $\chi^2_1$ random draw, such that $\Omega$ is diagonal dominant and hence postive definite. For each graph, we generate three Gaussian data sets of size $n \in \{10^2, 10^3, 10^4\}$. For each realization of the greedy search, we restart with 300 random graphs and also compare the result of starting from the true graph. We consider both standard BIC and extended BIC score. The maximum number of iterations of greedy search is $10^4$.

| BIC | n | Start | Dim | Skel | Skel & Coll | SHD* |
|---|---|---|---|---|---|---|
| BIC | $10^2$ | **R** | **0.39** | **0.13** | **0.07** | **3.79** |
| | | TG | 0.8 | 0.8 | 0.25 | 1.15 |
| | $10^3$ | **R** | **0.63** | **0.43** | **0.26** | **2.44** |
| | | TG | 0.88 | 0.88 | 0.53 | 0.63 |
| | $10^4$ | **R** | **0.76** | **0.59** | **0.45** | **2.29** |
| | | TG | 0.92 | 0.92 | 0.74 | 0.34 |
| eBIC | $10^2$ | **R** | **0.24** | **0.14** | **0.1** | **3.55** |
| | | TG | 0.92 | 0.92 | 0.36 | 1.03 |
| | $10^3$ | **R** | **0.48** | **0.34** | **0.21** | **2.78** |
| | | TG | 0.9 | 0.9 | 0.52 | 0.65 |
| | $10^4$ | **R** | **0.71** | **0.61** | **0.38** | **2.02** |
| | | TG | 0.93 | 0.93 | 0.71 | 0.42 |

**Table 3.1:** Proportion of estimated graphs that share the dimension (Dim), skeleton (Skel) and both skeleton and set of collider triples (Skel & Coll) with the true graph, and minimal structural hamming distance (SHD*) averaged over simulations. Estimates use BIC with standard (1) and increased penalty (2), and search initialized at random (R) or at the true graph (TG).

Table 3.1 shows the frequencies of obtaining a model of correct dimension, correct graph skeleton, and both correct collider triples and skeleton when $p = 5$. The frequency of having the same dimension gives an upper bound of the frequency of getting equivalent model, while the frequency of having the same skeleton and collider triples gives a lower bound. According to the tables, the standard BIC slightly outperforms the extended BIC when $p \in \{5, 6\}$. Indeed, for small $p$'s like these the graphs are not really sparse.

We also record the structural hamming distance (SHD), which counts the number of edge addition, edge deletion and edge reversal needed to transfrom one graph to another. Here the SHD* is the minimum SHD over pairs $(\overline{G}_1, \overline{G}_2)$ such that $\overline{G}_1$ has the same skeleton and collider triples as $G_1$ and $\overline{G}_2$ has the same skeleton and collider triples as $G_2$. We actually apply the sufficeint condition in Theorem 3.4 and obtain an upper bound of the minimum SHD between two equivalence classes.

Further, in Table 3.2 we present the difference between the dimensions of the true graph and of the estimated graph, with standard BIC score.

| n | Start | Dim(EST) - Dim(TG) | | | | | |
|---|---|---|---|---|---|---|---|
| | | **-3** | **-2** | **-1** | **0** | **1** | **2** |
| $10^2$ | **R** | **5** | **14** | **35** | **39** | **7** | **0** |
| | TG | 0 | 0 | 0 | 80 | 20 | 0 |
| $10^3$ | **R** | **1** | **3** | **25** | **63** | **8** | **0** |
| | TG | 0 | 0 | 0 | 88 | 12 | 0 |
| $10^4$ | **R** | **0** | **2** | **14** | **76** | **8** | **0** |
| | TG | 0 | 0 | 0 | 97 | 7 | 1 |

**Table 3.2:** Absolute frequency distribution of the difference between the dimension of the true graph (TG) and the dimension of the estimated graph (EST) in 100 simulations for p=5 and BIC with standard penalty.

### 3.4.2 Protein Expression Data

We apply our procedure to well-known protein expression data, namely, a collection of 14 data sets on expression of $p = 11$ proteins in human T-cells [Sachs et al., 2005]. Each data set is obtained under different experimental conditions (interventions), and the sample sizes range from 727 to 907. Figure 2 in Sachs et al. [2005] shows the conventionally accepted signaling molecule interactions. There are feedback loops and hidden variables.

To apply our linear Gaussian model to the data with unknown (very likely non-Gaussian) distribution, we consider a simple extension that accounts for marginal non-Gaussianity: The nonparanormal / Gaussian copula models of Liu et al. [2012], Harris and Drton [2013]. The main point is to replace the random vector $X = (X_1, \ldots, X_p)$ by the transformed version $f(X) = (f_1(X_1), \ldots, f_p(X_p))$, where each $f_i$ is a monotone univariate function and $f(X)$ is multivariate Gaussian. The linear structural equations then model the relations among standardized versions of $f_i(X_i)$'s. The bias-corrected Kendall's tau correlation matrix $\sin(\frac{\pi}{2}\hat{\tau}_{ij})$ is a consistent approximation of the correlation matrix of $f(X)$, where $\hat{\tau}_{ij}$ is Kendall's $\tau$ for the pair $(X_i, X_j)$ [Harris and Drton, 2013]. The standardization projects the original covariance matrix space of $\frac{p(p+1)}{2}$ dimension to the $\frac{p(p-1)}{2}$-dimensional correlation matrix space. As long as the graph has no more than $\frac{p(p-3)}{2}$ edges, the projection keeps the expected dimension.

We perform the greedy search on each dataset with 100 restarts from random graphs and choose the extended BIC score. Table 3.3 gives the counts of total number of edges, directed edges and bidirected edges among the 14 estimated graphs. Four of the 14 graphs have a directed cycle: the graphs for datasets 3, 6, 7 each contain a 3-cycle and that of dataset 4 contains a 4-cycle.

| Type of edges | Min | Median | Max |
|---|---|---|---|
| All | 10 | 13 | 13 |
| Directed | 2 | 8 | 11 |
| Bidirected | 2 | 4.5 | 8 |

**Table 3.3:** Summary statistics on the number of edges in the estimated graphs for the 14 protein expression datasets.

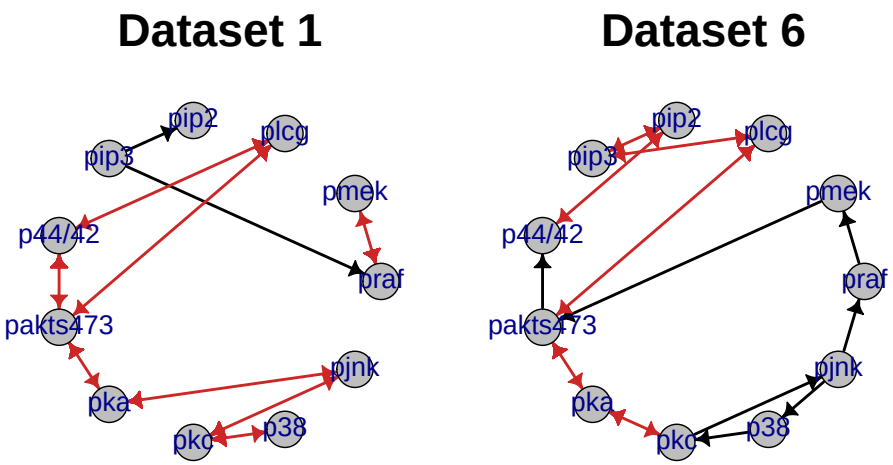



**Figure 3.2:** Estimated graphs corresponding to the case of minimum number of edges (10 edges, dataset 1) and maximum number of edges (13 edges, dataset 6).

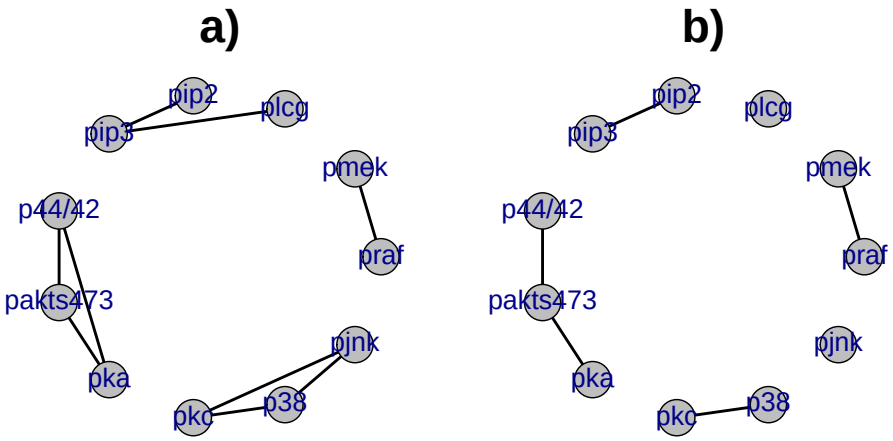



**Figure 3.3:** Edges appearing in at least 9 (a) and 13 (b) skeletons from the estimated graphs.

We display two of the selected graphs in Figure 3.2, one with minimum (dataset 1) and one with maximum (dataset 6) number of edges, the latter also displaying a 3-cycle. Although further work is needed to fully determine possible equivalences, there is no obvious reason (e.g., by Theorem 3.4) for a distributionally equivalent graph without a cycle to exist. We conjecture that this is indeed not the case. Considering all 14 graph estimates together it is reassuring to observe that some structure is shared. Figure 3.3 shows the (undirected) edges that appear in all/at least 11 of the skeletons of the estimated graphs.

Our selected graphs show good agreement with regulatory relationships described in Sachs et al. [2005], e.g., the interplay PLCG-PIP2-PIP3 appears in at least 12 of the inferred graphs; the connections PKC-P38-PJNK, P44/42 (named ERK in Sachs et al. [2005]) and PKA-PAKTS473 (named AKT in Sachs et al. [2005]) are in all 14 graphs. Our results suggest the possibility of feedback loops in the regulatory network

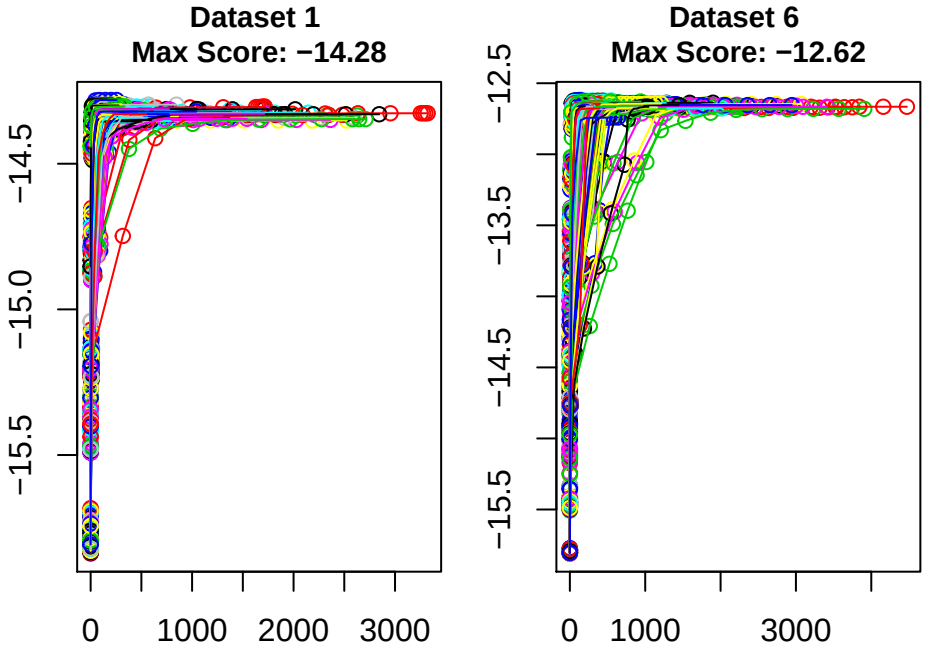



**Figure 3.4:** Curves of scores versus the time (seconds) in 300 random restarts greedy search for dataset 1 and dataset 6.

(e.g., PKC-P38-PJNK). Moreover, three expected relationships that are well-reported from the field-related literature emerge in our work but were undetected in Sachs et al. [2005].

Finally, in order to illustrate the behavior of the greedy search itself we focus again on datasets 1 and 6. Figure 3.4 shows the respective search paths in terms of the score achieved at each iteration. While local optima are possible, we observe that most search paths end with a score near the overall maximum.

## 3.5 Discussion

We considered structure learning for linear causal models with Gaussian errors that may exhibit feedback loops and correlation induced by latent variables. In order to gain tractability in this difficult problem, we restricted our attention to simple mixed graphs. Such graphs have the favorable property of always inducing a model whose dimension is as one expects from counting parameters. This property allows one to form meaningful model selection scores that we deployed in greedy algorithms. While a search over simple mixed graphs remains challenging, computationally and statistically, our experiments suggest that useful information can be learned from greedy search methods. This generalizes similar conclusions for acyclic simple graphs [Nowzohour et al., 2017].

We also showed that an existing sufficient condition for distributional equivalence admits a natural generalization from acyclic to cyclic simple mixed graphs. However,

the condition is very restrictive. It would be important to find more broadly applicable conditions for distributional equivalence.

# Chapter 4

# BCD Algorithm for Interventional Data in Directed Graphical Models

In Chapter 3, we use a block-coordinate descent (BCD) algorithm of Drton et al. [2019b] to compute the maximum likelihood estimates (MLE) of the parameters of a linear structural equation model given by a mixed graph. The algorithm works for a single dataset of observational data, but cannot handle the combination of different experimental (interventional) setups. It is of great interest to extend the algorithm to deal with both observational and interventional data. When considering such an extension here, we will restrict ourselves to directed cyclic graphs without bidirected edges.

Our starting point is a directed graph $G$ defining a linear structural equation model. We are then interested in data collected under different interventions. The type of interventions we consider are "hard" interventions that fix the values of the intervened variables in a randomized fashion so that they follow a controlled probability distribution Pearl [2009], Spirtes et al. [2000]. Throughout this chapter, we use $I \subseteq [V]$ to denote the intervention target in one interventional environment—in other words, the set $I$ indexes the intervened variables. Given an intervention target $I$, the manipulated graph $G_{\overline{I}}$ is obtained by deleting from $G$ all edges pointing to nodes in $I$, which represent the structure of the SEM after intervention. The collection $\mathcal{I} \subset 2^V$ is the family of intervention targets $I_k$'s across all the different environments for which data is available. To distinguish from the entry indices in data matrices, we use $(k)$-superscripts to specify data from different interventional environments: $Y^{(k)} \in \mathbb{R}^{V \times n^{(k)}}$ is the data matrix with intervention target $I_k$, in which each column is one sample and $n^{(k)}$ is the sample size.

The following example provides an illustration of the intervention and the manipulated graph under the intervention.

**Example 4.1.** *For the graph $G$ in Figure 4.1(a), the linear SEM in the observational environment is*

$$\begin{aligned} X_1 &= \varepsilon_1, \\ X_2 &= \lambda_{12} X_1 + \lambda_{42} X_4 + \varepsilon_2, \\ X_3 &= \lambda_{23} X_2 + \varepsilon_3, \\ X_4 &= \lambda_{14} X_1 + \lambda_{34} X_3 + \varepsilon_4, \end{aligned}$$

*where* $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4) \sim \mathcal{N}_4(0, \mathrm{diag}(\omega_{11}, \omega_{22}, \omega_{33}, \omega_{44}))$.

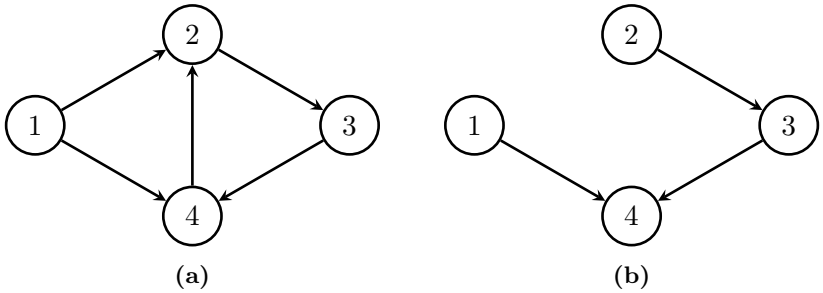



**Figure 4.1:** Original graph, and manipulated graph with intervention target $I = \{2\}$.

*When we intervene on variable $X_2$, i.e., the intervention target is $I = \{2\}$, the manipulated linear SEM becomes*

$$\begin{aligned} X_1 &= \varepsilon_1, \\ X_2 &= \varepsilon_2', \\ X_3 &= \lambda_{23} X_2 + \varepsilon_3, \\ X_4 &= \lambda_{14} X_1 + \lambda_{34} X_3 + \varepsilon_4, \end{aligned}$$

*with $(\varepsilon_1, \varepsilon_3, \varepsilon_4) \sim \mathcal{N}_3(0, \mathrm{diag}(\omega_{11}, \omega_{33}, \omega_{44}))$ and $\varepsilon_2'$ from (known or unknonw) intervention distribution. The manipulated graph is displayed in Figure 4.1(b).*

## 4.1 Optimization Problems for Gaussian Errors

Suppose that we are given a data matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$ that holds in its columns $n$ vector-valued observations from a single environment corresponding to the original unmanipulated graph $G$ with parameters $(\Lambda, \Omega)$. With Gaussian errors, the log-likelihood function takes the form

$$\ell_{G,\mathbf{X}}(\Omega, \Lambda) = -\log\det(\Omega) + \log\det(I - \Lambda)^2 - \mathrm{tr}\left\{(I - \Lambda)\Omega^{-1}(I - \Lambda)^T S\right\},$$

where $S = \mathbf{X}\mathbf{X}^T/n$ is the sample covariance matrix; recall (3.9).

To find the critical point(s), we can take derivatives and obtain the likelihood equations, see Proposition 1 in Drton et al. [2019b]. However, generally, the resulting likelihood equations have a high algebraic degree and a direct solution may be difficult. Following the approach taken in Drton and Richardson [2004], Drton et al. [2019b], we instead use a block-coordinate descent method and decompose the original problem into partial subproblems of lower degrees that are solved iteratively.

We focus on estimating $\Lambda$ and $\Omega$ with observational and interventional data. (The observational case corresponds to intervention target $\emptyset$.) Suppose that there are $K$ datasets with different interventions $\mathcal{I} = \{I_1, \ldots, I_K\}$ and sample sizes $\{n^{(1)}, \ldots, n^{(k)}\}$. A simple approach to estimation would be to run the original BCD algorithm on each dataset and aggregate the $K$ MLEs by computing the weighted average as the final estimate. The average is only taken over those environments, in which the considered node is not intervened upon, i.e., the considered edges are present and the error variance associated to the node is unchanged. A natural choice for the weighting scheme

is to assign weights proportional to the sample sizes. If the linear SEM is recursive, i.e., the associated graph is a directed acyclic graph (DAG), then the log-likelihood has the same formula in each environment, making this the optimal weighting scheme.

While averaging is easily done by applying existing software for optimization separately in each environment, a statistically better approach is to combine all available data into one stacked data matrix and estimate all parameters based on a likelihood function for the combined data. When considering DAGs, this amounts to running one regression for each node with all the data stacked together, as studied in Hauser and Bühlmann [2015]. Nevertheless, the task becomes challenging for graphs with feedback loops, as various interventions may yield diverse strongly connected component structures, and the log-likelihood function may exhibit varying forms. In this chapter, we study this challenge in computing the MLE from the stacked data in a BCD-type scheme with respect to each node.

For a general directed graph representing a linear SEM, different interventional environments correspond to different parameter matrices. Each manipulated graph is a subgraph of the original graph, and the coefficient matrix $\Lambda^{(k)}$ is obtained by masking some off-diagonal entries in $\Lambda$ with zeroes. Recall Example 4.1, where the coefficients $\lambda_{12}$ and $\lambda_{42}$ become zero under the intervention $\{2\}$. We assume that the data $\mathbf{X}^{(k)} \in \mathbb{R}^{p \times n^{(k)}}$ from each interventional environment is centered (with mean zero in each row). The log-likelihood of a single interventional dataset $k$ is then the following function of $(\Omega^{(k)}, \Lambda^{(k)})$:

$$\begin{aligned}\ell_{G,\mathbf{X}}(\Lambda^{(k)}, \Omega^{(k)}) = &- \log\det(\Omega^{(k)}) - \log\det(I - \Lambda^{(k)})^2 \\ &- \operatorname{tr}\left\{(I - \Lambda^{(k)})(\Omega^{(k)})^{-1}(I - \Lambda^{(k)})^T S^{(k)}\right\},\end{aligned}$$

where $S^{(k)} = \mathbf{X}^{(k)}(\mathbf{X}^{(k)})^T/n^{(k)}$ is the sample covariance matrix of the $k$'th environment.

The block update involves iteratively estimating the parental edge weights and error variance for each node. For a fixed node $i$, the parental edge weights $\Lambda_{\mathrm{pa}(i),i}$ (or $\Lambda^T_{i,\mathrm{pa}(i)}$) and error variance $\omega_{ii}$ appear among the arguments of likelihood function for every data set without intervention on node $i$. This is analogous to the situation in Hauser and Bühlmann [2015]. We define $\mathcal{I}_{-i} := \{I \in \mathcal{I},\ i \notin I\}$ as the set of intervention targets relevant for estimating $(\Lambda_{\mathrm{pa}(i),i}, \omega_{ii})$.

Notice that for each intervention target $I_k \in \mathcal{I}_{-i}$, the $i$'th column of $\Lambda^{(k)}$ remains the same as that in $\Lambda$. The log-likelihood function for environment $k$ can be written as

$$\begin{aligned}\ell_{G,\mathbf{X}^{(k)}}(\Lambda^{(k)}, \Omega^{(k)}) = &- \log\omega_{ii}^{(k)} - \frac{1}{n^{(k)}\omega_{ii}^{(k)}}\|\mathbf{X}^{(k)}_{i,\cdot} - \Lambda^T_{i,\mathrm{pa}(i)}\mathbf{X}^{(k)}_{\mathrm{pa}(i),\cdot}\|^2 - \log\det(\Omega^{(k)}_{-i,-i}) \\ &- \frac{1}{n^{(k)}}\operatorname{tr}((\Omega^{(k)}_{-i,-i})^{-1}\epsilon^{(k)}_{-i}(\epsilon^{(k)}_{-i})^T) + \log\det(I - \Lambda^{(k)})^2, \qquad (4.1)\end{aligned}$$

where the notation $-i$ is used when forming a subvector with coordinate $i$ dropped, and also when dropping the $i$th row or column from a matrix. For the sake of simplicity, we write $\mathbf{X}_i$ for $\mathbf{X}^{(k)}_{i,\cdot}$ and $\mathbf{X}^{(k)}_{\mathrm{pa}(i)}$ for $\mathbf{X}^{(k)}_{\mathrm{pa}(i),\cdot}$. The total log-likelihood for

all data sets combined is now

$$\ell_{G,\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(K)}}(\Omega,\Lambda) = \sum_{k:I_k\in\mathcal{I}_{-i}} n^{(k)} \cdot \ell_{G,\mathbf{X}^{(k)}}(\Lambda^{(k)},\Omega^{(k)}). \tag{4.2}$$

The likelihood equations are obtained by taking derivatives with respect to all parameters. (The error variance $\omega_{ii}^{(k)}$ is a nuisance parameter when $i \in I_k$, and it does not affect the equations involving the parameters of interest, which are the elements of $\Lambda$ and $\Omega$ associated to the original graph $G$.) Even in purely observational cases, the likelihood equation system can have a high degree; the degree could also increase largely due to multiple interventional environments; see Section 4.2 and Drton et al. [2019b]. To alleviate this issue, we optimize the joint log-likelihood using block-coordinate descent method. The parameters are partitioned into blocks based on nodes and optimization is performed iteratively over each block. In the $i$-th block update problem, the submatrices $\Omega_{-i,-i}$ and $\Lambda_{-i}^T$ are fixed. This type of sub-systems with a smaller number of equations usually has lower degrees.

To simplify the log-likelihood we do some algebra. By Lemma 2 in Drton et al. [2019b], $\det(I-\Lambda^{(k)})$ is a linear function of $\Lambda_{\mathrm{pa}(i),i}$ for each $k$:

$$\det(I-\Lambda^{(k)}) = c_{i,0}^{(k)} + (\Lambda^{(k)})_{i,\mathrm{pa}(i)}^T c_{i,\mathrm{pa}(i)}^{(k)}. \tag{4.3}$$

Fixing $\Omega_{-i,-i}$ and $B_{-i}$, the maximization of total log-likelihood is reduced to the maximizition of

$$\ell_{G,\mathbf{X}^1,\ldots,\mathbf{X}^K}(\omega_{ii},\Lambda_{\mathrm{pa}(i),i}) = \sum_{k:I_k\in\mathcal{I}_{-i}} \Bigg( -n^{(k)}\log\omega_{ii}^{(k)} - \frac{1}{\omega_{ii}^{(k)}}\|\mathbf{X}_i^{(k)} - \Lambda_{i,\mathrm{pa}(i)}^T\mathbf{X}_{\mathrm{pa}(i)}^{(k)}\|^2 + n^{(k)}\log[(c_{i,0}^{(k)} + \Lambda_{i,\mathrm{pa}(i)}^T c_{i,\mathrm{pa}(i)}^{(k)})^2] \Bigg). \tag{4.4}$$

If $\mathbf{X}_i^{(k)} - \Lambda_{i,\mathrm{pa}(i)}^T\mathbf{X}_{\mathrm{pa}(i)}^{(k)} \neq 0$ for each $k$ such that $i \notin I_k$, then we can optimize $\omega_{ii}$ for each value of $\Lambda_{\mathrm{pa}(i),i}$:

$$(\omega_{ii}^{(k)})^* = \frac{1}{\sum_{k:I_k\in\mathcal{I}_{-i}} n^{(k)}} \sum_{k:I_k\in\mathcal{I}_{-i}} n^{(k)}\|Y_i^{(k)} - \Lambda_{i,\mathrm{pa}(i)}^T Y_{\mathrm{pa}(i)}^{(k)}\|^2$$

maximizes the total log-likelihood with respect to $\omega_{ii}^{(k)}$. It leads to the following profile log-likelihood function for the parameter vector $\Lambda_{\mathrm{pa}(i),i}$:

$$\ell(\Lambda_{\mathrm{pa}(i),i}) = -\sum_{k:I_k\in\mathcal{I}_{-i}} n^{(k)}\log\frac{\sum_{k:i\notin I_k}\|\mathbf{X}_i^{(k)} - \Lambda_{i,\mathrm{pa}(i)}^T\mathbf{X}_{\mathrm{pa}(i)}^{(k)}\|^2}{(c_{i,0}^{(k)} + \Lambda_{i,\mathrm{pa}(i)}^T c_{i,\mathrm{pa}(i)}^{(k)})^2}. \tag{4.5}$$

The gradient of total log-likelihood with respect to $\Lambda_{\mathrm{pa}(i),i}$ is a sum of fractions with different denominators. If one tries to reduce these terms to a common denominator, the expression becomes exceedingly complicated. The (partial) derivative(s) are high order polynomial equation(s) with multiple variables. In general there is no closed-form solution for the critical point(s) of the profile log-likelihood function (4.4). However, in special cases with at most 2 different denominators, we can find the optimizer in closed form and perform the block-coordinate update; see Section 4.3.

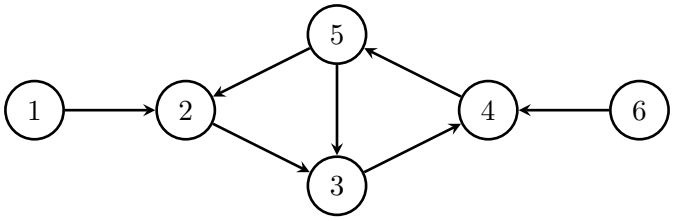



**Figure 4.2:** Graph of a linear SEM, with ML degree 23 for observational data.

## 4.2 Maximum likelihood degrees, a concrete example

We provide an example to illustrate the complexity in the critical equations of the MLE and the block-coordinate descent problems. We consider the directed graph in Figure 4.2. The linear SEM associated to this graph has generic parameters

$$\Lambda = \begin{pmatrix} 0 & \lambda_{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_{23} & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{34} & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{45} & 0 \\ 0 & \lambda_{52} & \lambda_{53} & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{64} & 0 & 0 \end{pmatrix}, \quad \Omega = \begin{pmatrix} \omega_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & \omega_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & \omega_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & \omega_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & \omega_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & \omega_{66} \end{pmatrix}.$$

Suppose that the collection of interventions is $\mathcal{I} = \{\emptyset, \{2\}, \{4\}\}$, with sample sizes $n^{(1)}, n^{(2)}, n^{(3)}$ and sample covariance matrix $S^{(1)}, S^{(2)}, S^{(3)}$ respectively. The corresponding true parameters are denoted by $\Lambda^{(k)}$ and $\Omega^{(k)}$ for $k \in \{1, 2, 3\}$.

The likelihood equations of the purely observational environment (i.e., intervention on $\emptyset$) has degree 23. When it comes to the likelihood equation with respect to the 3 different environments of different sample sizes, the ML degree is 73, largely increased. The mentioned ML degrees were obtained by computing Gröbner bases in the MATHEMATICA software. See Sullivant [2018] and Drton et al. [2009] for background on the computation of ML degrees of Gaussian models.

Then we check the degrees of block update problems, again with the help of Gröbner bases.

**Nodes 1 and 6:** For these two nodes, the block update problems are trivially of degree 1; we merely compute the empirical variances $\hat{\omega}_{11}$ and $\hat{\omega}_{66}$.

**Node 2:** Intervention $I_3 = \{4\}$ breaks both cycles $2 \to 3 \to 4 \to 5 \to 2$ and $3 \to 4 \to 5$. Therefore, our Theorem 4.2 from Section 4.3 asserts that the update problem for node 2 has degree 2.

**Node 4:** Node 4 has only one parent 5 in the same strongly connected component $\{2, 3, 4, 5\}$. However, our Theorem 4.2 does not apply, due to two different nontrivial $\log\det$ terms with different interventions.

**Node 3:** The block update with respect to node 3 is the most complicated. The strongly connected components containing node 3 in the three intervened graphs are all different, with node sets $\{2, 3, 4, 5\}, \{3, 4, 5\}, \{3\}$.

In more detail, the profile log-likelihood function with respect to node 4 is

$$\begin{aligned} g_4(\lambda_{34}, \lambda_{64}) &= -n^{(1)} \log\left(\frac{\|\mathbf{X}_4 - \lambda_{34}\mathbf{X}_3 - \lambda_{64}\mathbf{X}_6\|^2}{(1 - \lambda_{23}\lambda_{34}\lambda_{45}\lambda_{52} - \lambda_{34}\lambda_{45}\lambda_{53})^2}\right) \\ &\quad - n^{(2)} \log\left(\frac{\|\mathbf{X}_4 - \lambda_{34}\mathbf{X}_3 - \lambda_{64}\mathbf{X}_6\|^2}{(1 - \lambda_{34}\lambda_{45}\lambda_{53})^2}\right) \\ &= -\left(n^{(1)} + n^{(2)}\right) \log(\|\mathbf{X}_4 - \lambda_{34}\mathbf{X}_3 - \lambda_{64}\mathbf{X}_6\|^2) \\ &\quad + n^{(1)} \log((1 - \lambda_{23}\lambda_{34}\lambda_{45}\lambda_{52} - \lambda_{34}\lambda_{45}\lambda_{53})^2) + n^{(2)} \log((1 - \lambda_{34}\lambda_{45}\lambda_{53})^2). \end{aligned}$$

It has derivatives

$$\begin{aligned} \frac{\partial g_4(\lambda_{34}, \lambda_{64})}{\partial \lambda_{34}} &= n^{(1)} \frac{2(\lambda_{23}\lambda_{52} + \lambda_{53})\lambda_{45}\lambda_{34}}{(\lambda_{23}\lambda_{52} + \lambda_{53})\lambda_{45}\lambda_{34} - 1} + n^{(2)} \frac{2\lambda_{53}\lambda_{45}\lambda_{34}}{\lambda_{53}\lambda_{45}\lambda_{34} - 1} \\ &\quad - \left(n^{(1)} + n^{(2)}\right) \frac{2(\mathbf{X}_3\mathbf{X}_6^T\lambda_{64} + \mathbf{X}_3\mathbf{X}_3^T\lambda_{34} - \mathbf{X}_3\mathbf{X}_4^T)}{\|\mathbf{X}_4 - \lambda_{34}\mathbf{X}_3 - \lambda_{64}\mathbf{X}_6\|^2}, \\ \frac{\partial g_4(\lambda_{34}, \lambda_{64})}{\partial \lambda_{64}} &= \left(n^{(1)} + n^{(2)}\right) \frac{2(\mathbf{X}_6\mathbf{X}_6^T\lambda_{64} + \mathbf{X}_3\mathbf{X}_6^T\lambda_{34} - \mathbf{X}_6\mathbf{X}_4^T)}{\|\mathbf{X}_4 - \lambda_{34}\mathbf{X}_3 - \lambda_{64}\mathbf{X}_6\|^2}. \end{aligned}$$

The critical points of $g_4(\lambda_{34}, \lambda_{64})$ are given by the solutions of

$$\frac{\partial g_4(\lambda_{34}, \lambda_{64})}{\partial \lambda_{34}} = 0, \quad \frac{\partial g_4(\lambda_{34}, \lambda_{64})}{\partial \lambda_{64}} = 0.$$

The equation system has generic degree 4.

