Linear Structural Equation Models with Non-Gaussian Errors: Estimation and Discovery

Y. Samuel Wang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Mathias Drton, Chair

Emily B. Fox

Thomas S. Richardson

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Linear Structural Equation Models with Non-Gaussian Errors:
Estimation and Discovery

Y. Samuel Wang

Chair of the Supervisory Committee:
Chair Mathias Drton
Department of Statistics

Linear structural equation models (SEMs) are multivariate models which encode direct causal effects. We focus on SEMs in which unobserved latent variables have been marginalized and only observed variables are explicitly modeled. In this thesis, we study three problems where the distribution of the stochastic errors in the SEMs, and thus the corresponding data, are non-Gaussian. Throughout, we utilize graphical models to represent the causal structure.

First, we consider estimation of model parameters using an empirical likelihood framework when the causal structure is known. Asymptotically, under very mild conditions on the error distributions, this approach yields normal estimators and well calibrated confidence intervals and hypothesis tests. However, the procedure can be computationally expensive and suffer from poor performance when the sample size is small. We propose several modifications to a naive procedure and show that empirical likelihood can be an attractive alternative to existing methods when the data is non-Gaussian. The models considered in this section correspond to general mixed graphs.

We then consider the problem of estimating the underlying structure. Most of the previous work on causal discovery focuses on estimating an equivalence class of graphs rather than a specific graph. However, Shimizu et al. (2006) show that under certain conditions, when the errors are non-Gaussian, the exact causal structure can be identified. We extend

these results in two ways.

In Chapter 3, we show that when there is no unobserved confounding and the causal structure is suitably sparse, the identification results can be extended to the high-dimensional setting where the number of variables exceed the number of observations. The models considered correspond to directed acyclic graphs (DAGs) with bounded in-degree.

In Chapter 4, we show that non-Gaussian errors also allow for identification of the specific graph when unobserved confounding occurs in a restricted way. In particular, we consider the case where the underlying model corresponds to a bow-free acyclic path diagram (BAP). The proposed method consistently estimates the underlying structure, and unlike previous results does not require the number of latent variables or distribution of the errors to be specified in advance.

TABLE OF CONTENTS

Page

List of Figures iii

Chapter 1: Introduction 1
1.1 Linear structural equation models 1
1.2 Graphical models 3
1.3 Thesis outline 5

Chapter 2: Empirical likelihood estimation for SEMs with dependent errors 7
2.1 Introduction 7
2.2 Review of empirical likelihood 9
2.3 Empirical likelihood for SEMs 11
2.4 Numerical results 18
2.5 Discussion 25

Chapter 3: Causal discovery of high-dimensional directed acyclic graphs 27
3.1 Introduction 27
3.2 Causal discovery setup 31
3.3 Graph estimation algorithm 37
3.4 Numerical results 49
3.5 Discussion 54

Chapter 4: Causal discovery of bow-free acyclic graphs 57
4.1 Introduction 57
4.2 Causal discovery setup 61
4.3 Graph estimation algorithm 75
4.4 Numerical results 80
4.5 Discussion 87

Chapter 5: Discussion 90

Bibliography 93

Appendix A: Appendix 102
A.1 Proof of Theorem 3.1 102
A.2 Proof of Lemma 3.1 111
A.3 Proof of Lemma 3.3 113

LIST OF FIGURES

Figure Number Page

2.1 Optimization convergence comparison for naive vs profiled empirical likelihood 20
2.2 Computational effort comparison for naive vs profiled empirical likelihood . . 21
2.3 MSE comparison for competing estimation methods 22
2.4 Optimization convergence comparison of competing estimation methods . . . 23
2.5 Coverage ratio comparison for joint confidence regions 24
2.6 Data example: mixed graph from Sachs et al. (2005) 25

3.1 Examples of DAG equivalence class 29
3.2 Example of parental faithfulness 33
3.3 Low dimensional DAG estimation comparison 50
3.4 Timing comparison of min-max vs max-min methods 51
3.5 High dimensional consistency of HDL 52
3.6 Data example: estimated ordering of S&P 500 by sector; 2016 - 2017 54
3.7 Data example: estimated ordering of S&P 500 by sector; 2007 - 2017 55

4.1 Examples of BAPs 57
4.2 Examples of MAGs 60
4.3 BAP identification example 62
4.4 Comparison of competing methods for MAG discovery 82
4.5 Adversarial BAPs used in simulations 83
4.6 Comparison of competing methods for random BAP discovery 84
4.7 Comparison of competing methods for adversarial BAP discovery 85
4.8 Data example: full model from Grace et al. (2016) 86
4.9 Data example: BAP respresentation from Grace et al. (2016) 87
4.10 Data example: model discovered by BANG 88
4.11 Data example: model discovered by GBS 88

ACKNOWLEDGMENTS

I have been blessed to be surrounded by a wonderful community during my time as a graduate student in Seattle. Thanking everyone by name would leave no additional room in my dissertation for statistical discussion. Nonetheless, I would like to highlight a few specific people.

I thank my advisor, Mathias Drton. You have been a wonderful teacher and mentor. I deeply appreciate your patient explanations, careful guidance, and continuous support. My time at UW has been made richer by your helpful insights, statistical and otherwise.

I also thank Elena Erosheva for your continual effort towards making me a better statistician. You have deeply influenced the way I think about data, connect statistics to substantive problems, and communicate complex ideas.

I also thank my committee members Thomas Richardson and Emily Fox for their helpful suggestions and support. In addition, I would like to acknowledge all the teachers at UW and Rice who have guided my learning. In particular, Galen Shorack was particularly helpful in developing my statistical intuition, and Mark Embree demonstrated that mathematics was not just useful but also deeply enjoyable.

Finally, I would like to thank my parents, David and Fen Wang. Both are responsible for all of my current or future achievements. My mother is an educational psychologist by training, who selflessly gave herself to her family. Though she did not work in an academic setting, she continued conducting many successful experiments–with a sample size of 3. My father was my first math teacher, but the most important lessons have come through the way he lives his life.

DEDICATION

To my parents, David and Fen.

Then in his joy he goes and sells all
that he has and buys that field

Chapter 1

INTRODUCTION

Structural equation models are popular multivariate statistical models which directly encode causal structure. They are used in a variety of scientific settings including biology (Liu et al., 2008), ecology (Grace et al., 2016), neuroscience (Burghy et al., 2012), political science (Thrien and Nol, 2000), psychology (van der Linden, 2018), public health (Calis et al., 2008), and sociology (Matsueda and Heimer, 1987). Much of the previous work in this area assumes that the data follow a Gaussian distribution; however, in this thesis we consider the case where this assumption does not hold. In this opening chapter we give a brief review of linear structural equation models, introduce notation and terminology for graphical models, and close with a road map for the remainder of the thesis.

1.1 Linear structural equation models

In this work, we consider *structural equation models* (SEMs) where each variable is a linear function of the other variables and a stochastic error term. Often, some of the variables are latent (Bollen, 1989); however, in this work, we focus on SEMs in which the effects of latent variables–if present–have been marginalized out. Adopting the dominant linear paradigm, we will thus be concerned with models in which linear functions relate only observed variables, but error terms may be dependent. Such models are sometimes referred to as semi-Markovian (Shpitser and Pearl, 2006). Avoiding any explicit specification of latent confounding, the models play an important role in exploration of cause-effect structures (Colombo et al., 2012; Pearl, 2009; Richardson and Spirtes, 2002; Spirtes et al., 2000; Wermuth, 2011).

Formally, let $Y_1, \ldots, Y_n \in \mathbb{R}^p$ be a multivariate sample with each observation indexed by a set V with $|V| = p$. So, $Y_i = (Y_{vi})_{v \in V}$ with each Y_{vi} real-valued. Now consider the system

of structural equations

$$Y_{vi} = \mu_v + \sum_{u \in V \setminus v} \beta_{vu} Y_{ui} + \varepsilon_{vi}, \quad v \in V, \quad i = 1, \ldots, n, \tag{1.1}$$

where the μ_v and β_{vu} are unknown parameters and the ε_{vi} are random errors. Define vectors $\varepsilon_i = (\varepsilon_{vi})_{v \in V}$ and $\mu = (\mu_v)_{v \in V}$, and a matrix $B = (\beta_{vu})_{v,u \in V}$ with $\beta_{vu} = 0$ if $v = u$. We assume that the error vectors ε_i are independent and identically distributed, have zero means, and have covariance matrix $\mathbb{E}(\varepsilon_i \varepsilon_i^t) = \Omega = (\omega_{vu})_{u,v \in V}$. However, we do not specify any parametric form for their distribution. For each i, the equations in (1.1) can be written as

$$Y_i = \mu + B Y_i + \varepsilon_i. \tag{1.2}$$

If $(I - B)$ is non-singular, then this system is solved uniquely by

$$Y_i = (I - B)^{-1}(\mu + \varepsilon_i). \tag{1.3}$$

This solution has mean vector $(I - B)^{-1}\mu$ and covariance matrix

$$\Sigma(B, \Omega) \; := \; (I - B)^{-1} \Omega (I - B)^{-T}. \tag{1.4}$$

Note that Equations 1.1 and 1.2 are not simply algebraic equivalences, but can also be viewed as assignment operators which define a causal structure. Specific models are now obtained by hypothesizing that a particular collection of coefficients β_{vu} and error covariances ω_{vu} are zero. Throughout this work, the primary parameters of interest will be B and Ω, or the zero/non-zero pattern in B and Ω, so we will typically assume that $\mu = 0$. If this is not the case, one could simply consider $\tilde{Y}_i = Y_i - \bar{Y}$ without loss of generality. We will let $Y = (Y_1^T, \ldots, Y_n^T)$ be an $n \times p$ matrix where each row corresponds to a single sample and each column corresponds to an observed variable.

Example 1.1. *Suppose each week, we observe for various graduate students, indexed by i,*

the cups of coffee consumed Y_{ci}, the pages written Y_{pi}, and the hours slept Y_{si}. If we believe that the sleeping less causes more coffee to be consumed, more coffee consumed causes more pages to be written, more pages written allows one to sleep better at night, and that there are no other causal effects between the observed set of variables, we might posit the following linear SEM:

$$\begin{aligned} Y_{ci} &= \beta_{cs} Y_{si} + \varepsilon_{ci} \\ Y_{pi} &= \beta_{pc} Y_{ci} + \varepsilon_{pi} \\ Y_{si} &= \beta_{sp} Y_{pi} + \varepsilon_{si}. \end{aligned} \tag{1.5}$$

such that $Y_i = (Y_{ci}, Y_{pi}, Y_{si}) = (I - B)^{-1}\varepsilon_i$ with

$$B = \begin{bmatrix} 0 & 0 & \beta_{cs} \\ \beta_{pc} & 0 & 0 \\ 0 & \beta_{sp} & 0 \end{bmatrix}. \tag{1.6}$$

Furthermore, if we believe that there are unobserved factors, such as miles hiked (or more likely, episodes watched), which may confound the relationship between pages written and hours slept, we might allow ε_{pi} and ε_{si} to be correlated such that $\mathbb{E}(\varepsilon_i \varepsilon_i^T) = \Omega$ with

$$\Omega = \begin{bmatrix} \omega_{cc} & 0 & 0 \\ 0 & \omega_{pp} & \omega_{ps} \\ 0 & \omega_{ps} & \omega_{ss} \end{bmatrix}. \tag{1.7}$$

1.2 Graphical models

Throughout this thesis, we will use a natural graphical representation of mixed graphs/path diagrams that originates in work of Wright (1921). In particular, the causal structure of an SEM can be conveniently represented by the path diagram/mixed graph triple $G = (V, E_{\rightarrow}, E_{\leftrightarrow})$. Here, the vertex set V yields a correspondence between the nodes of the graph and the observed variables. The set $E_{\rightarrow} \subset V \times V$ is a set of *directed edges* $u \rightarrow v$, also

denoted by the ordered pair (u, v), which encode that variable u may have a direct effect on variable v. The set $E_{\leftrightarrow} \subset V \times V$ comprises *bidirected edges* $u \leftrightarrow v$, also denoted by the unordered pair $\{u, v\}$, that indicate the errors ε_{ui} and ε_{vi} may be correlated possibly due to unobserved confounding. A sequence of directed edges $v_1 \to v_2 \to \ldots \to v_t$ is a *directed path* from v_1 to v_t and a directed path which originates and ends at the same node, i.e. $v_1 = v_t$, is called a *directed cycle*. If two nodes $u, v \in V$, have both a directed and bidirected edge between them, we say there is a *bow* between u and v.

Define the *parents* of node v as the set $\text{pa}(v) = \{u \in V : u \to v \in E_{\to}\}$ and the *ancestors* of node v as $\text{an}(v)$, the set of nodes u for which there exists a directed path from u to v. We let $\text{An}(v) = \text{an}(v) \cup \{v\}$. Similarly, define the *children* of node v as the set $\text{ch}(v) = \{u \in V : v \to u \in E_{\to}\}$ and the *descendants* of node v as $\text{de}(v)$, the set of nodes u for which there exists a directed path from v to u. Define the *siblings* of v as the set $\text{sib}(v) = \{u \in V : u \leftrightarrow v \in E_{\leftrightarrow}\}$. Bidirected edges have no orientation, and $v \in \text{sib}(u)$ if and only if $u \in \text{sib}(v)$. Note that we do not permit self edges, so $(v, v) \notin E_{\to}$ and $\{v, v\} \notin E_{\leftrightarrow}$.

If there are no directed cycles, such that there does not exist a pair $u, v \in V$ where $v \in \text{an}(u)$ and $u \in \text{an}(v)$, then we refer to the graph as *acyclic*. Furthermore, if there are no bows, we refer to the graph as *bow-free*. Finally, if there are no bidirected edges so that $E_{\leftrightarrow} = \emptyset$, we refer to the graph as a *directed graph*. Graphs which are acyclic and directed or acyclic and bow-free are of particular interest and referred to as *directed acyclic graphs* (DAGs) and *bow-free acyclic path diagrams* (BAPs) respectively.

Each graph G induces a model through the requirement that

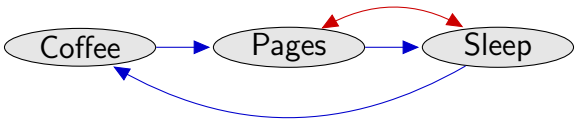
$$B \in \mathcal{B}(G) \;:=\; \left\{B \in \mathbb{R}^{V \times V} : \det(I - B) \neq 0,\ \beta_{vu} = 0 \text{ if } u \notin \text{pa}(v)\right\}, \tag{1.8}$$

$$\Omega \in \mathcal{W}(G) \;:=\; \left\{\Omega \in \mathbb{R}^{V \times V} : \Omega \text{ pos. def.},\ \omega_{vu} = \omega_{uv} = 0 \text{ if } v \notin \text{sib}(u) \text{ and } u \neq v\right\}. \tag{1.9}$$

We let $\mathcal{P}_G$ be the set of distributions consistent with a graph, such that for $Y_i = (I - B)^{-1}\varepsilon_i$

$$\mathcal{P}_G \;:=\; \{P : Y_i \sim P;\ \mathbb{E}(\varepsilon_i) = 0;\ \mathbb{E}(\varepsilon_i \varepsilon_i^T) = \Omega;\ B \in \mathcal{B}(G) \text{ and } \Omega \in \mathcal{W}(G)\}. \tag{1.10}$$

Example 1.1 (continued)**.** *We can represent the posited SEM between coffee, pages, and sleep with the following graph. Note that there is a bow between pages and sleep and all three nodes form a directed cycle.*



Given the correspondence between a node $v \in V$ and the random variable Y_v, we will at times let v stand in for Y_v; for instance, when stating stochastic independence relations. For a detailed exposition of graphical models, we refer readers to Lauritzen (1996).

1.3 Thesis outline

Typically, when working with SEMs, a researcher may be interested in some or all of the following tasks:

(i) Given graph G, estimate the free elements of B and Ω.

(ii) Given G and estimates of B and Ω, form confidence intervals or perform hypothesis tests.

(iii) Given observational data Y, posit a graph G which corresponds to the underlying causal structure.

In Chapter 2, we consider tasks (i) and (ii) when the errors, ε do not follow a Gaussian distribution. In particular, we consider general mixed graphs, which may have cycles or bows, and use an empirical likelihood estimation procedure to estimate the free elements of B and Ω, form asymptotically correct confidence intervals, and perform goodness of fit tests. The empirical likelihood procedure does not assume an explicit distribution for the errors, and yields consistent estimates and asymptotically correct procedures under very mild conditions. However, it is a computationally intensive procedure and can suffer from poor

performance at small sample sizes. Thus, we propose several modifications to a naive empirical likelihood approach including a profiled formulation, adjusted empirical likelihood to improve computational convergence at small sample sizes, and extended empirical likelihood to improve confidence interval coverage and goodness of fit calibration at small sample sizes. We show in simulations that the use of empirical likelihood may be an attractive alternative to existing methods when the data are non-Gaussian.

While the work in Chapter 2 develops a method which tolerates non-Gaussian errors, the work in Chapters 3 and 4 exploits the non-Gaussian errors for causal discovery described in Task (iii). In general, when the data are Gaussian, causal discovery methods are not able to discover a specific graph, but can only recover the underlying structure up to an equivalence class. However, it has been previously shown that when the errors are non-Gaussian, exact causal structure can be determined from observational data (Shimizu et al., 2006). We begin Chapter 3, with a brief review of these results in the case where there are no directed cycles or unobserved confounding. This corresponds to the case where the model can be represented by a directed acyclic graph (DAG). We show that under additional assumptions, the underlying causal structure can be discovered in the high dimensional case where the number of variables p may be comparable or larger than the number of observations n. This guarantee also applies to hub-graphs, graphs with large maximum out-degree.

In Chapter 4, we relax the assumption of no unobserved confounding, and consider the case of bow-free acyclic path diagrams (BAPs). These models relax the assumptions made in Chapter 3, and allow for latent confounding, albeit in a structured way. We show that the non-Gaussian identification results for DAGs also extend to BAPs. That is, when the errors are non-Gaussian, the exact causal graph, not simply an equivalence class, can be recovered from observational data. We propose a method that consistently estimates the underlying structure, which unlike previous work, does not require specification of the number of latent variables or parametric error distribution.

We close with Chapter 5, which discusses the work and presents open problems and possible avenues for future work.

Chapter 2

EMPIRICAL LIKELIHOOD ESTIMATION FOR SEMS WITH DEPENDENT ERRORS

2.1 Introduction

In this chapter, we assume that the the causal structure and confounded variables in a SEM are fixed, i.e. the graph structure $G = \{V, E_{\rightarrow}, E_{\leftrightarrow}\}$ is known. This could be due to prior scientific knowledge or the result of some discovery algorithm similar those discussed in subsequent chapters. In this context, a researcher might be interested in point estimates for the linear coefficients in B and covariances in Ω, creating confidence intervals for the point estimates, performing hypothesis tests, or comparing two competing models.

We do not restrict the structure of the graph and consider general mixed graphs (except for unidentifiable graphs discussed in Theorem 2.1), allowing for both bows and directed cycles.

2.1.1 Previous work

Often, the errors in a SEM, and consequently also the observations Y_i, are assumed to be multivariate Gaussian which yield maximum likelihood estimates (MLEs) and corresponding tests and confidence intervals calibrated by the Gaussian likelihood. Under this assumption, the Gaussian likelihood is typically maximized using generic optimization methods; as done in the popular packages `sem` (Fox et al., 2017) and `lavaan` (Rosseel, 2012) for `R` (R Core Team, 2017). The coordinate-descent methods proposed by Drton et al. (2009) and Drton et al. (2017) can be a useful computational alternative that largely avoids convergence issues.

As a less parametric method, *generalized least squares* (GLS) does not explicitly assume Gaussianity, but instead minimizes a discrepancy (which is weighted by the sample precision)

between the sample covariance and the covariance implied by the parameters. Although the estimates are slightly more robust to distributional misspecification, they are still asymptotically equivalent to the Gaussian MLEs (Olsson et al., 2000). When multivariate Gaussianity is inappropriate, MLEs and GLS generally lose statistical efficiency and yield incorrectly calibrated confidence intervals.

Weighted least squares methods (WLS)–also called *asymptotically distribution free*–weight the discrepancy between the observed and hypothesized covariance structure by explicitly estimated fourth moments. Although WLS estimates are consistent and produce asymptotically correct confidence intervals even with non-Gaussian errors, the estimation of higher order moments may come at a loss of statistical efficiency and cause convergence issues, which has limited their use (Muthen and Kaplan, 1992).

Chaudhuri et al. (2007) propose using the *empirical likelihood* (EL) of Owen (2001) to estimate a covariance matrix with structural zeros. In our setup, this corresponds to the special case of a mixed graph with no directed edges. Kolenikov and Yuan (2009) use EL to estimate the parameters of a linear SEM. In contrast to the mixed graph formulation, Kolenikov and Yuan (2009) consider the case where the latent variable structure is explicitly modeled and all errors are independent. The EL approach is appealing as it gives consistent estimates and asymptotically correct confidence intervals even when the errors are not multivariate Gaussian. However, EL can present numerous practical difficulties when the sample size is small relative to the number of parameters or estimating equations used. Moreover, standard implementation of EL methods is computationally feasible only for systems with a handful of variables. We believe that these issues have prevented application of EL to linear SEMs beyond what was done by Kolenikov and Yuan (2009).

2.1.2 Contribution

In this chapter, we apply the empirical likelihood framework to SEMs represented by mixed graphs and propose several modifications to a naive approach which address the most salient practical concerns:

(i) We show that in the mixed graph setting, the covariance parameters Ω can be profiled out. This greatly reduces the computational burden by reducing the number of estimating equations imposed and parameters directly estimated. It also naturally encodes the positive definite constraint on Ω and yields a positive definite estimate of Ω for any point B with a well defined empirical likelihood.

(ii) When maximizing the empirical likelihood, we leverage a recent insight and directly incorporate gradient information in a quasi-Newton procedure instead of the typical derivative-free approaches to empirical likelihood optimization. This again yields substantial computational savings.

(iii) We use the *adjusted empirical likelihood* (AEL), first proposed by Chen et al. (2008). This adjustment ensures that an empirical likelihood and corresponding gradient is well defined for every value in the parameter space.

(iv) We apply the idea of *extended empirical likelihood* (EEL), which furnishes drastically improved coverage of confidence intervals at small sample sizes (Tsao and Wu, 2014).

Our simulations show that with these proposed modifications, empirical likelihood becomes an attractive alternative for practitioners concerned with non-Gaussianity in structural equation modeling.

2.2 Review of empirical likelihood

We first give a review of the empirical likelihood framework which gives estimates and calibrates confidence intervals via maximization of the empirical likelihood function (Owen, 2001).

Let $Y = (Y_1^T, \ldots, Y_n^T)$ be a sample from a p-variate distribution P belonging to a non-/semiparametric statistical model $\mathcal{M}$. Let Δ_n be the $n-1$ dimensional probability simplex. For $\delta = (\delta_1, \ldots, \delta_n) \in \Delta_n$, define the log-empirical likelihood $\ell(\delta; Y) = \sum_{i=1}^n \log(\delta_i)$. This is the log-likelihood of the sample under the discrete distribution with mass δ_i at each point Y_i. Suppose we are interested in a parameter $\theta = \theta(P)$ taking values in $\Theta \subseteq \mathbb{R}^d$ such that

for a map $F : \mathbb{R}^p \times \mathbb{R}^d \mapsto \mathbb{R}^q$ we have $\mathbb{E}_P F(Y_i, \theta(P)) = 0$ for all $P \in \mathcal{M}$. The log-empirical likelihood at a given parameter value θ is then

$$\ell(\theta; Y) = \max_{\delta \in \Delta_\theta} \ell(\delta; Y) = \max_{\delta \in \Delta_\theta} \sum_{i=1}^{n} \log(\delta_i), \tag{2.1}$$

where the feasible set

$$\Delta_\theta = \left\{ \delta \in \Delta_n : \sum_{i=1}^{n} \delta_i F(Y_i, \theta) = 0 \right\} \tag{2.2}$$

reflects that the expectation of $F(\cdot; \theta)$ vanishes for distributions compatible with θ.