For node 3, we have the profile log-likelihood function

$$\begin{aligned} &g_3(\lambda_{23}, \lambda_{53}) \\ &= -n^{(1)} \log\left(\frac{\|\mathbf{X}_3 - \lambda_{23}\mathbf{X}_2 - \lambda_{53}\mathbf{X}_5\|^2}{(1 - \lambda_{23}\lambda_{34}\lambda_{45}\lambda_{52} - \lambda_{34}\lambda_{45}\lambda_{53})^2}\right) \\ &\quad - n^{(2)} \log\left(\frac{\|\mathbf{X}_3 - \lambda_{23}\mathbf{X}_2 - \lambda_{53}\mathbf{X}_5\|^2}{(1 - \lambda_{34}\lambda_{45}\lambda_{53})^2}\right) - n^{(3)} \log\left(\|\mathbf{X}_3 - \lambda_{23}\mathbf{X}_2 - \lambda_{53}\mathbf{X}_5\|^2\right) \\ &= -(n^{(1)} + n^{(2)} + n^{(3)}) \log(\|\mathbf{X}_3 - \lambda_{23}\mathbf{X}_2 - \lambda_{53}\mathbf{X}_5\|^2) \\ &\quad + n^{(1)} \log((1 - \lambda_{23}\lambda_{34}\lambda_{45}\lambda_{52} - \lambda_{34}\lambda_{45}\lambda_{53})^2) + n^{(2)} \log((1 - \lambda_{34}\lambda_{45}\lambda_{53})^2), \end{aligned}$$

with derivatives

$$\begin{aligned} \frac{\partial g_3(\lambda_{23}, \lambda_{53})}{\partial \lambda_{23}} &= n^{(1)} \frac{2\lambda_{52}\lambda_{45}\lambda_{34}\lambda_{23}}{(\lambda_{23}\lambda_{52} + \lambda_{53})\lambda_{45}\lambda_{34} - 1} \\ &\quad - \left(n^{(1)} + n^{(2)} + n^{(3)}\right) \frac{2(\mathbf{X}_2\mathbf{X}_2^T\lambda_{23} + \mathbf{X}_2\mathbf{X}_5^T\lambda_{53} - \mathbf{X}_2\mathbf{X}_3^T)}{\|\mathbf{X}_3 - \lambda_{23}\mathbf{X}_2 - \lambda_{53}\mathbf{X}_5\|^2}, \\ \frac{\partial g_3(\lambda_{23}, \lambda_{53})}{\partial \lambda_{53}} &= n^{(1)} \frac{2\lambda_{45}\lambda_{34}\lambda_{53}}{(\lambda_{23}\lambda_{52} + \lambda_{53})\lambda_{45}\lambda_{34} - 1} + n^{(2)} \frac{2\lambda_{45}\lambda_{34}\lambda_{53}}{\lambda_{45}\lambda_{34}\lambda_{53} - 1} \\ &\quad - \left(n^{(1)} + n^{(2)} + n^{(3)}\right) \frac{2(\mathbf{X}_2\mathbf{X}_5^T\lambda_{23} + \mathbf{X}_5\mathbf{X}_5^T\lambda_{53} - \mathbf{X}_5\mathbf{X}_3^T)}{\|\mathbf{X}_3 - \lambda_{23}\mathbf{X}_2 - \lambda_{53}\mathbf{X}_5\|^2}. \end{aligned}$$

A Gröbner basis computation shows that the equation system has degree 7.

## 4.3 Block-coordinate Descent

### 4.3.1 Arbitrary directed graphs, special interventions

If we do not restrict the type of graph that defines the model, a closed-form block update can only be possible under special conditions on the interventions (recall the example in Section 4.2).

Let $G = (V, D)$ be the directed graph in observational environment and $\mathcal{I} \subseteq 2^V$ be the collection of intervention targets. We bind all the data matrices $\mathbf{X}^{(k)}$, $I_k \in \mathcal{I}_{-i}$ by columns and obtain the stacked data matrix $\mathbf{X}$, which holds the relevant data for estimating $(\Lambda_{\mathrm{pa}(i),i}, \omega_{ii})$, as described in Section 4.1. A sufficient condition for a closed-form block update on node $i$ is:

$$\exists\ G' \subseteq G, \quad \text{s.t. } \forall\ I \in \mathcal{I}_{-i},\ \mathcal{C}(i, G_{\overline{I}}) = G' \text{ or } V[\mathcal{C}(i, G_{\overline{I}})] = \{i\}, \tag{4.6}$$

where the operator $V[\cdot]$ returns the set of the nodes in a (sub)graph. In other words, the condition requires that there can be at most two different structures of the strongly connected component containing node $i$, and one of which is the singleton set $\{i\}$. Under this condition, we can derive the closed-form block update formula for (4.4).

Given a fixed node $i$, if $V[\mathcal{C}(i, G_{\overline{I}})] = \{i\}$ holds for each intervention target $I \in \mathcal{I}_{-i}$, the profile log-likelihood function after optimizing over $\omega_{ii}$ is

$$g_i(\Lambda_{\mathrm{pa}(i),i}) = -n_1 \log \left( \frac{\|\mathbf{X}_i - \Lambda_{i,\mathrm{pa}(i)}^T \mathbf{X}_{\mathrm{pa}(i)}\|^2}{(c_{i,0} + \Lambda_{i,\mathrm{pa}(i)}^T c_{i,\mathrm{pa}(i)})^2} \right) + C,$$

where $c_{i,0} + \Lambda_{i,\mathrm{pa}(i)}^T c_{i,\mathrm{pa}(i)}$ is the formula of $\det(I - \Lambda)$ with variable $\Lambda_{\mathrm{pa}(i),i}$, and $c_{i,\mathrm{pa}(i)} \neq 0$. It is reduced to a quadratic ratio optimization problem, which is the same as that in the purely observational case.

If the set $V[\mathcal{C}(i, G_{\overline{I}})]$'s take values in $\{i\}$ and some $G'$ across all the intervention targets $I$'s, the profile log-likelihood function after optimizing over $\omega_{ii}$ is

$$\begin{aligned} g_i(\Lambda_{\mathrm{pa}(i),i}) = &- n_1 \log \left( \frac{\|\mathbf{X}_i - \Lambda_{i,\mathrm{pa}(i)}^T \mathbf{X}_{\mathrm{pa}(i)}\|^2}{(c_{i,0} + \Lambda_{i,\mathrm{pa}(i)}^T c_{i,\mathrm{pa}(i)})^2} \right) \\ &- n_2 \log \left( \|\mathbf{X}_i - \Lambda_{i,\mathrm{pa}(i)}^T \mathbf{X}_{\mathrm{pa}(i)}\|^2 \right) + C. \end{aligned} \tag{4.7}$$

To maximize the profile log-likelihood function, we may consider the following minimization problem:

$$\min_{\alpha \in \mathbb{R}^{|\,\mathrm{pa}(i)|}} n_1 \log \frac{\|\mathbf{X}_i^T - \mathbf{X}_{\mathrm{pa}(i)}^T \alpha\|^2}{(c_{i,0} + c_{i,\mathrm{pa}(i)}^T \alpha)^2} + n_2 \log \left( \|\mathbf{X}_i^T - \mathbf{X}_{\mathrm{pa}(i)}^T \alpha\|^2 \right). \tag{4.8}$$

With the reparameterization tricks from Drton et al. [2019b], the minimization problem (4.8) can be reduced to solving a quadratic equation.

**Theorem 4.2.** *Given a node $i$, let $n_1, n_2 > 0$ be the total numbers of data corresponding to strongly connected components $G'$ and $\{i\}$, respectively, and let $r = n_1/n_2$. Suppose that the stacked partial data matrix $\mathbf{X}_{\mathrm{pa}(i)\cup\{i\}}$ has full rank $|\,\mathrm{pa}(i)| + 1 \leq n_1 + n_2$.*

*Let* $\hat{\alpha} = (\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_i^T$ *be the minimizer of* $\|\mathbf{X}_i^T - \mathbf{X}_{\mathrm{pa}(i)}^T\alpha\|^2$. *We define* $n := n_1 + n_2, m := |\,\mathrm{pa}(i)|, c_0 := c_{i,0}$ *and* $c_1 := c_{i,\mathrm{pa}(i)} \neq 0$, $y_0^2 := \|\mathbf{X}_i^T - \mathbf{X}_{\mathrm{pa}(i)}^T\hat{\alpha}\|^2$ *and* $l^2 := c_1^T(\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1$. *Then the solution of the optimization problem in* (4.8) *satisfies*

$$\alpha^* = \hat{\alpha} + \delta \cdot (\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1, \tag{4.9}$$

*where* $\delta$ *is a solution to the quadratic equation*

$$l^2\delta^2 + (c_1^T\hat{\alpha} + c_0)(k+1)\delta - ry_0^2 = 0. \tag{4.10}$$

*Proof.* We adopt the orthogonal transformation method in Drton et al. [2019b] and also derive further auxiliary properties for our problem. First, we start with finding an orthogonal $m \times m$ matrix $Q_1$ such that $Q_1c_1 = (0, \ldots, 0, \|c_1\|)^T$. Then we compute a QR decomposition $\mathbf{X}_{\mathrm{pa}(i)}^TQ_1^T = Q_2^TR$ with $Q_2 \in \mathbb{R}^{n\times m}$ orthogonal and $R \in \mathbb{R}^{n\times m}$ upper triangular. Noting that $\mathbf{X}_{\mathrm{pa}(i)}$ and $\mathbf{X}_{\mathrm{pa}(i)}^TQ_1^T$ have full ranks, we can postulate that all diagonal entries of $R$ are positive, and then the matrix $R$ is unique for any given $Q_1$. After reparameterizing to $\alpha' = Q_1\alpha$, the common $L_2$-norm term is transformed to

$$\begin{aligned} y_0^2 &= \|\mathbf{X}_i^T - \mathbf{X}_{\mathrm{pa}(i)}^T\alpha\|^2 = \|Q_2\mathbf{X}_i^T - R\alpha'\|^2 \\ &= \sum_{j=1}^{m}[(Q_2\mathbf{X}_i^T)_j - (R\alpha')_j]^2 + \sum_{j=m+1}^{N}(Q_2\mathbf{X}_i^T)_j^2, \end{aligned}$$

and the denominator is transformed to $(c_0 + \|c_1\|\alpha'_m)^2$.

Since $R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$ with $R_1 \in \mathbb{R}^{m\times m}$, we reparameterize again with $\alpha'' = R_1\alpha'$ and the original minimization problem is equivalent to minimizing

$$\begin{aligned} &n_1 \log \frac{\sum_{j=1}^{m}[(Q_2\mathbf{X}_i^T)_j - \alpha''_j]^2 + \sum_{j=m+1}^{n}(Q_2\mathbf{X}_i^T)_j^2}{(c_0 + \|c_1\|R_{mm}^{-1}\alpha''_m)^2} \\ &+ n_2 \log \left( \sum_{j=1}^{m}[(Q_2\mathbf{X}_i^T)_j - \alpha''_j]^2 + \sum_{j=m+1}^{n}(Q_2\mathbf{X}_i^T)_j^2 \right). \end{aligned} \tag{4.11}$$

Any solution of (4.11) must satisfies that $\alpha''_j = (Q_2\mathbf{X}_i^T)_j$ for $j \in [m-1]$, and the optimal value of $\alpha''_m$ is given by minimizing

$$\begin{aligned} &n_1 \log \frac{[(Q_2\mathbf{X}_i^T)_m - \alpha''_m]^2 + \sum_{j=m+1}^{n}(Q_2\mathbf{X}_i^T)_j^2}{(c_0 + \|c_1\|R_{mm}^{-1}\alpha''_m)^2} \\ &+ n_2 \log \left( [(Q_2\mathbf{X}_i^T)_m - \alpha''_m]^2 + \sum_{j=m+1}^{N}(Q_2\mathbf{X}_i^T)_j^2 \right), \end{aligned} \tag{4.12}$$

i.e., maximizing

$$\begin{aligned} g_i(x) :=& n_1 \log \left( x + \frac{c_0R_{mm}}{\|c_1\|} \right)^2 \\ &- (n_1 + n_2) \log \left( x^2 - 2(Q_2\mathbf{X}_i^T)_m x + \sum_{j=m}^{n}(Q_2\mathbf{X}_i^T)_j^2 \right) + C. \end{aligned}$$

The univariate function $g_i$ has derivative

$$g_i'(x) = \frac{2n_1}{x + c_0 R_{mm}/\|c_1\|} - 2(n_1+n_2)\frac{x - (Q_2\mathbf{X}_i^T)_m}{x^2 - 2(Q_2\mathbf{X}_i^T)_m x + \sum_{j=m}^N (Q_2\mathbf{X}_i^T)_j^2}.$$

We can apply Lemma 4.5 to give the two possible optimal values of $x$, with

$$a = 1,\ b = -(Q_2\mathbf{X}_i^T)_m,\ c = \sum_{j=m}^N (Q_2\mathbf{X}_i^T)_j^2, d = c_0,\ \lambda = \|c_1\|/R_{mm} \neq 0.$$

By the full rank assumption for $\mathbf{X}_{\mathrm{pa}(i)\cup i}$, we know that

$$c - b^2 = \sum_{j=m+1}^N (Q_2\mathbf{X}_i^T)_j^2 = \|\mathbf{X}_i^T - \mathbf{X}_{\mathrm{pa}(i)}^T\hat{\alpha}\|^2 > 0.$$

The $b^2 - ac < 0$ condition for case (ii) of Lemma 4.5 is fulfilled.

Finally, we need to simplify the formula using original parameters. Since $r = n_1/n_2$, the two solutions has the form

$$\begin{aligned}\alpha_m'' &= \frac{b\lambda(r-1) - ac_0(r+1) \pm \sqrt{(b\lambda(r-1) - ac_0(r+1))^2 + 4a\lambda(c\lambda r - bc_0(r+1))}}{2a\lambda} \\ &= -\frac{b}{a} + \frac{(b\lambda - ac_0)(r+1) \pm \sqrt{(b\lambda - ac_0)^2(r+1)^2 + 4(ac-b^2)\lambda^2 r}}{2a\lambda}.\end{aligned}$$

The optimal solution in original coordinates is $\alpha = Q_1^T R_1^{-1}\alpha''$. Since $R_1^{-1}(Q_2\mathbf{X}_i^T)$ is the linear regression coefficient vector of $\mathbf{X}_{\mathrm{pa}(i)}^T Q_1^T$ on $\mathbf{X}_i^T$, we have

$$Q_1^T R_1^{-1}(Q_2\mathbf{X}_i^T)_{[m]} = (\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_i^T := \hat{\alpha}.$$

Let $e_{m,m} = (0,\ldots,0,1)$ be the $m$-th canonical basis vector of $\mathbb{R}^m$ and $e_{m,N} = (0,\ldots,0,1,0,\ldots,0)$ be the $m$-th canonical basis vector of $\mathbb{R}^N$ ($m$-th entry is 1). Noticing that $R_{mm}^{-1}(Q_2\mathbf{X}_i^T)_m$ is the $m$-th entry of $R_1^{-1}(Q_2\mathbf{X}_i^T)_{[m]}$, and the last column of $R_1^{-T}$ is $R_{mm}^{-1}e_{m,m}$, we can derive that

$$\begin{aligned}\lambda \cdot b &= -\|c_1\|R_{mm}^{-1}(Q_2\mathbf{X}_i^T)_m = -\langle Q_1 c_1, R_1^{-1}(Q_2\mathbf{X}_i^T)_{[m]}\rangle \\ &= -\langle c_1, Q_1^T R^{-1}(Q_2\mathbf{X}_i^T)_{[m]}\rangle = -c_1^T\hat{\alpha},\end{aligned}$$

and

$$\begin{aligned}Q_1^T R_1^{-1}\|c_1\|R_{mm}^{-1}e_{m,m} &= Q_1^T R_1^{-1}R_1^{-T}Q_1 c_1 = (Q_1^T R^T R Q_1)^{-1}c_1 \\ &= (Q_1^T R^T Q_2 Q_2^T R Q_1)^{-1}c_1 = (\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1.\end{aligned}$$

The matrices $Q_1, Q_2, R$ may change, while the value of $R_{mm}$ (or equivalently, $\lambda$) is uniquely determined by $\mathbf{X}_{\mathrm{pa}(i)}$ and $c_1$. To see this, we have

$$\begin{aligned}\mathbf{X}_{\mathrm{pa}(i)}^T(\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1 &= \mathbf{X}_{\mathrm{pa}(i)}^T Q_1^T R_1^{-1}\|c_1\|R_{mm}^{-1}e_{m,m} \\ &= Q_2^T\begin{pmatrix}R_1\\0\end{pmatrix}R_1^{-1}\|c_1\|R_{mm}^{-1}e_{m,m} \\ &= Q_2^T\begin{pmatrix}I_m\\0\end{pmatrix}\|c_1\|R_{mm}^{-1}e_{m,m} = Q_2^T\|c_1\|R_{mm}^{-1}e_{m,N}.\end{aligned}$$

Since $Q_2$ is orthogonal, the Euclidean norms of both sides must be equal. That is,

$$l = \sqrt{c_1^T(\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1} = \|\mathbf{X}_{\mathrm{pa}(i)}^T(\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1\| = \|c_1\|R_{mm}^{-1} = \lambda.$$

Then we can compute

$$c = b^2 + \sum_{j=m+1}^{N}(Q_2\mathbf{X}_i^T)_j^2 = b^2 + \|\mathbf{X}_i^T - \mathbf{X}_{\mathrm{pa}(i)}^T\hat{\alpha}\|^2 = b^2 + y_0^2,$$

and

$$\begin{aligned}\alpha_m'' &= -b + \frac{(-c_1^T\hat{\alpha} - c_0)(r+1) \pm \sqrt{(c_1^T\hat{\alpha} + c_0)^2(r+1)^2 + 4rl^2y_0^2}}{2l} \\ &:= (Q_2\mathbf{X}_i^T)_{[m]} + \frac{(-c_1^T\hat{\alpha} - c_0)(r+1) \pm \sqrt{\Delta_{r,\hat{\alpha}}(l)}}{2l}.\end{aligned}$$

Therefore, the two possible optimal vectors are

$$\begin{aligned}\alpha &= Q_1^TR_1^{-1}(Q_2\mathbf{X}_i^T)_{[m]} + \frac{-(c_1^T\hat{\alpha} + c_0)(r+1) \pm \sqrt{\Delta_{r,\hat{\alpha}}(l)}}{2l^2}\|c_1\|R_{mm}^{-1} \cdot Q_1^TR_1^{-1}e_{m,m} \\ &= \hat{\alpha} + \frac{-(c_1^T\hat{\alpha} + c_0)(r+1) \pm \sqrt{\Delta_{r,\hat{\alpha}}(l)}}{2l^2}(\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1.\end{aligned}$$

Each possible solution is the simple linear regression coefficient vector $\hat{\alpha}$ adding a multiple of $(\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1$. The coefficient of the second term is exactly the solution to the quadratic equation

$$l^2t^2 + (c_1^T\hat{\alpha} + c_0)(r+1)t - ry_0^2 = 0,$$

with

$$\begin{aligned}\hat{\alpha} &= \mathbf{X}_{\mathrm{pa}(i)}^T(\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}\mathbf{X}_i^T, \\ y_0^2 &= \|\mathbf{X}_i^T - \mathbf{X}_{\mathrm{pa}(i)}^T\hat{\alpha}\|^2, \\ l^2 &= \|\mathbf{X}_{\mathrm{pa}(i)}^T(\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1\|^2.\end{aligned}$$

□

In practice, we can first compute the two possible $\Lambda_{\mathrm{pa}(i),i}^*$'s by Theorem 4.2. The update of $\omega_{ii}$ is given by

$$\omega_{ii}^* = \frac{1}{n_1+n_2}\|\mathbf{X}_i - \Lambda_{i,\mathrm{pa}(i)}^T\mathbf{X}_{\mathrm{pa}(i)}\|^2 \tag{4.13}$$

for the two possible $\Lambda_{\mathrm{pa}(i),i}^*$'s. We compare the two profile log-likelihood values

$$n_1\log((c_{i,0} + (\Lambda^*)_{i,\mathrm{pa}(i)}^Tc_{i,\mathrm{pa}(i)})^2) - (n_1+n_2) * \log(\omega_{ii}^*) + C, \tag{4.14}$$

and the choice of $\Lambda_{\mathrm{pa}(i),i}^*$ corresponds to the larger log-likelihood value of the two candidates.

**Remark 4.3.** *The ratio $r = n_1/n_2$ affects the possible weights on the direction of $(\mathbf{X}_{\mathrm{pa}(i)}\mathbf{X}_{\mathrm{pa}(i)}^T)^{-1}c_1$. If $n_2 = 0$ or equivalently $r = \infty$, the problem degenerates to the purely observational case. The equation* (4.10) *becomes linear: $(c_1^T\hat{\alpha} + c_0)t - y_0^2 = 0$, which gives the result in Drton et al. [2019b].*

We outline the computations to be done in the overall block-coordinate descent in Algorithm 1.

**Algorithm 1** Block-coordinate descent, for directed graph and special type of interventions.

**Require:** $\omega^0$, $\Lambda^0$; $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}$; $I_1, \ldots, I_k$ and $n^{(1)}, \ldots, n^{(K)}$

**repeat**
  **for** $i \in V$ **do**
    **if** the condition (4.6) do not hold **then**
      The block update cannot be solved in closed-form, stop
    **end if**
    Find $S_2 = \{k : I_k \in \mathcal{I}_{-i},\ \mathcal{C}(i, G_{\overline{I_k}}) = \{i\}\}$ and $S_1 = \{k : I_k \in \mathcal{I}_{-i}\} \backslash S_2$
    Set $Y = [Y^{(k_1)}, \ldots Y^{(k_l)}]$ with $k_1, \ldots, k_l \in S_1 \cup S_2$
    Compute $n_1 = \sum_{k \in S_1} n^{(k)}$ and $n_2 = \sum_{k \in S_2} n^{(k)}$
    **if** $n_1 = 0$ **then**
      Compute $\hat{\Lambda}_{\mathrm{pa}(i),i}$ by solving least square problem
        $\arg\min_\alpha \|\mathbf{X}_i^T - \mathbf{X}_{\mathrm{pa}(i)}^T \alpha\|^2$
    **else if** $n_2 = 0$ **then**
      Compute $\hat{\Lambda}_{\mathrm{pa}(i),i}$ as the block-coordinate update for observational data
    **else**
      Compute the two possible $\hat{\Lambda}_{\mathrm{pa}(i),i}$'s using (4.9) and (4.10)
      Compute corresponding $\hat{\omega}_{ii}$'s and log-likelihood's using (4.13)-(4.14)
      Choose the larger log-likelihood value and corresponding $(\hat{\Lambda}_{\mathrm{pa}(i),i}, \hat{\omega}_{ii})$
    **end if**
    Update $\omega$ and $\Lambda_{\cdot,i}$ using $\hat{\omega}_{ii}$ and $\hat{\Lambda}_{\mathrm{pa}(i),i}$
  **end for**
**until** Convergence criterion is met

### 4.3.2 Special directed graphs under arbitrary interventions

In Section 4.3.1 we show that the block-coordinate problem has a closed-form solution when the interventions satisfy certain condition (4.6). Indeed, if we only consider the directed graphs in which every strongly connected component contains **at most one cycle**, i.e., any two cycles in the graph are disjoint, then (4.6) always holds for arbitrary interventions. For this type of graphs, a strongly connected component is either a cycle or a singleton set. The determinant of $I - \Lambda$ can be decomposed into the product of the sub-determinant of the strongly connected components (cycles in this setup) $\mathcal{C}_m$'s: $\det(I - \Lambda) = \prod_{m=1} \det(I - \Lambda_{\mathcal{C}_m})$.

For the parameters corresponding to one node $i$, there are two different cases. If the node is not in any cycle, then $(\Lambda_{\mathrm{pa}(i),i}, \omega_{ii})$ can be estimated by linear regression on parental nodes, because $\Lambda_{\mathrm{pa}(i),i}$ does not appear in $\det(I - \Lambda)^2$. Otherwise the node is in a cycle $\mathcal{C}(i, G)$. Each intervention target satisfies either $I \cap V[\mathcal{C}(i, G)] = \emptyset$ or $I \cap V[\mathcal{C}(i, G)] \neq \emptyset$. The former keeps the cycle $\mathcal{C}(i, G)$ in the manipulated graph $G_{\overline{I}}$ after intervention, and it contributes to a non-constant $\log\det(I - \Lambda_{\mathcal{C}(i,G)})$ term in the log-likelihood. It is worth noting that the potential breaking of **other** cycles $\mathcal{C}$ does not affect the updating of $\Lambda_{\mathrm{pa}(i),i}$, since those other entries in $\Lambda$ only appear in the $\log\det$ term as a constant with respect to $\Lambda_{\mathrm{pa}(i),i}$. The latter leads to a singleton-set component, and the log determinant term is 1. We distinguish the data from the two intervention types by superscript (1) or (2). Again, we assume that the sample sizes are $n_1$ and $n_2$, respectively.

For the feasibility of the closed-form block coordinate update on a specific node $i$ with any interventions, the **local** condition of the strongly connected component $\mathcal{C}(i, G)$ being a cycle is sufficient. Under this condition, there is only one parent of $i$ in the cycle: $\mathrm{pa}(i) \cap V[\mathcal{C}(i, G)] = \{j\}$. The block update with respect to $i$ is equivalent to optimizing the univariate profile log-likelihood function of $\lambda_{ji}$, and the optimal values of other entries in $\Lambda_{\mathrm{pa}(i),i}$ are unique determined by the optimal value of $\lambda_{ji}$.

We should clarify that the weaker condition $|\,\mathrm{pa}(i) \cap V[\mathcal{C}(i, G)]| = 1$ is **not** sufficient for degree 2 updating with arbitrary interventions, since a complicated strongly connected component can be modified in different ways with interventions while still keeping the parental edge $j \to i$. The **global** single cycle condition is a combination of all those local conditions on a single node. Actually, we can relax the local condition of each node to the weaker version (intersection having size 1), while still obtain the same global condition of strongly connected components in the graph. This is shown in the following proposition.

**Proposition 4.4.** *In a directed graph $G$, if every node $i$ has at most one parent in its strongly connected component $\mathcal{C}(i)$: $|\,\mathrm{pa}(i) \cap V[\mathcal{C}(i)]| = 1, \forall i \in V$, then each strongly connected component of $G$ is either a singleton set or a directed cycle.*

*Proof.* We only need to prove the result for strongly connected components with more than one node. Pick an arbitrary node $i_1$ in the strongly connected component $\mathcal{C}(i, G)$, the node $i_1$ has at least one parent in $V[\mathcal{C}(i, G)]$ by the strongly connected property. Combining with the at most one parent condition, the node $i_1$ has exactly one parent in $V[\mathcal{C}(i, G)]$, denoted by $i_2$. We can repeat this procedure to find the next node in $\mathcal{C}(i, G)$, until $i_{r+1} = i_s$ is the first duplicate node.

We assert that $s = 1$. Indeed, $i_{r+1} = i_s$ means that there is a cycle $i_s \to i_r \to i_{r-1} \to \cdots \to i_{s+1} \to i_s$. The strongly connected property implies that there exists a directed path from $i_{s-1}$ to $i_r$, and all nodes in the path are in $V[\mathcal{C}(i, G)]$. Suppose the first node that the path intersects the cycle is $i_t$, then $i_t$ has a parent $i_{t+1}$ in the cycle, and another parent in the directed path. The node $i_t$ has two parents in the strongly connected component $\mathcal{C}(i, G)$, which contradicts the parent condition.

Next we show that the strongly connected component $\mathcal{C}(i, G)$ is exactly the cycle $i_1 \to \cdots \to i_r \to i_1$ by contradiction. First, there cannot exist other edges among $i_1, \ldots, i_r$ except the cycle, otherwise some nodes will have two parents in $V[\mathcal{C}(i, G)]$. Second, the component $\mathcal{C}(i, G)$ cannot contain extra nodes. If so, there exists a node $j \in V[\mathcal{C}(i, G)] \backslash \{i_1, \ldots, i_r\}$ that is a parent of some node $i_q$. Then we have $\{i_{q+1}, j\} \subseteq \mathrm{pa}(i_q) \cap V[\mathcal{C}(i, G)]$, which again contradicts the parent condition. □

### 4.3.3 Properties

At each update, the block coordinate descent algorithm finds a local maximum of log-likelihood function with respect to the error variance of one node and its parental edge weights. The value of the log-likelihood function is non-decreasing throughout the iterations. To ensure the algorithm is well-defined, each block update should have an optimal solution with positive $\omega_{ii}$. This condition is equivalent to $\|\mathbf{X}_i - \Lambda_{i,\mathrm{pa}(i)}^T \mathbf{X}_{\mathrm{pa}(i)}\|$ being positive at the update for every $i$, which in turn means $\mathbf{X}_i$ is not in the span of row space of $\mathbf{X}_{\mathrm{pa}(i)}$, i.e., the matrix $\mathbf{X}_{\mathrm{pa}(i)\cup\{i\}}$ has linearly independent rows. This is exactly the condition $(\mathrm{A1})_i$ in Drton et al. [2019b] for directed graphs.

To provide the uniqueness analysis of our block coordinate descent algorithm (Theorem 4.6), we still need a preliminary lemma on the property of a special rational function. The lemma is also used in the proof of Theorem 4.2.

**Lemma 4.5.** *For constants $a, b, c, d, \lambda \in \mathbb{R}$ with $\lambda \neq 0$, $a, c \geq 0$, $b^2 - ac \leq 0$, and sample sizes $n_1, n_2 \in \mathbb{Z}^+$, the function*

$$f(x) = n_1 \frac{1}{x + \frac{d}{\lambda}} - (n_1 + n_2) \frac{ax + b}{ax^2 + 2bx + c}, \quad x \in \mathbb{R}\backslash\{-d/\lambda\},\ ax^2 + 2bx + c \neq 0$$

*has the properties:*

*(i) If $a = 0$, $f(x)$ has no roots. $f(x) < 0$ for $x \in (-\infty, -d/\lambda)$ and $f(x) > 0$ for $x \in (-d/\lambda, \infty)$.*

*(ii) If $a > 0$ and $b^2 - ac < 0$, $f(x)$ has 2 different roots given by*

$$an_2x^2 - \left(b(n_1 - n_2) - \frac{ad}{\lambda}(n_1 + n_2)\right)x - \left(cn_1 - \frac{bd}{\lambda}(n_1 + n_2)\right) = 0,$$

*i.e.,*

$$x_1 = \frac{b\lambda(n_1 - n_2) - ad(n_1 + n_2) - \sqrt{(b\lambda(n_1 - n_2) - ad(n_1 + n_2))^2 + 4a\lambda n_2\,(c\lambda n_1 - bd(n_1 + n_2))}}{2a\lambda n_2},$$

$$x_2 = \frac{b\lambda(n_1 - n_2) - ad(n_1 + n_2) + \sqrt{(b\lambda(n_1 - n_2) - ad(n_1 + n_2))^2 + 4a\lambda n_2\,(c\lambda n_1 - bd(n_1 + n_2))}}{2a\lambda n_2},$$

*where $x_1 < -d/\lambda < x_2$. The function satisfies that $f(x) < 0$ for $x \in (x_1, -d/\lambda) \cup (x_2, \infty)$ and $f(x) > 0$ for $x \in (-\infty, x_1) \cup (-d/\lambda, x_2)$.*

*(iii) If $a > 0$ and $b^2 - ac = 0$ and $\lambda \neq ad/b$, $f(x)$ has a unique root*

$$x^* = \frac{b\lambda n_1 - ad(n_1 + n_2)}{a\lambda n_2}.$$

*If $-d/\lambda < -b/a$, then $f(x) < 0$ in $(x^*, -d/\lambda) \cup (-b/a, \infty)$ and $f(x) > 0$ in $(-\infty, x^*) \cup (-d/\lambda, -b/a)$. If $-d/\lambda > -b/a$, then $f(x) < 0$ in $(-b/a, -d/\lambda) \cup (x^*, \infty)$ and $f(x) > 0$ in $(-\infty, -b/a) \cup (-d/\lambda, x^*)$.*

*(iv) If $a > 0$, $b^2 - ac = 0$ and $\lambda = ad/b$, $f(x)$ has no roots, then $f(x) < 0$ for $x \in (-d/\lambda, \infty)$ and $f(x) > 0$ for $x \in (-\infty, -d/\lambda)$.*

*Proof.* If $a = 0$, then $b$ is also zero and the second term of $f(x)$ cancels. $f(x)$ degenerates to a reciprocal function

$$f(x) = \frac{n_1}{x + \frac{d}{\lambda}}.$$

It is readily apparent that $f(x) < 0$ for $x \in (-\infty, -d/\lambda)$ and $f(x) > 0$ for $x \in (-d/\lambda, \infty)$.

In the following parts we always assume that $a > 0$. We can rewrite $f(x)$ by reduction to the common denominator.

$$f(x) = \frac{-an_2x^2 + \left(b(n_1 - n_2) - \frac{ad}{\lambda}(n_1 + n_2)\right)x + cn_1 - \frac{bd}{\lambda}(n_1 + n_2)}{(x + \frac{d}{\lambda})(ax^2 + 2bx + c)}. \tag{4.15}$$

We denote the numerator by $h(x)$. It is a quadratic function of $x$, with coefficient of $x^2$ smaller than 0.