The empirical likelihood (EL) from (2.1) provides a basis for statistical inference. Maximizing it over $\theta \in \Theta$ yields the *maximum empirical likelihood estimator*

$$\check{\theta} = \arg\max_{\theta} \ell(\theta; Y) \tag{2.3}$$

that we refer to as MELE. Ratios of the EL yield *empirical likelihood ratio* statistics. Owen (1988) derives an EL analogue of Wilk's Theorem, and the result was expanded to the general estimating equation framework by Qin and Lawless (1994). The specific regularity conditions needed are discussed in Section 2.3.3, and the results imply under very general conditions that the MELE is consistent and asymptotically normal. In addition, EL ratio statistics have limiting χ^2 distributions that can be used to calibrate statistical tests and create confidence intervals or regions. For a detailed exposition of these ideas, we refer readers to Owen (2001).

The nice theoretical properties for EL, however, come at a high practical cost. The practical issues become particularly pressing for applications to linear SEMs, for which the number of parameters and estimating equations generally grow on the order of p^2, where $p = |V|$ is the number of variables considered. We describe three difficulties that complicate the direct use of EL for SEMs:

(i) For some values θ, the origin may be outside the convex hull of $\{F(Y_i, \theta) : i = 1, \ldots, n\}$, in which case the feasible set Δ_θ from (2.2) is empty and the EL at θ is zero. This

"convex hull problem" occurs more often when the sample size is small relative to the number of estimating equations or when the data is skewed. As discussed by Grendár and Judge (2009), it is possible that $\Delta_\theta = \emptyset$ for all parameter vectors θ, which is known as the "empty set problem". In addition, the log-EL is typically not a convex function (Chaudhuri et al., 2017), and finding an initial point that has well-defined EL and is in the basin of attraction of the MELE can be difficult.

(ii) The optimization problem defining the log-EL $\ell(\theta; Y)$ from (2.1) is typically solved iteratively through its dual. Although this problem is convex, it can be computationally burdensome when the number of estimating equations, which corresponds to the number of dual variables, is large.

(iii) Confidence intervals based on the asymptotic normal variance and χ^2 likelihood ratio calibration have been shown to often undercover at small sample sizes (Tsao and Wu, 2014).

2.3 Empirical likelihood for SEMs

We now turn to the application of EL to SEMs. For expository simplicity, we assume throughout that our observations are centered. In other words, the intercept parameter vector μ for (1.2) is zero, so that $\mathbb{E}(Y_i) = 0$. However, our ideas extend straightforwardly to the case where we also make inference about $\mu \neq 0$.

2.3.1 Profiled formulation

Consider the linear SEM given by a mixed graph $G = (V, E_{\rightarrow}, E_{\leftrightarrow})$. The general framework laid out in Section 2.2 can be applied directly to such a model by taking the covariance matrix of the observations Y_i as the general parameter θ. We may then define an EL at a pair of parameter matrices (B, Ω) as the EL at the covariance matrix $\Sigma(B, \Omega)$ from (1.4). In such a direct application to the linear SEM, the log-EL function $\ell(B, \Omega; Y)$ is the maximum

of the log-EL $\ell(\delta; Y)$ over the set

$$\Delta_{\Sigma(B,\Omega)} \;=\; \left\{ \delta \in \Delta_n : \sum_{i=1}^{n} \delta_i Y_i = 0,\; \sum_{i=1}^{n} \delta_i \left[\operatorname{vech}\left(Y_i Y_i^T\right) - \operatorname{vech} \Sigma(B,\Omega) \right] = 0 \right\}. \qquad (2.4)$$

Here, vech is the half-vectorization operator for symmetric matrices. Under this formulation, there are p constraints for the mean and $p(p+1)/2$ covariance constraints, and the MELE is computed by optimization with respect to the pair of $p \times p$ matrices (B, Ω), with Ω restricted to be positive definite.

Inspection of the covariance constraints reveals that a great simplification is possible by profiling out Ω. Indeed, the covariance constraint yields an explicit solution for Ω given B, Y, and δ. Specifically, with $D = \operatorname{diag}(\delta_1, \ldots, \delta_n)$, we have

$$Y^T D Y = \Sigma(B,\Omega) = (I-B)^{-1}\Omega(I-B)^{-T} \iff (I-B)Y^T D Y (I-B)^T = \Omega. \qquad (2.5)$$

The entries of Ω are either constrained to be zero or freely varying. No constraints arise from the freely varying entries, and we may base estimation of B on only the structural zeros in Ω, that is,

$$\left\{ (I-B)Y^T D Y (I-B)^T \right\}_{uv} = 0 \quad \forall \{u,v\} \notin E_{\leftrightarrow}.$$

Once a solution for B is found, we may simply compute $\Omega = \Omega(B)$ by setting

$$\omega_{uv} = \left\{ (I-B)Y^T D Y (I-B)^T \right\}_{uv}$$

for $u = v$ or $\{u,v\} \in E_{\leftrightarrow}$. The profile log-EL in this approach is the function

$$\ell(B; Y) \;=\; \max_{\delta \in \Delta_B} \ell(\delta; Y) \qquad (2.6)$$

obtained from the set of weight vectors

$$\Delta_B \;=\; \Big\{\, \delta \in \Delta_n : \quad \sum_{i=1}^n \delta_i Y_i = 0, \\ \sum_{i=1}^n \delta_i \left(Y_{vi} - \sum_{s \in \mathrm{pa}(v)} \beta_{vs} Y_{si} \right) \left(Y_{ui} - \sum_{t \in \mathrm{pa}(u)} \beta_{ut} Y_{ti} \right) = 0 \quad \forall \{v,u\} \notin E_{\leftrightarrow} \Big\}. \tag{2.7}$$

The MELE $\check{B}$ is found by maximizing $\ell(B;Y)$ over the set $\mathcal{B}(G)$ from (1.8), and then $\check{\Omega} = \Omega(\check{B})$. We emphasize that there are now only $p(p-1)/2 - |E_{\leftrightarrow}|$ covariance constraints, and only the matrix B needs to be optimized. This leads to the somewhat odd observation that adding parameters (specifically bidirected edges) actually decreases the computational burden. Even when $|E_{\leftrightarrow}| = 0$, there is still a benefit due to the profiled variance parameters.

Following a standard strategy, we perform an outer and inner maximization (Owen, 2001) as

$$\max_{B \in \mathcal{B}} \ell(B;Y) \;=\; \max_{B \in \mathcal{B}} \max_{p \in \Delta_B} \ell(\delta;Y). \tag{2.8}$$

First, we evaluate $\ell(B;Y)$, that is, solve the "inner maximization" in (2.6) at a fixed B, through the dual problem. Strong duality holds because the constraints in (2.7) are linear in the weights δ_i. Let $F(Y_i,B)$ be the map with coordinates $F_v(Y_i,B) = Y_{iv}$ for $v \in V$ and $F_{uv}(Y_i,B) = f_u(Y_i,B) f_v(Y_i,B)$ for each nonedge $\{u,v\} \notin E_{\leftrightarrow}$, where $f_v(Y_i,B) = Y_{vi} - \sum_{s \in \mathrm{pa}(v)} \beta_{vs} Y_{si}$. With dual variables $\alpha \in \mathbb{R}$ and $\lambda \in \mathbb{R}^{p+p(p-1)/2-|E_{\leftrightarrow}|}$, the Lagrangian for the inner optimization over Δ_B is

$$L_B(\delta,\alpha,\lambda) = -\sum_{i=1}^n \log(\delta_i) + \alpha \left(\sum_{i=1}^n \delta_i - 1 \right) + n \sum_{i=1}^n \delta_i \left(\sum_{v \in V} \lambda_v F_v(Y_i,B) + \sum_{\{u,v\} \notin E_{\leftrightarrow}} \lambda_{uv} F_{uv}(Y_i,B) \right). \tag{2.9}$$

Combining the first order conditions for δ_i with the constraint that $\sum_i \delta_i = 1$ implies that

$\alpha = n$ at the global maximum. Maximizing over the weights, with $\alpha = n$, we find

$$\check{\delta}_i = \frac{1}{n} \frac{1}{1 + \sum_{v \in V} \lambda_v F_v(Y_i, B) + \sum_{\{u,v\} \notin E_{\leftrightarrow}} \lambda_{uv} F_{uv}(Y_i, B)}, \tag{2.10}$$

and substitution into L_B yields a convex dual function of λ. We optimize it via Newton-Raphson with a backtracking line search to ensure $0 \leq \delta_i \leq 1$.

In the "outer maximization", we optimize $\ell(B; Y)$ with respect to B using a gradient based quasi-Newton method. Although we can only evaluate $\ell(B; Y)$ numerically, once we have the optimal dual variables λ and the corresponding weights from (2.10), we can analytically compute the gradient of $\ell(B; Y)$ as

$$\nabla \ell(B; Y) = -\lambda^T \sum_{i=1}^{n} \check{\delta}_i \nabla F(Y_i, B); \tag{2.11}$$

see Chaudhuri et al. (2017) for further discussion of the properties of the gradient of the log empirical likelihood. The Hessian, however, cannot be computed in closed form, so we use BFGS which builds an approximate Hessian via the gradient.

Although both formulations yield the same MELE, the profile approach from (2.6) and (2.7) drastically eases difficulties (i) and (ii) discussed in Section 2.2 as the number of estimating equations for the covariance is reduced to $p + p(p-1)/2 - |E_{\leftrightarrow}|$. This reduces the number of dual variables to optimize in the inner maximization. Moreover, when profiled, the outer maximization searches over only $B \in \mathcal{B}(G)$ while the naive direct formulation from (2.4) requires a search over both $B \in \mathcal{B}(G)$ and $\Omega \in \mathcal{W}(G)$; in particular, positive definiteness of Ω needs to be respected in the naive optimization. Finally, satisfying the convex hull condition for the error covariances typically requires a simultaneous good choice of B and Ω. The directed edge weights can be easily initialized with regression estimates, but the covariance parameters are typically more difficult to specify. In Section 2.4, we show that the computational advantages produce substantial gains in computation time and converge to a valid stationary point at a much higher proportion of the time even when the sample

size is small.

2.3.2 Small sample improvements

In addition to reformulating the optimization problem, we make two modifications to improve the performance of EL for SEMs. We apply *adjusted empirical likelihood* (AEL) to improve the search for a MELE and use *extended empirical likelihood* (EEL) to improve the coverage of confidence intervals.

Chen et al. (2008) proposed AEL to alleviate the convex hull problem mentioned in difficulty (i) above. The adjustment amounts to adding a pseudo-observation whose contribution to the estimating equations is $F_{n+1}(B) = -a_n \bar{F}(B) = -a_n \frac{1}{n} \sum_{i=1}^{n} F(Y_i, B)$ for a choice of $a_n > 0$. Adding this term ensures that no matter the value of B, the set of feasible weight vectors, now in Δ_{n+1}, is non-empty. Hence, the log-AEL $\ell^a(B; Y)$ and its gradient

$$\nabla \ell^a(B; Y) = -\lambda^T \sum_{i=1}^{n} \left[\check{\delta}_i + \left(-\frac{a_n}{n} \right) \check{\delta}_{n+1} \right] \nabla F(Y_i, B) \tag{2.12}$$

are well defined across the entire parameter space. Chen et al. (2008) show that AEL retains the asymptotic properties of the original EL when $a_n = o(n^{2/3})$, and suggest $a_n = \log(n)/2$. We adopt this choice.

The terms in our covariance constraint are products, $F_{uv}(Y_i, B) = g_v(Y_i, B) g_u(Y_i, B)$. This is generally not true for the added term $F_{n+1}(B)$ and it is not straightforward how to define an appropriately sparse and positive definite matrix $\Omega(B)$ using AEL weights. Thus, we propose finding an estimate $\check{B}$ that maximizes the AEL and computing $\check{\Omega} = \Omega(\check{B})$ based on weights from recalculating the original EL at $\check{B}$. As demonstrated in our numerical experiments, this approach alleviates some convergence issues but, of course, the original EL may be zero at the AEL maximizer $\check{B}$, in which case we do not have an estimate of Ω and say that the AEL procedure has not converged.

To address undercoverage of confidence regions for smaller samples, as described in difficulty (iii), we adopt the EEL of Tsao and Wu (2014) who show that their χ^2-calibrated EEL

confidence regions outperform those from the original EL. Assuming the MELE $\check{B}$ exists, a positive EEL may be defined for any matrix $B \in \mathcal{B}(G)$ by taking the original EL at a convex combination of B and $\check{B}$. Specifically, the log-EEL suggested by Tsao and Wu (2014) is

$$\ell^e(B;Y) = \ell(h^{-1}(B,Y),Y) \tag{2.13}$$

for $h(B,Y) = \check{B} + \gamma(n,\ell(\theta;Y))(B-\check{B})$ with $\gamma(n,\ell(B;Y)) = \left(1 + \frac{2(-n\log(n)-\ell(B;Y))}{2n}\right)$.

2.3.3 Asymptotic distribution of empirical likelihood estimators

It follows from Qin and Lawless (1994, Theorem 1) that under the following assumptions, MELEs are asymptotically normal and empirical likelihood ratios converge to χ^2 limits. The same is true for the modifications from Section 2.3.2.

Proposition 2.1. *Let $G = (V, E_{\rightarrow}, E_{\leftrightarrow})$ be a mixed graph, let $B_0 \in \mathcal{B}(G)$ and $\Omega_0 \in \mathcal{W}(G)$. Let ϵ be a zero-mean random vector with covariance matrix Ω_0. Assume that:*

- *(a) The Jacobian of the parametrization $(B,\Omega) \mapsto \Sigma(B,\Omega)$ defined on $\mathcal{B}(G) \times \mathcal{W}(G)$ has full rank at (B_0,Ω_0).*
- *(b) The joint distribution of ϵ and $\epsilon^{(2)} = (\epsilon_v\epsilon_u : v,u \in V)$ is non-degenerate and has finite third moments.*

If $Y_1,\ldots,Y_n$ is an i.i.d. sample from the distribution determined by (B_0,Ω_0,ϵ), i.e., the distribution of $(I-B_0)^{-1}\epsilon$, then the MELE $\check{\theta} = \left(\text{vech}\left[\check{B}\right], \text{vech}\left[\Omega(\check{B})\right]\right)$ is asymptotically normal with

$$\sqrt{n}\left(\check{\theta} - \theta_0\right) \to N(0,V), \qquad V^{-1} = \mathbb{E}\left(\frac{\partial F(Y,\theta_0)}{\partial\theta}\right)^T \mathbb{E}[F(Y,\theta_0)F(Y,\theta_0)^T]^{-1}\mathbb{E}\left(\frac{\partial F(Y,\theta_0)}{\partial\theta}\right). \tag{2.14}$$

Here, F is given by the estimating equations corresponding to the naive formulation in (2.4). Furthermore, EL ratio statistics have χ^2 limits. In particular, for $q = p + p(p+1)/2$ and

$d = |E_{\rightarrow}| + |E_{\leftrightarrow}| + p$, *we have*

$$2\left(-n\log(n) - \ell(\check{\theta}; Y)\right) \to \chi^2_{(q-d)}, \qquad 2\left[\ell(\check{\theta}; Y) - \ell(\theta_0; Y)\right] \to \chi^2_d. \tag{2.15}$$

Proof. We recall our notation $\theta = (B, \Omega)$ and $\theta_0 = (B_0, \Omega_0)$. Based on the right-most expression in (2.5), the considered naive/direct estimating equations may be based on the function $F(y, B)$ with coordinates

$$F_v(y, B) = y_v, \qquad F_{uv}(y, B) = \left(y_v - \sum_{s \in \mathrm{pa}(v)} \beta_{vs} y_s\right)\left(y_u - \sum_{t \in \mathrm{pa}(u)} \beta_{ut} y_t\right) - \omega_{uv},$$

for $v \in V$ and $\{u, v\} \in V \times V$, respectively.

Our claim follows from Theorem 1 of Qin and Lawless (1994) under the following conditions:

(1) $\mathbb{E}(F(Y_i, \theta_0)F(Y_i, \theta_0)^T)$ is positive definite.

(2) In a neighborhood of the d-dimensional parameter θ_0, the derivative $\frac{\partial F(y,\theta)}{\partial \theta}$ is continuous, and $\left\|\frac{\partial F(y,\theta)}{\partial \theta}\right\|$ and $\|F(y, \theta)\|^3$ are bounded by an integrable function $M_1(y)$.

(3) $\mathbb{E}\left(\frac{\partial F(Y_i,\theta_0)}{\partial \theta}\right)$ has rank d.

(4) $\partial^2 F(y, \theta)/\partial\theta\theta^T$ is continuous and $\left\|\partial^2 F(y, \theta)/\partial\theta\theta^T\right\|$ is bounded by an integrable function $M_2(y)$ in a neighborhood of the true parameter θ_0.

Here, $\|\cdot\|$ denotes the Euclidean norm.

Noting that $F_{uv}(Y_i, B_0) = \epsilon_v\epsilon_u - \omega_{uv}$, condition (1) is an immediate consequence of assumption (b) in our proposition. Condition (3) is implied by assumption (a). With polynomial estimating equations, all derivatives in conditions (2) and (4) exist. Now, F and its first and second partial derivatives are at most quadratic functions of Y_i, which in turn is a linear function of a realization of the error vector ϵ. Local bounds on the concerned quantities are easily obtained and assumption (b) ensures their integrability. □

If the rank condition from (a) holds, then the rational map $(B, \Omega) \mapsto \Sigma(B, \Omega)$ has full

rank Jacobian at almost all choices of (B, Ω), and the map is generically finite-to-one. There is thus a connection to local/finite identifiability of (B, Ω) from the covariance matrix. For state-of-the-art methods for determining identifiability see Foygel et al. (2012); Chen (2016); Drton and Weihs (2016).

2.4 Numerical results

We now show a series of numerical experiments to evaluate the effectiveness of the proposed methods and compare the results to existing methods.

2.4.1 Convergence of optimizers for naive vs profile formulation

We first compare the naive/direct procedure which explicitly estimates B and Ω to the profiled procedure which only involves B. For both procedures, we use the original EL and adjusted EL. We also consider a hybrid method, which first finds the maximum AEL point to initialize a search which then uses original EL. We randomly generate acyclic mixed graphs with 8 nodes, 10 directed edges, and 6 bidirected edges. We randomly select directed edges $u \to v$ from all pairs such that $u < v$ and then select bidirected edges $u \leftrightarrow v$ from the remaining unselected pairs. This setup ensures that (B, Ω) are generically identifiable from $\Sigma(B, \Omega)$ by the result of Brito and Pearl (2002).

We generate random true parameter matrices $B = (\beta_{uv})$ and $\Omega = (\omega_{uv})$ as follows. The coefficients β_{uv} are drawn uniformly from $(-1, .-.2) \cup (.2, 1)$. For Ω, we draw off-diagonal elements $\omega_{uv} = \omega_{vu}$, $u \neq v$, uniformly from $(-.8, .-.3) \cup (.3, .8)$. We then use exponential draws to set $\omega_{vv} = \sum_{u \neq v} |\omega_{uv}| + 1 + \exp(1)$.

We consider errors from four distributions. First, we generate centered multivariate Gaussian errors with covariance matrix Ω. Second, we generate them from a multivariate T-distribution with 4 degrees of freedom, which we denote by T_4, again with expectation zero and covariance matrix Ω. Third, we consider log-normal errors. In this case, we simulate a multivariate Gaussian vector Z, centered and with covariance matrix equal to the correlation matrix C that corresponds to Ω. We then set the error vector to $\epsilon = \exp(Z) - \sqrt{e}$, which

yields covariance matrix $e(\exp(C) - 1)$. Finally, in order to draw a multivariate distribution with recentered gamma marginals and covariance Ω, we follow the steps:

1. Draw $\epsilon_v \sim \text{gamma}(\text{shape} = \omega_{vv} - \sum_{v \neq u} |\omega_{uv}|, \text{scale} = 1)$.
2. For each $\{u, v\} \in E_{\leftrightarrow}$, generate $\delta_{uvi} \sim \text{gamma}(\text{shape} = |\omega_{uv}|, \text{scale} = 1)$ and a random sign $\xi_{uv} \in \{-1, 1\}$.
3. If $\omega_{uv} > 0$, add $\xi_{uv}\delta_{uvi}$ to ϵ_{ui} and ϵ_{vi}. If $\omega_{uv} < 0$, add $\xi_{uv}\delta_{uvi}$ to ϵ_{ui} and $-\xi_{uv}\delta_{uvi}$ to ϵ_{vi}.
4. Subtract the true mean from each error term so that it has mean 0.

All optimizations are initialized with a procedure from Drton et al. (2017), where the free elements of B are calculated via least squares. The resulting residuals are used to initialize the non-zero values ω_{uv}. If a row is not diagonally dominant, the off-diagonal elements are scaled so that $\sum_{j \neq i} |\omega_{ij}| < .9 \times \omega_{ii}$ to ensure Ω is positive definite.

Figure 2.1 shows that in all cases the profiled formulation converges at least as often as the naive formulation. AEL converges more often than original EL, and the hybrid procedure converges the most often. Even at a sample size of $n = 100$, the profiled problem converges nearly every single time, except in the case of log-normal errors. Figure 2.2 shows that the profiled form can be up to 40 times faster on average than the naive form.

2.4.2 Estimation error

We now explore the estimation errors resulting from different approaches. We compare both original EL and AEL to the Gaussian MLE computed as in Drton et al. (2017), GLS, and WLS. The latter two estimates are computed using the R package `lavaan` (Rosseel, 2012). We also include a hybrid procedure that finds the Gaussian MLE $\hat{B}$ and then uses the resulting residuals and the maximum EL weights at $\hat{B}$ to form an estimate $\check{\Omega} = (I - \hat{B})^T Y D Y^t (I - \hat{B})^t$. Note that the T_4 distribution does not have finite 6th moments, so the limiting distributions from Proposition 2.1 may not hold; however, all estimation procedures still appear to be consistent.

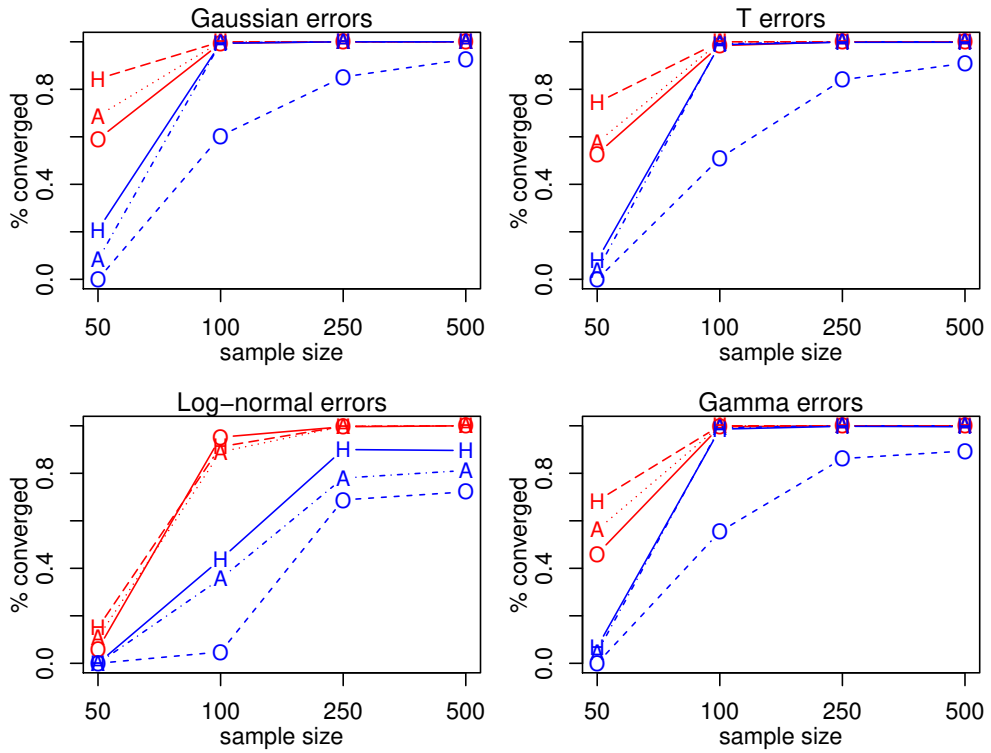


Figure 2.1: Proportion of 500 simulations which converge to a valid stationary point, plotted versus the sample size. O- original EL; A- adjusted EL; H- hybrid EL. Red points indicate the profile formulation; blue points indicate the naive formulation.