If $b^2 - ac < 0$ then $ax^2 + 2bx + c > 0$ for all $x \in \mathbb{R}$. We note that $h(-d/\lambda) = n_1(a(-d/\lambda)^2 + 2b(-d/\lambda) + c) > 0$ and $h(x)$ is a downward opening parabola, which has two different roots $x_1 < -d/\lambda < x_2$. We know that $h(x) < 0$ for $x \in (-\infty, x_1) \cup (x_2, \infty)$ and $h(x) > 0$ for $x \in (x_1, -d/\lambda) \cup (-d/\lambda, x_2)$. Combining the fact that the sign of the denominator changes from negative to positive at $-d/\lambda$, we obtain the positive/negative intervals of $f(x)$.

If $b^2 - ac = 0$, we have

$$ax^2 + 2bx + c = a(x + \frac{b}{a})^2,$$

and

$$f(x) = \frac{n_1}{x + \frac{d}{\lambda}} - (n_1 + n_2)\frac{ax + b}{a(x + \frac{b}{a})^2} = \frac{n_1}{x + \frac{d}{\lambda}} - \frac{n_1 + n_2}{x + \frac{b}{a}}.$$

The function $f(x)$ has a unique root

$$x^* = \frac{b\lambda n_1 - ad(n_1 + n_2)}{a\lambda n_2},$$

and $x^* < -d/\lambda < -b/a$ or $x^* > -d/\lambda > -b/a$. If $-d/\lambda < -b/a$, then $f(x) < 0$ in $(x^*, -d/\lambda) \cup (-b/a, \infty)$ and $f(x) > 0$ in $(-\infty, x^*) \cup (-d/\lambda, -b/a)$. If $-d/\lambda > -b/a$, then $f(x) < 0$ in $(-b/a, -d/\lambda) \cup (x^*, \infty)$ and $f(x) > 0$ in $(-\infty, -b/a) \cup (-d/\lambda, x^*)$.

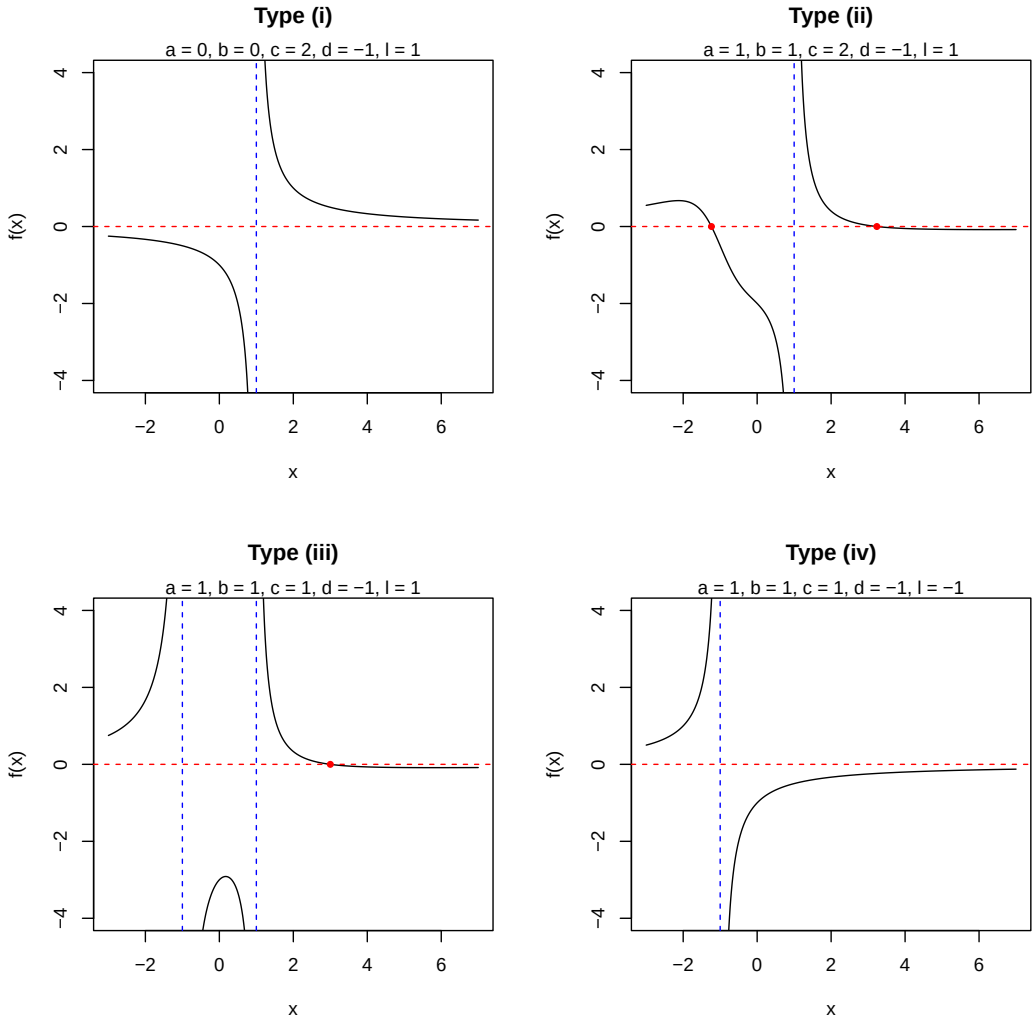



**Figure 4.3:** Examples of functions from the 4 cases in Lemma 4.5. Two roots in case (ii) and one root in case(iii). In case (i) and (iv) the function has no roots.

Finally, if $b^2 - ac = 0$ and $d/\lambda = b/a$, we have that

$$f(x) = -\frac{n_2}{x + \frac{d}{\lambda}}.$$

It has no roots and the positive/negative intervals are obvious. □

The condition for a unique solution is as follows.

**Theorem 4.6.** *Let $G = (V, D)$ be an arbitrary directed graph and $\mathcal{I} = \{I_1, \ldots, I_k\}$ contain the intervention targets. Let $\mathbf{X} = \mathbb{R}^{p \times n}$ be the stacked data matrix of full rank $|V| = p \leq n$. For every node $i \in V$, let the graphical condition (4.6) in Section 4.3.1 hold, and let the data without intervention on $i$ be of size $n_{i,0} \geq p$. We further denote that the total sample sizes for the two types of interventions by $n_{i,1}, n_{i,2}$, both positive, with $n_{i,1} + n_{i,2} = n_{i,0}$. Then the problem (4.8) has a unique solution $(\Lambda_{\mathrm{pa}(i),i}, \omega_{ii})$ if*

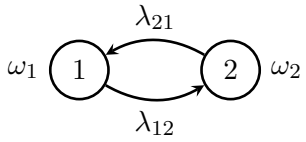



**Figure 4.4:** A 2-cycle with parameters.

*and only if the matrix* $\mathbf{X}_{\mathrm{pa}(i)\cup\{i\}} \in \mathbb{R}^{|\,\mathrm{pa}(i)|\times n_{i,0}}$ *has linearly independent rows. In this case, the solution has the property* $\|\mathbf{X}_i - \Lambda_{i,\mathrm{pa}(i)}^T \mathbf{X}_{\mathrm{pa}(i)}\| > 0$.

*Proof.* We use the notations in Theorem 4.2:

$$a = 1,\ b = (Q_2\mathbf{X}_i^T)_m,\ c = \sum_{j=m}^{n} (Q_2\mathbf{X}_i^T)_j^2, d = c_{i,0},\ \lambda = \|c_{i,\mathrm{pa}(i)}\|/R_{mm}.$$

When $n_{i,1}, n_{i,2} \neq 0$, linearly independent rows of $\mathbf{X}_{\mathrm{pa}(i)\cup i}^{n_{i,0}}$ implies that $ax^2+2bx+c > 0$ for all $x \in \mathbb{R}$ and $b^2 - ac < 0$.

If $\lambda = 0$, i.e., $c_{i,\mathrm{pa}(i)} = 0$, the optimization problem with respect to node $i$ is actually a linear regression. It has a unique solution $\Lambda_{\mathrm{pa}(i),i}^* = \hat{\alpha}$ since $\mathbf{X}_{\mathrm{pa}(i)\cup\{i\}}$ has full rank. Otherwise, Lemma 4.5 states that $g_i$ has two critical points $x_1 < -d/\lambda < x_2$. Both of them are local maximum points, and they are in the two branches of $g_i(x)$ respectively. In practice, the two possible optimal vectors are computed by (4.9) and (4.10). The updated value of $\Lambda_{\mathrm{pa}(i),i} = \alpha^*$ is determined by comparing the two local maxima of the log-likelihood function. □

**Remark 4.7** (2-cycle)**.** *A 2-cycle is a very special structure as it is locally overparameterized. If the 2-cycle, as a strongly connected component, is positioned first in the topological ordering of all components, then after marginalizing all other variables, the cycle contains 4 parameters but with only 3 free entries in the marginal covariance matrix. When each intervention target contains either both variable or neither, the update problem is the same as when considering only observational data. We thus require the intervention on exact one variable for parameter identifiability.*

*Example 3 in Drton et al. [2019b] demonstrates that the update problem is not well-defined for certain values of* $\lambda_{12}$ *(*$\beta_{21}$ *in Drton et al. [2019b], with a different paramterization). While the problematic value may not commonly occur, we have discovered a more significant issue: the update problem has infinite number of solutions and the solutions are determined by the initial value of edge weights within 2 iterations, which essentially arises from the non-identifiability of parameters.*

*Let* $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{1\times n}$ *be the data, the block update problem for node 2 is described in Drton et al. [2019b]. It takes the form*

$$\min_{\lambda_{12}\in\mathbb{R}} \frac{\|\mathbf{X}_2 - \lambda_{12}\mathbf{X}_1\|^2}{(1-\lambda_{21}\lambda_{12})^2},$$

*and the solution is (when* $1 - \lambda_{21}\hat{\lambda}_{12} \neq 0$*)*

$$\lambda_{12}^* = \hat{\lambda}_{12} - \frac{(\mathbf{X}_2\mathbf{X}_2^T - \mathbf{X}_2\mathbf{X}_1^T(\mathbf{X}_1\mathbf{X}_1^T)^{-1}\mathbf{X}_1\mathbf{X}_2^T)(\mathbf{X}_1\mathbf{X}_1^T)^{-1}\lambda_{21}}{1-\lambda_{21}\hat{\lambda}_{12}}$$

*where* $\hat{\lambda}_{12} = (\mathbf{X}_1\mathbf{X}_1^T)^{-1}\mathbf{X}_1\mathbf{X}_2^T$. *Defining* $a = \mathbf{X}_2\mathbf{X}_2^T, b = \mathbf{X}_1\mathbf{X}_2^T, c = \mathbf{X}_1\mathbf{X}_1^T$, *we can simplify the formula above as*

$$\lambda_{12}^* = \frac{b - a\lambda_{21}}{c - b\lambda_{21}}.$$

*It can also be written as*

$$\lambda_{21} = \frac{b - c\lambda_{12}^*}{a - b\lambda_{12}^*},$$

*and this form is exactly the update formula of* $\lambda_{21}$. *Consequently, if we choose the update ordering* $2 \prec 1$, *the new value* $\lambda_{12}^*$ *is determined by the initial value* $\lambda_{21}$ *and the new value* $\lambda_{21}^*$ *is the same as* $\lambda_{21}$. *The algorithm will stop after 2 iterations because the values of edge weights and error variances remain unchanged after the first round of update.*

*Furthermore, the estimated marginal covariance matrix* $\hat{\Sigma}$ *is the same for arbitrary initialization under a fixed update ordering, and its value is exactly the sample marginal covariance matrix.*

## 4.4 Simulation Studies

We compare the performance of the simple aggregation method and our BCD-type algorithms using observational and interventional synthetic data. The measure used for comparison is the root mean square error (RMSE) of the estimate. We adopt the scheme and hyperparameters selection in Drton et al. [2019b] for generating graphs, with some modifications for our setups.

We use 24 different configurations of parameters $(p, n, k, d)$, where $p$ is the number of nodes, $n$ is the sample size of **observational** data, $k$ denotes the length of the unique cycle, and $d$ controls the sparsity. For each graph, we randomly select the number of interventional environments, such that $|\mathcal{I}| \in \{1, 2, 3\}$. Each random intervention target is of size 2 or 3. We then compute the intervened model and simulate data of sample size $\max\{n^{(k)}, p+1\}$, $n^{(k)} \sim U[\lfloor (n+1)/2 \rfloor, n]$ for each intervention target. As a result, the data for each graph consists of both observational and interventional data, **with the total size ranging from approximately** $3n/2$ **to** $4n$. We have made slight changes to the setup of graph size, sample size, and graph sparsity compared to Drton et al. [2019b]. Specifically, we set $p \in \{10, 20\}$, $n \in \{5p/2, 10p\}$ and $k \in \{0, p/10 + 1, p/5 + 2\}$, $d \in \{0.1, 0.2\}$.

In the simulations, we focus on special directed graphs that contain at most one unique cycle. We start with an empty graph and add a $k$-cycle, denoted as $1 \to 2 \to \cdots \to k \to 1$, which forms the first strongly connected component. Next, we consider the remaining $(i, j)$ pairs with $i < j$, which amounts to $p(p-1)/2 - k(k-1)/2$ pairs.

For each pair, we introduce a nonzero edge weight based on independent uniform random variables $U_{ij} \sim U(0, 1)$. Specifically, if $U_{ij} < d$, we add the edge $i \to j$. The sparsity parameter $d \in (0, 1)$ controls the average number of edges in the graph. To ensure randomness, after adding the edges according to the specified topological ordering, we randomly permute the node labels. This construction procedure guarantees that the graph has a unique cycle of length $k$.

For the parameter values, we draw all the free entries of the matrix $\Lambda$ independently from a uniform distribution on the interval $[-2, -0.5] \cup [0.5, 2]$. Similarily, the diagonal entries of the matrix $\Omega$ are randomly drawn from uniform distribution on the interval $[0.3, 1]$. When it comes to intervention, if a target set $I$ is selected, we mask the corresponding rows in $\Lambda$ by setting them to zero, i.e., $\Lambda_{\cdot,I} = 0$. To ensure parameter identifiability, the unique cycle is broken by at least one intervention target. Additionally, the interventional errors $\varepsilon_I$ are sampled from $|I|$ independent standard normal distributions.

| | | | | RMSE | | Converged | | Running time | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | $n$ | $k$ | $d$ | **Agg** | **MLE** | **Agg** | **MLE** | **Agg** | **MLE** |
| 10 | 25 | 0 | 0.1 | 0.1495 | 0.1448 | 1000 | 1000 | 9.122 | 4.997 |
| 10 | 25 | 0 | 0.2 | 0.1569 | 0.1480 | 1000 | 1000 | 12.05 | 6.275 |
| 10 | 25 | 2 | 0.1 | 342.6 | 37.92 | 997 | 1000 | 30.21 | 18.51 |
| 10 | 25 | 2 | 0.2 | 552.5 | 16.44 | 997 | 1000 | 35.92 | 19.96 |
| 10 | 25 | 4 | 0.1 | 18.74 | 0.1741 | 998 | 1000 | 81.55 | 27.74 |
| 10 | 25 | 4 | 0.2 | 826.2 | 0.1704 | 995 | 1000 | 79.07 | 28.41 |
| 10 | 100 | 0 | 0.1 | 0.07241 | 0.07166 | 1000 | 1000 | 9.316 | 5.326 |
| 10 | 100 | 0 | 0.2 | 0.07301 | 0.07191 | 1000 | 1000 | 11.86 | 6.413 |
| 10 | 100 | 2 | 0.1 | 58.82 | 18.45 | 1000 | 1000 | 27.31 | 18.24 |
| 10 | 100 | 2 | 0.2 | 29.00 | 9.924 | 1000 | 1000 | 34.93 | 20.23 |
| 10 | 100 | 4 | 0.1 | 0.1379 | 0.09264 | 997 | 1000 | 101.00 | 30.20 |
| 10 | 100 | 4 | 0.2 | 0.1194 | 0.08530 | 996 | 1000 | 93.66 | 30.85 |
| 20 | 50 | 0 | 0.1 | 0.0967 | 0.09362 | 1000 | 1000 | 27.23 | 13.91 |
| 20 | 50 | 0 | 0.2 | 0.1035 | 0.09595 | 1000 | 1000 | 35.84 | 19.10 |
| 20 | 50 | 3 | 0.1 | 0.1779 | 0.1146 | 997 | 1000 | 104.47 | 38.94 |
| 20 | 50 | 3 | 0.2 | 116.1 | 0.1163 | 993 | 1000 | 113.67 | 46.84 |
| 20 | 50 | 6 | 0.1 | 0.1219 | 0.1004 | 997 | 1000 | 121.35 | 49.08 |
| 20 | 50 | 6 | 0.2 | 0.1160 | 0.09876 | 996 | 1000 | 128.43 | 56.51 |
| 20 | 200 | 0 | 0.1 | 0.04677 | 0.04624 | 1000 | 1000 | 25.24 | 16.10 |
| 20 | 200 | 0 | 0.2 | 0.04868 | 0.04705 | 1000 | 1000 | 36.97 | 22.24 |
| 20 | 200 | 3 | 0.1 | 9.465 | 0.06828 | 999 | 1000 | 111.80 | 44.73 |
| 20 | 200 | 3 | 0.2 | 0.07109 | 0.05826 | 984 | 1000 | 118.50 | 53.46 |
| 20 | 200 | 6 | 0.1 | 0.07196 | 0.05243 | 995 | 1000 | 127.72 | 52.74 |
| 20 | 200 | 6 | 0.2 | 0.06239 | 0.05136 | 990 | 1000 | 134.71 | 59.81 |

**Table 4.1:** Statistics for randomly generated directed graphs with at most one unique cycle. Each row summarizes 1000 simulations. "RMSE" represents the average root mean square error of the estimate for a single parameter among the total 1000 simulations. "Converged" means whether all runnings of algorithms in one single simulation converge or not. Running time is the average CPU time (in milliseconds).

For each configuration $(p, n, k, d)$, we randomly generate 1000 graphs with their corresponding parameters and data. Then we apply the aggregation method and our BCD-type algorithm. We set the maximum number of iterations for each run to be 5000. In each run of the algorithms, we employ various initialization schemes, in-

cluding random initialization, default $(O, I)$ initialization, and adaptive initialization based on the data.

In the simulations, each running of our MLE algorithm converges, whereas some runs of the BCD algorithm for a single observational or interventional environment diverge. Table 4.1 summarizes the simulation results for directed graphs with one unique cycle. The columns "Agg" and "MLE" correspond to the aggregation method and Algorithm 1. The "RMSE" column represents the average root mean square error of the estimate for a single parameter among the 1000 simulations. Our algorithm consistently outperforms the aggregation method in all simulation settings. The "Converged" column records the number of simulations that our algorithm converges for all environments.

The average CPU running times in milliseconds are also recorded (on a 1.6GHz processor). For the aggregation method, the time is the sum of the time taken by the BCD algorithm performed on each observational or interventional dataset, as well as the aggregation steps. In terms of efficiency, Algorithm 1 is faster than the aggregation method, as expected. Although the running time of our algorithm is longer than the running time of BCD algorithm for one single environment, the combination of multiple environments and the post-processing step make the aggregation method slower.

## 4.5 Discussion

### 4.5.1 Mixed graphs

If the graph is directed, intervention on $X_i$ does not influence any other random errors. However, when a node/variable incident to bidirected edges is intervened, the maximum likelihood estimation problem becomes very complicated. An example is shown in Figure 4.5. In the original graph, the BCD algorithm in Drton et al. [2019b] will give $\hat{\omega}_1$ as the MLE of $\omega_1$ because the extra variance from hidden variables is dealt with by the coefficient $\omega_{12}$. However, in the manipulated graph, the estimator $\hat{\omega}_1'$ estimates a value larger than $\omega_1$. There is an implicit inequality constraint: $\hat{\omega}_1 \leq \hat{\omega}_1'$. These types of inequalities appear at every node that has bidirected edges adjacent to a intervened node. Currently, there is no good method to incorporate these constraints in the maximum likelihood estimation procedure when combining observational and interventional data. Therefore, in this chapter we only consider directed graphs.

### 4.5.2 Stability

We do not explicitly discuss parameter identifiability in this chapter, with the exception of the last Remark 4.7 and the description of our simulation settings. In our simulations, we ensure that every cycle is broken by at least one intervention, which guarantees identifiability. Now, we will provide some further analysis on the parameter identifiability of directed cyclic graphs with only observational data.

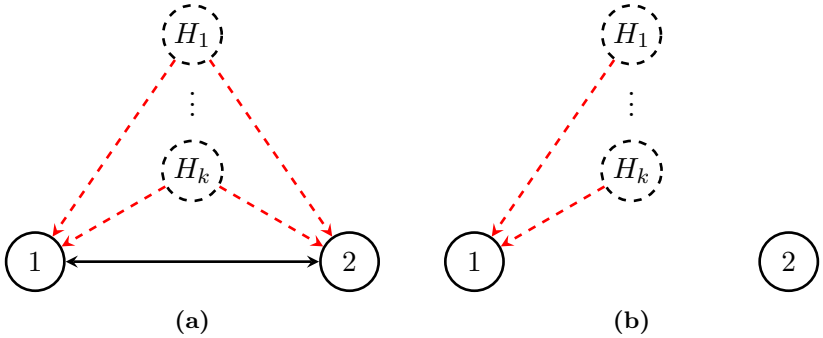



**Figure 4.5:** Original graph with bidirected edge, and manipulated graph with intervention target $I = \{2\}$.

Consider the series expansion

$$(I - \Lambda)^{-1} = I + \Lambda + \Lambda^2 + \cdots,$$

in which $\Lambda^m_{ij}$ is the sum of all directed path from $i$ to $j$ with length $m$ in $G$. The covariance between $i$ and $j$ equals the sum of all treks from $i$ and $j$. One entry, $\Sigma_{ij}$, in the covariance matrix is the sum of products of three entries $(I - \Lambda)^{-T}_{ik}$, $\Omega_{kl}$ and $(I - \Lambda)^{-1}_{lj}$. Each product term corresponds to the concatenation of two directed path and a common source node. If $G$ is acyclic, the series expansion of $(I - \Lambda)^{-1}$ has only finitely many nonzero terms since $\Lambda^p = 0$. But when $G$ contains a cycle, the series has infinitely many nonzero terms.We can restrict ourselves to "stable" models, where the spectral radius satisfies $\rho(\Lambda) < 1$. Under this condition, the series converges, and the model is well-posed.

We start with a directed cycle $G = (V, D)$ with length $p \geq 3$, like that in Proposition 16.2.4 of Sullivant [2018], i.e., the $p$-cycle $1 \to 2 \to 3 \to \cdots \to p - 1 \to p \to 1$. We use the same notations $K = \Sigma^{-1}$, $\Delta = \Omega^{-1}$, $\Delta_{ii} = \delta_{ii}$. Then the parameterization is

$$\begin{aligned} K &= (I - \Lambda)^T \Delta (I - \Lambda) \\ &= \begin{pmatrix} \delta_{11} + \delta_{22}\lambda_{12}^2 & -\delta_{22}\lambda_{12} & 0 & \cdots \\ -\delta_{22}\lambda_{12} & \delta_{22} + \delta_{33}\lambda_{23}^2 & -\delta_{33}\lambda_{23} & \cdots \\ 0 & -\delta_{33}\lambda_{23} & \delta_{33} + \delta_{44}\lambda_{34}^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \end{aligned}$$

and the entrywise equations are

$$K_{ii} = \delta_{ii} + \delta_{i,i+1}\lambda^2_{i,i+1}, \tag{4.16}$$

$$K_{i,i+1} = -\delta_{i+1,i+1}\lambda_{i,i+1}. \tag{4.17}$$

Section 7 in Drton et al. [2011] and Proposition 16.2.4 of Sullivant [2018] consider the same parameterization and state that the equation system has generically two solutions. However, we have the following theorem focusing on the property of the solutions. We claim that at most one of the two solutions is stable.

**Theorem 4.8.** *If the graph $G$ is a $p$-cycle, the equation $\phi_G(\Lambda, \Omega) = K$ has generically two solutions, but at most one of the solutions gives a stable model with $\rho(\Lambda) < 1$.*

*Proof.* Let $(\Lambda_1, \Omega_1)$ and $(\Lambda_2, \Omega_2)$ denote the solutions. To examine the claim, we will derive an invariant of the two possible $\Lambda$'s via (4.16) and (4.17). By Lemma 4.9, we obtain that

$$\begin{aligned}\det(\Lambda_1) \cdot \det(\Lambda_2) &= \prod_{i=1}^{p} \lambda_{i,i+1}^{(1)} \cdot \prod_{i=1}^{p} \lambda_{i,i+1}^{(2)} = \prod_{i=1}^{p} \lambda_{i,i+1}^{(1)} \lambda_{i,i+1}^{(2)} \\ &= \prod_{i=1}^{p} \frac{\det(K_{-(i+1),-(i+1)})}{\det(K_{-i,-i})} = 1. \end{aligned} \tag{4.18}$$

Since the spectral radius of $\Lambda$ is its maximal eigenvalue:

$$\rho(\Lambda) = \left| \prod_{i=1}^{p} \lambda_{i,i+1} \right|^{\frac{1}{p}} = |\det(\Lambda)|,$$

we derive that at most one solution can satisfy the stablility condition $\rho(\Lambda) < 1$ via the determinant product invariant (4.18). □

**Lemma 4.9.** *The two values of $\lambda_{i,i+1}$ in the two solutions satisfy the quadratic equation:*

$$\begin{aligned} K_{i,i+1} \det(K_{-i,-i}) x^2 + \left(\det(K) + 2K_{i,i+1} \det(K_{-i,-(i+1)})\right) x \\ + K_{i,i+1} \det(K_{-(i+1),-(i+1)}) = 0, \end{aligned} \tag{4.19}$$

*where $K_{i,j}$ is the $(i,j)$ entry of $K$, and $K_{-i,-j}$ is the submatrix of $K$ after deleting the $i$'th row and $j$'th column.*

*Proof.* First, if there exist some $i$ such that $K_{i,i+1} = 0$, then (4.17) implies that $\lambda_{i,i+1} = 0$ for all possible solutions. That means the solutions are actually the same as the solutions for an acyclic graph in which there is not edge between $i$ and $i+1$. However, we know that $\Phi_G$ for an acyclic graph is injective, i.e., the equation system has a unique solution.

Then we assume that $K_{i,i+1} \neq 0$ for all $i$, which implies that $\lambda_{i,i+1} \neq 0$ for all $i$. We will prove the result for $i = 1$ by induction on $m$, and then the result holds for general $i$.

Computing $\delta_{ii}$ from (4.17) and plugging into (4.16) for each $i$ produces the equation system

$$K_{ii} + \frac{K_{i-1,i}}{\lambda_{i-1,i}} + K_{i,i+1}\lambda_{i,i+1} = 0. \tag{4.20}$$

If $p = 3$, we can manually eliminate $\lambda_{23}$ and $\lambda_{31}$, and obtain the quadratic equation for $\lambda_{12}$:

$$\begin{aligned} K_{12}(K_{22}K_{33} - K_{23}^2)\lambda_{12}^2 + (K_{11}K_{22}K_{33} + K_{12}^2K_{33} - K_{13}^2K_{22} - K_{23}^2K_{11})\lambda_{12} \\ + K_{12}(K_{11}K_{33} - K_{13}^2) = 0. \end{aligned}$$

Now suppose that (4.19) holds for a general $p \geq 3$, we consider the $(p+1)$-cycle with $(p+1) \times (p+1)$ concentration matrix $K$.

$$K = \begin{pmatrix} K_{11} & K_{12} & 0 & \dots & 0 & K_{p+1,1} \\ K_{12} & K_{22} & K_{23} & \dots & 0 & 0 \\ 0 & K_{23} & K_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ K_{p+1,1} & 0 & \dots & \dots & K_{p,p+1} & K_{p+1,p+1} \end{pmatrix}_{(p+1)\times(p+1)}$$

We eliminate the last variable $\lambda_{m+1,1}$ and reduce the $m+1$ equations to a equation system with $m$ equations, which has the same structure as the equation system for the $m$-cycle.

The equation (4.20) with $i = p+1$ gives that

$$\lambda_{p,p+1} = -\frac{K_{p,p+1}}{K_{p+1,p+1} + K_{p+1,1}\lambda_{p+1,1}}. \tag{4.21}$$

Note that the denominator is nonzero by (4.17). Plugging (4.21) in (4.20) with $i = p$, we obtain that

$$(K_{p+1,p+1} - \frac{K_{p,p+1}^2}{K_{p+1,p+1}}) + \frac{K_{p-1,p}}{\lambda_{p-1,p}} + \frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}} \cdot \frac{K_{p,p+1}\lambda_{p+1,1}}{K_{p+1,p+1} + K_{p+1,1}\lambda_{p+1,1}} = 0. \tag{4.22}$$

We introduce a new variable

$$\lambda' := -\frac{K_{p,p+1}\lambda_{p+1,1}}{K_{p+1,p+1} + K_{p+1,1}\lambda_{p+1,1}},$$

then the edge weight of $(p+1) \to 1$ can be written as

$$\lambda_{p+1,1} = -\frac{K_{p+1,p+1}\lambda'}{K_{p,p+1} + K_{p+1,1}\lambda'}.$$

We can substitute $\lambda_{p+1,1}$ with this expression in (4.20) when $i = 1$. After some calculation and simplification the first equation becomes

$$(K_{11} - \frac{K_{p+1,1}^2}{K_{p+1,p+1}}) - \frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}\lambda'} + K_{12}\lambda_{12} = 0. \tag{4.23}$$

Following the procedures above, we successfully eliminate the variable $\lambda_{p+1,1}$. The new equation system has variables $\{\lambda_{12}, \dots, \lambda_{p-1,p}, \lambda'\}$ and $p$ equations:

$$\begin{cases} (K_{11} - \frac{K_{p+1,1}^2}{K_{p+1,p+1}}) - \frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}\lambda'} + K_{12}\lambda_{12} = 0, \\ K_{22} + \frac{K_{12}}{\lambda_{12}} + K_{23}\lambda_{23} = 0, \\ \vdots \\ K_{p-1,p-1} + \frac{K_{p-2,p-1}}{\lambda_{p-2,p-1}} + K_{p-1,p}\lambda_{p-1,p} = 0, \\ (K_{pp} - \frac{K_{p,p+1}^2}{K_{p+1,p+1}}) + \frac{K_{p-1,p}}{\lambda_{p-1,p}} - \frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}} \cdot \lambda' = 0. \end{cases}$$

Hence, we can apply the inductive hypothesis for $p$-cycle with concentration matrix

$$K' = \begin{pmatrix} K_{11} - \frac{K_{p+1,1}^2}{K_{p+1,p+1}} & K_{12} & 0 & \dots & 0 & -\frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}} \\ K_{12} & K_{22} & K_{23} & \dots & 0 & 0 \\ 0 & K_{23} & K_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}} & 0 & 0 & \dots & K_{p-1,p} & K_{pp} - \frac{K_{p,p+1}^2}{K_{p+1,p+1}} \end{pmatrix}_{p\times p},$$

with respect to variables $\{\lambda_{12}, \lambda_{23}, \dots, \lambda_{p-1,p}, \lambda'\}$.

By our induction hypothesis, $\lambda_{12}$ satisfies the quadratic equation

$$K'_{12}\det(K'_{-1,-1})x^2 + \big(\det(K') + 2K'_{12}\det(K'_{-1,-2})\big)\,x + K'_{12}\det(K'_{-2,-2}) = 0. \quad (4.24)$$

To reveal the connection between $K$ and $K'$, we apply row and column transformations on $K$ as follows:

$$\xrightarrow[C_p - \frac{K_{p,p+1}}{K_{p+1,p+1}}C_{p+1}]{R_p - \frac{K_{p,p+1}}{K_{p+1,p+1}}R_{p+1}} \begin{pmatrix} K_{11} & K_{12} & 0 & \dots & -\frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}} & K_{p+1,1} \\ K_{12} & K_{22} & K_{23} & \dots & 0 & 0 \\ 0 & K_{23} & K_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}} & 0 & \dots & K_{p-1,p} & K_{pp} - \frac{K_{p,p+1}^2}{K_{p+1,p+1}} & 0 \\ K_{p+1,1} & 0 & \dots & 0 & 0 & K_{p+1,p+1} \end{pmatrix}$$

$$\xrightarrow[C_1 - \frac{K_{p,p+1}}{K_{p+1,p+1}}C_{p+1}]{R_1 - \frac{K_{p,p+1}}{K_{p+1,p+1}}R_{p+1}} \begin{pmatrix} K_{11} - \frac{K_{p+1,1}^2}{K_{p+1,p+1}} & K_{12} & 0 & \dots & -\frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}} & 0 \\ K_{12} & K_{22} & K_{23} & \dots & 0 & 0 \\ 0 & K_{23} & K_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{K_{p,p+1}K_{p+1,1}}{K_{p+1,p+1}} & 0 & \dots & K_{p-1,p} & K_{pp} - \frac{K_{p,p+1}^2}{K_{p+1,p+1}} & 0 \\ 0 & 0 & \dots & 0 & 0 & K_{p+1,p+1} \end{pmatrix}$$

$$= \begin{pmatrix} K' & \mathbf{0} \\ \mathbf{0} & K_{p+1,p+1} \end{pmatrix}.$$

Hence, we have

$$\det(K_{-i,-i}) = K_{p+1,p+1}\det(K'_{-i,-i}),\ i = 1, 2.$$

$$\det(K_{-1,-2}) = K_{p+1,p+1}\det(K'_{-1,-2}),$$

and

$$\begin{aligned} \frac{\det(K) + 2K_{12}\det(K_{-1,-2})}{K_{p+1,p+1}} =& K'_{11}\det(K'_{-1,-1}) - K_{12}\det(K'_{-1,-2}) \\ &+ (-1)^{p+1}K'_{1p}\det(K'_{-1,-p}) + 2K_{12}\det(K_{-1,-2}) \\ =& K'_{11}\det(K'_{-1,-1}) - K_{12}\det(K'_{-1,-2}) \\ &+ (-1)^{p+1}K'_{1p}\det(K'_{-1,-p}) + 2K_{12}\det(K'_{-1,-2}) \\ =& \det(K') + 2K'_{12}\det(K'_{-1,-2}). \end{aligned}$$

We immediately obtain that

$$\frac{\det(K') + 2K'_{12}\det(K'_{-1,-2})}{K'_{12}\det(K'_{-1,-1})} = \frac{\det(K) + 2K_{12}\det(K_{-1,-2})}{K_{12}\det(K_{-1,-1})},$$

and

$$\frac{K'_{12}\det(K'_{-2,-2})}{K'_{12}\det(K'_{-1,-1})} = \frac{K_{12}\det(K_{-2,-2})}{K_{12}\det(K_{-1,-1})}.$$

The equation (4.24) is actually the same as

$$K_{12}\det(K_{-1,-1})x^2 + (\det(K) + 2K_{12}\det(K_{-1,-2}))\,x + K_{12}\det(K_{-2,-2}) = 0, \quad (4.25)$$

which is the desired equation for $\lambda_{12}$ in the $(p+1)$-cycle. □

# Chapter 5

# Identifiability of Linear SEMs using Algebraic Matroids

Starting from this chapter we discuss the structural identifiability problems of linear SEMs with error variance assumptions.