Proceeding as in Section 2.4.1, we generate 1000 graphs for each error distribution and sample size. To measure estimation accuracy, we average the relative error $\|\text{vech}(\check{\Sigma}) - \text{vech}(\Sigma)\|^2/\|\text{vech}(\Sigma)\|^2$ for $\Sigma(B,\Omega)$ across each of the simulation runs in which all methods converge; recall Figure 2.1. The results are shown in Figure 2.3.

In general, there is no substantial difference in accuracy between the adjusted and original empirical likelihood methods. For the Gaussian case, MLE and GLS perform better than the methods which do not assume Gaussianity, but the improvement is slight. In the T_4 and log-normal case, the EL procedures perform substantially better than the other methods. Finally, for the gamma case, the hybrid method seems to outperform the other methods, followed closely by the EL methods; however, the differences between the methods are not substantial.

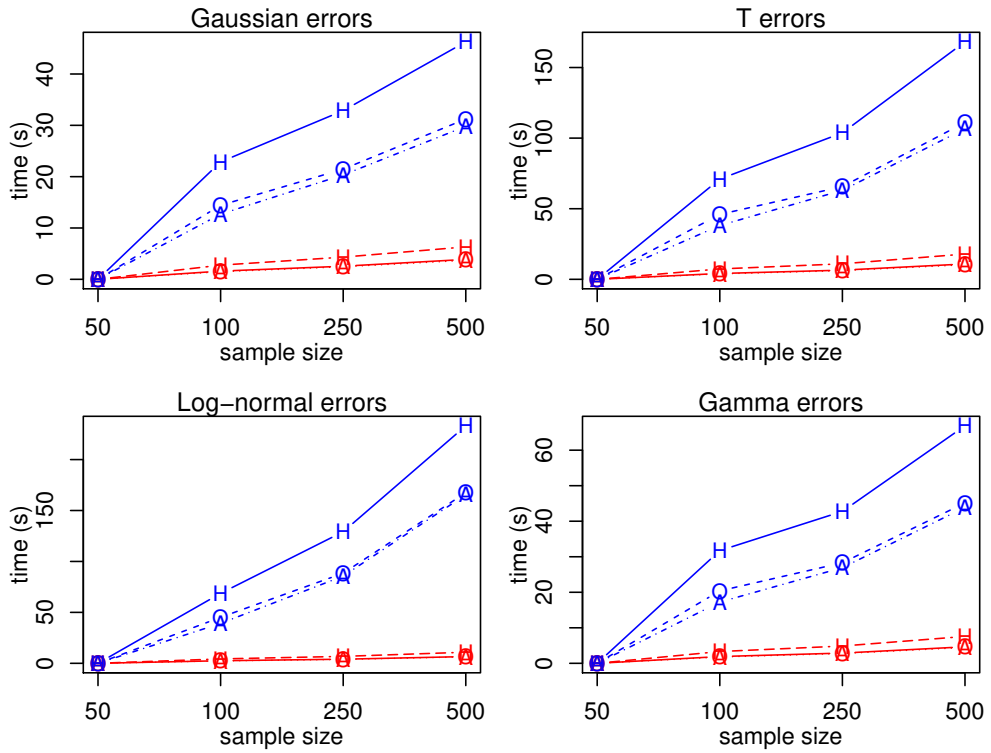


Figure 2.2: The average run time in seconds among the simulations in which all methods converge to a valid stationary point, plotted versus the sample size. O- original EL; A- adjusted EL; H- hybrid EL. Red points indicate the profile formulation; blue points indicate the naive formulation.

In Figure 2.4, all methods converge more than 95% of the time in all distributions, except for the log-normal case. In this case, the WLS procedure still only converges roughly 90% at $n = 1000$.

2.4.3 Confidence regions

We examine the coverage frequencies of joint confidence regions for the parameters β_{uv} and ω_{uv}. We construct Wald regions using the estimates of $\text{Var}(\hat{\theta})$ from the Gaussian MLE, GLS, and WLS. We also calculate a sandwich variance estimator using the Gaussian likelihood as the estimating equations and the asymptotic EL variance via Qin and Lawless (1994).

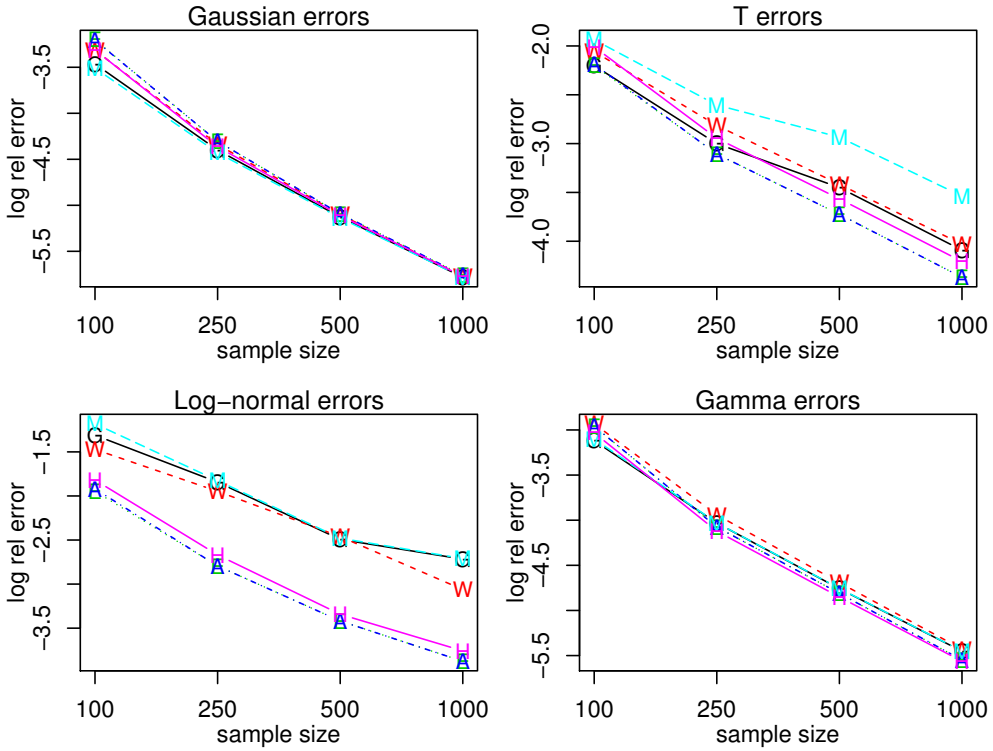


Figure 2.3: Log mean relative squared estimation error in Σ over 1000 simulations, plotted versus the sample size. Average is only taken on simulations in which all methods converged. A- adjusted EL; E- empirical likelihood; G- generalized least squares; H- hybrid Gauss/EL; M- Gaussian MLE; W- weighted least squares.

Alternatively, we calculate the EL at (B_0, Ω_0) using original EL, EEL, AEL. We then compare the resulting EL ratio to its asymptotic χ^2 distribution. If a method does not converge, we count this as a case in which the confidence region does not cover the true parameters.

At each sample size and error distribution, we construct 1000 graphs with 6 nodes, 8 directed edges and 4 bidirected edges from the procedure described in Section 2.4.1. For the T distribution, we increase the degrees of freedom to 7 to ensure Proposition 2.1 applies. The coverage rates for 90% confidence intervals are shown in Figure 2.5. Based on the displayed results, regions obtained from the Gaussian MLE and GLS can only be recommended when the errors are (close to) Gaussian. The EEL method performs the best, staying close to

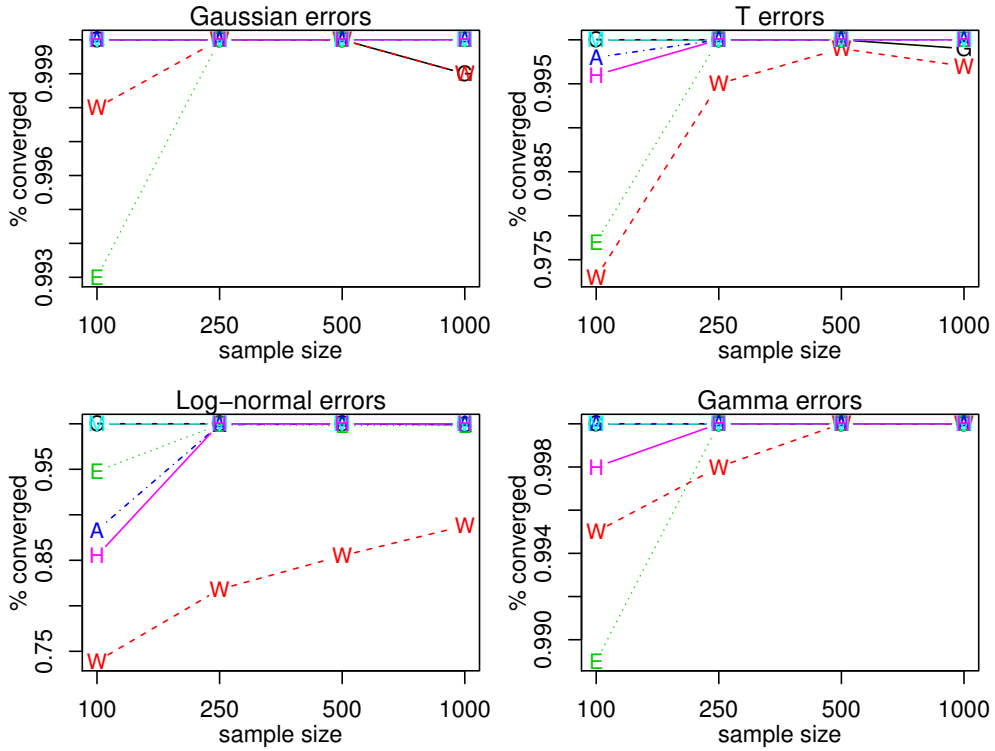


Figure 2.4: Proportion of times (over 1000 simulations) the method converged to a local maximum of the objective function, plotted versus the sample size. A- adjusted EL; E- empirical likelihood; G- generalized least squares; H- hybrid Gauss/EL; M- Gaussian MLE; W- weighted least squares.

the parametric methods in the Gaussian case and doing the best in most non-Gaussian scenarios. The sandwich method is another good choice. However, we also observe that in order to achieve nominal coverage levels very large sample sizes may be required.