Recursive linear structural equation models and the associated directed acyclic graphs (DAGs) play an important role in causal discovery. Without any extra assumption, the classic identifiability result states that DAGs can be identified only up to Markov equivalence classes, when only observation data is available. Recent work proved that, the DAG can be uniquely identified if the errors in the model are **homoscedastic**, i.e., all have the same error variance [Peters and Bühlmann, 2014, Chen et al., 2019]. This result has become a well-known fact and models with equal variances play an important role as 'test beds' for causal discovery. The unique identifiability can be shown to arise from variance accumulation along the topological ordering.

When the space of graphs is expanded to include directed graphs that may contain cycles, the entire mechanism undergoes a change, rendering previous arguments for DAGs impossible. To address this challenge, we define the **linear Gaussian precision model** and propose certification rules for different models using **algebraic matroids**. It is important to note that the notations and definitions in this chapter differ slightly from those in the setup chapter. Furthermore, we need to clarify several new terminologies that are exclusively introduced for this chapter.

## 5.1 Matroid Approach: Preliminaries

**Definition 5.1.** *The **linear Gaussian precision model** given by directed graph $G = (V, D)$ is the family of all multivariate normal distributions on $\mathbb{R}^V$ with a precision matrix (inverse covariance matrix) in the set*

$$M_G = \left\{ K : K = \psi_G(\Lambda, s),\ \Lambda \in \mathbb{R}^D_{\text{reg}} \text{ and } s \in \mathbb{R}^+ \right\}.$$

*The precision matrix parameterization of the model is the map*

$$\begin{aligned} \psi_G : \mathbb{R}^D \times \mathbb{R}^+ &\mapsto PD, \\ (\Lambda, s) &\mapsto s(I - \Lambda)(I - \Lambda)^T, \end{aligned}$$

*where $PD$ is the cone of positive definite symmetric $p \times p$ matrices (recall $|V| = p$).*

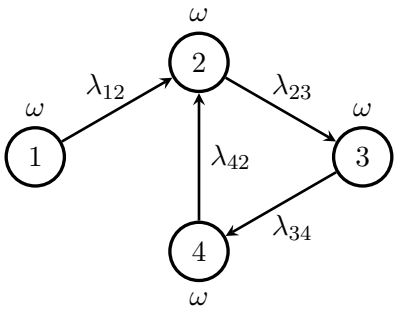



**Figure 5.1:** A directed cyclic graph, with edge weights and equal error variances.

**Example 5.2.** *Consider the following 4-node directed graph in Figure 5.1 and the linear structural equation model, which is non-recursive.*

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= \lambda_{12}X_1 + \lambda_{42}X_4 + \varepsilon_2 \\ X_3 &= \lambda_{23}X_2 + \varepsilon_3 \\ X_4 &= \lambda_{34}X_3 + \varepsilon_4 \end{aligned}$$

*All errors $\varepsilon_i$ follows the same normal distribution $\mathcal{N}(0, \omega)$ so the associated parameters of the model are $(\Lambda, s)$ which have the form*

$$\Lambda = \begin{pmatrix} 0 & \lambda_{12} & 0 & 0 \\ 0 & 0 & \lambda_{23} & 0 \\ 0 & 0 & 0 & \lambda_{34} \\ 0 & \lambda_{42} & 0 & 0 \end{pmatrix}, \qquad s = \frac{1}{\omega}.$$

*Then the model $M_G$ associated to the graph $G$ above consists of precision matrices $K$ of the form*

$$K = s(I-\Lambda)(I-\Lambda)^T = \begin{pmatrix} s(1+\lambda_{12}^2) & -s\lambda_{12} & 0 & s\lambda_{12}\lambda_{42} \\ -s\lambda_{12} & s(1+\lambda_{23}^2) & -s\lambda_{23} & -s\lambda_{42} \\ 0 & -s\lambda_{23} & s(1+\lambda_{34}^2) & -s\lambda_{34} \\ s\lambda_{12}\lambda_{42} & -s\lambda_{42} & -s\lambda_{34} & s(1+\lambda_{42}^2) \end{pmatrix}.$$

Suppose we have a family of models $\{M_i\}_{i=1}^k$ (each corresponding to graph $G_i$) for $p = |V|$ variables, such that all models sit in the same cone $PD$. If $M_{i_1} \cap M_{i_2} = \emptyset$ for each distinct pair $(i_1, i_2)$, we say that the discrete parameter $i$, or the graph collection $\{G_i\}_{i=1}^k$ is **globally identifiable**. This requirement is typically too restrictive and cannot be fulfilled in many cases. The following weaker notion of identifiability in Hollering and Sullivant [2021] is often used instead.

**Definition 5.3.** *Let $\{M_i\}_{i=1}^k$ be a finite set of algebraic models which sit in the same ambient space, the discrete parameter $i$ is **generically identifiable** if for each pair of $(i_1, i_2)$,*

$$\dim(M_{i_1} \cap M_{i_2}) < \min\left(\dim(M_{i_1}), \dim(M_{i_2})\right).$$

When only talking about two graphs, generic identifiablity is also called **model distinguishability** [Sullivant, 2018, Section 16]. The geometric interpretation is that

the intersection of any two models in the family is a Lebesgue measure zero subset of both the models. However, this definition of generic identifiability is not appropriate for our settings. If there are two graphs $G_1$ and $G_2$ with models $M_1 \subsetneq M_2$, they should be distinguishable in practice by regularization, which favors the simpler model with same fitting ability. We instead refer to the notion of **quasi equivalence** in Ghassami et al. [2020], Ng et al. [2020], which requires the intersection of two models has nonzero measure under the Lebesgue measure defined over the **union** of both models. This leads to a definition of generic identifiability that is suitable for directed graphs in our graphical modeling context.

**Definition 5.4.** *Let $\{M_i\}_{i=1}^k$ be a finite set of algebraic models which lie in the cone PD, the parameter i (or the model family) is **generically identifiable** if for each pair of $(i_1, i_2)$,*

$$\dim(M_{i_1} \cap M_{i_2}) < \max\left(\dim(M_{i_1}), \dim(M_{i_2})\right).$$

From the definition we immediately know that two models of different dimensions are generically identifiable. Hence we can focus on the identifiability of models of the same dimension. When $\dim(M_1) = \dim(M_2)$, the min and max functions are actually the same. Two irreducible models of the same dimension must either be equal or have lower dimensional intersections.

The linear SEMs are polynomially parameterized models and hence algebraic models. Each algebraic model $M$ has a **vanishing ideal** $\mathcal{I}(M)$, defined by

$$\mathcal{I}(M) = \{f \in \mathbb{R}[x] : \ f(x) = 0 \text{ for all } x \in M\},$$

which is unique and characterizes the model.

The following well-known proposition illustrates how vanishing ideals can help to certify generic identifiability.

**Proposition 5.5.** *[Sullivant, 2018, Proposition 16.1.12] Let $M_1$ and $M_2$ be two irreducible algebraic models (e.g., parameterized models) which sit inside the same ambient space. If there exist polynomials $f_1$ and $f_2$ such that*

$$f_1 \in \mathcal{I}(M_1) \setminus \mathcal{I}(M_2) \text{ and } f_2 \in \mathcal{I}(M_2) \setminus \mathcal{I}(M_1)$$

*then* $\dim(M_1 \cap M_2) < \min(\dim(M_1), \dim(M_2))$.

However, this proposition is far from an ideal tool since it requires expensive Gröbner basis computations, which is not manageable for models with lots of parameters. In this chapter we follow the approach of using algebraic matroids, which describes the projection of the model $M$ onto all coordinate subspaces. This approach is delineated for phylogenetic models in Hollering and Sullivant [2021], and we summarize the key points here.

**Definition 5.6.** *A **matroid** $\mathcal{M} = (E, \mathcal{I})$ is a pair where $E$ is a finite set and $\mathcal{I} \subseteq 2^E$ satisfies*

*(1) $\emptyset \in \mathcal{I}$,*

*(2) If $I' \subseteq I \in \mathcal{I}$, then $I' \in \mathcal{I}$,*

*(3) If $I_1, I_2 \in \mathcal{I}$ and $|I_2| > |I_1|$, then there exists $e \in I_2 \backslash I_1$ such that $I_1 \cup e \in \mathcal{I}$.*

**Definition 5.7.** *Let $W \subset k^n$ be an irreducible variety over the field $k$ and for $S \subseteq [n]$ let $\pi_S : k^n \to k^{|S|}$ be the projection onto the coordinates in $S$. Let $\overline{\pi_S(W)}$ be the Zariski closure of the projection of $W$. Then the pair $([n], \mathcal{I}_W)$ defines a matroid where*

$$\mathcal{I}_W = \{S \subseteq [n] : \overline{\pi_S(W)} = k^{|S|}\},$$

*which is called the **coordinate projection matroid** of $W$ and denoted by $\mathcal{M}(W)$.*

**Proposition 5.8.** *[Hollering and Sullivant, 2021, Proposition 3.1] Let $M_1$ and $M_2$ be two irreducible algebraic models sit in the same ambient space. Without loss of generality assume that $\dim(M_1) \geq \dim(M_2)$. If there exists a subset $S$ of the coordinates such that*

$$S \in \mathcal{M}(\overline{M_2}) \setminus \mathcal{M}(\overline{M_1}),$$

*then $\dim(M_1 \cap M_2) < \min(\dim(M_1), \dim(M_2))$.*

In our settings, the model $M$ (and variety $\overline{M}$) is parameterized and thus irreducible. Under this condition, there exists another equivalent representation of the coordinate projection matroid $\mathcal{M}(W)$. Since we focus on the case of $\dim(M_1) = \dim(M_2)$, the roles of $M_1$ and $M_2$ are indeed symmetric. Either $S \in \mathcal{M}(\overline{M_2}) \setminus \mathcal{M}(\overline{M_1})$ or $S \in \mathcal{M}(\overline{M_1}) \setminus \mathcal{M}(\overline{M_2})$ implies generic identifiability.

**Proposition 5.9.** *[Hollering and Sullivant, 2021, Proposition 2.8][Rosen, 2014] Suppose that $\phi(\theta_1, \ldots, \theta_d) = (\phi_1(\theta), \ldots, \phi_n(\theta))$ parameterizes $W$ (i.e., $W = \overline{\phi(k^d)}$). Let*

$$J(\phi) = \left(\frac{\partial \phi_j}{\partial \theta_i}\right), \ 1 \leq i \leq d, \ 1 \leq j \leq n$$

*be the transpose of the Jacobian matrix of $\phi$. Then the matroid defined by the columns of the matrix $J(\phi)$ using linear independence over the fraction field $Frac(k[\theta]) = k(\theta)$ gives the same matroid as the coordinate projection matroid $\mathcal{M}(\overline{\phi(k^d)})$. We call it the **Jacobian matroid** of $W$.*

## 5.2 Jacobian

Throughout this chapter we use the term "Jacobian" to refer to the transpose of the usual Jacobian matrix. The (transposed) Jacobian matrix, denoted by $J(\psi_G)$, is of size $(|D|+1) \times \frac{p(p+1)}{2}$. Each column of $J(\psi_G)$ corresponds to one entry $K_{ij}$ $(i \leq j)$ of the precision matrix and each row corresponds to one edge weight $\lambda_{kl}$ or the inverse of common error variance $s$. If there is no ambiguity of index, we will write $J$ for the Jacobian $J(\psi_G)$. When there are Jacobians of different graphs, we use superscript to distinguish them.

**Lemma 5.10.** *Let $G = (V, D)$ be a directed graph. Then the entries of $J = J(\psi_G)$ are given by*

*(i) For $i \in V$ and $k \to l \in D$,*

$$J_{\lambda_{kl}, K_{ii}} = \begin{cases} 2s\lambda_{il}, & k = i, \\ 0, & else. \end{cases}$$

*(ii) For $i, j \in V, i \neq j$ and $k \to l \in D$,*

$$J_{\lambda_{kl}, K_{ij}} = \begin{cases} -s, & \{k,l\} = \{i,j\}, \\ s\lambda_{jl}, & k = i \text{ and } j \to l \in D, \\ s\lambda_{il}, & k = j \text{ and } i \to l \in D, \\ 0, & \text{else.} \end{cases}$$

*(iii) The partial derivatives w.r.t. the inverse error variance $s$ are*

$$J_{s,K_{ii}} = \left(1 + \sum_{l \in \mathrm{ch}(i)} \lambda_{il}^2\right),$$
$$J_{s,K_{ij}} = \left(-\lambda_{ij} 1_{\{(i,j)\in D\}} - \lambda_{ji} 1_{\{(j,i)\in D\}} + \sum_{l \in \mathrm{ch}(i) \cap \mathrm{ch}(j)} \lambda_{il}\lambda_{jl}\right).$$

The lemma gives the nonzero patterns of $J(\psi_G)$. The columns of $K_{ii}$ have nonzero entries in the row of its outgoing edges and inverse variance. In the columns of $K_{ij}$, $i \neq j$ (in general we do not know whether $i$ or $j$ is larger, but $K_{ij}$ and $K_{ji}$ refer to the same column), the entry in the row $\lambda_{ij}$ (or $\lambda_{ji}$) is nonzero if and only if $i \to j \in D$ (or $j \to i \in D$); other nonzero entries are the row pairs with those edges pointing to a common child of $i$ and $j$. To make the matrix simpler without changing the column independence relations, we perform some row transformations (left multiplications) and have the following lemma.

**Lemma 5.11.** *Let $G = (V, D)$ be a graph, and let $J$ be the corresponding Jacobian. Let $R_{\lambda_{ij}}$ be the row of $J$ corresponding to the egde $\lambda_{ij} \in D$, and let $R_s$ be the row corresponding to $s$. Then after performing the sequence of row operations $R_s \to R_s - \frac{\lambda_{ij}}{2s} R_{\lambda_{ij}}$ and $R_s \to 2R_s$, the new entries of row $R_s$ are given by*

$$J_{s,K_{i,j}} = \begin{cases} 2, & \text{if } i = j, \\ -\lambda_{ij}, & \text{if } i \to j \in D, \\ 0, & \text{otherwise.} \end{cases}$$

*Proof.*

(i) First we assume that $i = j$, by Lemma 5.10 the entry $J_{s,K_{ii}}$ is $1 + \sum_{l \in \mathrm{ch}(i)} \lambda_{il}^2$ and the entries $J_{\lambda_{ij},K_{ii}}$ are either 0 or $2s\lambda_{ij}$. After the sequence of operations, all the square terms are cancelled and the constant 1 becomes 2.

(ii) Now we consider the case that $i \to j \in D$. By Lemma 5.10, the entry $J_{s,K_{ij}}$ before this sequence of row operations is

$$J_{s,K_{ij}} = -\lambda_{ij} + \sum_{l \in \mathrm{ch}(i) \cap \mathrm{ch}(j)} \lambda_{il}\lambda_{jl}.$$

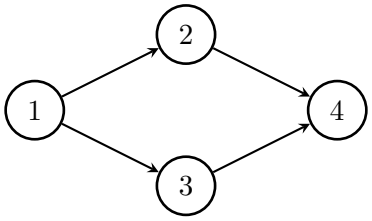



**Figure 5.2:** The diamond graph $G$ used in Example 5.12.

After applying this sequence of row operations means we must add the quantity

$$\sum_{(k,l)\in D} -\frac{\lambda_{kl}}{-2s} J_{\lambda_{kl},K_{ij}} = -\frac{\lambda_{ij}}{-2s}(-s) + \sum_{l\in \mathrm{ch}(i)\cap \mathrm{ch}(j)} -\frac{\lambda_{il}}{-2s} s\lambda_{jl} + \sum_{l\in \mathrm{ch}(i)\cap \mathrm{ch}(j)} -\frac{\lambda_{jl}}{-2s} s\lambda_{il}$$
$$= \frac{\lambda_{ij}}{2} - \sum_{l\in \mathrm{ch}(i)\cap \mathrm{ch}(j)} \lambda_{il}\lambda_{jl}$$

to $J_{s,K_{ij}}$. The entry becomes $\frac{-\lambda_{ij}}{2}$ and the factor 2 is multiplied in the last operation.

(iii) The last case is $i \neq j$ and $i, j$ are not connected. It is similar to case (ii) but without the $-\lambda_{ij}$ term. And for the added quantity in the operations, the $J_{\lambda_{ij},K_{ij}}$ entry does not exist. All terms cancels out and the result is 0.

□

The following example illustrates Lemma 5.10 and 5.11.

**Example 5.12.** *Consider the graph $G = (V, D)$ where $V = \{1, 2, 3, 4\}$ and $D = \{(1, 2), (2, 4), (1, 3), (3, 4)\}$, which is also pictured in Figure 5.2. The original Jacobian $J(\psi)$ is given by*

$$\begin{array}{c} \begin{array}{cccccccccc} K_{11} & K_{22} & K_{33} & K_{44} & K_{12} & K_{23} & K_{34} & K_{13} & K_{24} & K_{14} \end{array} \\ \begin{pmatrix} 2s\lambda_{12} & 0 & 0 & 0 & -s & 0 & 0 & 0 & 0 & 0 \\ 2s\lambda_{13} & 0 & 0 & 0 & 0 & 0 & 0 & -s & 0 & 0 \\ 0 & 2s\lambda_{24} & 0 & 0 & 0 & s\lambda_{34} & 0 & 0 & -s & 0 \\ 0 & 0 & 2s\lambda_{34} & 0 & 0 & s\lambda_{24} & -s & 0 & 0 & 0 \\ 1+\lambda_{12}^2+\lambda_{13}^2 & 1+\lambda_{24}^2 & 1+\lambda_{34}^2 & 1 & -\lambda_{12} & \lambda_{24}\lambda_{34} & -\lambda_{34} & -\lambda_{13} & -\lambda_{24} & 0 \end{pmatrix} \end{array} \begin{array}{c} \\ \begin{pmatrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{24} \\ \lambda_{34} \\ s \end{pmatrix} \end{array},$$

*and the transformed version $(Q \cdot J(\psi))$ is*

$$\begin{array}{c} \begin{array}{cccccccccc} K_{11} & K_{22} & K_{33} & K_{44} & K_{12} & K_{23} & K_{34} & K_{13} & K_{24} & K_{14} \end{array} \\ \begin{pmatrix} 2s\lambda_{12} & 0 & 0 & 0 & -s & 0 & 0 & 0 & 0 & 0 \\ 2s\lambda_{13} & 0 & 0 & 0 & 0 & 0 & 0 & -s & 0 & 0 \\ 0 & 2s\lambda_{24} & 0 & 0 & 0 & s\lambda_{34} & 0 & 0 & -s & 0 \\ 0 & 0 & 2s\lambda_{34} & 0 & 0 & s\lambda_{24} & -s & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & -\lambda_{12} & 0 & -\lambda_{34} & -\lambda_{13} & -\lambda_{24} & 0 \end{pmatrix} \end{array} \begin{array}{c} \\ \begin{pmatrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{24} \\ \lambda_{34} \\ s \end{pmatrix} \end{array}.$$

The Jacobian matroid contains all sets whose corresponding columns in the Jacobian are linearly independent. Its size can be very large even for small graphs. Every matroid is uniquely determined by its **bases**, which are the maximal independent sets with respect to inclusion (any subset of a maximal independent set is still an independent set). Each basis has the same cardinality which is called the **rank** of

the matroid, and in this case it is the rank of the Jacobian evaluated at a generic point, or equivalently, the dimension of the model. For a simple directed graph, a very favorable property is that the associated models have the dimension equaling to the number of parameter counts.

**Theorem 5.13.** *Let $G = (V, D)$ be a simple directed graph and $\mathcal{M}(\psi_G)$ be the Jacobian matroid of $\psi_G$. Then the Jacobian $J$ has full rank, thus any basis of the $\mathcal{M}(\psi_G)$ is of size $|D| + 1$.*

*Proof.* By Lemma 5.10, we know that each edge $(i, j)$ in $G$ leads to a $-s$ term in column $K_{ij}$. We use the special paramaterization $\lambda_{ij} = 0$ and $s = 1$ and consider the $(|D| + 1) \times (|D| + 1)$ submatrix $J_S$ with columns corresponding to the set $S = \{K_{ij} \mid i \to j \text{ or } j \to i \in D\} \cup \{K_{ii}\}$ for any choice of $i \in V$. It is an identity matrix up to row and permutations, hence it has rank $|D| + 1$. The submatrix has generic full rank, and the Jacobian itself also has generic full row rank (over the fraction field $k(\lambda, s)$). □

**Remark 5.14.** *This theorem shows that the model of a simple directed graph is of expected dimension. However, a similar result does not hold for non-simple directed graphs, even if the number of edges is smaller than that of the complete graph. In general, we only know that the model of a non-simple directed graph has dimension no larger than $|D| + 1$. If a simple graph $G_1$ and a non-simple graph $G_2$ have the same model dimension, then we know that $|D_2| \geq |D_1|$.*

We conclude this subsection by stating some trivial necessary conditions for two graphs to have the same matroid.

**Lemma 5.15.** *Let $G_1 = (V, D_1)$, $G_2 = (V, D_2)$ be two directed graphs with the same Jacobian matroid $\mathcal{M}$. If two node $i$ and $j$ are adjacent or have common children in one graph, then they must be adjacent or have common children in the other graph.*

*Proof.* If $i$ and $j$ are adjacent or have common children in $G_1$, then $K_{ij}$ is an independent set in matroid $\mathcal{M}_1$, and also in $\mathcal{M}_2$. By Lemma 5.10 we know that $i$ and $j$ are adjacent or have common children in $G_2$. □

**Lemma 5.16.** *Let $G_1 = (V, D_1)$, $G_2 = (V, D_2)$ be two directed graphs with the same Jacobian matroid $\mathcal{M}$. If the node $i$ is a sink node in both graphs, then $\mathrm{pa}_1(i) = \mathrm{pa}_2(i)$.*

*Proof.* For every node $k \in \mathrm{pa}_1(i)$, by Lemma 5.15 we know that $k$ and $i$ are also adjacent or have common children in $G_2$. But $i$ is a sink node in $G_2$ and have no children. Hence the only possible case is $k \in \mathrm{pa}_2(i)$. □

## 5.3 Graphical Conditions for Distinguishing SEMs with Matroids

This section presents some nontrivial sufficient conditions for two graphs to have different matroids. We begin with two preliminary lemmas, followed by the core out-degree theorem which states that two non-complete graphs with different out-degree

sequences have different matroids. Finally, we discuss some corollaries and extensions involving the concept of parentally closed sets.

In the previous section, we prove that the Jacobian of a simple directed graph has full rank by considering all columns corresponding to the edges. An intuitive idea is that the edge structure provides information about independent sets, and we can search for a maximal independent set for one of the two matroids, by selecting special columns of edges. However, the presence of collider triples presents significant challenges. Being a maximal independent set in one matroid is too restrictive, and it does not provide desired information for the same induced submatrix in the other Jacobian. Instead, we aim to find a column set that yields different rank submatrices in the two Jacobians. This approach leads to the following two lemmas and the outdegree theorem.

**Lemma 5.17.** *Let $G = (V, D)$ be a directed graph whose corresponding model has dimension $|D| + 1$, and let $J$ be the associated Jacobian. If $G$ is not complete, then for every node $i$ and any subset of the columns $S$ of size $|D| + 1$ such that $S \cap \{K_{i1}, K_{i2}, \dots, K_{i(i-1)}, K_{ii}, K_{i(i+1)}, \dots\} = \emptyset$, it holds that* $\operatorname{rank}(J_S) \leq |D| - |\operatorname{ch}(i)| + 1$.

*Proof.* For any choice of $k, l \neq i$, Lemma 5.10 implies that $J_{\lambda_{ij}, K_{kl}} = 0$. We notice that every column $K_{kl}$ satisfies the condition by the construction of $S$. Hence all the rows in $J_S$ corresponding to $\lambda_{ij}$ with $j \in \operatorname{ch}(i)$ are zero. The submatrix $J_S$ has at least $|\operatorname{ch}(i)|$ zero rows and has rank $\operatorname{rank}(J_S) \leq |D| - |\operatorname{ch}(i)| + 1$. □

**Lemma 5.18.** *Let $G = (V, D)$ be a simple directed graph, and let $J$ be the associated Jacobian. If $G$ is not complete, then for every node $i$, there exists a set of columns $S$ of size $|D| + 1$ such that $S \cap \{K_{i1}, K_{i2}, \dots, K_{i(i-1)}, K_{ii}, K_{i(i+1)}, \dots\} = \emptyset$ and* $\operatorname{rank}(J_S) \geq |D| - |\operatorname{ch}(i)| + 1$.

*Proof.* We construct a set $S$ satisfying the restriction and prove the rank property of $J_S$. Let $S_E = \{K_{kl} \mid (k, l) \in D \text{ or } (l, k) \in D\}$, and let $S_- \subseteq S_E$ be the subset consisting of columns of the form $K_{ki}$. Therefore, $S_-$ corresponds to the adjacent pairs $\{i, k\}$ and $|S_-| = \deg(i)$. We remove the subset $S_-$ from $S_E$ and add some other columns to make the size to be $|D| + 1$. There are two cases based on the value of $\deg(i)$. For the matrix computation in the proof, we present the large intermediate matrix in Section 5.5.

First consider the case where $\deg(i) < p - 1$. There is a node $j_0$ not adjacent to $i$. Let $S_+$ be a subset of size $(\deg(i) + 1) \leq p - 1$ of set $\{K_{11}, K_{22}, \dots, K_{i-1,i-1}, K_{i+1,i+1}, \dots\}$ that contains all $K_{jj}$ such that $(j, i) \in D$ or $(i, j) \in D$ and $K_{j_0 j_0}$. The set $S = (S_E \backslash S_-) \cup S_+$ is of size $|D| + 1$. Then the submatrix $J_S$ evaluated at $s = 1, \lambda_{ji} = \varepsilon > 0$ and all other edge weights 0 is

$$J_S = \cdots = \begin{array}{c} \begin{matrix} K_{j_0j_0} & \cdots & K_{j_1j_1} & \cdots & K_{j_qj_q} & K_{k_1l_1} & \cdots & K_{k_ml_m} \end{matrix} \\ \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 & -1 & \cdots & O(\varepsilon) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & O(\varepsilon) & \cdots & -1 \\ 0 & \cdots & 2\varepsilon & \cdots & 0 & \times & \cdots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & 2\varepsilon & \times & \cdots & \times \\ 0 & \cdots & \times & \cdots & \times & \times & \cdots & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \times & \cdots & \times & \times & \cdots & \times \\ 1 & \cdots & 2 & \cdots & 2 & 0 & \cdots & 0 \end{pmatrix} \end{array} \begin{array}{c} \\ \begin{pmatrix} \lambda_{k_1l_1} \\ \vdots \\ \lambda_{k_ml_n} \\ \lambda_{j_1i} \\ \vdots \\ \lambda_{j_qi} \\ \lambda_{in_1} \\ \vdots \\ \lambda_{in_p} \\ s \end{pmatrix} \end{array} . \quad (5.1)$$

After a reshuffling of the rows, the submatrix of $J_S$ corresponding to the rows labelled by $\{\lambda_{k_1l_1}, \ldots, \lambda_{k_ml_m}, \lambda_{j_1i}, \ldots, \lambda_{j_qi}, s\}$ is a block upper triangular matrix. The diagonal blocks correspond to the rows whose edges labels do not involve $i$ (labelled by the $\lambda_{k_\alpha,l_\alpha}$), the parents of $i$, and $s$. The block corresponding to the edges which do not involve $i$ is strictly diagonally dominant for a small enough choice of $\varepsilon$, and the other two blocks are scalar multiples of the identity matrix. This submatrix of $J_S$ is of full rank, and we have that $\operatorname{rank}(J_S) \geq (|D| - \deg(i)) + |\operatorname{pa}(i)| + 1 = |D| - |\operatorname{ch}(i)| + 1$.

If $\deg(i) = |V| - 1$ and $i$ is not a sink node, there exists some $j_0 \in \operatorname{ch}(i)$. Let $S_+ = \{K_{11}, K_{22}, \ldots, K_{(i-1)(i-1)}, K_{(i+1)(i+1)}, \ldots\}$ and consider the set $S = (S_E \setminus S_-) \cup S_+$. The set $S$ is of size $|D|$, while the submatrix $J_S$ is of the same form as that in the first case (equation (5.1)) and $\operatorname{rank}(J_S) \geq |D| - |\operatorname{ch}(i)| + 1$. Adding a column $K_{xy}$, $x, y \neq i$ to the set $S$ will not decrease the rank.

Lastly suppose that $\deg(i) = p-1$ and $i$ is a sink node in $G$. We take $S = (S_E \setminus S_-) \cup S_+$ with $S_+ = \{K_{11}, K_{22}, \ldots, K_{(i-1)(i-1)}, K_{(i+1)(i+1)}, \ldots\} \cup \{K_{xy}\}$ such that $x, y \neq i$ and $x, y$ are not adjacent in $G$. Since $G$ is not complete, such a pair of vertices is guaranteed to exist. The submatrix $J_S$ with $s = 1, \lambda_{ji} = \varepsilon > 0$ and all other edge

weights 0 is

$$J_S = \cdots =$$

$$\begin{array}{c} \begin{matrix} K_{xy} & \cdots & K_{xx} & K_{yy} & K_{j_1 j_1} & \cdots & K_{j_q j_q} & K_{k_1 l_1} & \cdots & K_{k_m l_m} \end{matrix} \\ \begin{pmatrix} 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 & \cdots & O(\varepsilon) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & O(\varepsilon) & \cdots & -1 \\ 1 & \cdots & 2\varepsilon & 0 & 0 & \cdots & 0 & \times & \cdots & \times \\ 1 & \cdots & 0 & 2\varepsilon & 0 & \cdots & 0 & \times & \cdots & \times \\ 0 & \cdots & 0 & 0 & 2\varepsilon & \cdots & 0 & \times & \cdots & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 2\varepsilon & \times & \cdots & \times \\ 0 & \cdots & \times & \times & \times & \cdots & \times & \times & \cdots & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \times & \times & \times & \cdots & \times & \times & \cdots & \times \\ 1 & \cdots & 2 & 2 & 2 & \cdots & 2 & 0 & \cdots & 0 \end{pmatrix} \end{array} \begin{matrix} \\ \lambda_{k_1 l_1} \\ \vdots \\ \lambda_{k_m l_m} \\ \lambda_{xi} \\ \lambda_{yi} \\ \lambda_{j_1 i} \\ \vdots \\ \lambda_{j_q i} \\ \lambda_{in_1} \\ \vdots \\ \lambda_{in_p} \\ s \end{matrix} . \quad (5.2)$$

The two 1's of column $K_{xy}$ in rows $\lambda_{xi}$ and $\lambda_{yi}$ can be eliminated by subtracting a multiple of columns $\{K_{xx}, K_{yy}\}$. This may introduce nonzero entries in column $K_{xy}$ in the rows corresponding to the children of $i$, but which can also be eliminated using the row corresponding to $s$. These procedures make $J_S$ the same form as in the first case (equation (5.1)) and give the same rank inequality. □

**Remark 5.19.** *Lemma 5.18 also holds for non-simple graphs which satisfy* $\deg(i) < p$, $|D| \leq \binom{|V|}{2}$, *and have expected dimension* $|D| + 1$.

The following example illustrates the proofs of Lemma 5.17 and Lemma 5.18.

**Example 5.20.** *We again consider the graph pictured in Example 5.2 and set* $i = 3$ *which satisfies* $\deg(3) = 2 < 3 = |V| - 1$. *The Jacobian* $J(\psi)$ *is:*

$$\begin{array}{c} \begin{matrix} K_{11} & K_{22} & K_{33} & K_{44} & K_{12} & K_{23} & K_{34} & K_{13} & K_{24} & K_{14} \end{matrix} \\ \begin{pmatrix} 2s\lambda_{12} & 0 & 0 & 0 & -s & 0 & 0 & 0 & 0 & 0 \\ 0 & 2s\lambda_{23} & 0 & 0 & 0 & -s & 0 & 0 & 0 & 0 \\ 0 & 2s\lambda_{24} & 0 & 0 & 0 & s\lambda_{34} & 0 & 0 & -s & 0 \\ 0 & 0 & 2s\lambda_{34} & 0 & 0 & s\lambda_{24} & -s & 0 & 0 & 0 \\ 1+\lambda_{12}^2 & 1+\lambda_{23}^2+\lambda_{24}^2 & 1+\lambda_{34}^2 & 1 & -\lambda_{12} & -\lambda_{23}+\lambda_{24}\lambda_{34} & -\lambda_{34} & 0 & -\lambda_{24} & 0 \end{pmatrix} \end{array} \begin{matrix} \\ \lambda_{12} \\ \lambda_{23} \\ \lambda_{24} \\ \lambda_{34} \\ s \end{matrix}$$

*Lemma 5.18 guarantees that there exists a set* $S \cap \{K_{13}, K_{23}, K_{33}, K_{34}\} = \emptyset$ *and* $\operatorname{rank}(J_S) \geq |D| - |\operatorname{ch}(3)| + 1 = 4 - 1 + 1 = 4$. *From the proof of the lemma, the set* $S$ *which is constructed is* $S = (S_E \setminus S_-) \cup S_+$ *where*

$$S_E = \{K_{12}, K_{23}, K_{34}, K_{24}\},$$
$$S_- = \{K_{23}, K_{34}\},$$

*and* $S_+$ *can be any subset of size* $\deg(3)+1=3$ *from* $\{K_{11}, K_{22}, K_{44}\}$ *which means it must be the entire set. Thus* $S = \{K_{11}, K_{22}, K_{44}, K_{12}, K_{24}\}$ *and the submatrix* $J_S$ *is:*

$$\begin{array}{c} \begin{array}{ccccc} K_{11} & K_{22} & K_{44} & K_{12} & K_{24} \end{array} \\ \begin{pmatrix} 2s\lambda_{12} & 0 & 0 & -s & 0 \\ 0 & 2s\lambda_{24} & 0 & 0 & -s \\ 0 & 2s\lambda_{23} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1+\lambda_{12}^2 & 1+\lambda_{23}^2+\lambda_{24}^2 & 1 & -\lambda_{12} & -\lambda_{24} \end{pmatrix} \end{array} \begin{array}{c} \\ \lambda_{12} \\ \lambda_{24} \\ \lambda_{23} \\ \lambda_{34} \\ s \end{array} .$$

*Observe that if we substitute in* $s=1$, $\lambda_{ji}=\epsilon$, *and let all other edge weights be zero, then the submatrix becomes*

$$\begin{array}{c} \begin{array}{ccccc} K_{11} & K_{22} & K_{44} & K_{12} & K_{24} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 2\varepsilon & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1+\varepsilon & 1 & 0 & 0 \end{pmatrix} \end{array} \begin{array}{c} \\ \lambda_{12} \\ \lambda_{24} \\ \lambda_{23} \\ \lambda_{34} \\ s \end{array} ,$$

*and we can now see that columns* $K_{11}, K_{22}, K_{12}, K_{44}$ *are linearly independent thus* $\operatorname{rank}(J_S) \geq 4$. *On the other hand, observe that the row corresponding to* $\lambda_{34}$ *is a zero row and so* $\operatorname{rank}(J_S) \leq 4$ *which is guaranteed by Lemma 5.17.*

Lemma 5.17 and Lemma 5.18 immediately lead to the following outdegree theorem, which is the main theorem in this chapter.

**Theorem 5.21. (Outdegree theorem)** *Let* $G_1 = (V, D_1)$, $G_2 = (V, D_2)$ *be two simple directed graphs. If one of the graphs is not complete and there exists a node* $i \in V$ *such that* $|\operatorname{ch}_1(i)| \neq |\operatorname{ch}_2(i)|$, *then* $G_1$ *and* $G_2$ *have different Jacobian matroids.*

*Proof.* If $|D_1| \neq |D_2|$, the two models must have different dimension and hence different matroids. We only need to consider the case $|D_1| = |D_2|$.

Without loss of generality we assume that $|\operatorname{ch}_1(i)| > |\operatorname{ch}_2(i)|$. By Lemma 5.18, there exists a column set $S = (S_E \backslash S_-) \cup S_+$ such that $\operatorname{rank}(J_S^{(2)}) \geq |D_2| - |\operatorname{ch}_2(i)| + 1$ but by Lemma 5.17 we know that $\operatorname{rank}(J_S^{(1)}) \leq |D_1| - |\operatorname{ch}_1(i)| + 1$. Hence, we have that

$$\operatorname{rank}(J_S^{(2)}) \geq |D_2| - |\operatorname{ch}_2(i)| + 1 > |D_1| - |\operatorname{ch}_1(i)| + 1 \leq \operatorname{rank}(J_S^{(1)}),$$

and thus $J^{(1)}$ and $J^{(2)}$ have different matroids. □

Indeed, the conditions can be extended to non-simple graphs with expected dimension $|D| + 1$ and degree of each node smaller than $p$. But we will omit the details since we are primarily focused on simple graphs. When we consider the sink nodes in the graph, i.e., nodes with empty children set, we immediately have the corollary.

**Corollary 5.22.** *If two simple directed graphs have the same Jacobian matroid, and at least one of them is not complete, then they must have the same sink nodes.*


**(a)** $G_1$ **(b)** $G_2$

**Figure 5.3:** Two 4-cycle of different directions, the key edges for selecting column set are highlighted in red.

A large proportion of the possible pairs of graphs (Tables 5.1 and 5.2) can be certified to give a different Jacobian matroid by Theorem 5.21. Indeed, we can compute the ratio of same outdegree graph pairs and all pairs for small size graphs. The values for $p = 5$ or 6 are smaller than 1/100. However, in terms of absolute amount there also remain many pairs for which the matroids are different but Theorem 5.21 cannot be applied to recognize this difference. The following example is a trivial and typical one.

**Example 5.23.** *(The n-cycles of two directions) For $n = 4$, we cannot use the approach as that in the proof of Theorem 5.21 to find different independence set for 4-cycles $1 \to 2 \to 3 \to 4 \to 1$ and $1 \leftarrow 2 \leftarrow 3 \leftarrow 4 \leftarrow 1$. Specifically, the selected column set $S = (S_E \backslash S_-) \cup S_+$ is $\{22, 33, 44, 23, 34\}$. The two corresponding submatrices indexed by $S$ (denoted by $M_1$ and $M_2$) both have rank 4.*

$$\begin{array}{c} \begin{matrix} K_{22} & K_{33} & K_{44} & K_{23} & K_{34} \end{matrix} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 2s\lambda_{23} & 0 & 0 & -s & 0 \\ 0 & 2s\lambda_{34} & 0 & 0 & 0 \\ 0 & 0 & 2s\lambda_{41} & 0 & 0 \\ 1+\lambda_{23}^2 & 1+\lambda_{34}^2 & 1+\lambda_{41}^2 & -\lambda_{23} & -\lambda_{34} \end{pmatrix} \begin{matrix} \lambda_{12} \\ \lambda_{23} \\ \lambda_{34} \\ \lambda_{41} \\ s \end{matrix} \end{array},$$

$$\begin{array}{c} \begin{matrix} K_{22} & K_{33} & K_{44} & K_{23} & K_{34} \end{matrix} \\ \begin{pmatrix} 2s\lambda_{21} & 0 & 0 & 0 & 0 \\ 0 & 2s\lambda_{32} & 0 & -s & 0 \\ 0 & 0 & 2s\lambda_{43} & 0 & -s \\ 0 & 0 & 0 & 0 & 0 \\ 1+\lambda_{21}^2 & 1+\lambda_{32}^2 & 1+\lambda_{43}^2 & -\lambda_{32} & -\lambda_{43} \end{pmatrix} \begin{matrix} \lambda_{21} \\ \lambda_{32} \\ \lambda_{43} \\ \lambda_{14} \\ s \end{matrix} \end{array}.$$

*The submatrix $M_1$ has a zero row of $\lambda_{12}$ and the submatrix $M_2$ has a zero row of $\lambda_{14}$. Noticing that there exists no v-structure in 4-cycles, if we substitute some column $K_{ii}$ by any $K_{ij}$ in the selected column set, the zero row of $M_1$ remains unchanged unless $i = 1, j = 2$. We can select the column set $\{22, 33, 23, 34, 14\}$, then the old zero row of $M_2$ has an nonzero entry, which is not the case for $M_1$. The two submatrices have different ranks.*


**Figure 5.4:** The two forbidden subgraph structures for transitive triangle (shielded collider) free graphs.

*The selected* $5 \times 5$ *submatrices are*

$$\begin{array}{c} \begin{matrix} K_{22} & K_{33} & K_{23} & K_{34} & K_{14} \end{matrix} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 2s\lambda_{23} & 0 & -s & 0 & 0 \\ 0 & 2s\lambda_{34} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -s \\ 1+\lambda_{23}^2 & 1+\lambda_{34}^2 & -\lambda_{23} & -\lambda_{34} & -\lambda_{41} \end{pmatrix} \end{array} \begin{matrix} \lambda_{12} \\ \lambda_{23} \\ \lambda_{34} \\ \lambda_{41} \\ s \end{matrix} ,$$

$$\begin{array}{c} \begin{matrix} K_{22} & K_{33} & K_{23} & K_{34} & K_{14} \end{matrix} \\ \begin{pmatrix} 2s\lambda_{21} & 0 & 0 & 0 & 0 \\ 0 & 2s\lambda_{32} & -s & 0 & 0 \\ 0 & 0 & 0 & -s & 0 \\ 0 & 0 & 0 & 0 & -s \\ 1+\lambda_{21}^2 & 1+\lambda_{32}^2 & -\lambda_{32} & -\lambda_{43} & -\lambda_{14} \end{pmatrix} \end{array} \begin{matrix} \lambda_{21} \\ \lambda_{32} \\ \lambda_{43} \\ \lambda_{14} \\ s \end{matrix} .$$

□

Our next theorem is actually another corollary of Theorem 5.21. It gives a general approach to construct column set corresponding to submatrices of different ranks, when the two graphs are **transitive triangle-free**.

**Definition 5.24.** *Let* $G = (V, D)$ *be a directed graph. A* ***transitive triangle*** *or* ***shielded collider*** *in* $G$ *is a triple of nodes* $i, j, k \in V$ *such that the induced subgraph of* $G$ *on* $i, j, k$ *is of one of the forms pictured in Figure 5.4. Alternatively, it means that for all* $j \in V$ *and for all* $i \in \mathrm{ch}(j)$ *it holds that* $\mathrm{ch}(j) \cap \mathrm{ch}(i) = \emptyset$*. If a graph has no transitive triangles then we say it is* ***transitive triangle free****.*

**Theorem 5.25.** *Let* $G_1 = (V, D_1)$*,* $G_2 = (V, D_2)$ *be different, transitive triangle-free, non-complete, simple directed graphs with node set* $V$*. Then* $G_1$ *and* $G_2$ *have different Jacobian matroids.*

*Proof.* If $G_1$ and $G_2$ satisfy the condition in Theorem 5.21, the conclusion obviously holds. Otherwise every node has the same out-degree in two graphs. Since $G_1$ and $G_2$ are different, there must exist a node $i$ such that $\mathrm{ch}_1(i) \neq \mathrm{ch}_2(i)$. We can assume that $i \to j \in D_1$ but $i \to j \notin D_2$. By Lemma 5.15 we know that either $j \to i \in D_2$ or $\mathrm{ch}_2(i) \cap \mathrm{ch}_2(j) \neq \emptyset$.

For the case of $j \to i \notin D_2$, it must hold that $\deg(j) \leq p - 1$ and $\mathrm{ch}_2(i) \cap \mathrm{ch}_2(j) \neq \emptyset$. We can use the same construction as that in Lemma 5.18 ($i$ replaced with $j$ and $j_0$

with $i$): Let $S = (S_E \setminus S_-) \cup S_+$ where

$$S_E = \{K_{km} \mid k \to m) \in D_2 \text{ or } m \to k \in D_2\},$$
$$S_- = \{K_{kj} \mid k \to j \in D_2 \text{ or } j \to k \in D_2\},$$

and $S_+$ is a subset of size $(\deg(j)+1) \leq p-1$ of the set

$$\{K_{ii}\} \cup \{K_{mm} : m \to j \in D_2 \text{ or } j \to m \in D_2\}.$$

By Lemma 5.17 and 5.18 we know that

$$\operatorname{rank}(J_S^{(2)}) \geq |D_2| - |\operatorname{ch}_2(j)| + 1 = |D_1| - |\operatorname{ch}_1(j)| + 1 \geq \operatorname{rank}(J_S^{(1)}).$$

The construction rules indicate that $K_{ll} \in S$ but $K_{ij} \notin S$. We can replace the column $K_{ll}$ by $K_{ij}$ to obtain a new set $S'$, such that the rank of the corresponding submatrix in $J^{(2)}$ increases by 1 and that for $J^{(1)}$ is kept unchanged. To see the rank increment, we take $s = 1, \lambda_{il} = 1, \lambda_{mj} = 1$ where $m \in \operatorname{pa}(j)$ and all other edge weights 0:

$$J_S^{(2)} = \begin{array}{c} \begin{array}{ccccc} K_{ii} & K_{m_1m_1}, \ldots, K_{m_pm_p} & K_{ij} & K_{j_1j_1}, \ldots, K_{j_qj_q} & K_{k_1l_1} \ldots K_{k_nl_n} \end{array} \\ \begin{pmatrix} \mathbf{a} & \mathbf{0} & 0 & \mathbf{0} & -I_{|D|-\deg(j)} \\ \mathbf{0} & \mathbf{0} & 0 & 2I_{|\operatorname{pa}(j)|} & \mathbf{0} \\ 0 & \mathbf{0}^T & 1 & \mathbf{0}^T & \mathbf{0}^T \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \times & \times & \times & \mathbf{0}^T \end{pmatrix} \end{array} \begin{array}{c} \lambda_{k_\alpha l_\alpha},\ k_\alpha, l_\alpha \neq j \\ \lambda_{j_\alpha, j},\ j_\alpha \in \operatorname{pa}(j) \\ \lambda_{jl} \\ \vdots \\ s \end{array} .$$

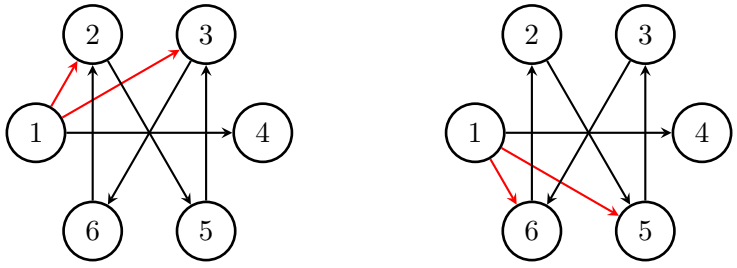
(5.3)

In Lemma 5.18 the rightmost pivot block is of the form $I + \varepsilon A$, where the second terms comes from common children of $k_\alpha$ and $l_\alpha$. The transitive triangle-free assumption now makes the second term zero. The zeros in the last column of blocks is also from that assumption. The vector $\mathbf{a}$ at the upper left block has only one nonzero entry $\partial K_{ii}/\partial \lambda_{ij}$. It can be eliminated by subtracting a multiple of columns of the rightmost $K_{k_\alpha l_\alpha}$ block. Hence, we have $\operatorname{rank}(J_{S'}^{(2)}) \geq |D_2| - \deg(j) + |\operatorname{pa}(j)| + 2 = |D_2| - |\operatorname{ch}(j)| + 2$. On the other hand, the $|\operatorname{ch}_1(j)|$ zero rows in $J_S^{(1)}$ correspond to the edges $\lambda_{jx}, x \in \operatorname{ch}_1(j)$. Since $i, j$ are not adjacent and $\operatorname{ch}_1(i) \cap \operatorname{ch}_1(j) = \emptyset$ (transitive triangle-free), we know that $(J_{S'}^{(1)})_{\lambda_{jx}, K_{ij}} = 0$ and $\operatorname{rank}(J_{S'}^{(1)}) \leq |D_1| - |\operatorname{ch}_1(i)| + 1 < |D_2| - |\operatorname{ch}_2(i)| + 2 \leq \operatorname{rank}(J_{S'}^{(2)})$.

If $j \to i \in D_2$, then the same construction of $S$ which is used in Theorem 5.21 can be used here, since it must be that either $\deg_2(j) < p-1$ or $\deg_2(j) = p-1$ but $j$ is not a sink node. In either case, one can take $S$ and then replace $K_{ii}$ with $K_{ij}$ to obtain $S'$. The argument then proceeds identically to the previous case so we omit the details. □

The following example illustrates the construction process of $S$ in Theorem 5.25.

**Example 5.26.** *Consider the graphs $G_1$ and $G_2$ in Figure 5.5 which are the same except for the change in the outgoing edges from node 1. These two graphs clearly have the same out-degree sequence which is $(3, 1, 1, 0, 1, 1)$. This means that Theorem 5.21 cannot be applied to certify that the $G_1$ and $G_2$ have different Jacobian matroids. We can instead use the construction in Theorem 5.25 to find a suitable set $S$ which will distinguish the matroids.*




**(a)** $G_1$ **(b)** $G_2$

**Figure 5.5:** Two graphs with the same out-degree sequence but which have no transitive triangles (shielded colliders).

*In this case we have the* $\mathrm{ch}_1(1) \neq \mathrm{ch}_2(1)$ *(note that* $|\mathrm{ch}_1(1)| = |\mathrm{ch}_2(1)|$*!) so* $i = 1$*. Furthermore, the edge* $(1,2) \in D_1$ *but* $(1,2) \notin D_2$ *so* $j = 2$*. This means that* $S = (S_E \setminus S_-) \cup S_+$ *where*

$$S_E = \{K_{km} \mid (k,m) \in D_2 \text{ or } (m,k) \in D_2\} = \{K_{14}, K_{15}, K_{16}, K_{25}, K_{26}, K_{35}, K_{36}\},$$
$$S_- = \{K_{kj} \mid (k,j) \in D_2 \text{ or } (j,k) \in D_2\} = \{K_{25}, K_{26}\},$$
$$S_+ = \{K_{55}, K_{66}\}.$$

*Note that in the language of Theorem 5.25, the common child of* $i = 1$ *and* $j = 2$ *is* $l = 5$*. Now according to Theorem 5.25, the set which will actually yield a rank difference is the set* $S'$ *which is obtained by replacing* $K_{ll} = K_{55}$ *with* $K_{ij} = K_{12}$*. This yields the set*

$$S' = \{K_{11}, K_{12}, K_{66}, K_{14}, K_{15}, K_{16}, K_{35}, K_{36}\}$$

*which indeed has* $|D_2| + 1 = 8$ *elements as is desired. Then* $J_S^{(2)}$ *has the form*

$$J_S^{(2)} = \begin{array}{c} \begin{matrix} K_{11} & K_{12} & K_{66} & K_{14} & K_{15} & K_{16} & K_{53} & K_{36} \end{matrix} \\ \begin{pmatrix} 2s\lambda_{14} & 0 & 0 & -s & 0 & 0 & 0 & 0 \\ 2s\lambda_{15} & s\lambda_{25} & 0 & 0 & -s & 0 & 0 & 0 \\ 2s\lambda_{16} & 0 & 0 & 0 & 0 & -s & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -s \\ 0 & 0 & 0 & 0 & 0 & 0 & -s & 0 \\ 0 & 0 & 2s\lambda_{62} & 0 & 0 & 0 & 0 & 0 \\ 0 & s\lambda_{15} & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 2 & -\lambda_{14} & -\lambda_{15} & -\lambda_{16} & -\lambda_{53} & -\lambda_{36} \end{pmatrix} \end{array} \begin{matrix} \lambda_{14} \\ \lambda_{15} \\ \lambda_{16} \\ \lambda_{36} \\ \lambda_{53} \\ \lambda_{62} \\ \lambda_{25} \\ s \end{matrix}.$$

*It is clear that this matrix has rank* 8 *while* $J_S^{(1)}$ *only has rank* 7*.*

We can rewrite Theorems 5.21 and 5.25 as the following theorem, to give some generically identifiable subclasses.

**Theorem 5.27.** *Let* $\mathcal{G}$ *be the collection of non-complete simple directed graphs. If the collection satisfies one of the following conditions, then the graph parameter indexing the graphs in* $\mathcal{G}$ *is generically identifiable under equal variance assumption:*

*(i) Every graph* $G \in \mathcal{G}$ *has a unique outdegree sequence.*

*(ii) Every graph $G \in \mathcal{G}$ is transitive triangle-free.*

Based on sink nodes analysis, we can obtain a partial identifiability result, which says that a DAG and a cyclic graph **cannot** generate the same distribution generically under the equal error variances assumption.

**Theorem 5.28.** *Let $G_1 = (V, D_1)$ and $G_2 = (V, D_2)$ be different directed graphs for models with equal error variances. If $G_1$ is a DAG and $G_2$ contains a cycle, then $\{M_{G_1}, M_{G_2}\}$ are generically distinguishable.*

*Proof.* We prove the result by contradiction. Suppose that $G_1$ and $G_2$ can generate the same distributions generically, then they must have the same matroid. There exists an open ball $B$ of $|D_1| + 1$ dimension in $M_{G_1} \cap M_{G_2}$, in which every value of the vectorized form $\mathrm{vec}(K)$ can be generated by a parameterization of $G_1$ and of $G_2$. By Theorem 5.13 and Lemma 5.16, the two graphs $G_1$ and $G_2$ must have the same sink nodes $V^{(1)}$ and edges to sink nodes.

For a precision matrix $K$ with $\mathrm{vec}(K) \in B$, all the edges from $V \backslash V^{(1)}$ to $V^{(1)}$ and the variance inverse $s$ can be uniquely determined from $K$, which are the same in both graphs. Then we consider the subgraphs induced by $V \backslash V^{(1)}$ in $G_1$ and $G_2$, denoted by $G_1^{(1)}$ and $G_2^{(1)}$. We can subtract the the contributions of edges from $V \backslash V^{(1)}$ to $V^{(1)}$ in $K$. This procedure gives the common precision matrix $K_1^{(1)} = K_2^{(1)}$ for the subgraphs. Since the inverse variance parameter has been determined to be common, the two subgraphs must have the same sink nodes again. Indeed, if node $i$ is a sink node in $G_1^{(1)}$ but not in $G_2^{(1)}$, then the $(i, i)$ entry in $K_1^{(1)}$ and $K_2^{(1)}$ cannot be equal.

We can perform sink nodes deletion iteratively. Since $G_1$ is a DAG and $G_2$ is cyclic, there must exists some $k$, such that after $k$ steps the obtained subgraph $G_2^{(k)}$ have no sink nodes while $G_1^{(k)}$ have some sink nodes. Thus $K_1^{(k)}$ and $K_2^{(k)}$ are not the same. Two different original precision matrices $K_1$ and $K_2$ can be recovered from $K_1^{(k)}$ and $K_2^{(k)}$ with the edge weights and inverse variance that have been uniquely determined to be common, which contradicts to the fact that the original precision matrix $K$ is arbitrary in an full-dimension open ball in $M_{G_1} \cap M_{G_2}$. □

Now we are ready to discuss the extension of the outdegree theorem. In Theorem 5.21 we consider the whole neighborhood of a node, i.e., $\mathrm{ne}(i) = \mathrm{pa}(i) \cup \mathrm{ch}(i)$. One sufficient condition for different matroids is different sizes of children set $|\mathrm{ch}_1(i)| \neq |\mathrm{ch}_2(i)|$, which is still too restrictive. Indeed, we can focus on a subset of the neighborhood and compare the numbers of children in the subset for the two graphs. This approach leads to our last major theorem, which is a much stronger generalization of Theorem 5.21, but also more difficult to check. Before stating the theorem we need a new definition.

**Definition 5.29.** *Let $G = (V, D)$ be a directed graph and $i \in V$. A subset $L \subseteq \mathrm{ne}(i)$ is **parentally closed** set with respect to $i$ if $\mathrm{pa}(L) \cap \mathrm{ne}(i) \subseteq L$. We denote the collection of all parentally closed sets with respect to $i$ by $\mathcal{L}_i$.*

**Theorem 5.30.** *Let $G_1 = (V, D_1)$, $G_2 = (V, D_2)$ be two simple directed graphs which are not complete. Let $i \in V$, and let $\mathcal{L}_i^k$ be the collection of parentally closed sets with respect to $i$ for graph $G_k$. If there exists a set $L \in \mathcal{L}_i^k$ such that $|\mathrm{ch}_k(i) \cap L| > |\mathrm{ch}_{3-k}(i) \cap L|$, $k \in \{1, 2\}$, then $G_1$ and $G_2$ have different matroids.*

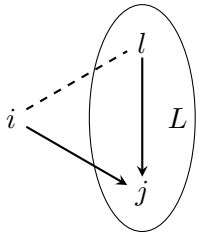



**Figure 5.6:** This displays the subgraph relating $i, j$, and $l$ in the proof of Theorem 5.30. Since $L$ is parentally closed and $j \in L$ and $l \in \mathrm{pa}(j) \cap \mathrm{ne}(i)$, it must be that $l \in L$.

*Proof.* The proof is similar to that of Theorem 5.21. We select a special column set $S$, such that the submatrix in one Jacobian has a lower bound of rank, and that in the other Jacobian has a smaller upper bound of rank.

Suppose that the inequality $|\operatorname{ch}_1(i) \cap L| > |\operatorname{ch}_2(i) \cap L|$ holds for some $L \in \mathcal{L}_i^1$. If $L = V \backslash \{i\}$, the case has been discussed in the proof of Proposition 5.21. Otherwise we can assume that $|L| < p - 1$. Under this condition, we set

$$\begin{aligned} S_E &= \{K_{mn} \mid m \to n \text{ or } n \to m \in D_2\}, \\ S_- &= \{K_{im} \mid m \in L\}, \\ S_+ &= \{K_{mm} \mid m \to i \text{ or } i \to m \in D_2, m \in L\} \cup \{K_{j_0 j_0}\}, \end{aligned}$$

where $j_0$ is an arbitrary node in $V \backslash (\{i\} \cup L)$.

The size of the column set $S = (S_E \backslash S_-) \cup S_+$ is $|D_2| - |L \cap \mathrm{ne}_2(i)| + |L \cap \mathrm{ne}_2(i)| + 1 = |D_2| + 1$, hence the submatrices $J_S^{(1)}, J_S^{(2)}$ are of size $(|D_2| + 1) \times (|D_2| + 1)$. For each node $j \in \operatorname{ch}_1(i) \cap L$, nonzero entries of $J^{(1)}$ in the row $\lambda_{ij}$ may only appear in columns $K_{ii}, K_{ij}$ or $K_{il}$, for $l \in \mathrm{pa}_1(j)$. We claim that none of these columns are contained in $S$.

By the construction rule, we know that $K_{ii} \notin S, S_+$ and $K_{ij} \in S_-$ since the parental closure of $L$ implies that $j \in L$. If $K_{il} \in S_E$ then $l \in \mathrm{ne}(i)$ and thus $l \in \mathrm{ne}(i) \cap \mathrm{pa}_1(j) \subseteq L$ so $l \in L$ (this is pictured in Figure 5.6), which implies that $K_{il} \in S_-$ and so it holds that $K_{il} \notin S$. Thus $J_S^1$ contains at least $|\operatorname{ch}_1(i) \cap L|$ zero rows, and $\mathbf{rank}(J_S^1) \leq |D_1| + 1 - |\operatorname{ch}_1(i) \cap L|$.

Next, we construct one specific paramaterization of $K$ and $J_S^2$, to show that the latter has generic rank greater or equal than $|D_2| + 1 - |\operatorname{ch}_2(i) \cap L|$. For each $l_q \in \mathrm{pa}_2(i) \cap L$, we set $\lambda_{il_q} = \varepsilon$. Then we set $s = 1$ and all other edge weights zero. The submatrix $J_S^2$ takes similar form as that in (5.1), and it has $|\operatorname{ch}_2(i) \cap L|$ instead of $|\operatorname{ch}_2(i)|$ uncertain rows. Hence we have

$$\operatorname{rank}(J_S^{(2)}) \geq |D_2| - |\operatorname{ch}_2(i) \cap L| + 1 > |D_1| - |\operatorname{ch}_1(i) \cap L| + 1 \geq \operatorname{rank}(J_S^{(1)}),$$

which completes the proof. □

**Remark 5.31.** *Theorems 5.21 and 5.25 are special cases of Theorem 5.30. The former only checks the trivial parent closed set $L = \mathrm{ne}(i)$, while the latter considers $\mathcal{L}_1 = \mathcal{P}(\operatorname{ch}_1(i))$ and $\mathcal{L}_2 = \mathcal{P}(\operatorname{ch}_1(i))$, where $\mathcal{P}$ is the powerset operator.*


**(a)** $G_1$ **(b)** $G_2$

**Figure 5.7:** Example: necessity of checking unions of minimal parentally closed sets.

*The minimal parentally closed sets w.r.t a node $i$ can be constructed by starting from one single child and closing the parents in* $\mathrm{ch}(i)$ *sequentially. The collection $\mathcal{L}$ consists of the union of multiple minimal parentally closed sets. The union operation increase the complexity, but it is necessary. For example, if $j \in \mathrm{pa}_1(i) \cap \mathrm{ch}_2(i)$ lies in the intersection of two different parentally closed sets ($\{j, l_1\}$ and $\{j, l_2\}$) in $\mathcal{L}_1$, the node $j$ contributes in both sets as a child of $i$ in $G_2$, with equalities $|\mathrm{ch}_1(i) \cap \{j, l_1\}| = |\mathrm{ch}_2(i) \cap \{j, l_1\}|$, $|\mathrm{ch}_1(i) \cap \{j, l_2\}| = |\mathrm{ch}_2(i) \cap \{j, l_1\}|$. Then the union really makes a difference: $2 = |\mathrm{ch}_1(i) \cap \{j, l_1, l_2\}| > |\mathrm{ch}_2(i) \cap \{j, l_1, l_2\}| = 1$.*

## 5.4 Computational Study

### 5.4.1 Summary

We perform computational checks for simple directed graphs with $p \leq 6$ (except complete graphs with 6 nodes). All of these graphs are certified to have different models (and most of them have different matroids). Computing ranks of submatrices determines the maximal independent sets and gives the Jacobian matroid of a graph. We store the matroids in a list for comparisons.

The results for $p \leq 3$ can be checked manually. For $p = 4, 5, 6$, there are 729, 59409, 14348907 simple directed graphs, and we must rely on software. The following Lemma 5.32 can slightly reduce the number of graphs that need to be checked. We also apply the outdegree theorem to avoid some redundant checks.

**Lemma 5.32.** *If the graph parameter indexing the simple directed graphs up to $k$ nodes is generically identifiable, then the graph parameter indexing all simple directed graphs with $k+1$ nodes and up to $k$ edges is also generically identifiable.*

*Proof.* Let $G_1$ and $G_2$ be two arbitrary simple directed graphs with $k+1$ nodes and $k$ edges. If both graphs are connected, then they are polytrees. The identifiability is from the result for DAGs. Otherwise at least one graph is not connected and has a node $i$ of degree 0. Suppose the two graphs have the same matroid, Lemma 5.16 and Theorem 5.21 imply that the node $i$ is a sink node and have the same adjacencies in both graph. That is, the node $i$ has degree 0 in both graphs. Since the assumption ensures the identifiability of graphs with $k$ nodes, the two subgraphs with nodes in $V \backslash \{i\}$ cannot generate the same distribution generically. That marginal distribution is independent to the distribution of $X_i$, thus the original graphs $G_1$ and $G_2$ are distinguishable. □

**Theorem 5.33.** *Let $p \leq 6$. Consider the set of graphical models given by simple non-complete graphs with $p$ nodes. Then the graph parameter is generically identifiable.*

*Proof.* The details and example graphs for this proof are in Sections 5.4.2,5.4.3 ,5.4.4,5.4.5 and 5.4.6.

The case of $p = 2$ is trivial. Then we compute and compare all matroids for the 27 graphs with 3 nodes. Any pair of graphs are distinguishable via matroid or covariance matrix condition. Indeed, there are 5 different matroids among 8 graphs with 3 edges. The two 3-cycles correspond to two matroids, and other 6 graphs are 2-to-1. Each pair are transitive triangles with the edge between last two nodes in two directions. Only the node with outdegree zero corresponds to the minimum of all diagonal entries in $K$; see Chen et al. [2019].

By Lemma 5.32, if we have certified the identifiability of all graphs with $p$ nodes, then for graphs with $p+1$ nodes, we only need to test those that have at least $p+1$ edges. In this way, we can perform complete symbolic checks via MATHEMATICA for graphs with 4 or 5 nodes and conduct random checks for graphs with $p = 6$ nodes.

There are 496 different graphs with at least 4 edges for $p = 4$. The computation yields 484 different Jacobian matroids, of which 472 matroids have a 1-to-1 correspondence to graphs and other 12 matroids have a 1-to-2 correspondence. The 12 pairs of graphs with the same matroids happen to be the 24 complete DAGs. Each pair contains one DAG of the 24, and another one with the edge between last two nodes reversed. Like that in $p = 3$ case, the graph pairs with the same matroid can be identified via minimum value in $K$.