2.4.4 Protein signaling network

Sachs et al. (2005, Figure 2) present a signaling network of 11 observed molecules and 13 unobserved molecules. The black edges in Figure 2.6 give a plausible mixed graph representation of that network and was also considered by Drton et al. (2017). A log-transformation of the available protein expression data improves Gaussianity but leaves the distribution

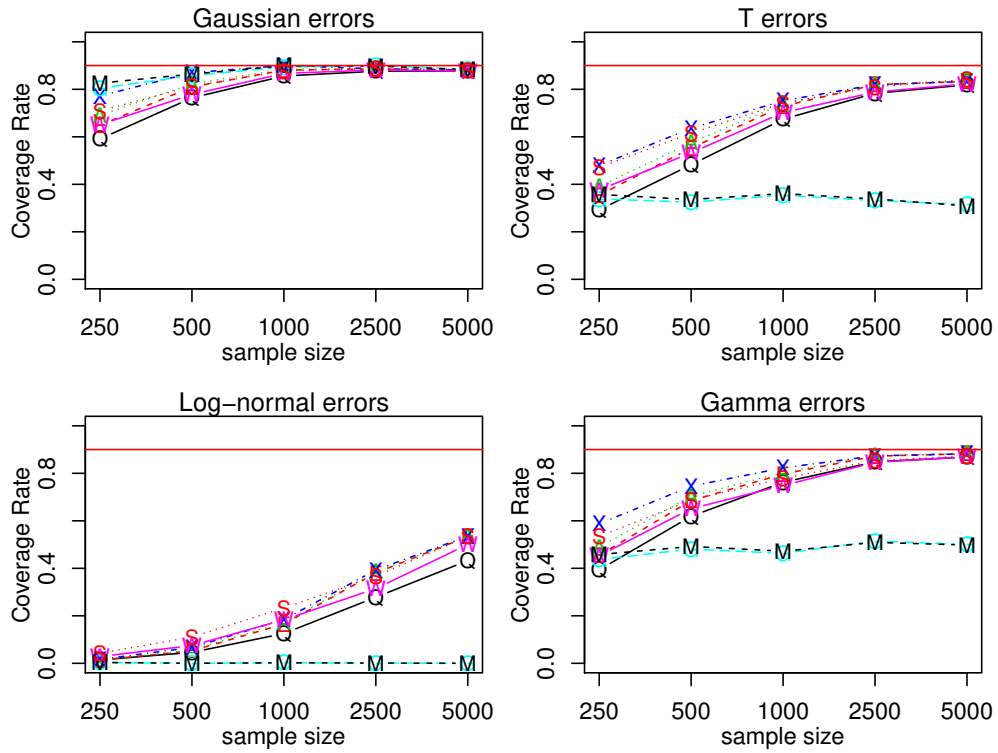


Figure 2.5: Coverage frequencies of joint confidence intervals. A- adjusted EL; X- extended EL; G- generalized least squares; Q- Qin and Lawless asymptotic EL variance; M- Gaussian MLE; S- sandwich estimator; W- weighted least squares; E- EL direct χ^2 calibration.

of some of the variables skewed and/or multimodal. We consider two separate tests; each compares the SEM sub-model corresponding to the graph of black edges against a full model which adds one of the two red edges also shown in Figure 2.6. Note that the added red edge from Mek $\rightarrow$ PKA induces a directed cycle. For the log-transformed data, we perform a Gaussian as well as an empirical likelihood ratio test. For the test involving the directed edge Mek $\rightarrow$ PKA, the Gaussian LR is .416 (p-value = .52) and the ELR is 4.379 (p-value = .04). For the test involving the bidirected edge Akt $\leftrightarrow$ PIP2, the Gaussian LR is 15.216 (p-value $<$ .001) and the ELR is .782 (p-value = .37). While we do not have a certified gold standard network, and the implicit assumption of linearity may not be appropriate for all postulated relationships, these examples present situations in which the Gaussian assumption

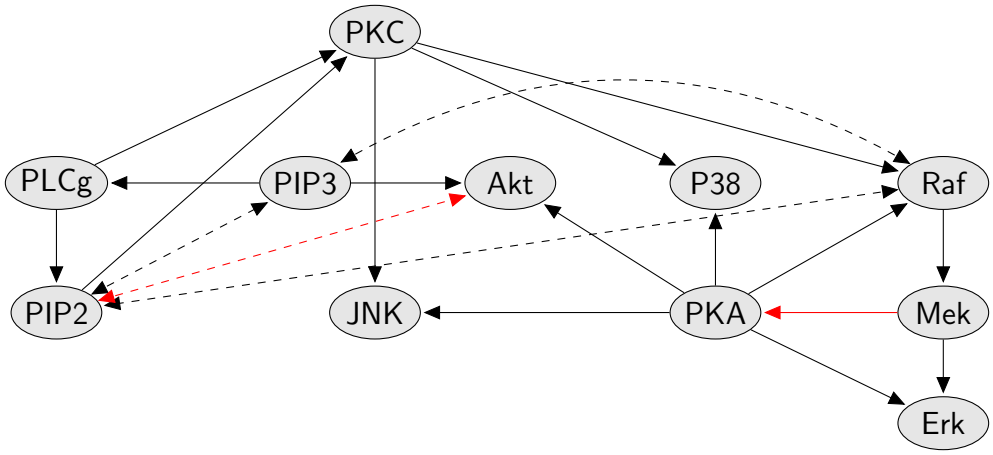


Figure 2.6: Plausible mixed graph for the protein-signaling network dataset. The relevant sub-model can be formed by removing the red bidirected edge between PIP2 and Akt and the red edge from MEK to PKA.

is particularly inappropriate and may cause concerns for a practitioner.

2.5 Discussion

In this chapter, we showed that EL methods are an attractive alternative for estimation and testing of non-Gaussian linear SEMs. Our approach of profiling out the error covariance matrix Ω drastically reduces computational effort and creates a far more tractable and reliable estimation procedure. Furthermore, we showed that the use of AEL may further improve convergence of optimizers, particularly, when the sample size is small and the errors are skewed. EEL was seen to drastically improve the coverage rate of the joint confidence intervals.

Our EL methods are applicable under very few distributional assumptions, all the while allowing statistical inference in close to analogy to parametric modeling. When the data is non-Gaussian, the modified EL methods outperform the other methods we considered in almost all scenarios we explored. This concerns the proportion of times a valid estimate is returned, statistical efficiency, and also confidence region coverage. While there remains significant room for improvement in the design of confidence regions, we conclude that EL

methods are a valuable tool for applications of linear SEMs to non-Gaussian data.

Chapter 3

CAUSAL DISCOVERY OF HIGH-DIMENSIONAL DIRECTED ACYCLIC GRAPHS

3.1 Introduction

Randomized experiments are the gold standard for inferring causal relationships. However, experiments can be expensive and time consuming, unethical, or simply impossible given available technology. Observational studies thus remain an important source of information for many problems, and proposing causal relationships based on analysis of observational data is valuable for hypothesis generation and accelerating scientific discovery. In this chapter, we consider the case where the underlying causal structure is unknown and is actually the target of estimation. In this framework, positing causal structure is equivalent to selecting an appropriate graph, and we propose a method which can recover the exact causal structure from observational data even when the number of variables, p, exceeds the number of samples, n.

We restrict the focus of this chapter to *directed acyclic graphs* (DAGs). When compared to the previous setting, there are two main restrictions. First, we assume $E_{\leftrightarrow} = \emptyset$, so that there is no unobserved confounding. This assumption is sometimes referred to as *causal sufficiency* (Spirtes et al., 2000). Under this assumption, the elements of the error vectors ε_i are mutually independent and jointly follow the product distribution $\otimes_{v \in V} P_v$ where P_v is the univariate distribution of ε_{vi}. We also assume that the graph G corresponding to the true model does not contain any directed cycles. This implies that there exists some ordering of the variables, σ, such that β_{vu} is constrained to be zero unless $\sigma(u) < \sigma(v)$. Note that in this setting, $u \in \text{an}(v)$ implies $v \notin \text{an}(u)$.

3.1.1 Previous work

In general, discovery of causal structure from observational data is difficult because of the super-exponential set of possible models, some of which may be indistinguishable from others. This can be especially complicated in settings where the number of variables, p, is comparable or even exceeds the number of samples, n. Despite the inherent difficulty of inferring the structure of a causal model from observational data, many methods for causal discovery have been developed and have seen fruitful applications; see Drton and Maathuis (2017) for an overview of general structure learning approaches. In particular, the celebrated PC algorithm (Spirtes et al., 2000) is a constraint-based method which first discovers a set of conditional independence relationships and then identifies the associated Markov equivalence class. Such an equivalence class contains all acyclic digraphs that are compatible with given conditional independence relationships. Kalisch and Bühlmann (2007) show if the maximum total degree of the graph is controlled and the data is Gaussian, then in the high-dimensional setting the PC algorithm can consistently recover the Markov equivalence class of the true generating model. Harris and Drton (2013) extend the result to Gaussian copula models using rank correlations.

Aside from constraint based methods, Chickering (2002) proposes Greedy Equivalent Search (GES), which consistently identifies the global maximum under certain conditions on the scoring function. Nandy et al. (2015) show that GES is also consistent in the high dimensional setting if the underlying structure is suitably sparse.

The previously mentioned methods are able to consistently identify a Markov equivalence class. Within a Markov equivalence class, the presence or absence of an edge between any pair of nodes is the same for all graphs; however, in general, the orientation of the edges may differ, resulting in a set of possible graphs which grows exponentially with the number of nodes; see Steinsky (2013) and He et al. (2015) for results on the size and number of Markov equivalence classes. Figure 3.1 shows all 3 node graphs with two edges such that nodes 1 and 2 as well as nodes 2 and 3 are adjacent. In model (a), variables 1 and 3 are

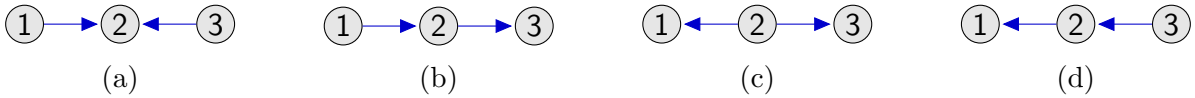


Figure 3.1: The Markov equivalence class of graph (a) is a singleton. However, graph (b), (c) and (d) are Markov equivalent.

marginally independent, while in models (b), (c), and (d), 1 and 3 are only conditionally independent given 2. Hence, model (a) can be distinguished from models (b), (c), and (d) using conditional independence tests; however, models (b), (c), and (d) are mutually indistinguishable. Although Maathuis et al. (2009) provide a procedure for bounding the size of a causal effect over graphs within an equivalence class, scientific interpretation of the set of possibly conflicting graphs can be difficult.

In contrast, it has been shown that under various additional assumptions, the exact graph structure, not simply the equivalence class, can be recovered from observational data. Loh and Bühlmann (2014) and Peters and Bühlmann (2014) show that when the variances of ε are equal (or known up to ratios), the exact graph can be recovered.

Shimizu et al. (2006) show when the data follow a *linear non-Gaussian acyclic model* (LiNGAM), which corresponds exactly to our DAG setting with non-Gaussian errors, the exact graph structure, not just the equivalence class, can be uniquely identified from observational data. In this setting, models (a), (b), (c), and (d) from Figure 3.1 would all be mutually distinguishable. Shimizu et al. (2006) appeal to results on independent component analysis (ICA), a procedure which finds a linear transformation of the data that minimizes the mutual information (Comon, 1994). In subsequent work, Shimizu et al. (2011) propose the DirectLiNGAM method which iteratively selects a causal ordering by computing pairwise statistics. Hyvärinen and Smith (2013) extend this work by proposing new pairwise test statistics. However, all of the proposed methods are inconsistent in high-dimensional settings that allow the number of variables to scale as fast or faster than the sample size.

In contrast to the ICA proof of Shimizu et al. (2006), we consider a more algebraic

approach and show that specific polynomials of the moments of Y vanish for some of the graphs, but in general, not for others. For example, suppose the errors of SEM ε have mean 0, are mutually independent, and $\mathbb{E}(\varepsilon_v^3) \neq 0$ for $v = 1, 2, 3$. Then model (b) implies

$$\frac{\mathbb{E}(Y_1^2 Y_2)}{\mathbb{E}(Y_1^3)} = \frac{\mathbb{E}\left(\varepsilon_1^2(\beta_{21}\varepsilon_1 + \varepsilon_2)\right)}{\mathbb{E}\left(\varepsilon_1^3\right)} = \beta_{21}\frac{\mathbb{E}(\varepsilon_1^3)}{\mathbb{E}(\varepsilon_1^3)} = \frac{\mathbb{E}(Y_1 Y_2)}{\mathbb{E}(Y_1^2)},$$

so that

$$\mathbb{E}(Y_1^2 Y_2)\mathbb{E}(Y_1^2) - \mathbb{E}(Y_1^3)\mathbb{E}(Y_1 Y_2) = 0.$$

However, models (c) implies

$$\begin{aligned}
\frac{\mathbb{E}(Y_1^2 Y_2)}{\mathbb{E}(Y_1^3)} &= \frac{\mathbb{E}\left((\varepsilon_1 + \beta_{12}\varepsilon_2)^2\varepsilon_2)\right)}{\mathbb{E}\left((\varepsilon_1 + \beta_{12}\varepsilon_2)^3\right)} = \frac{\beta_{12}^2\mathbb{E}\left(\varepsilon_2^3\right)}{\mathbb{E}\left(\varepsilon_1^3 + (\beta_{12}\varepsilon_2)^3\right)} \\
\frac{\mathbb{E}(Y_1 Y_2)}{\mathbb{E}(Y_1^2)} &= \frac{\mathbb{E}\left((\varepsilon_1 + \beta_{12}\varepsilon_2)\varepsilon_2)\right)}{\mathbb{E}\left((\varepsilon_1 + \beta_{12}\varepsilon_2)^2\right)} = \frac{\beta_{12}\mathbb{E}\left(\varepsilon_2^2\right)}{\mathbb{E}\left(\varepsilon_1^2 + (\beta_{12}\varepsilon_2)^2\right)},
\end{aligned}$$

so that in general

$$\mathbb{E}(Y_1^2 Y_2)\mathbb{E}(Y_1^2) - \mathbb{E}(Y_1^3)\mathbb{E}(Y_1 Y_2) \neq 0.$$

The analysis of similar polynomials of other moments could be used to mutually distinguish all other graphs.