For $p = 5$, the number of graphs with at least 5 edges increases to 54528. The computational result shows similar patterns as for smaller $p'$s. Most of the graphs have unique matroid, while others have a 2-to-1 correspondence. The two graphs $G_1$ and $G_2$ have the same Jacobian matroid if and only if the following conditions are all satisfied:

(1) Both $G_1$ and $G_2$ are complete,

(2) The subgraphs induced by node $\{i_1, i_2, i_3\}$ are the same in both $G_1$ and $G_2$,

(3) $\mathrm{pa}_1(i_4) = \{i_1, i_2, i_3\}$, $\mathrm{pa}_1(i_5) = \{i_1, i_2, i_3, i_4\}$ and $\mathrm{pa}_2(i_5) = \{i_1, i_2, i_3\}$, $\mathrm{pa}_2(i_4) = \{i_1, i_2, i_3, i_5\}$.

Thus, $G_1$ and $G_2$ are complete and have $p-2$ nodes that serve as common parents of the last 2 nodes. They also share the same edges except the edge between the last 2 nodes. These graph pairs are distinguished by diagonal entries in $K$.

When considering graphs with 6 nodes, the total number of which have at least 6 edges is approximately $10^7$. For each fixed set size, the number of column sets is also much larger compared to graphs with 5 nodes. Consequently, performing exhaustive symbolic computation is impossible. To simplify the computation, We apply Theorem 5.21 and use random integers parameterization. According to Theorem 5.21, we only need to compute and compare the matroids within every collection of graphs sharing the same outdegree sequence. We consider all sequence of length 6 with every

value not exceeding 5 and the sum ranging from 6 to 14 (excluding complete graphs), which is a superset of all valid outdegree sequences.

Obviously, outdegree sequences from different unordered set are different. If two different sequences are from the same unordered set, there must still exist a node with different outdegrees in the two graphs. Thus, we can focus on all the sequences ordered from largest to smallest along the nodes $1, 2, 3, 4, 5, 6$. We generate all possible outdegree sequences and, for each sequence, list all possible graphs via depth first search. For the sequences with sum smaller than 15, we only need to compare the matroids within all graphs corresponding to the same sequence, which can greatly reduce the computation time. In practice, we substitue random integers into the parameters and compute the independent sets. The computation result shows that every graph has a unique matroid.

Alternatively we also apply the parentally closed set condition in Theorem 5.30, and it successfully distinguish all pairs of graphs among non-complete graphs with 6 nodes. The two criteria yield compatible results, and the comparison is recorded in Table 5.3. □

Based on Theorem 5.33 and the computation results presented in the following subsections, we can propose a promising conjecture.

**Conjecture 5.34.** *Consider the infinite and countable set of graphical models given by simple directed graphs with $p \in \{1, 2, \dots\}$ nodes. Then the graph indexing parameter is generically identifiable.*

### 5.4.2 p = 2

In the case of $p = 2$, we have 3 different graph structures: {}, {(1,2)}, {(2,1)}.

- {}:

  With edge set {}, the map $\psi$ has only 1 parameter, that is the variance of noise $\omega$. $K = sI$ and $J(\psi) = (1, 0, 1)$. The independent sets in the matroid are $\emptyset, \{1\}, \{3\}$.

- {(1,2)}:

$$K = s \begin{pmatrix} 1 & -\lambda_{12} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\lambda_{12} & 1 \end{pmatrix} = s \begin{pmatrix} 1+\lambda_{12}^2 & -\lambda_{12} \\ -\lambda_{12} & 1 \end{pmatrix},$$

$$J(\psi) = \left( \frac{\partial K}{\partial(\lambda_{12}, s))} \right) = \begin{pmatrix} 2s\lambda_{12} & 0 & -s \\ 1+\lambda_{12}^2 & 1 & -\lambda_{12} \end{pmatrix}.$$

  The independent sets of the matroid are

$$\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{1,3\}.$$

  We can also consider the covariance matrix of $X$. That is

$$\Sigma = \omega \begin{pmatrix} 1 & \lambda_{12} \\ 0 & 1+\lambda_{12}^2 \end{pmatrix},$$

and it satisfies the equation

$$\Sigma_{11} = \Sigma_{22} - \Sigma_{12}^2/\Sigma_{11}.$$

- {(2,1)}:

$$K = s \begin{pmatrix} 1 & 0 \\ -\lambda_{21} & 1 \end{pmatrix} \begin{pmatrix} 1 & -\lambda_{21} \\ 0 & 1 \end{pmatrix} = s \begin{pmatrix} 1 & -\lambda_{21} \\ -\lambda_{21} & 1+\lambda_{21}^2 \end{pmatrix},$$

$$J(\psi) = \left( \frac{\partial K}{\partial(\lambda_{21}, s))} \right) = \begin{pmatrix} 0 & 2s\lambda_{21} & -s \\ 1 & 1+\lambda_{21}^2 & -\lambda_{21} \end{pmatrix}.$$

The independent sets of the matroid are the same as that of $\{(1,2)\}$. However, we can derive similar but different covariance matrix elements relationship:

$$\Sigma_{22} = \Sigma_{11} - \Sigma_{12}^2/\Sigma_{22},$$

and which could be used to identify the two different models.

The matroid structure can be used to determine if there exists an edge between node 1 and 2. The edge direction is identified by the equation of covariance matrix elements. Specifically, the node with larger variance corresponds to the child in the two nodes.

### 5.4.3 p = 3

In the case of $p = 3$, since we do not allow 2-cycle, the graph has at most 4 parameters (3 edge weights and 1 inverse error variance). The number of possible graph structures (assuming connected) is 20. We will first consider the 5 distinct graph structures with descendant relationship in topological order, then examine other structures via permutation of nodes.

Two matroids are different if and only if there exist a pair of different independent sets. Since every independent set is contained in some maximal independent set, two matroid are different if and only if there exist a pair of different maximal independent sets. We will only consider maximal independent sets with respect to the indexes of columns in this section.

- $\{(1,2),(1,3)\}$:

$$\begin{aligned} K &= s \begin{pmatrix} 1 & -\lambda_{12} & -\lambda_{13} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -\lambda_{12} & 1 & 0 \\ -\lambda_{13} & 0 & 1 \end{pmatrix} \\ &= s \begin{pmatrix} 1+\lambda_{12}^2+\lambda_{13}^2 & -\lambda_{12} & -\lambda_{13} \\ -\lambda_{12} & 1 & 0 \\ -\lambda_{13} & 0 & 1 \end{pmatrix}. \end{aligned}$$

$$\begin{aligned} J(\psi) &= \left( \frac{\partial(K_{11}, K_{22}, K_3 3, K_{12}, K_{23}, K_{13})}{\partial(\lambda_{12}, \lambda_{13}, s))} \right) \\ &= \begin{pmatrix} 2s\lambda_{12} & 0 & 0 & -s & 0 & 0 \\ 2s\lambda_{13} & 0 & 0 & 0 & 0 & -s \\ 1+\lambda_{12}^2+\lambda_{13}^2 & 1 & 1 & -\lambda_{12} & 0 & -\lambda_{13} \end{pmatrix}. \end{aligned}$$

The maximal independent sets are

$$\{11,22,12\},\{11,22,13\},\{11,33,12\},\{11,33,13\},$$
$$\{11,23,13\},\{22,12,13\},\{33,12,13\},$$

or with column indices

$$\{1,2,4\},\{1,2,6\},\{1,3,4\},\{1,3,6\},\{1,4,6\},\{2,4,6\},\{3,4,6\}.$$

In this case, the positions of node 2 and node 3 are symmetric. There are another 2 graph structures through nodes permutation: $\{(2,1),(2,3)\}$ (via (12)) and $\{(3,1),(3,2)\}$ (via (13)). The two graph structure corresponds to column permutation (12) and (13), respectively, in the Jacobian matrix.

In the maximal independent sets of graph structure $\{(1,2),(1,3)\}$, column 1 appears 5 times, column 2 appears 3 times, and column 3 appears 3 times. Hence, (12) and (13) yield distinct maximal independent sets.

- $\{(1,3),(2,3)\}$

$$\begin{aligned} K &= s\begin{pmatrix} 1 & 0 & -\lambda_{13} \\ 0 & 1 & -\lambda_{23} \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\lambda_{13} & -\lambda_{23} & 1 \end{pmatrix} \\ &= s\begin{pmatrix} 1+\lambda_{13}^2 & \lambda_{13}\lambda_{23} & -\lambda_{13} \\ \lambda_{13}\lambda_{23} & 1+\lambda_{23}^2 & -\lambda_{23} \\ -\lambda_{13} & -\lambda_{23} & 1 \end{pmatrix}. \end{aligned}$$

$$\begin{aligned} J(\psi) &= \left(\frac{\partial(K_{11},K_{22},K_{33},K_{12},K_{23},K_{13})}{\partial(\lambda_{23},\lambda_{13},s))}\right) \\ &= \begin{pmatrix} 0 & 2s\lambda_{23} & 0 & s\lambda_{13} & -s & 0 \\ 2s\lambda_{13} & 0 & 0 & s\lambda_{23} & 0 & -s \\ 1+\lambda_{13}^2 & 1+\lambda_{23}^2 & 1 & \lambda_{13}\lambda_{23} & -\lambda_{23} & -\lambda_{13} \end{pmatrix}. \end{aligned}$$

The maximal independent sets are

$$\{11,22,33\},\{11,22,12\},\{11,22,23\},\{11,22,13\},\{11,33,12\},\{11,33,23\},$$
$$\{11,12,23\},\{11,12,13\},\{11,23,13\},\{22,33,12\},\{22,33,13\},\{22,12,23\},$$
$$\{22,12,13\},\{22,23,13\},\{33,12,23\},\{33,12,13\},\{33,23,13\},\{12,23,13\}.$$

or with column indices

$$\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,2,6\},\{1,3,4\},\{1,3,5\},$$
$$\{1,4,5\},\{1,4,6\},\{1,5,6\},\{2,3,4\},\{2,3,6\},\{2,4,5\},$$
$$\{2,4,6\},\{2,5,6\},\{3,4,5\},\{3,4,6\},\{3,5,6\},\{4,5,6\}.$$

In this case, the maximal independent sets contain 4 occurrences of $(1,2)$ pairs, 2 occurrences of $(1,3)$ pairs, and 2 occurrences of $(2,3)$ pairs.

The possible permutations on nodes are (13) and (23), resulting the graph structures $\{(2,1),(3,1)\}$ and $\{(1,2),(3,2)\}$ respectively. The corresponding premutations on the columns of jacobian matrix are also (13) and (23). These permutations yield distinct maximal independent sets.

- $\{(1,2),(2,3)\}$

$$K = s\begin{pmatrix} 1 & -\lambda_{12} & 0 \\ 0 & 1 & -\lambda_{23} \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ -\lambda_{12} & 1 & 0 \\ 0 & -\lambda_{23} & 1 \end{pmatrix}$$
$$= s\begin{pmatrix} 1+\lambda_{12}^2 & -\lambda_{12} & 0 \\ -\lambda_{12} & 1+\lambda_{23}^2 & -\lambda_{23} \\ 0 & -\lambda_{23} & 1 \end{pmatrix}.$$

$$J(\psi) = \left(\frac{\partial(K_{11},K_{22},K_{33},K_{12},K_{23},K_{13})}{\partial(\lambda_{12},\lambda_{23},s))}\right)$$
$$= \begin{pmatrix} 2s\lambda_{12} & 0 & 0 & -s & 0 & 0 \\ 0 & 2s\lambda_{23} & 0 & 0 & -s & 0 \\ 1+\lambda_{12}^2 & 1+\lambda_{23}^2 & 1 & -\lambda_{12} & -\lambda_{23} & 0 \end{pmatrix}.$$

The maximal independent sets are

$$\{11,22,33\},\{11,22,12\},\{11,22,23\},\{11,33,23\},$$
$$\{11,12,23\},\{22,33,12\},\{22,12,23\},\{33,12,23\},$$

or with column indices

$$\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,3,5\},\{1,4,5\},\{2,3,4\},\{2,4,5\},\{3,4,5\}.$$

In this case, there are 5 possible permutations in both the nodes and the Jacobian matrix column space: (12), (23), (13), (123), and (132).

In the maximal independent sets of the original graph structure $(1,2),(2,3)$, there are 3 occurrences of $(1,2)$ pairs, 2 occurrences of $(2,3)$ pairs, and 2 occurrences of $(1,3)$ pairs. Therefore, (12), (23), and (13) yield distinct independent sets. Additionally, noticing that there are 3 occurrences of $(2,4)$ pairs and 2 occurrences of $(3,4)$ pairs, the permutations (123) and (321) can also produce distinct independent sets.

- $\{(1,2),(1,3),(2,3)\}$ (and $\{(1,2),(1,3),(3,2)\}$ also has the same matroid)

$$K = s\begin{pmatrix} 1 & -\lambda_{12} & -\lambda_{13} \\ 0 & 1 & -\lambda_{23} \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ -\lambda_{12} & 1 & 0 \\ -\lambda_{13} & -\lambda_{23} & 1 \end{pmatrix}$$
$$= s\begin{pmatrix} 1+\lambda_{12}^2+\lambda_{13}^2 & -\lambda_{12}+\lambda_{13}\lambda_{23} & -\lambda_{13} \\ -\lambda_{12}+\lambda_{13}\lambda_{23} & 1+\lambda_{23}^2 & -\lambda_{23} \\ -\lambda_{13} & -\lambda_{23} & 1 \end{pmatrix}.$$

$$J(\psi) = \left(\frac{\partial(K_{11},K_{22},K_{33},K_{12},K_{23},K_{13})}{\partial(\lambda_{12},\lambda_{23},\lambda_{13},s))}\right)$$
$$= \begin{pmatrix} 2s\lambda_{12} & 0 & 0 & -s & 0 & 0 \\ 0 & 2s\lambda_{23} & 0 & s\lambda_{13} & -s & 0 \\ 2s\lambda_{13} & 0 & 0 & s\lambda_{23} & 0 & -s \\ 1+\lambda_{12}^2+\lambda_{13}^2 & 1+\lambda_{23}^2 & 1 & -\lambda_{12}+\lambda_{13}\lambda_{23} & -\lambda_{23} & -\lambda_{13} \end{pmatrix}.$$

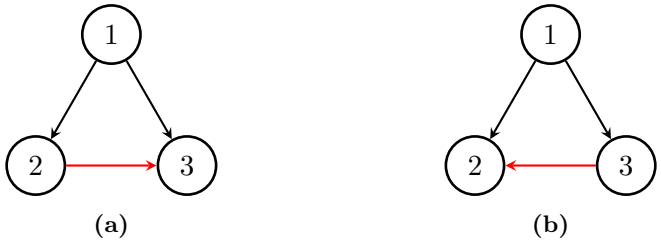



**Figure 5.8:** Two graphs with the same Jacobian matroid.

The maximal independent sets are

$$\{11, 22, 33, 12\}, \{11, 22, 33, 13\}, \{11, 22, 12, 23\},$$
$$\{11, 22, 12, 13\}, \{11, 22, 23, 13\}, \{11, 33, 12, 23\},$$
$$\{11, 33, 12, 13\}, \{11, 33, 23, 13\}, \{11, 12, 23, 13\},$$
$$\{22, 33, 12, 13\}, \{22, 12, 23, 13\}, \{22, 12, 23, 13\},$$

or with column indices

$$\{1, 2, 3, 4\}, \{1, 2, 3, 6\}, \{1, 2, 4, 5\}, \{1, 2, 4, 6\}, \{1, 2, 5, 6\}, \{1, 3, 4, 5\},$$
$$\{1, 3, 4, 6\}, \{1, 3, 5, 6\}, \{1, 4, 5, 6\}, \{2, 3, 4, 6\}, \{2, 4, 5, 6\}, \{3, 4, 5, 6\}.$$

Similar to the previous case, there are 5 possible permutations on nodes: (12), (23), (13), (123) and (132). The corresponding permutations on the columns are (12)(56), (23)(46), (13)(45), (123)(456) and (321)(654).

After examining all the permutations, it is observed that all permutations except (23) yield different maximal independent sets.

Although these two graphs $\{(1,2),(1,3),(2,3)\}$ and $\{(1,2),(1,3),(3,2)\}$ have the same Jacobian matroid, they can still be distinguished by other method. In Figure 5.8 (a) the graph structure $\{(1,2),(1,3),(2,3)\}$, the smallest entry in $\{K_{11}, K_{22}, K_{33}\}$ is $K_{33}$. However, in Figure 5.8 (b), the smallest diagonal entry is $K_{22}$.

- $\{(1,2),(2,3),(3,1)\}$

$$K = s \begin{pmatrix} 1 & -\lambda_{12} & 0 \\ 0 & 1 & -\lambda_{23} \\ -\lambda_{31} & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -\lambda_{13} \\ -\lambda_{12} & 1 & 0 \\ 0 & -\lambda_{23} & 1 \end{pmatrix}$$
$$= s \begin{pmatrix} 1+\lambda_{12}^2 & -\lambda_{12} & -\lambda_{31} \\ -\lambda_{12} & 1+\lambda_{23}^2 & -\lambda_{23} \\ -\lambda_{31} & -\lambda_{23} & 1+\lambda_{31}^2 \end{pmatrix}.$$

$$J(\psi) = \left( \frac{\partial(K_{11}, K_{22}, K_{33}, K_{12}, K_{23}, K_{13})}{\partial(\lambda_{12}, \lambda_{23}, \lambda_{31}, s))} \right)$$
$$= \begin{pmatrix} 2s\lambda_{12} & 0 & 0 & -s & 0 & 0 \\ 0 & 2s\lambda_{23} & 0 & 0 & -s & 0 \\ 0 & 0 & 2s\lambda_{31} & 0 & 0 & -s \\ 1+\lambda_{12}^2 & 1+\lambda_{23}^2 & 1+\lambda_{31}^2 & -\lambda_{12} & -\lambda_{23} & -\lambda_{13} \end{pmatrix}.$$

The maximal independent sets are

$$\{11,22,33,12\},\{11,22,33,23\},\{11,22,33,13\},$$
$$\{11,22,12,13\},\{11,22,23,13\},\{11,33,12,23\},$$
$$\{11,33,23,13\},\{11,12,23,13\},\{22,33,12,23\},$$
$$\{22,33,12,13\},\{22,12,23,13\},\{33,12,23,13\},$$

or with column indices

$$\{1,2,3,4\},\{1,2,3,5\},\{1,2,3,6\},\{1,2,4,6\},\{1,2,5,6\},\{1,3,4,5\},$$
$$\{1,3,5,6\},\{1,4,5,6\},\{2,3,4,5\},\{2,3,4,6\},\{2,4,5,6\},\{3,4,5,6\}.$$

All 5 possible permutations will preserve the cycle order or reverse it. The reversed graph structure is $\{(1,3),(3,2),(2,1)\}$, induced by (23) on nodes and (23)(46) on precision matrix entries. In the maximal independent sets, there are 3 occurrences of the $(1,2,3)$ triple, 2 occurrences of the $(1,3,4)$ triple, and 3 occurrences of the $(2,3,4)$ triple. Therefore, the combinations of these sets are different from the combinations of sets for the graph structure $(1,2),(2,3),(1,3)$.

We can compute the precision matrix and find

$$K = s \begin{pmatrix} 1+\lambda_{13}^2 & -\lambda_{21} & -\lambda_{13} \\ -\lambda_{21} & 1+\lambda_{21}^2 & -\lambda_{32} \\ -\lambda_{13} & -\lambda_{32} & 1+\lambda_{32}^2 \end{pmatrix}.$$

The transposed Jacobian $J$ for $\{(1,3),(3,2),(2,1)\}$ is

$$J(\psi) = \begin{pmatrix} 2s\lambda_{13} & 0 & 0 & 0 & 0 & -s \\ 0 & 2s\lambda_{23} & 0 & -s & 0 & 0 \\ 0 & 0 & 2s\lambda_{32} & 0 & -s & 0 \\ 1+\lambda_{13}^2 & 1+\lambda_{21}^2 & 1+\lambda_{32}^2 & -\lambda_{21} & -\lambda_{32} & -\lambda_{13} \end{pmatrix}.$$

The maximal independent sets are

$$\{11,22,33,12\},\{11,22,33,23\},\{11,22,33,13\},$$
$$\{11,22,12,23\},\{11,22,23,13\},\{11,33,12,23\},$$
$$\{11,33,12,13\},\{11,12,23,13\},\{22,33,12,13\},$$
$$\{22,33,23,13\},\{22,12,23,13\},\{33,12,23,13\},$$

or with column indices

$$\{1,2,3,4\},\{1,2,3,5\},\{1,2,3,6\},\{1,2,4,5\},\{1,2,5,6\},\{1,3,4,5\},$$
$$\{1,3,4,6\},\{1,4,5,6\},\{2,3,4,6\},\{2,3,5,6\},\{2,4,5,6\},\{3,4,5,6\},$$

which are different from those of the graph $\{(1,2),(2,3),(3,1)\}$.

### 5.4.4 p = 4

As described in the proof of Theorem 5.33, the 12 pairs of graphs with the same matroids correspond to the 24 complete DAGs.


**Figure 5.9:** Two graphs with the same Jacobian matroid, p=4.

| number of edges | $\|D\| = 4$ | $\|D\| = 5$ |
|---|---|---|
| n.d. by outdeg seq | 1443 | 708 |
| n.d. by outdeg seq & closures | 0 | 0 |
| total number of pairs | 28680 | 18336 |

**Table 5.1:** Number of pairs of 4-node simple directed graphs, with same number of edges, that **cannot** be distinguished by Theorem 5.21 (outdegree sequence method) or Theorem 5.30 (outdegree sequence and parentally closed sets method).

### 5.4.5 p = 5

For $p = 5$, the computation of Jacobian matroid for all graphs takes more than three days. However, the result follows the same pattern: complete graphs with $p-2$ nodes being common parents of the last 2 nodes, and the same edges except for the edge between the last 2 nodes, exhibit the same matroid. Figure 5.10 provides an example of two graphs with the same Jacobian matroid.

### 5.4.6 p = 6

It is not possible to traverse all simple directed graphs with 6 nodes due to their large number. Therefore, we approach the problem by considering subclasses of graphs that have the same numbers of edges. Within each subclass, we further analyze outdegree sequences. To expedite the computation, we assign random integers as parameter values, which accelerates the rank computations. The result demonstrate that all non-complete simple directed graphs with 6 nodes possess different matroids. Additionally, the parentally closed set condition functions effectively and successfully

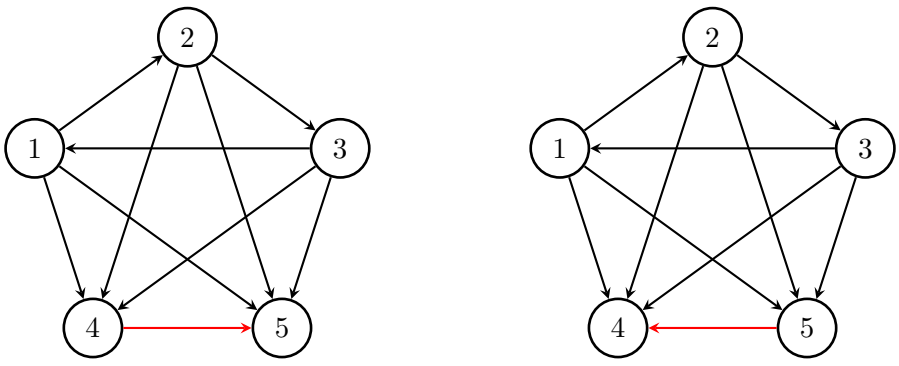



**Figure 5.10:** Two graphs with the same Jacobian matroid, $p$=5.

| number of edges | $\|D\| = 5$ | $\|D\| = 6$ | $\|D\| = 7$ |
|---|---|---|---|
| n.d. by outdeg seq | 567006 | 1215940 | 1292870 |
| n.d. by outdeg seq & closures | 0 | 0 | 0 |
| total number of pairs | 32510016 | 90310080 | 117957120 |

| number of edges | $\|D\| = 8$ | $\|D\| = 9$ |
|---|---|---|
| n.d. by outdeg seq | 615060 | 104920 |
| n.d. by outdeg seq & closures | 0 | 0 |
| total number of pairs | 66349440 | 13104640 |

**Table 5.2:** Number of pairs of 5-node simple directed graphs, with same number of edges, that **cannot** be distinguished by Theorem 5.21 (outdegree sequence method) or Theorem 5.30 (outdegree sequence and parentally closed sets method).

distinguishes all pairs of those graphs. The comparison of two graphical criteria is presented in Table 5.3.

| number of edges | $\|D\| = 6$ | $\|D\| = 7$ | $\|D\| = 8$ |
|---|---|---|---|
| outdeg seq | 282621720 | 1391117760 | 4359482730 |
| outdeg seq & closures | 0 | 0 | 0 |
| total number of pairs | 51302291040 | 339223959360 | 1356896661120 |

| number of edges | $\|D\| = 9$ | $\|D\| = 10$ | $\|D\| = 11$ |
|---|---|---|---|
| outdeg seq | 8597383980 | 10414049394 | 7430794740 |
| outdeg seq & closures | 0 | 0 | 0 |
| total number of pairs | 3283355595520 | 4728032365056 | 3907464637440 |

| number of edges | $\|D\| = 12$ | $\|D\| = 13$ | $\|D\| = 14$ |
|---|---|---|---|
| outdeg seq | 2914540765 | 558189990 | 41578860 |
| outdeg seq & closures | 0 | 0 | 0 |
| total number of pairs | 1736650639360 | 369937182720 | 30198865920 |

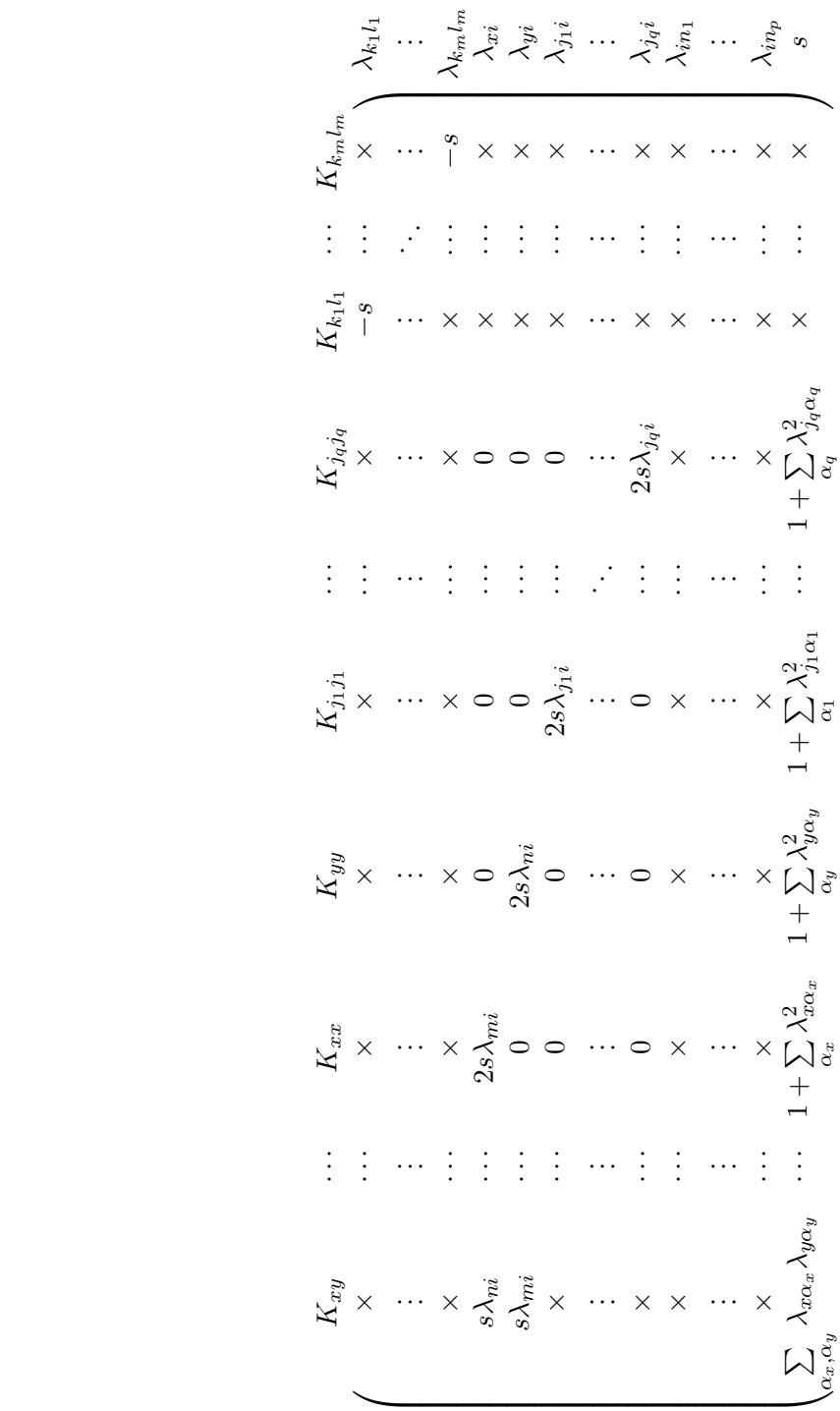
**Table 5.3:** Number of pairs of 6-node simple directed graphs, with same number of edges, that **cannot** be distinguished by Theorem 5.21 (outdegree sequence method) or Theorem 5.30 (outdegree sequence and parentally closed sets method).

## 5.5 Long Matrices in Intermediate Steps

- The matrix in (5.1).

$$
\begin{array}{c}
\overbrace{\begin{array}{ccccccccc}
K_{j_0j_0} & \cdots & K_{j_1j_1} & \cdots & K_{j_qj_q} & \cdots & K_{k_1l_1} & \cdots & K_{k_ml_m}
\end{array}}^{1+\sum_{\alpha_0}\lambda^2_{j_0\alpha_0}} \\
\begin{array}{ccccccccc}
\times & \cdots & \times & \cdots & \times & \cdots & -s & \cdots & \times \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\times & \cdots & \times & \cdots & \times & \cdots & \times & \cdots & \times \\
\times & \cdots & 2s\lambda_{j_1i} & \cdots & 0 & \cdots & \times & \cdots & \times \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\times & \cdots & 0 & \cdots & 2s\lambda_{j_qi} & \cdots & \times & \cdots & \times \\
\times & \cdots & \times & \cdots & \times & \cdots & \times & \cdots & \times \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\times & \cdots & \times & \cdots & \times & \cdots & \times & \cdots & \times
\end{array}
\end{array}
\begin{array}{c}
\overbrace{\begin{array}{c}
\lambda_{k_1l_1} \\
\vdots \\
\lambda_{k_ml_m} \\
\lambda_{j_1i} \\
\vdots \\
\lambda_{j_qi} \\
\lambda_{im_1} \\
\vdots \\
\lambda_{im_p}
\end{array}}^{1+\sum_{\alpha_1}\lambda^2_{j_1\alpha_1}} \\
\begin{array}{c}
-s \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times
\end{array}
\begin{array}{c}
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots
\end{array}
\begin{array}{c}
\times \\
\vdots \\
\times \\
0 \\
\vdots \\
2s\lambda_{j_qi} \\
\times \\
\vdots \\
\times
\end{array}
\begin{array}{c}
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots
\end{array}
\begin{array}{c}
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times
\end{array}
\begin{array}{c}
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots
\end{array}
\begin{array}{c}
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times
\end{array}
\end{array}
\begin{array}{c}
\overbrace{\begin{array}{c}
\lambda_{k_1l_1} \\
\vdots \\
\lambda_{k_ml_m} \\
\lambda_{j_1i} \\
\vdots \\
\lambda_{j_qi} \\
\lambda_{im_1} \\
\vdots \\
\lambda_{im_p} \\
s
\end{array}}^{1+\sum_{\alpha_q}\lambda^2_{j_q\alpha_q}} \\
\begin{array}{c}
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times
\end{array}
\begin{array}{c}
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots \\
\cdots
\end{array}
\begin{array}{c}
-s \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times
\end{array}
\begin{array}{c}
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots \\
\cdots
\end{array}
\begin{array}{c}
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times
\end{array}
\begin{array}{c}
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots \\
\cdots \\
\vdots \\
\cdots \\
\cdots
\end{array}
\begin{array}{c}
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times \\
\vdots \\
\times \\
\times
\end{array}
\end{array}
$$

- The matrix in (5.2).



# Chapter 6

# Partial Homoscedasticity in Causal Discovery with Linear Models

The common theme in the previous chapters was the consideration of graphical models with feedback loops, which appear as directed cycles in the graphs underlying the models. In this chapter we consider a different extension of DAG models that takes up the theme of models with homoscedastic errors. In a fully homoscedastic model, the errors appearing in the structural equations all have equal variance (compare Chapter 5). Our extension considers a more nuanced approach, in which we keep with graphs that are DAGs but allow the error variances in the model specification to be equal **only in within blocks** of variables. This assumption is called **groupwise equal variance** or **partial homoscedasticity**.

Formally, our models are now given by a pair of a DAG and an associated partition, where the partition blocks indicate which groups of errors are assumed to have equal variance. The finest possible partition then corresponds to the classical model that makes no assumptions about error variance equality, whereas the coarsest partition recovers the fully homoscedastic case.

## 6.1 Setup

We consider linear structural equation models with partial knowledge about equality among the error variances. The partial knowledge is given by a partition of the node set $V$, and all nodes in the same block of the partition share the same error variance. The following two definitions give the details.

**Definition 6.1.** *Let $\Pi = \{\pi_1, \dots, \pi_K\}$ be a family of non-empty subsets of $V$. Then $\Pi$ is a partition of $V$ if $\pi_1, \dots, \pi_K$ are pairwise disjoint and $\cup_{k=1}^{K} \pi_k = V$. The sets $\pi_1, \dots, \pi_K$ are the **blocks** of the partition. Corresponding to $\Pi$ is the equivalence relation that has $i, j \in V$ equivalent if $i, j$ are in the same block of $\Pi$; we then write $i \sim_\Pi j$.*

**Definition 6.2.** *Let $G = (V, D)$ be a DAG, and let $\Pi$ be a partition of $V$. The **partially homoscedastic linear Gaussian model** given by the pair $(G, \Pi)$ is the family of all multivariate normal distributions on $\mathbb{R}^V$ with covariance matrix in the set*

$$M_{G,\Pi} = \left\{ \Sigma : \Sigma = \phi_G(\Lambda, \boldsymbol{\omega}),\ \Lambda \in \mathbb{R}^D,\ \boldsymbol{\omega} \in (0, \infty)^V \text{ with } \omega_{ii} = \omega_{jj} \text{ if } i \sim_\Pi j \right\}.$$


**(a)** $G_1$ **(b)** $G_2$

**Figure 6.1:** Under the constraint $\omega_{11} = \omega_{22}$, $G_1$ and $G_2$ generate different models.

*Given a partition* $\Pi$*, we call two DAGs* $G_1$ *and* $G_2$ ***model equivalent*** *if they induce the same partially homoscedastic linear Gaussian model, i.e., if* $M_{G_1,\Pi} = M_{G_2,\Pi}$*.*

The extra constraints on error variances lead to a refinement of the classic Markov equivalence classes obtained from only conditional independence relations. Indeed, as we will prove in Theorem 6.12 below, for every pair of nodes $i, j$ in the same block, the respective parents can be uniquely determined by the equal variance constraint. We exemplify this point in a three-variable problem.

**Example 6.3.** *Let* $G_1$ *and* $G_2$ *be the two DAGs in Figure 6.1. Under the finest partition* $\Pi_{\min} := \{\{1\}, \{2\}, \{3\}\}$*, they are in the same Markov equivalence class since they encode the same conditional independence relations. Without the assumption* $\omega_{11} = \omega_{22}$*, their models are characterized by the common conditional independence* $X_1 \perp\!\!\!\perp X_2 \mid X_3$*, and defined by the same semi-algebraic set* $\{\Sigma \mid \sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13} = 0,\ \Sigma \in PD_3\}$*, i.e.,* $M_{G_1,\Pi_{\min}} = M_{G_1} = M_{G_2} = M_{G_2,\Pi_{\min}}$*.*

*Now consider the partition* $\Pi = \{\{1, 2\}, \{3\}\}$*, i.e., the error variances satisfy* $\omega_{11} = \omega_{22}$*. The equal variance constraint gives one more equation for each model. We have*

$$M_{G_1,\Pi} = \{\Sigma \mid \sigma_{11}\sigma_{33} = \sigma_{22}\sigma_{33} - \sigma_{23}^2,\ \sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13} = 0,\ \Sigma \in PD_3\},$$

*and*

$$M_{G_2,\Pi} = \{\Sigma \mid \sigma_{22}\sigma_{33} = \sigma_{11}\sigma_{22} - \sigma_{12}^2,\ \sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13} = 0,\ \Sigma \in PD_3\}.$$

*The two models are characterized by different polynomials and are indeed different. Both models have dimension 4, but their intersection is of lower dimension.*

## 6.2 Equal Variance Constraints and Model Characterization

### 6.2.1 Equal Variance Constraints

As shown in Example 6.1, the key point of partially homoscedastic models is that they feature new constraints due to the groupwise equal variance constraints, and that these constraints alter model equivalence when compared to the traditional case without variance assumptions. We now give a general description of these constraints, using the formula for conditional covariances.

**Theorem 6.4.** *Let* $G = (V, D)$ *be a DAG, and let* $\Sigma = \phi_G(\Lambda, \boldsymbol{\omega})$ *for* $\Lambda \in \mathbb{R}^D$ *and* $\boldsymbol{\omega} \in (0, \infty)^V$*. Then for any* $i \in V$*, the error variance* $\omega_{ii}$ *can be computed from the*

*covariance matrix* $\Sigma$ *as*

$$\omega_{ii} = \sigma_{ii} - \Sigma_{i,A}(\Sigma_{A,A})^{-1}\Sigma_{A,i}, \tag{6.1}$$

*where* $A$ *may be taken to be any subset with* $\mathrm{pa}(i) \subseteq A \subseteq V \setminus \mathrm{de}(i)$.

*Proof.* We adapt the proof of Theorem 7.1 in Drton [2018], where $A = \mathrm{pa}(i)$. If a trek between $i$ and $j$ ends at $i$ with an edge of the form $k \leftarrow i$, then the trek is a directed path from $i$ to $j$ and $j \in \mathrm{de}(i)$. Now since $A \subseteq V \setminus \mathrm{de}(i)$, every trek between $i$ and a node in $A$ must end with an edge $k \to i$. By Theorem 2.3 we have

$$\Sigma_{A,i} = \Sigma_{A,\mathrm{pa}(i)}\Lambda_{\mathrm{pa}(i),i} = \Sigma_{A,A}\Lambda_{A,i},$$

where the second equality comes from the fact that $\mathrm{pa}(i) \subseteq A$ and $\Lambda_{ki} = 0$ for $k \notin \mathrm{pa}(i)$. In addition, these zeroes in $\Lambda_{A,i}$ imply that

$$\sigma_{ii} = \omega_{ii} + \Lambda^T_{\mathrm{pa}(i),i}\Sigma_{\mathrm{pa}(i),\mathrm{pa}(i)}\Lambda_{\mathrm{pa}(i),i} = \omega_{ii} + \Lambda^T_{A,i}\Sigma_{A,A}\Lambda_{A,i}. \qquad \square$$

An immediate corollary is the equation for an equal variance assumption.

**Corollary 6.5.** *If two random errors* $\epsilon_i$ *and* $\epsilon_j$ *have equal variances, i.e.,* $i$ *and* $j$ *are in the same block of a considered partition* $\Pi$*, then all covariance matrices in* $M_{G,\Pi}$ *satisfy that*

$$\sigma_{ii} - \Sigma_{i,A_i}(\Sigma_{A_i,A_i})^{-1}\Sigma_{A_i,i} = \sigma_{jj} - \Sigma_{j,A_j}(\Sigma_{A_j,A_j})^{-1}\Sigma_{A_j,j} \tag{6.2}$$

*for all subsets* $A_i$ *and* $A_j$ *such that* $\mathrm{pa}(i) \subseteq A_i \subseteq V \setminus \mathrm{de}(i)$ *and* $\mathrm{pa}(j) \subseteq A_j \subseteq V \setminus \mathrm{de}(j)$.

Theorem 6.4 admits the following converse, which is fundamental for our model characterization results.

**Theorem 6.6.** *Let* $G = (V, D)$ *be a DAG, and let* $i \in V$ *be one of its nodes. Let* $A \subseteq V \setminus \{i\}$*. Fix any vector of positive error variances* $\boldsymbol{\omega} \in (0, \infty)^V$*. If for all* $\Lambda \in \mathbb{R}^D$ *the matrix* $\Sigma = \phi_G(\Lambda, \boldsymbol{\omega})$ *satisfies equation* (6.1)*, then it must hold that* $\mathrm{pa}(i) \subseteq A \subseteq V \setminus \mathrm{de}(i)$.

*Proof.* First we suppose that there exists a node $k \in \mathrm{pa}(i) \setminus A$. Choose $\Lambda$ to have all entries zero except for $\lambda_{ki}$. For this choice, the trek rule in Theorem 2.3 implies that $\Sigma_{i,A} = 0$ and, thus, the right hand side of (6.1) is equal to $\sigma_{ii}$. But the trek rule also yields that $\sigma_{ii} = \omega_{ii} + \lambda^2_{ki}\omega_{kk} > \omega_{ii}$, which contradicts the assumption that (6.1) holds. We conclude that $\mathrm{pa}(i) \subseteq A$.

On the other hand, suppose that there exists a node $k \in A \setminus (V \setminus \mathrm{de}(i)) = \mathrm{de}(i) \cap A$. Then $G$ contains a (non-trivial) directed path from $i$ to $k$. Without loss of generality, we may assume that all interior nodes on the path between $i$ and $k$ are not in $A$. Indeed, we can always pick $k$ to be the first node in $A$ that lies on the path. So the path is of the form $i \to m_1 \to \cdots \to m_t \to k$ with $m_1, \ldots, m_t \notin A$. Now, take $\Lambda$ with all entries zero except $\lambda_{im_1}, \lambda_{m_1m_2}, \ldots, \lambda_{m_{t-1}m_t}, \lambda_{m_tk}$. The trek rule in Theorem 2.3

asserts that $\sigma_{ii} = \omega_{ii}$ under this parameterization (every trek between $i$ and $i$ has at least one edge with zero edge weight). But then equation (6.1) becomes

$$\begin{aligned}\omega_{ii} &= \sigma_{ii} - \Big(\lambda_{im_1}\lambda_{m_tk}\prod_{s=2}^{t}\lambda_{m_{s-1}m_s}\Big)^2[(\Sigma_{A,A})^{-1}]_{kk} \\ &= \sigma_{ii} - \Big(\lambda_{im_1}\lambda_{m_tk}\prod_{s=2}^{t}\lambda_{m_{s-1}m_s}\Big)^2\frac{1}{\sigma_{kk}} < \sigma_{ii} = \omega_{ii},\end{aligned}$$

which is again a contradiction. We conclude that $A \subseteq V \setminus \mathrm{de}(i)$. □

Combining Theorems 6.4 and 6.6, we can characterize the equal variance constraints by equations among functions of the covariance matrix.

**Theorem 6.7.** *Let $G = (V, D)$ be a DAG, and let $\Pi$ be a partition of the node set $V$. Suppose $i \sim_\Pi j$ are two distinct nodes that lie in the same block of $\Pi$, and let $A_i \subseteq V \setminus \{i\}$ and $A_j \subseteq V \setminus \{j\}$. Then the equation (6.2) holds for all matrices $\Sigma \in M_{G,\Pi}$ if and only if $\mathrm{pa}(i) \subseteq A_i \subseteq V \setminus \mathrm{de}(i)$ and $\mathrm{pa}(j) \subseteq A_j \subseteq V \setminus \mathrm{de}(j)$.*

*Proof.* The "if" direction is given by Corollary 6.5. For the "only if" direction, we distinguish several cases for the set $A_i$. The arguments for the corresponding different cases of $A_j$ are analogous. In each case, we construct a set of parameters such that the considered rational equation in (6.2) does not hold.

a) $\exists\, k \in \mathrm{pa}(i) \setminus A_i$: We choose $\lambda_{ki} \neq 0$ and set all other edge weights equal to zero. Then since $k \notin A_i$, the trek rule implies that $\Sigma_{i,A_i} = 0$. Hence (6.2) yields that

$$\sigma_{ii} = \sigma_{jj} - \Sigma_{j,A_j}(\Sigma_{A_j,A_j})^{-1}\Sigma_{A_j,j} \leq \sigma_{jj}.$$

By the trek rule, it further holds that $\sigma_{jj} = \omega_{jj}$ and $\sigma_{ii} = \omega_{ii} + \lambda_{ki}^2\omega_{kk}$. We arrive at the following contradiction:

$$\sigma_{ii} \leq \sigma_{jj} = \omega_{jj} = \omega_{ii} < \omega_{ii} + \lambda_{ki}^2\omega_{kk} = \sigma_{ii}.$$

We conclude that $\mathrm{pa}(i) \subseteq A_i$.

b) $\exists\, k \in \mathrm{de}(i) \cap A_i$: There is then a directed path from $i$ to $k$ and as in the proof of Theorem 6.6, we assume that $k$ was chosen such that this path is "minimial". In other words, the directed path is of the form $i \to m_1 \to \cdots \to m_t \to k$ with $m_1, \ldots, m_t \notin A_i$. We proceed by distinguishing three subcases (illustrated in Figure 6.2):

   (i) Suppose $k$ can be chosen such that there exists a directed path from $i$ to $k$ that is minimal in the above sense and does not intersect $j$. Then we can set all edge weights zero except those on the path. As in the proof of Theorem 6.6, we have $\sigma_{ii} = \omega_{ii} = \omega_{jj} = \sigma_{jj}$ and find a contradiction because under equation (6.2),

$$\omega_{ii} = \sigma_{ii} - \Big(\lambda_{im_1}\lambda_{m_tk}\prod_{s=2}^{t}\lambda_{m_{s-1}m_s}\Big)^2\frac{1}{\sigma_{kk}} < \sigma_{jj} = \omega_{jj}.$$

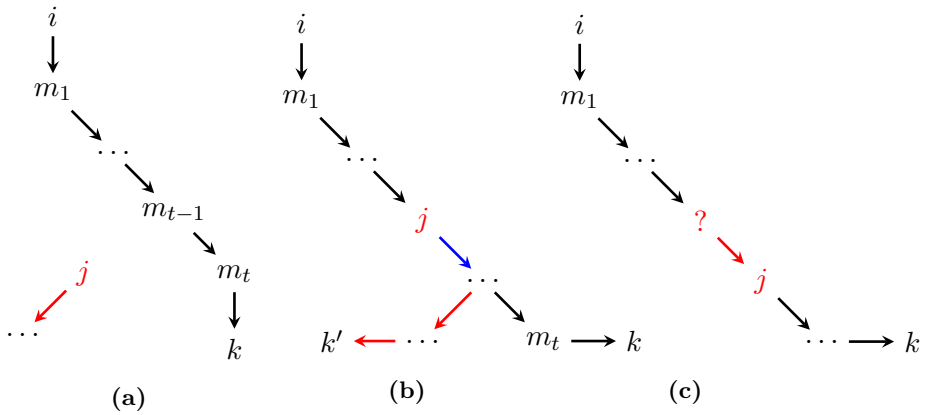



**Figure 6.2:** The three subcases when there exists a node $k \in \mathrm{de}(i) \cap A_i$.

(ii) Next, consider the case where every minimal directed path from $i$ to a node $k \in \mathrm{de}(i) \cap A_i$ contains the node $j$ and where in addition $A_j \cap \mathrm{de}(j) \neq \emptyset$. Let $k' \in \mathrm{de}(j) \cap A_j$. Then there exists a directed path from $j$ to $k'$. It follows that in this subcase $j$ must be in $\mathrm{de}(i)$. Since the graph is a DAG, the considered directed path from $j$ to $k'$ may not contain $i$. Hence, we encounter exactly the situation of subcase (i), but with the role of $i$ and $j$ switched. Hence, also in this case we can construct a counterexample to equation (6.2).

(iii) The remaining subcase is that every minimal directed path from $i$ to a node $k \in \mathrm{de}(i) \cap A_i$ contains the node $j$, and that these paths intersect $A_j$ only after they have visited $j$. Select one such minimal directed path. If the node preceding $j$ on the path is not in $A_j$, we can reduce the problem to case (a) by switching $i$ and $j$ ($\mathrm{pa}(j) \backslash A_j \neq \emptyset$). Otherwise, we set all edge weights zero except those on the considered minimal path. Let $A'_j$ be the intersection of $A_j$ and the nodes on the path. In the new DAG with only edges in the directed path, the set $A'_j$ satisfies that $\mathrm{pa}(j) \subseteq A'_j \subseteq V \backslash \mathrm{de}(j)$, and thus

$$\omega_{jj} = \sigma_{jj} - \Sigma_{j,A'_j}(\Sigma_{A'_j,A'_j})^{-1}\Sigma_{A'_j,j} = \sigma_{jj} - \Sigma_{j,A_j}(\Sigma_{A_j,A_j})^{-1}\Sigma_{A_j,j}.$$

However, computing the left hand side of (6.2) leads to a strict inequality.

$$\begin{aligned}\sigma_{ii} - \Sigma_{i,A_i}(\Sigma_{A_i,A_i})^{-1}\Sigma_{A_i,i} &= \omega_{ii} - \Big(\lambda_{im_1}\lambda_{m_tk}\prod_{s=2}^{t}\lambda_{m_{s-1}m_s}\Big)^2\frac{1}{\sigma_{kk}} < \omega_{ii} \\ &= \omega_{jj} = \sigma_{jj} - \Sigma_{j,A_j}(\Sigma_{A_j,A_j})^{-1}\Sigma_{A_j,j}. \qquad \square\end{aligned}$$

Every equal variance condition corresponds to equations of conditional variances, where the conditioning sets can be selected from a range of sets. Different conditioning sets seem to give different constraint equations, but once conditional independence constraints are taken into account, all valid conditioning sets lead to equivalent constraints.

To deal with those equivalent constraints, we partially order sets by set inclusion and extend the ordering lexicographically to pairs of sets: $(A_i, A_j) \leq (B_i, B_j)$ if $A_i \subsetneq B_i$

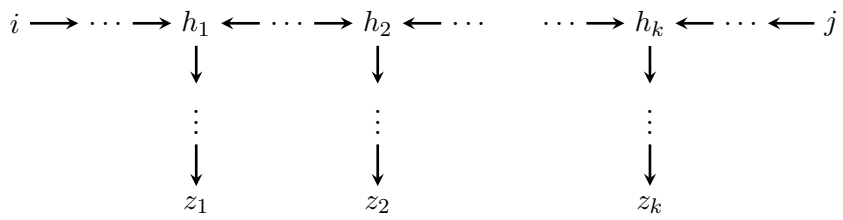



**Figure 6.3:** The active path $q$.

or if $A_i = B_i$ and $A_j \subseteq B_j$. With the help of partial ordering, we can define the minimal and maximal set pairs that match a equal variance constraint.

**Corollary 6.8.** *Let $G$ be a DAG, and let $\Pi$ be a partition of $V$ such that $i \sim_\Pi j$ are in the same block of the partition. Let $\mathcal{A}_{ij}$ be the family of all pairs $(A_i, A_j)$ with $A_i \subseteq V \setminus \{i\}$ and $A_j \subseteq V \setminus \{j\}$ for which equation* (6.2)*, i.e.,*

$$\sigma_{ii} - \Sigma_{i,A_i}(\Sigma_{A_i,A_i})^{-1}\Sigma_{A_i,i} = \sigma_{jj} - \Sigma_{j,A_j}(\Sigma_{A_j,A_j})^{-1}\Sigma_{A_j,j},$$

*holds for all covariance matrices $\Sigma \in M_{G,\Pi}$. Then*

*(i) $\mathcal{A}_{ij}$ contains a unique minimal pair, namely, $A_i = \mathrm{pa}(i)$ and $A_j = \mathrm{pa}(j)$, and*

*(ii) $\mathcal{A}_{ij}$ contains a unique maximal pair, namely, $B_i = V \setminus \mathrm{de}(i)$ and $B_j = V \setminus \mathrm{de}(j)$.*

### 6.2.2 Characterization of the Models

To give the characterization of partially homoscedastic linear models, we still need the classic conditional independence constraints from $d$-separations. We start the discussion with a proposition relating $d$-separation and conditional independence, which was proved in Geiger and Pearl [1990].

**Proposition 6.9.** *Let $G = (V, E)$ be a DAG, and let $\Pi$ be a partition of $V$. Let $i, j$ be two distinct nodes, and let $S \subseteq V \setminus \{i, j\}$. Then the conditional independence $X_i \perp\!\!\!\perp X_j \mid X_S$ holds for all multivariate normal random vectors $X$ with covariance matrix in $M_{G,\Pi}$ if and only if the d-separation $i \perp_d j \mid S$ holds in $G$.*

*Proof.* The "if" follows from Theorem 2.4 because $M_{G,\Pi} \subseteq M_G$.

For the "only if", suppose that $i$ and $j$ are not $d$-separated by $S$. We then have to construct an example of $\Sigma \in M_{G,\Pi}$ in which the conditional independence does not hold, i.e., $\det(\Sigma_{iS,jS}) \neq 0$. To this end, we may slightly modify an example of Geiger and Pearl [1990]. The modification uses equal error variances to ensure $\Sigma$ is in $M_{G,\Pi}$ and not merely in $M_G$. If $i$ and $j$ are $d$-connected given $S$, then there exists a path $q$ between $i$ and $j$, on which every collider is in $S$ (recall that our convention allows a path to visit the same node more than once). We denote the set of all these colliders by $S' = \{z_1, z_2, \ldots, z_k\} \subseteq S$; see Figure 6.3 for an illustration. In order to form a covariance matrix in $M_{G,\Pi}$, we assign the same weight $\rho \in (0, 1)$ to all edges of the path $q$ and set all other edge weights zero. We set all error variances $\omega_{ii} = 1$. Let $\Lambda$ and $\boldsymbol{w}$ be the resulting choice of parameters, and let $\Sigma = \phi_G(\Lambda, \boldsymbol{\omega})$ the associated covariance matrix.

By the trek rule, the diagonal entries of $\Sigma = (\sigma_{kl})$ satisfy that

$$\sigma_{ii} = \sigma_{jj} = 1 \quad \text{and} \quad \sigma_{kk} = 1 \, \forall \, k \notin S \setminus S',$$

because the fact that $i$ and $j$ are $d$-connected given $S$ implies that the only nodes that are both in $S$ and on the path $q$ are the colliders in the set $S'$. Next, notice that there exists a unique nonzero trek between each pair of consecutive nodes in the sequence $i \equiv z_0, z_1, z_2, \ldots, z_k, z_{k+1} \equiv j$. Let $r_t$ be the number of edges on the segment of $q$ that goes from $z_t$ to $z_{t+1}$. By the trek rule, for all $t = 0, \ldots, k$,

$$\sigma_{z_t, z_{t+1}} = \rho^{r_t}.$$

Ordering the nodes as $i, z_1, \ldots, z_k, j$ followed by the nodes in $S \setminus S'$, we obtain that

$$\Sigma_{ijS,ijS} = \left(\begin{array}{cccccc|c} 1 & \rho^{r_0} & 0 & \cdots & 0 & 0 & \\ \rho^{r_0} & \sigma_{z_1,z_1} & \rho^{r_1} & \cdots & 0 & 0 & \\ 0 & \rho^{r_1} & \sigma_{z_2,z_2} & \ddots & 0 & 0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & O \\ 0 & 0 & 0 & \ddots & \sigma_{z_k,z_k} & \rho^{r_k} & \\ 0 & 0 & 0 & \cdots & \rho^{r_k} & 1 & \\ \hline & & & O & & & I_{S\setminus S'} \end{array}\right). \tag{6.3}$$

Now observe that $\det(\Sigma_{iS,jS}) = \rho^{\sum_{t=0}^{k} r_t} \neq 0$. □

Every SEM encoded by a fixed DAG and partition satisfies some equal variance polynomial constraints, and at least one SEM does not fulfill the constraints that are not implied by the DAG and partition. Indeed, these two types of constraints do not affect each other, and we have the following characterization theorem which gives a full algebraic characterization of partially homoscedastic linear Gaussian models.

**Theorem 6.10.** *Let $G = (V, D)$ be a DAG, and let $\Pi$ be a partition of $V$. Then a covariance matrix $\Sigma \in PD$ is in the partially homoscedastic linear model $M_{G,\Pi}$ if and only if $\Sigma$ satisfies all conditional independence constraints given by d-separations* ***and*** *all equal variance constraints from Corollary 6.5.*

*Proof.* The "only if" follows from Proposition 6.9 and Corollary 6.5. For the "if" part, let $\Sigma$ satisfy all conditional independence and equal variance constraints associated to $G$. By Theorem 2.4(iii), a covariance matrix that satisfies all conditional independence constraints given by $d$-separation has to be an element of $M_G$. Hence, there exist $\Lambda \in \mathbb{R}^D$ and $\boldsymbol{\omega} \in (0, \infty)^V$ such that $\Sigma = \phi_G(\Lambda, \boldsymbol{\omega}) \in M_G$. But then, by Theorem 6.4, the equalities among conditional variances imply that $\omega_{ii} = \omega_{jj}$ for $i \sim_\Pi j$. Therefore, $\Sigma \in M_{G,\Pi}$. □

## 6.3 Equivalence Classes and CPDAG

Let $G_1 = (V, D_1)$ and $G_2 = (V, D_2)$ be two DAGs with the same given node set. We have the following definition of model equivalence.

**Definition 6.11.** *Let $\Pi$ be a fixed partition of the index set $V$. Two DAGs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are* $\Pi$***-model equivalent*** *if $M_{G_1,\Pi} = M_{G_2,\Pi}$. In this case, we write $G_1 \approx_\Pi G_2$.*

Under the finest partition $\Pi_{\min} = \{\{i\} : i \in V\}$, the Markov equivalence theory says that two DAGs with the same skeleton and unshielded colliders are distributionally equivalent/model equivalent [Studený, 2019]. According to our model characterization results, we can now state the equivalence theorem for partially homoscedastic models for general fixed partitions.

**Theorem 6.12.** *Let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two DAGs, and let $\Pi = \{\pi_1, \dots, \pi_K\}$ be a partition of the index set $V$. Then $G_1$ and $G_2$ are $\Pi$-model equivalent if and only if the following two conditions hold:*

*(i) $G_1$ and $G_2$ have the same skeleton and unshielded colliders, and*

*(ii) $\mathrm{pa}_1(i) = \mathrm{pa}_2(i)$ for all nodes $i$ that belong to a partition block $\pi_k$ of size $|\pi_k| \geq 2$.*

*Proof.* For the "if" direction, suppose that conditions (i) and (ii) hold. By the standard Markov equivalence theory, condition (i) implies that $G_1$ and $G_2$ have the same $d$-separation relations and, thus, $M_{G_1} = M_{G_2}$. Now, let $\Sigma$ be an arbitrary element of $M_{G_1,\Pi}$. Since $M_{G_1,\Pi} \subseteq M_{G_1} = M_{G_2}$, there is a (unique) choice of $\Lambda^{(2)} \in \mathbb{R}^{E_2}$ and $\boldsymbol{\omega}^{(2)} \in (0,\infty)^V$ such that $\Sigma = \phi_{G_2}(\Lambda^{(2)}, \boldsymbol{\omega}^{(2)})$. Let $i \neq j$ be any two nodes with $i \sim_\Pi j$, i.e., there is a partition block $\pi_k$ of size $|\pi_k| \geq 2$ that contains both $i, j$. By Corollary 6.5, since $\Sigma \in M_{G_1,\Pi}$, we have

$$\begin{aligned}&\sigma_{ii} - \Sigma_{i,\mathrm{pa}_1(i)}(\Sigma_{\mathrm{pa}_1(i),\mathrm{pa}_1(i)})^{-1}\Sigma_{\mathrm{pa}_1(i),i} \\ &= \sigma_{jj} - \Sigma_{j,\mathrm{pa}_1(j)}(\Sigma_{\mathrm{pa}_1(j),\mathrm{pa}_1(j)})^{-1}\Sigma_{\mathrm{pa}_1(j),j}.\end{aligned}$$

By condition (ii), $\mathrm{pa}_1(i) = \mathrm{pa}_2(i)$ and $\mathrm{pa}_1(j) = \mathrm{pa}_2(j)$. Therefore, we have

$$\begin{aligned}\omega_{ii}^{(2)} &= \sigma_{ii} - \Sigma_{i,\mathrm{pa}_2(i)}(\Sigma_{\mathrm{pa}_2(i),\mathrm{pa}_2(i)})^{-1}\Sigma_{\mathrm{pa}_2(i),i} \\ &= \sigma_{jj} - \Sigma_{j,\mathrm{pa}_2(j)}(\Sigma_{\mathrm{pa}_2(j),\mathrm{pa}_2(j)})^{-1}\Sigma_{\mathrm{pa}_2(j),j} = \omega_{jj}^{(2)}.\end{aligned}$$

We conclude that $\Sigma \in M_{G_2,\Pi}$ and, thus, $M_{G_1,\Pi} \subseteq M_{G_2,\Pi}$. Swapping the role of $G_1$ and $G_2$, we conclude that $M_{G_1,\Pi} = M_{G_2,\Pi}$ and $G_1 \approx_\Pi G_2$.

For the "only if" direction, suppose $M_{G_1,\Pi} = M_{G_2,\Pi}$. Theorem 6.10 implies that $G_1$ and $G_2$ induce the same conditional independence constraints and the same set of equal variance constraints (as specified in Corollary 6.5). We deduce that $G_1$ and $G_2$ have the same $d$-separation relations and, thus, condition (i) holds. Let $i, j$ be any two distinct nodes in the same partition block $\pi_k$. Since $G_1$ and $G_2$ induce the same set of equal variance constraints, the set $\mathcal{A}_{ij}$ defined in Corollary 6.8 is the same for $G_1$ as for $G_2$. Corollary 6.8 now implies that the unique minimal element of $\mathcal{A}_{ij}$ must be comprised of the parent sets of node $i$ and $j$ in both $G_1$ and $G_2$. But this means that $\mathrm{pa}_1(i) = \mathrm{pa}_2(i)$ and $\mathrm{pa}_1(j) = \mathrm{pa}_2(j)$. Therefore, condition (ii) holds. □

**Remark 6.13.** *The two extreme cases of our setup are the classic setting in which all variances are freely varying ($|\Pi| = |V|$, i.e., $\Pi = \Pi_{\min} = \{\{i\} : i \in V\}$) and the previously studied case with all variances equal ($|\Pi| = 1$, i.e., $\Pi = \Pi_{\max} = \{V\}$).*

*When $\Pi = \Pi_{\min}$, condition (ii) in Theorem 6.12 never applies and the theorem is just the classic Markov equivalence theorem. When $\Pi = \Pi_{\max}$, condition (ii) applies to all nodes, and every equivalence class only contains one DAG. We recover and extend the known results that the DAG is always identifiable under the assumption of equal error variances.*

**Remark 6.14.** *Another interesting special case arises in the context of two-sample problems, in which we observe each one of $p$ variables under two different experimental conditions. In this setting, one important problem is to estimate the difference between the two DAGs for the two samples. This problem is greatly simplified by assuming equality of the two error variances that arise in the structural equations for the two independent copies of the $k$th random variables, $k = 1, \ldots, p$ [Wang et al., 2018]. We can accommodate the two-sample problem in our framework by grouping all $2p$ random variables together. The combined graph is of size $|V| = 2p$, and the equal variance assumption in Wang et al. [2018] corresponds to a partition $\Pi = \{\pi_1, \ldots, \pi_p\}$ with $|\pi_1| = \cdots = |\pi_p| = 2$. Theorem 6.12 ensures that under this partition the combined DAG is uniquely determined by the joint distribution of the observations from the two samples. Although the purpose is different, our result is compatible with the difference identifiability result (Theorem 4.4 and Corollary 4.5) in Wang et al. [2018].*

We have different equivalence theory compared to the classic setup (finest partition). It is thus of interest to provide a representation of each equivalence class, as it will also no longer be the same as in the classic setup. However, we can still represent the equivalence class by a **completed partially directed acyclic graph** (CPDAG) [Andersson et al., 1997].

**Definition 6.15.** *Let $\Pi$ be a partition of the vertex set of a DAG $G = (V, D)$. The **completed partially directed acyclic graph** (CPDAG) of the DAG $G$ **under partition** $\Pi$ is the graph obtained by forming the union of all DAGs equivalent to $G$:*

$$G^*_\Pi := \cup\, (G' \mid G' \approx_\Pi G). \tag{6.4}$$

*So, $G^*_\Pi$ contains edge $i \to j$ if the edge is contained in some DAG $G' \approx_\Pi G$. It is customary to draw $G^*_\Pi$ as a mixed graph with an undirected edge between nodes $i$ and $j$ for which both $i \to j$ and $j \to i$ are in $G^*_\Pi$.*

We emphasize that an undirected edge in a CPDAG indicates that there exist two DAGs in the equivalence class in which the edge appears with opposite directions. Moreover, a CPDAG contains a directed edge $i \to j$ precisely when all DAGs in the equivalence class of $G$ contain this edge.

For the classic heteroscedastic setup (i.e., $\Pi = \Pi_{\min} = \{\{i\} : i \in V\}$), the CPDAG may be constructed using an algorithm described in Meek [1995]. In addition, Meek [1995] shows how to construct a CPDAG in a setting where there is background knowledge about some of the edges. The background knowledge is of the form $\mathcal{K} = \langle \mathbf{F}, \mathbf{R} \rangle$, where $\mathbf{F}$ contains the edges not in the DAG and $\mathbf{R}$ contains the edges in the DAG. The algorithm first translates conditional independence statements into adjacencies and unshielded collider triples. Then the first 3 of the 4 orientation rules in Verma and Pearl [1992] (Figure 6.4) are applied to obtain the CPDAG without background knowledge, which is exactly the CPDAG under $\Pi_{\min}$. The last phase incorporates background knowledge and checks whether a compatible CPDAG exists

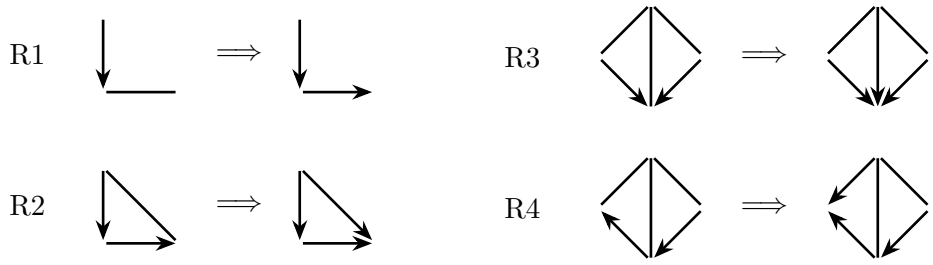



**Figure 6.4:** The four orientation rules.

or not. The following is the procedure, in which background knowledge is inserted edge by edge, and the CPDAG at the current step is denoted by $G^*$:

S1 If there is an edge $i \to j$ in $\mathbf{F}$ such that $i \to j$ in $G^*$ then FAIL.