3.1.2 Contribution

In this chapter, we consider the LiNGAM case, as in Shimizu et al. (2006). We propose a modified DirectLiNGAM algorithm and show under certain conditions and choice of tuning parameter, that it consistently recovers a single graph corresponding to the generating mechanism, even in the high-dimensional setting and under slight model misspecification. For the method, we propose a new test statistic which encodes causal direction and is suitable for

the high-dimensional case. Most notably, our theoretical analysis considers restricting the maximum in-degree of the graph and assumes that the data follow a log-concave distribution. Our theoretical guarantees also apply to hub graphs where the maximum out-degree may grow with the size of the graph. This corresponds to many observed biological networks (Hao et al., 2012) which do not satisfy the conditions needed for high-dimensional consistency of the PC algorithm (Kalisch and Bühlmann, 2007).

3.2 Causal discovery setup

Before we introduce the discovery algorithm, we make a few useful definitions, and, in particular, define our test statistic.

3.2.1 Parental faithfulness

For $v_j \in V$, let $l = (v_1, v_2, \ldots, v_z)$ be a directed path in G, so $(v_j, v_{j+1}) \in E$ for all j. Given coefficients $(\beta_{vu})_{(u,v)\in E}$, the path weight of l is

$$W(l) = \prod_{j=1}^{z-1} \beta_{v_{j+1},v_j}.$$

Let $\mathcal{L}_{v,u}$ be the set of all directed paths from u to v. The total effect of u on v is

$$\pi_{vu} = \sum_{l\in\mathcal{L}_{v,u}} W(l). \tag{3.1}$$

The total direct effect also characterizes the conditional mean where

$$\pi_{vu} = \mathbb{E}(Y_v|Y_u = y+1) - \mathbb{E}(Y_v|Y_u = y)$$

and $\pi_{vu} = 0$ unless $u \in \mathrm{An}(v)$. By convention, we let $\pi_{vv} = 1$. We can also conveniently calculate π_{vu} without enumerating all paths by inverting $I - B$. Indeed, the matrix of total effects $\Pi = (\pi_{vu})_{u,v\in V}$ is $\Pi = (I - B)^{-1}$.

An important approach to causal discovery begins by inferring relations such as conditional independence and then determines graphs compatible with empirically found relations. For this approach to succeed it is important that the considered relations correspond to structure in the graph G as opposed to special choice of parameters such as the coefficients β_{vu}. In the context of conditional independence, the assumption that any relation present in an underlying joint distribution $P \in \mathcal{P}(G)$ corresponds to the absence of certain paths in G is known as the *faithfulness* assumption; see Uhler et al. (2013) for a detailed discussion of this assumption. For the purpose of our work, we define a weaker condition which we refer to as *parental faithfulness*. In particular, if $u \in \mathrm{pa}(v)$, we require that the total effect of u on v does not vanish when we modify the considered distribution by regressing v onto any set of its non-descendants.

Formally, let $\Sigma = \mathbb{E}_P(Y_i Y_i^t)$ be the covariance matrix of the, for convenience, centered random vector $Y_i \sim P$. Let Σ_{CC} be the principal sub-matrix for a choice of indices $C \subseteq V$, and for given $v \in V \setminus C$, let Σ_{Cv} be the sub-vector comprised of the entries in places (c, v) for $c \in C$. Let

$$\beta_{vC} \;:=\; (\beta_{vc.C})_{c \in C} = (\Sigma_{CC})^{-1}\Sigma_{Cv} \tag{3.2}$$

be the population regression coefficients when v is regressed onto C. The quantity $\beta_{vc.C}$ is defined even if $(c, v) \notin E$, and in general $\beta_{vc.C} \neq \beta_{vc}$ even if $(c, v) \in E$. A pair $(u, v) \in E$ is *parentally faithful* if for any set $C \subseteq V \setminus \{\mathrm{de}(v), v, u\}$, the residual total effect

$$\pi_{vu.C} \;:=\; \pi_{vu} - \sum_{c \in C} \beta_{vc.C}\pi_{cu} \neq 0. \tag{3.3}$$

If this holds for every pair $(u, v) \in E$, we say that the joint distribution P is *parentally faithful* with respect to G. The condition also trivially implies that $\beta_{vu} \neq 0$ if $(u, v) \in E$. Clearly, parental faithfulness only pertains to the linear coefficients and errors variances, and the choices for which parental faithfulness fails form a set of (Lebesgue) measure zero. The concept is exemplified in Figure 3.2.

(a) (b)

Figure 3.2: In (a), the choice $\beta_{31} = \beta_{21} = 1$ and $\beta_{32} = -1$ results in parental unfaithfulness because $\pi_{31.\emptyset} = 0$. In (b), the choice $\beta_{31} = \beta_{32} = \beta_{42} = 1$, $\beta_{41} = 2$, and $\mathbb{E}(\varepsilon_1^2) = \mathbb{E}(\varepsilon_2^2) = \mathbb{E}(\varepsilon_3^2) = 1$ results in parental unfaithfulness because $\pi_{42.3} = 0$.

3.2.2 Test statistic

In general, reliable determination of the causal direction between u and v requires all confounding to be removed. The DirectLiNGAM method of Shimizu et al. (2011) achieves this by adjusting v and u for all x such that $\sigma(x) < \sigma(v)$ and $\sigma(x) < \sigma(u)$. However, this results in adjusting for a growing set of variables and propagates error proportional to the number of variables, rendering high-dimensional estimation inconsistent, or impossible when the size of the adjustment set exceeds the sample size. On the other hand, if error propagation is limited by restricting the size of the adjustment sets, this may result in cases where confounding is not completely removed. The method we present below provides a solution to this problem via the use of a test statistic that can simultaneously certify causal direction and a lack of confounding.

Shimizu et al. (2011) calculate the kernel based mutual information between v and the residuals of u when it is regressed onto v. Under the suitable version of faithfulness, the corresponding population information is positive if and only if $v \in \mathrm{de}(u)$ or there is uncontrolled confounding between v and u, that is, u and v have a common ancestor even when certain edges are removed from the graph. Hence, the mutual information can be used to test the hypothesis that $v \notin \mathrm{de}(u)$ versus the hypothesis that $v \in \mathrm{de}(u)$ or there is confounding between u and v. Unfortunately, calculating the mutual information can be quite computationally burdensome, so Hyvärinen and Smith (2013) propose a different parameter

R_{vu}. Without confounding, R_{vu} is positive if v is an ancestor of u and negative if u is an ancestor of v. With confounding, however, the parameter can take either sign, so it cannot be reliably used if we remain uncertain about whether or not confounding occurs.

We propose a parameter and corresponding statistic which possesses the same properties as the mutual information used by Shimizu et al. (2011). However, it is computationally inexpensive and can be used tractably, even in very large graphs, similar to the statistic of Hyvärinen and Smith (2013). Our statistic is also a rational function of the sample moments of Y which facilitates analysis of error propagation.

Using the population regression coefficients $\beta_{vc.C}$ from (3.2), define for any i the residual

$$Y_{vi.C} = Y_{vi} - \sum_{c \in C} \beta_{vc.C} Y_{ci}$$

from the regression of v onto the set C.

Theorem 3.1. *Let $P \in \mathcal{P}(G)$ be a distribution in the model given by an acyclic digraph G, and let $Y_i \sim P$. For $K > 2$, two distinct nodes u and v, and any set $C \subseteq V \setminus \{u, v\}$, define*

$$\tau^{(K)}_{v.C \to u} := \mathbb{E}_P(Y^{K-1}_{vi.C} Y_{ui}) \mathbb{E}_P(Y^2_{vi.C}) - \mathbb{E}_P(Y^K_{vi.C}) \mathbb{E}_P(Y_{vi.C} Y_{ui}). \tag{3.4}$$

(i) *If $u \notin \mathrm{pa}(v)$ and $\mathrm{pa}(v) \subseteq C \subseteq V \setminus \{\mathrm{de}(v), v, u\}$, then*

$$\tau^{(K)}_{v.C \to u} = 0.$$

(ii) *Suppose P is parentally faithful with respect to G. If $u \in \mathrm{pa}(v)$ and $C \subseteq V \setminus \{\mathrm{de}(v), v, u\}$, then for generic error moments up to order $K > 2$, we have*

$$\tau^{(K)}_{v.C \to u} \neq 0.$$

Estimators $\hat{\tau}^{(K)}_{v.C \to u}$ of the parameter from (3.4) are naturally obtained from empirical regression coefficients and empirical moments.

In Theorem 3.1(ii), the term *generic* indicates that the set of error moments for which this statement does not hold has measure zero. Given that there is a finite number of sets $C \subset V$, the union of all exceptional sets is also a null set. A detailed proof of Theorem 3.1 is included in the Appendix. Claim (i) can be shown via direct calculation, and we give a brief sketch of (ii) here. For fixed coefficients $(\beta_{vu})_{(u,v)\in E}$ and set $C \subset V$, $\tau^{(K)}_{v.C\to u}$ is a rational function of the error moments. Thus existence of a single choice of error moments for which $\tau^{(K)}_{v.C\to u} \neq 0$ is sufficient to show that the statement holds for generic error moments. As the argument boils down to showing that a certain polynomial is not the zero polynomial (Okamoto, 1973), the choice considered need not necessarily be realizable by a particular distribution. In particular, we choose all moments of order less than K equal to those of the centered Gaussian distribution with variance $\sigma_v^2 = \mathbb{E}(\varepsilon_v^2)$, but for the Kth moment we add an offset $\eta_v > 0$, so

$$\mathbb{E}(\varepsilon_v^K) = \begin{cases} \eta_v & \text{if } K \text{ is odd,} \\ (K-1)!!\sigma_v^K + \eta_v & \text{if } K \text{ is even,} \end{cases} \tag{3.5}$$

where $(K-1)!!$ is a double factorial. If there is no confounding between $Y_{v.C}$ and Y_u, that is, no ancestor of u is the source of a directed path to v that avoids $C \cup \{u\}$, then

$$\tau^{(K)}_{v.C\to u} \;=\; \pi_{vu.C}\left(\pi_{vu.C}^{K-2}\,\eta_u\sigma_v^2 - \eta_v\sigma_u^2\right) \tag{3.6}$$

with $\pi_{vu.C}$ being the residual total effect from (3.3). By the assumed parental faithfulness, $\pi_{vu.C} \neq 0$, and a choice of $\eta_u \neq \frac{\eta_v\sigma_u^2}{\pi_{vu.C}^{K-2}\sigma_v^2}$ thus implies $\tau^{(K)}_{v.C\to u} \neq 0$. A more involved but similar argument can be made in the case of confounding.

Corollary 3.1. *Let P_v and P_u be two distributions that each have all moments up to order K equal to those of some Gaussian distribution. Then there exists a graph G, for which $u \in \mathrm{pa}(v)$, and distributions P which are parentally faithful with respect to G, but*

$$\tau^{(K)}_{v.C\to u} = 0$$

for some set $C \subseteq V \setminus \{\mathrm{de}(v), v, u\}$.