S1′ If there is an edge $i \to j$ in $\mathbf{R}$ such that $j \to i$ in $G^*$ or $i, j$ are not adjacent in $G^*$ then FAIL.

S2 Randomly choose one edge $i \to j$ from $\mathbf{R}$, and let $\mathbf{R} = \mathbf{R} \backslash \{i \to j\}$.

S3 Orient $i \to j$ in $G^*$ and close orientations under rules R1, R2, R3 and R4 in Figure 6.4.

S4 If $\mathbf{R} \neq \emptyset$, then go to step S1.

When the partition is nontrivial and there are some equal variance constraints, the parents condition in Theorem 6.12 indicates that the neighborhood structure of nodes in the same partition is fixed in one equivalence class, which can be interpreted as background knowledge. In our setup, a CPDAG compatible to the background knowledge always exists, and we can use a simplified version of the general algorithm to construct the equivalence class.

Given a DAG $G$ and a partition $\Pi$, the equivalence class is obtained by the following algorithm (Algorithm 2). Theorem 6.16 below certifies the correctness of the algorithm.

**Algorithm 2** Constructing the equivalence class of a DAG, given the partition.

**Require:** A DAG $G$, the partition $\Pi$
1: Create an empty graph $G'$
2: Copy the skeleton and all edge orientations with unshielded colliders of $G$ to $G'$
3: Apply rules R1, R2 and R3 on $G'$ until no more edges can be oriented
4: **for** $i \in V$ with $i \in \pi_k$ and $|\pi_k| \geq 2$ **do**
5: Copy the orientation of edges in $G$ having one endpoint at $i$ to $G'$
6: **end for**
7: Apply rules R1 and R2 on $G'$ until no more edges can be oriented
8: **return** $G^*_\Pi = G'$

**Theorem 6.16.** *Given a DAG $G$ and partition $\Pi$, Algorithm 2 outputs the CPDAG $G^*_\Pi$.*

*Proof.* Algorithm 2 builds upon the work of Meek [Meek, 1995] who shows how to construct the CPDAG of an equivalence class when provided a set of conditional independence relations and arbitrary background knowledge about the edge orientations. His general algorithm first constructs the classical CPDAG by reading off unshielded colliders and propagating rules R1, R2, R3. Next, the general algorithm iteratively adds each edge from background knowledge and applies all rules R1, R2, R3, R4 to the 1-edge changes. Theorems 2-4 in Meek [1995] prove the correctness of the general algorithm.

The application of R1-R3 before inserting background knowledge creates the classical CPDAG for known conditional independence relations and without extra information (it is the CPDAG under partition $\Pi_{\min} = \{\{i\} : i \in V\}$). In our setup, we start with a DAG $G$ in the equivalence class and determine directly the skeleton and unshielded colliders and the classical CPDAG via rules R1-R3.

The partial homoscedasticity encoded in the given partition $\Pi$ now provides special "background knowledge" that fixes the orientation of all the edges with one endpoint at special nodes. As we show in the remainder of this proof, when we insert this special knowledge into the classical CPDAG, the situations of R3 and R4 in Meek [1995] cannot arise. It thus suffices to apply only R1 and R2, and we can insert all the background knowledge simultaneously, because we know that all extra information is compatible and the desired CPDAG always exists.

For our proof of the correctness of the simplifications in Algorithm 2 over Meek's general procedure, recall that the equal variance constraints give the adjacency directions of all nodes whose block has size at least 2. The set $\mathbf{R}$ consists of edges incident to these nodes, and the set $\mathbf{F}$ consists of the reversal of the edges in $\mathbf{R}$. We then argue as follows.

(i) First, we know there is at least one DAG in the equivalence class, so the general algorithm will not fail. That means the background knowledge check S1 and S1′ are redundant. We can just iteratively perform S2, S3 and S4 and obtain the same result.

(ii) Next, notice that we can add all edges in $\mathbf{R}$ simultaneously and close the orientations sequentially. Indeed, every newly oriented edge is dependent on some of the background knowledge. As long as all dependencies are added, the edge will be oriented without conflicts. Either adding edges sequentially or simultaneously would finally cover all dependencies of each orientable edge, and results in the same final output.

(iii) Finally, we claim that only the rules R1 and R2 become applicable in the orientation propagation step S3 of our algorithm. Indeed, there is an unshielded collider triple in R3, but the propagation with background knowledge does not make any new collider triples, otherwise the output CPDAG cannot have same conditional independence statements as the DAG itself. Hence, any pattern of R3 must have been obtained in the initial phase of constructing the classical CPDAG, and will not appear in the last propagation phase.

For R4, consider the first time that its pattern appears in the propagation phase. The orientation $i_3 \to i_4$ is not obtained in the classical CPDAG phase

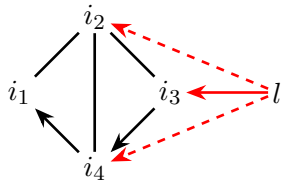



**Figure 6.5:** $i_3 \to i_4$ from R1.

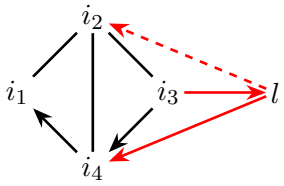
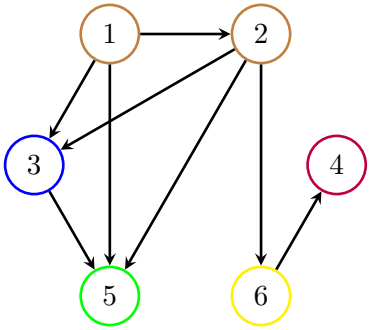
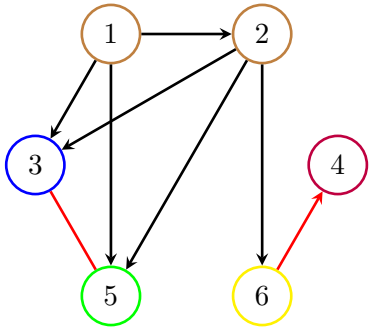



**Figure 6.6:** $i_3 \to i_4$ from R2.




**(a)** DAG $G$




**(b)** CPDAG $G'$

**Figure 6.7:** A DAG and the corresponding CPDAG, under a fixed partition.

as otherwise $i_4 \to i_1$ would have also been oriented and the pattern of R4 appears in the classic CPDAG phase, which is a contradiction. If $i_3 \to i_4$ results from background knowledge directly, then we know the orientations of all adjacencies of either $i_3$ or $i_4$, which will orient $i_2 - i_3$ or $i_2 - i_4$. This is a contradiction. Figure 6.5 depicts the case of $i_3 \to i_4$ obtained from R1: unshielded triple $l \to i_3 - i_4$. The edge $l \to i_2$ must exist to keep $i_2 - i_3$ not oriented, consequently the undirected edge $i_2 - i_4$ implies the adjacency between $l$ and $i_4$. The triple $(l, i_3, i_4)$ is shielded, contradicting the pattern of R1. Figure 6.6 illustrates the case of $i_3 \to i_4$ obtained from R2. To keep $i_2 - i_4$ not oriented, the edge $l \to i_2$ must exist. But then $i_2 - i_3$ can be oriented as $i_3 \to i_2$, which is again a contradiction.

In conclusion, we have proved that our modification to the general algorithm for equal variance constraints background knowledge is correct. □

At the end of this part, we provide an example to illustrate the construction of the CPDAG.

**Example 6.17.** *Consider the DAG $G$ in Figure 6.7 with node set $V = \{1, 2, 3, 4, 5, 6\}$ and the partition $\Pi = \{\{1, 2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$. In other words, the partition sequence is $(1, 1, 2, 3, 4, 5)$, where the $i'$th element of the sequence indicates the block that node $i$ belongs to. To determine the equivalence class of $G$, we first keep the skeleton and unshielded colliders. Then those edges containing node 1 or 2 (partition block size $\geq 2$) are oriented the same way as they are in $G$. Next, we propagate the edge orientation by rules R1 and R2, and we find that the edge between 4 and 6 is oriented as $4 \to 6$. Finally, the remaining edge $3 - 5$ can have both direction and is kept undirected in the final CPDAG that represents the equivalence class of $G$.*

## 6.4 Greedy Search and Simulation Studies

### 6.4.1 Greedy Search Scheme

Let $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^{p \times n}$ be a data matrix drawn from a multivariate normal distribution, with $n$ observations (columns) of the $p$ variables. Without loss of generality, we may assume the mean vector of the normal distribution to be zero. For a fixed DAG $G = (V, D)$ and partition $\Pi$ of $V$, the partially homoscedastic linear Gaussian model given by $(G, \Pi)$ has log-likelihood function

$$\begin{aligned} &\ell_G(\Lambda, \boldsymbol{\omega}) \\ &= \frac{n}{2}\left(-\log\det(\operatorname{diag}(\boldsymbol{\omega})) + \log\det(I-\Lambda)^2 - \operatorname{tr}\left\{(I-\Lambda)\operatorname{diag}(\boldsymbol{\omega})^{-1}(I-\Lambda)^T S\right\}\right), \end{aligned} \tag{6.5}$$

where $S = \mathbf{X}\mathbf{X}^T/n$ is the sample covariance matrix.

Let $\Pi = \{\pi_1, \ldots, \pi_K\}$ be the partition of nodes. The log-likelihood function can be rewritten as the sum of log-likelihood values of the $K$ blocks.

$$\begin{aligned} \ell_G(\Lambda, \boldsymbol{\omega}) &= \frac{n}{2}\sum_{k=1}^{K}\left(-|\pi_k|\log\omega_k - \frac{1}{n\omega_k}\left(\sum_{i\in\pi_k}\left\|X_i - \Lambda_{i,\mathrm{pa}(i)}^T X_{\mathrm{pa}(i)}\right\|^2\right)\right) \\ &:= \frac{n}{2}\sum_{k=1}^{K}\ell_{G,\pi_k}(\Lambda, \omega_k). \end{aligned} \tag{6.6}$$

The maximum likelihood estimates $(\widehat{\Lambda}, \widehat{\boldsymbol{\omega}})$ can be computed by linear regression inside each block, following the decomposition in (6.6):

$$\begin{aligned} \widehat{\Lambda}_{\mathrm{pa}(i),i} &= \operatorname*{argmin}_{\beta\in\mathbb{R}^{|\,\mathrm{pa}(i)|}} \|X_i - \beta^T X_{\mathrm{pa}(i)}\|^2, \\ \widehat{\omega}_k &= \frac{\sum_{i\in\pi_k}\left\|X_i - \widehat{\Lambda}_{\mathrm{pa}(i),i}^T X_{\mathrm{pa}(i)}\right\|^2}{n|\pi_k|}. \end{aligned}$$

To have a trade-off between model fit and model complexity, we adopt the Bayesian information criterion (BIC) score as the selection rule. We want to maximize the BIC score over the space of DAGs. Notice that the score also decomposes into the sum of score of each block:

$$\begin{aligned} s_{\mathrm{BIC}}(G) &= \frac{1}{n}\left(\ell_G(\widehat{\Lambda}, \widehat{\boldsymbol{\omega}}) - \frac{\log(n)}{2}|E|\right) \\ &= \frac{1}{2}\sum_{k=1}^{K}\left(-|\pi_k|\log\hat{\omega}_k - |\pi_k| - \frac{\log(n)}{n}\sum_{i\in\pi_k}|\,\mathrm{pa}(i)|\right). \end{aligned} \tag{6.7}$$

The greedy search scheme starts at some initial random or empty DAG and selects the DAG with highest BIC score in the local neighborhood at each step, same as that in Section 3.3. The procedure terminates when the current DAG has higher BIC score than all other DAGs in the local neighborhood. The neighborhood of a DAG is the collection of all DAGs that can be obtained from G by one edge addition, removal or reversal. To accelerate the search, we select a random subset of size up to 300 in the whole neighborhood and find the maximal score in this subset. The greedy search algorithm for groupwise equal variance models is abbreviated by GEV.

### 6.4.2 Simulation Studies

We compare the performance of our GEV algorithm against the greedy equivalence search (GES) of Chickering [2003] and the PC-algorithm [Spirtes et al., 2000]. The former tries to find the structure with the maximum $l_0$-penalized log-likelihood and the default penalty multiple is $\log(n)/2n$, corresponding to the BIC score. The latter has a significance level $\alpha$ for conditional independence tests that determine adjacencies. To make the score-based and the constraint-based methods comparable, we consider a grid of values for $\alpha$ from $10^{-5}$ to 0.8, increasing by the ratio 1.1 [Harris and Drton, 2013]. Then we can choose the value of $\alpha$ according to the maximum BIC score.

Both PC and GES algorithm return a classic Markov equivalence class, while our GEV needs the input of a fixed partition and returns the CPDAG of the final DAG in the search process. The parental information (edge directions) is the main difference between these two types of output. So we use the modified structural Hamming distance (SHD) from Peters and Bühlmann [2014] as the error measurement. The classic SHD (see Section 3.4) counts every edge mistake by 1, while the modified version assigns a distance of 2 on each pair of reversed edges.

The experiment includes 24 different configurations of $(p, n, prob)$. For these we consider $p \in \{5, 10, 20, 40\}$ as the number of nodes, $n \in \{100, 500, 1000\}$ as the sample size and $prob \in \{3/(2p-2), 0.3\}$ as the probability of one edge existing at a position $(i, j)$, which controls the sparsity of the randomly generated DAGs. The first choice of $prob$ corresponds to sparse graphs, and the latter leads to dense graphs. Every edge weight is uniformly drawn from $[-1, -0.3] \cup [0.3, 1]$, and the error variance of each partition block is uniformly drawn from $[0.3, 1]$. After sampling the adjacency between every node pair, we randomly permute the node labels.

The following box-plots summarize the result of 100 simulations, in which the $y$-axis represents the SHD between the true CPDAG and the estimated CPDAG obtained by the considered methods. In the case of 2 partition blocks (Figures 6.8 and 6.9), our GEV algorithm can successfully exploit the homoscedasticity for all setup configurations. As expected, the SHDs are lower if extra equal error variances information is utilized in the search. In the case of $\lceil p/3 \rceil + 1$ blocks (Figures 6.10 and 6.11), the GEV algorithm still performs the best for sparse graphs, while its accuracy decreases for dense graphs, especially with large number of nodes. The PC algorithm exhibits the lowest SHDs, which is also reasonable since the tuning of confidence level $\alpha$ can actually increase the accuracy.

## 6.5 Discussion

The framework of partially homoscedastic linear Gaussian models is a generalization of linear SEMs with equal error variances. It encodes equal variance assumption through a partition of the variables. The framework unifies the classical setting in which the error variances may be arbitrary and the equal error variance setup that has been studied in recent literature. These two cases are captured by the two extreme partitions, with a single block and all variables in separate blocks, respectively.

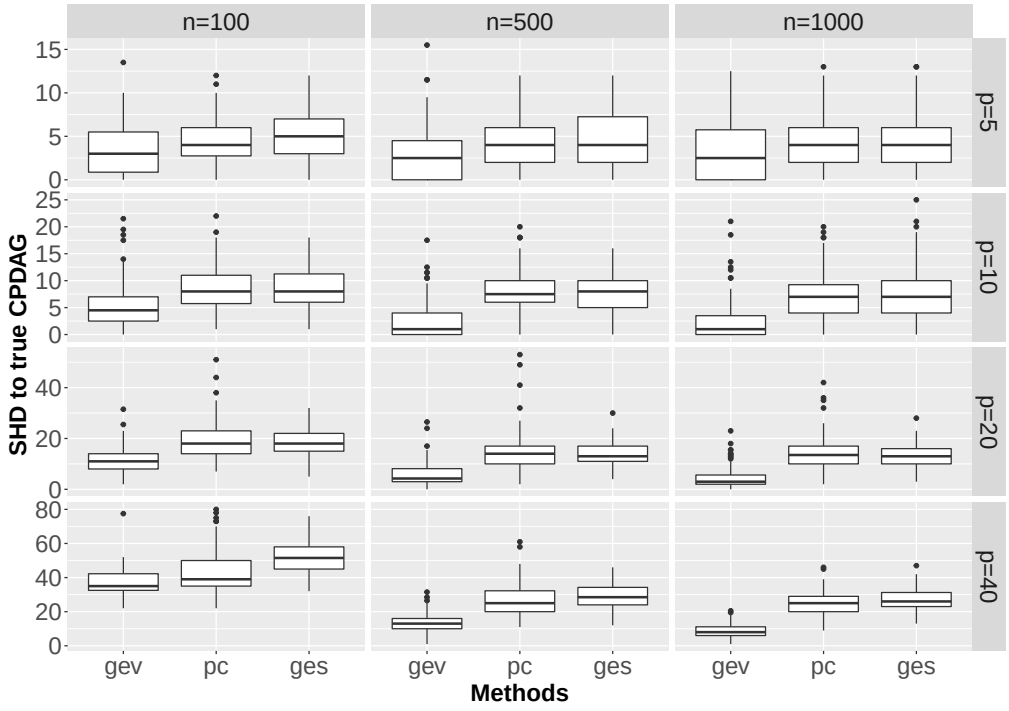



**Figure 6.8:** Box-plots of SHD by groups of $p$ and $n$, sparse graphs 2 blocks.

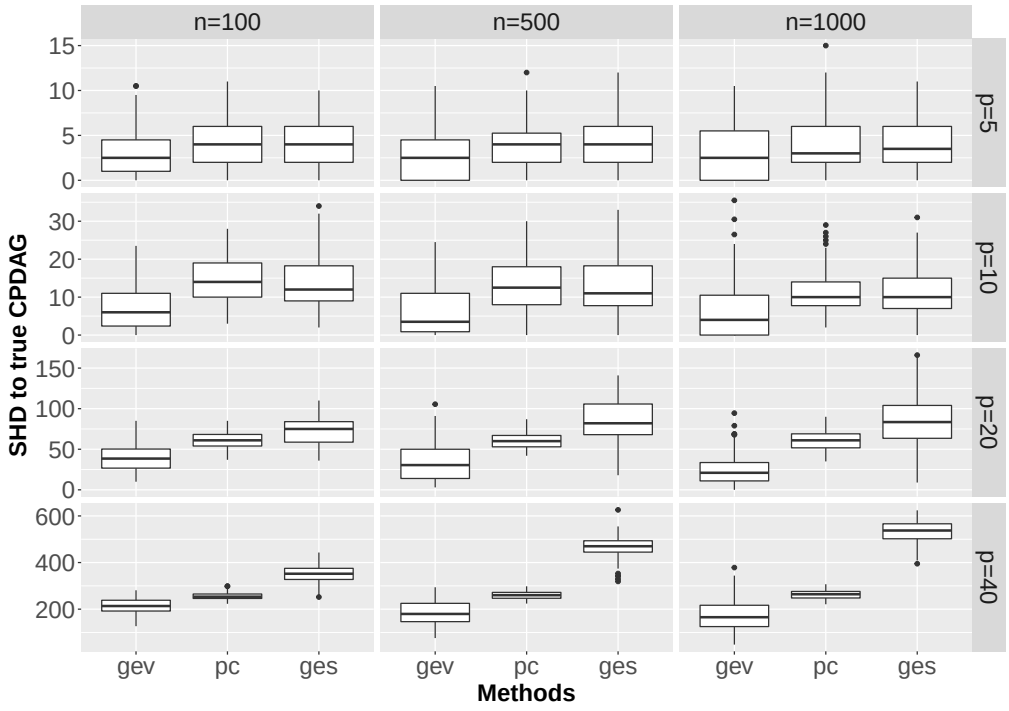



**Figure 6.9:** Box-plots of SHD by groups of $p$ and $n$, dense graphs, 2 blocks.

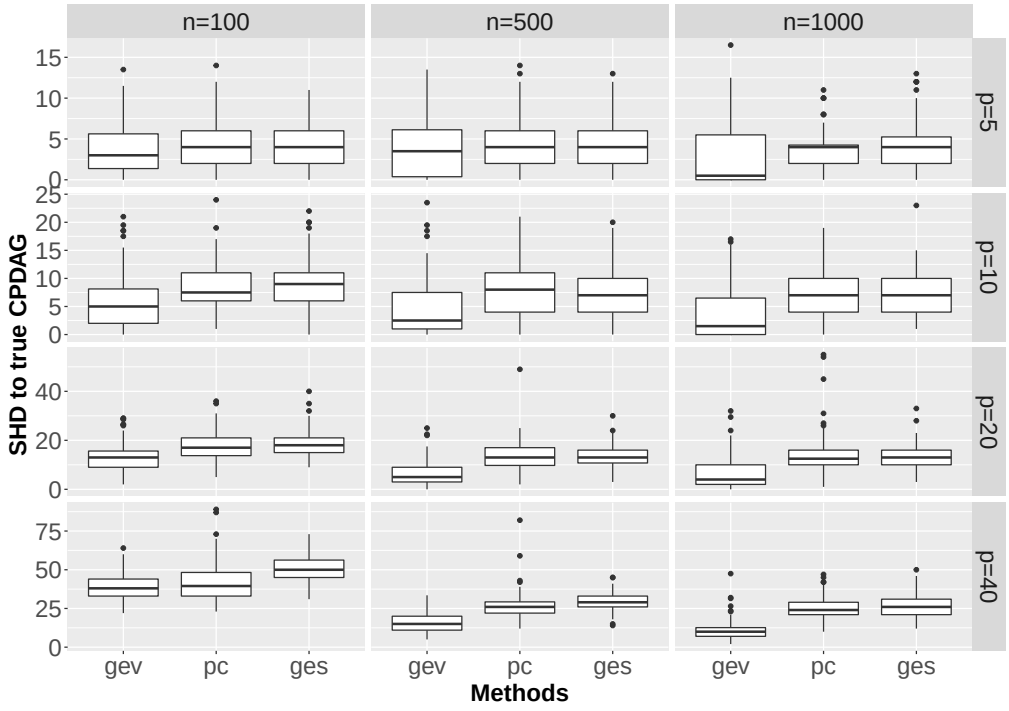



**Figure 6.10:** Box-plots of SHD by groups of $p$ and $n$, sparse graphs, $\lceil p/3 \rceil + 1$ blocks.

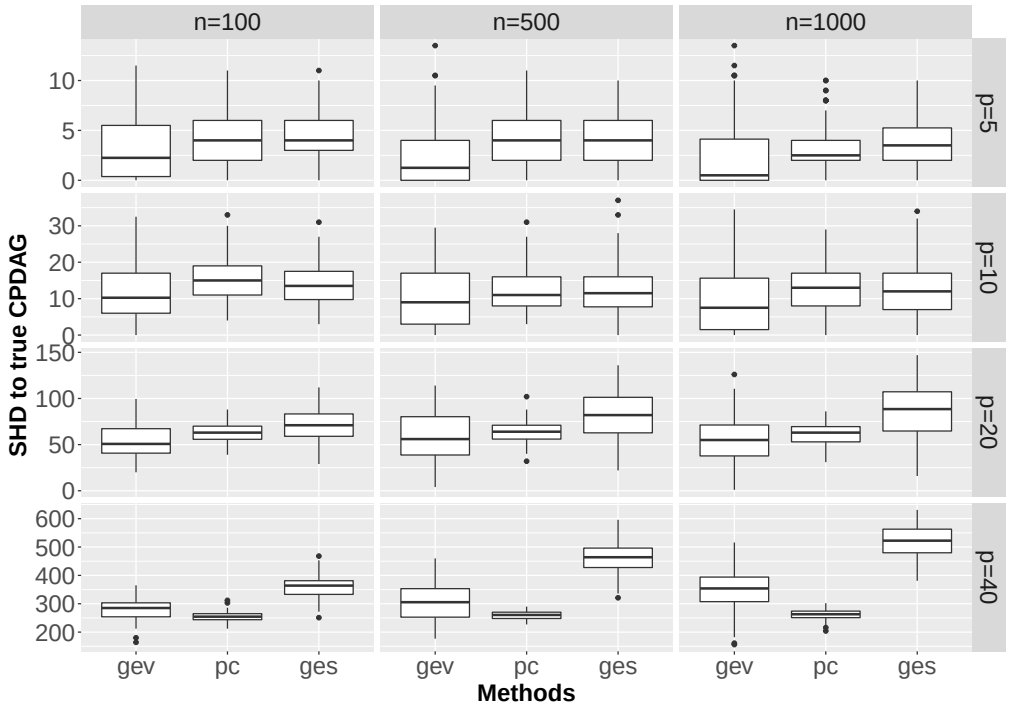



**Figure 6.11:** Box-plots of SHD by groups of $p$ and $n$, dense graphs, $\lceil p/3 \rceil + 1$ blocks.

Each partially homoscedastic linear model can be characterized algebraically via conditional independence constraints and equal variance constraints. The former are well known from the classical graphical model perspective on linear SEMs, and we explicitly derived the latter in this paper. The equal variance constraints reveal the essence of how equal variance assumptions lead to identifiability of edge orientations. This perspective differs from previous work on the equal variance assumption which, e.g., considered ordering variances [e.g. Chen et al., 2019]. We also show how equivalance classes in the partially homoscedastic setting are naturally represented by a refined CPDAG, which may be constructed efficiently with the help of existing results on CPDAGs in setting with background knowledge. For model selection, we demonstrated that greedy search provides an effective tool to exploit knowledge about partial homoscedasticity.

# Chapter 7

# Conclusions

This thesis presents advances in the research on cyclic graphical models, exploring the areas of structural identifiability, structure learning and parameter estimation.

Our first result pertains to simple mixed graphs that may contain cycles. We show that this type of models is of expected dimension, allowing for the application of model complexity penalties. We extend the distributional equivalence theory of bow-free acyclic graphs in Nowzohour et al. [2017] and develop an analogous greedy search algorithm to find the best-fitting equivalence class. This extension offers more flexibility in modeling self-regulatory networks. However, there is still a considerable gap between the available sufficient and necessary conditions for distributional equivalence (without extra identifying assumptions), and a complete characterization of equivalence classes remains unknown.

Next, we study the algorithm used in the first task to compute the log-likelihood of a graph with given data. The original algorithm, a Blockwise Coordinate Descent method described in Drton et al. [2019b], is designed only for observational data. In order to accommodate the combination of observational and interventional data from multiple environments, we have developed an algorithm with similar but different update steps. At present, our new algorithm is restricted to the case of directed cyclic graphs, and certain conditions must be satisfied by the intervention targets or the graph itself. While the original BCD algorithm solves a linear regression problem at each iteration, our algorithm for interventional data requires solving a quadratic sum-of-ratios fractional program. Under the specified conditions, this optimization problem is particularly tractable and admits a closed form solution. Extending the algorithm to handle arbitrary interventions or general mixed graphs raises interesting questions for future research.

The second half of the thesis focuses on the topics of structural identifiability and model equivalence characterization. The general question is challenging, and we introduce certain assumptions to facilitate the study of specific sub-problems. The consideration of directed cyclic graphs with equal error variances stems from the equal variance DAG model discussed in Peters and Bühlmann [2014] and numerous follow-work such as Chen et al. [2019]. To address the challenging cyclic case, we utilize the algebraic matroid approach from Hollering and Sullivant [2021] to establish sufficient graphical conditions for distinguishing the models of two graphs. The results of symbolic computation certify our criteria for small size graphs, and lead to the conjecture that the identifiability holds for the graph index parameter of all simple directed graphs.

In another generalization of the equal variance DAG setup, we propose the class of partially homoscedastic models, where variables are partitioned into blocks and all variables within the same block share the same equal error variance value. We observe that the equal variance constraints can be formulated as rational equations with entries in the covariance matrix. This allows us to provide an algebraic characterization of the model represented by a given DAG and partition. On the basis of the algebraic characterization and classic results for DAG models without variance assumptions, we are able to fully solve the model equivalence problem. For model selection we propose and study a greedy search scheme. The partial homoscedasticity offers new flexibility compared the strict equality of all error variances. A number of research opportunities emerge for future work, including testing and learning the partition.

# Bibliography


C. Améndola, P. Dettling, M. Drton, F. Onori, and J. Wu. Structure learning for cyclic linear causal models. In *Proceedings of the 36th conference on Uncertainty in artificial intelligence*, pages 999–1008, 2020.

S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

W. Chen, M. Drton, and Y. S. Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

D. M. Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2(3):445–498, 2002.

D. M. Chickering. Optimal structure identification with greedy search. volume 3, pages 507–554. 2003. Computational learning theory.

M. Drton. Algebraic problems in structural equation modeling. In *The 50th anniversary of Gröbner bases*, volume 77 of *Advanced Studies in Pure Mathematics*, pages 35–86. Math. Soc. Japan, Tokyo, 2018.

M. Drton and M. H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

M. Drton and T. S. Richardson. Iterative conditional fitting for Gaussian ancestral graph models. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 130–137, 2004.

M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*, volume 39 of *Oberwolfach Seminars*. Birkhäuser Verlag, Basel, 2009.

M. Drton, R. Foygel, and S. Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

M. Drton, C. Fox, A. Käufl, and G. Pouliot. The maximum likelihood threshold of a path diagram. *The Annals of Statistics*, 47(3):1536–1553, 2019a.

M. Drton, C. Fox, and Y. S. Wang. Computation of maximum likelihood estimates in cyclic structural equation models. *The Annals of Statistics*, 47(2):663–690, 2019b.

M. Drton, E. Robeva, and L. Weihs. Nested covariance determinants and restricted trek separation in Gaussian graphical models. *Bernoulli*, 26(4):2503–2540, 2020.

R. Evans. Markov properties for mixed graphical models. In *Handbook of graphical models*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, pages 39–60. CRC Press, Boca Raton, FL, 2019.

R. J. Evans. Model selection and local geometry. *The Annals of Statistics*, 48(6): 3513–3544, 2020.

P. Forré and J. M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th conference on Uncertainty in artificial intelligence*, pages 269–278, 2018.

R. Foygel and M. Drton. Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, 23:2020–2028, 2010.

D. Geiger and J. Pearl. On the logic of causal models. In *Uncertainty in artificial intelligence, 4*, volume 9 of *Machine Intelligence and Pattern Recognition*, pages 3–14. North-Holland, Amsterdam, 1990.

D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: graphical models and model selection. *The Annals of Statistics*, 29(2):505–529, 2001.

A. Ghassami, A. Yang, N. Kiyavash, and K. Zhang. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3494–3504. PMLR, 13–18 Jul 2020.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, pages 1–12, 1943.

N. Harris and M. Drton. PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(1):3365–3383, 2013.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society Series B*, 77(1):291–318, 2015.

B. Hollering and S. Sullivant. Identifiability in phylogenetics using algebraic matroids. *Journal of Symbolic Computation*, 104:142–158, 2021.

R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–3439, 2012.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: conflict resolution with answer set programming. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 340–349, 2014.

D. Koller and N. Friedman. *Probabilistic graphical models.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2009. Principles and techniques.

S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series.* The Clarendon Press, Oxford University Press, New York, 1996.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4): 2293–2326, 2012.

M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, editors. *Handbook of graphical models.* Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410, 1995.

J. M. Mooij and T. Claassen. Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Conference on Uncertainty in Artificial Intelligence*, pages 1159–1168, 2020.

I. Ng, A. Ghassami, and K. Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17943–17954. Curran Associates, Inc., 2020.

C. Nowzohour, M. H. Maathuis, R. J. Evans, and P. Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11(2):5342–5374, 2017.

M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, pages 763–765, 1973.

J. Pearl. *Causality.* Cambridge University Press, Cambridge, second edition, 2009. Models, reasoning, and inference.

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

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020.

K. Rantanen, A. Hyttinen, and M. Järvisalo. Discovering causal graphs with cycles and latent confounders: an exact branch-and-bound approach. *International Journal of Approximate Reasoning*, 117:29–49, 2020.

T. Richardson. A polynomial-time algorithm for deciding Markov equivalence of directed cyclic graphical models. In *Uncertainty in artificial intelligence (Portland, OR, 1996)*, pages 462–469. Morgan Kaufmann, San Francisco, CA, 1996a.

T. Richardson. A discovery algorithm for directed cyclic graphs. In *Uncertainty in artificial intelligence (Portland, OR, 1996)*, pages 454–461. Morgan Kaufmann, San Francisco, CA, 1996b.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

Z. Rosen. Computing algebraic matroids. *arXiv preprint arXiv:1403.8148*, 2014.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308 (5721):523–529, 2005.

L. Solus, Y. Wang, and C. Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

P. Spirtes. Directed cyclic graphical representations of feedback models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence: Proceedings of the 11th Conference*, pages 491–498, San Francisco, 1995. Morgan Kaufmann.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000.

M. Studený. Conditional independence and basic Markov properties. In *Handbook of Graphical Models*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, pages 3–38. CRC Press, Boca Raton, FL, 2019.

S. Sullivant. *Algebraic statistics*, volume 194 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2018.

S. van de Geer and P. Bühlmann. $\ell_0$-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

T. van Ommen and J. M. Mooij. Algebraic equivalence class selection for linear structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017.

T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the 8th Annual Conference on Uncertainty in Artificial Intelligence, Stanford University, Stanford, CA, USA, July 17-19, 1992*, pages 323–330. Morgan Kaufmann, 1992.

Y. Wang, C. Squires, A. Belyaeva, and C. Uhler. Direct estimation of differences in causal graphs. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3774–3785, 2018.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

S. Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.