Proof. The moments of P_v and P_u satisfy (3.5) with $\eta_v = \eta_u = 0$. Consequently, if there exists a set C such that there is no confounding between $Y_{v.C}$ and Y_u, then $\tau_{v.C \to u}$ satisfies (3.6), the right-hand side of which is zero when $\eta_v = \eta_u = 0$. □

Corollary 3.1 confirms that indeed the null set to be avoided in Theorem 3.1(ii) contains any point for which all error moments are consistent with some Gaussian distribution. Thus, our identification of causal direction requires that the error moments of order at most K be inconsistent with all Gaussian distributions. In practice, we consider the case $K = 3, 4$. We refer readers to Hoyer et al. (2008a) for a full characterization of when graphs with both Gaussian and non-Gaussian errors are identifiable.

We now define $\mathcal{P}_{F_K}(G)$ to be the subset of $\mathcal{P}(G)$ comprised of those joint distributions P for which $\tau^{(K)}_{v.C \to u} \neq 0$ whenever $u \in \mathrm{pa}(v)$ and $C \subseteq V \setminus \{\{u, v\} \cup \mathrm{de}(v)\}$. For fixed graph G, the set of linear coefficients and error moments up to order K that induce an element of $\mathcal{P}(G) \setminus \mathcal{P}_{F_K}(G)$ has measure zero. This set difference includes distributions which are not parentally faithful with respect to G and distributions for which there exist a parent/child pair for which both error distributions have Gaussian moments up to order K.

We now state a corollary about a parameter which can be used to test whether some node v has a parent in some set $V_2 \subset V$.

Corollary 3.2. *Let* $P \in \mathcal{P}(G)$*, let* $v \in V$*, and consider two disjoint sets* $V_1, V_2 \subseteq V \setminus \{v\}$*. For a chosen non-negative integer* J*, define*

$$T^{(K)}_1(v, V_1, V_2) := \min_{C \in V_1(J)} \max_{u \in V_2} |\tau^{(K)}_{v.C \to u}|, \tag{3.7}$$

$$T^{(K)}_2(v, V_1, V_2) := \max_{u \in V_2} \min_{C \in V_1(J)} |\tau^{(K)}_{v.C \to u}|, \tag{3.8}$$

where $V_1(J) = \{C \subseteq V_1 : |C| = J\}$ *if* $J \leq |V_1|$ *and* $V_1(J) = V_1$ *if* $J \geq |V_1|$.

(i) *If* $|\mathrm{pa}(v)| \leq J$ *and* $\mathrm{pa}(v) \subseteq V_1 \subseteq V \setminus \mathrm{de}(v)$*, then* $T^{(K)}_1(v, V_1, V_2) = T^{(K)}_2(v, V_1, V_2) = 0$.

(ii) *Suppose* $\beta_{vu} \neq 0$ *for all* $u \in \text{pa}(v)$. *If* $V_2 = \{\text{de}(V_2) \cup V_2\}$ *and* $\text{pa}(v) \cap V_2 \neq \emptyset$, *then for generic error moments of order up to* K, *we have* $T_1^{(K)}(v, V_1, V_2) > 0$ *and* $T_2^{(K)}(v, V_1, V_2) > 0$.

Proof. Statement (i) follows immediately from Theorem 3.1. Now consider Statement (ii). Since, $\text{pa}(v) \cap V_2 \neq \emptyset$, but $V_2 = \{\text{de}(V_2) \cup V_2\}$, there must exist some $u \in \text{pa}(v) \cap V_2$ such that $\text{de}(u) \cap \text{pa}(v) = \emptyset$. For that u and any $C \subseteq V_1$, the residual total effect is

$$\pi_{vu.C} = \beta_{vu} - \sum_{c \in C} \beta_{vc.C}\pi_{cu} = \beta_{vu} \tag{3.9}$$

because the assumed fact $\text{de}(V_2) \cap V_1 = \emptyset$ implies that $\pi_{cu} = 0$ for all $c \in C$. In addition, $\text{de}(v) \cap \text{pa}(v) = \emptyset$ implies that $\pi_{vu} = \beta_{vu}$ and we have assumed $\pi_{vu.C} = \beta_{vu} \neq 0$. Now, $\pi_{vu.C} = \beta_{vu} \neq 0$ by assumption. Thus, as shown in Theorem 3.1, generic error moments will ensure that $|\tau^{(K)}_{v.C \to u}| > 0$, which in turn implies that $T_1^{(K)}(v, V_1, V_2) > 0$ and $T_2^{(K)}(v, V_1, V_2) > 0$. □

Note that Theorem 3.1(ii) requires parental faithfulness since we consider arbitrary $u \in \text{pa}(v)$, whereas Corollary 3.2(ii) only requires that $\beta_{uv} \neq 0$ since we take the max over a set which contains all of its own descendants. Statistics $\hat{T}_j^{(K)}(v, V_1, V_2)$ estimating $T_j^{(K)}(v, V_1, V_2)$ for $j = 1, 2$ can be calculated from the sample moment based estimates $\hat{\tau}_{v.C \to u}$. In the remainder of the chapter, when making statements which apply to both parameters or corresponding statistics, we drop the subscript 1 or 2 and simply write T or $\hat{T}$. In addition, as we will always fix K, in later sections we also lighten notation by omitting the superscript, so writing $T(v, V_1, V_2)$, $\tau_{v.C \to u}$ and $\hat{\tau}_{v.C \to u}$.

3.3 Graph estimation algorithm

We now present a modified DirectLiNGAM algorithm which estimates the underlying causal structure. At each step, we identify a root, a node without a parent; then recur on the sub-graph induced by removing the identified root. After step z, we will then have a z-tuple,

$\Theta^{(z)}$, which is the topological ordering of the roots identified so far, and $\Psi^{(z)} = V \setminus \Theta^{(z)}$, which are the remaining unordered nodes. We say that $\Theta^{(z)}$ is consistent with a valid ordering of G if for every $s, t \leq z$, $s < t$ only if $\Theta_t^{(z)}$ is not an ancestor of $\Theta_s^{(z)}$ and $\Theta^{(z)} \cap \mathrm{de}(\Psi^{(z)}) = \emptyset$. For each step, we select the next node in a causal ordering by selecting the node with the smallest test statistic $\hat{T}(v, \mathcal{C}_v^{(z-1)}, \Psi^{(z-1)})$ for some $\mathcal{C}_v^{(z-1)} \subseteq \Theta^{(z-1)}$.

Algorithm 1 Estimate Graph

$\Theta^{(0)} = \emptyset$; $\Psi^{(0)} = [p]$;
for $z = 1, \ldots, p$ **do**
  for $v \in \Psi^{(z-1)}$ **do**
    Select the set of possible parents $\mathcal{C}_v^{(z)} \subseteq \Theta^{(z-1)}$
    Compute $\hat{T}(v, \mathcal{C}_v^{(z)}, \Psi^{(z-1)})$
  end for
  $r = \arg\min_{v \in \Psi^{(z-1)}} \hat{T}(v, \mathcal{C}_v^{(z)}, \Psi^{(z-1)})$
  $\Theta^{(z)} = \left(\Theta^{(z-1)}, \{r\}\right)$
  $\Psi^{(z)} = \Psi^{(z-1)} \setminus \{r\}$
end for
Prune ancestors to form parents $\mathcal{C}_v^\star$ for all $v \in V$
return $\Theta^{(p)}$ as the topological ordering; $\{\mathcal{C}_v^\star\}_{v \in V}$ as the set of parents

In contrast to the DirectLiNGAM method, the proposed algorithm does not adjust for all non-descendants, but only for subsets of limited size. This is required for meaningful regression residuals when the number of variables exceeds the sample size and also limits error propagation from the estimated linear coefficients; however, it results in higher computational burden. If we naively allow $\mathcal{C}_v^{(z)} = \Theta^{(z-1)}$, the number of subsets $C \subset \mathcal{C}^{(z)}$ such that $|C| = J$ grow at $\mathcal{O}(z^J)$, presenting an enormous computational effort even for moderate values of p and J. Thus, we reduce computation by pruning nodes which are not parents of v by letting

$$\mathcal{C}_v^{(z)} = \left\{ p \in \Theta^{(z-1)} : \min_{\substack{C \subseteq \Theta^{(z-1)} \setminus \{p\} \\ |C| \leq J}} \hat{\tau}_{v.C \to p} > g^{(z)} \right\} \tag{3.10}$$

for some cut-off $g^{(z)}$. In practice, specifying an explicit value for $g^{(z)}$ a priori can be difficult,

so we use a data-driven rising cut-off $g^{(z)} = \max(g^{(z-1)}, \alpha\hat{T}(r, \mathcal{C}_r^{(z)}, \Psi^{(z-1)}))$ where r is the root selected at step z and α is a non-negative tuning parameter. A larger value of α prunes more aggressively, decreasing the computational effort. However, setting α too large could result in incorrect estimates if some parent of v is incorrectly pruned from $\mathcal{C}_v^{(z)}$. Section 3.3.1 further discusses selecting an appropriate α.

For fixed $\mathcal{C}_v^{(z)}$, $\Psi^{(z-1)}$, and v, the signs of T_1 (min-max) and T_2 (max-in) will always agree, so both could be used to certify whether $\text{pa}(v) \cap \Psi^{(z-1)} = \emptyset$. However, when the parameters are positive, $T_1 \geq T_2$, so the min-max statistic may be more robust to sampling error when testing if the parameter is non-zero. However, it comes at a greater computational cost. At each step, $\Psi^{(z)}$ decreases by a single node and $\mathcal{C}_v^{(z)}$ may grow by a single node. If the $|\Psi|^2$ values of $\min_{C \in \mathcal{C}_v^{(z-1)}} \hat{\tau}_{v.C \to u}$ values have been stored, updating $\hat{T}_2$, the max-min, only requires testing the $\binom{\mathcal{C}^{(Z-1)}}{J-1}$ subsets of $\mathcal{C}_v^{(z)}$ which include $\mathcal{C}_v^{(z-1)} \setminus \mathcal{C}_v^{(z)}$, the variable selected at the previous step. Updating the min-max statistic without redundant computation would require storing the $\mathcal{O}\left(|\Psi|^2 z^J\right)$ values of $\tau_{v.C \to u}$. In practice, we completely recompute $\hat{T}_1$ at each step. Section 3.4 demonstrates this trade-off between computational burden and robustness on simulated data.

3.3.1 Deterministic statement

In Theorem 3.2 we make a deterministic statement about sufficient conditions for which Algorithm 1 will output a specific graph G when given data $Y \in \mathbb{R}^{n \times p}$. We assume each row of Y, $Y_i \sim P_Y$, but we allow model misspecification so that P_Y may not be in $\mathcal{P}(G)$ for any acyclic digraph G. However, we require that the sample moments of Y are close enough to the population moments for some distribution $P \in \mathcal{P}(G)$. For notational convenience, for $H \subseteq V$ and $\alpha \in \mathbb{R}^{|H|}$, let $\hat{m}_{H,\alpha} = \frac{1}{n}\sum_i^n \left(\prod_{v \in H} Y_{vi}^{\alpha_v}\right)$ denote a sample moment estimated from data Y, and let $m_{H,\alpha} = \mathbb{E}_P\left(\prod_{v \in H} Z_v^{\alpha_v}\right)$ denote a population moment for $Z \sim P$.

Condition (C1)**.** *For some p-variate distribution P, there exists a DAG G with $|\text{pa}(v)| \leq J$ for all $v \in V$ such that:*

(a) *For all* $v, u \in V$ *and* $C \subseteq V \setminus \{u, v\}$ *with* $|C| \leq J$ *and* $C \cap \mathrm{de}(v) = \emptyset$*; if* $u \in \mathrm{pa}(v)$ *then the population quantities for* P

$$\left|\tau^{(K)}_{v.C \to u}\right| > \gamma,$$

for some constant $\gamma > 0$.

(b) *For all* $v, u \in V$ *and* $C \subseteq V \setminus \{v, u\}$ *with* $|C| \leq J$ *and* $\mathrm{pa}(v) \subseteq C \subseteq V \setminus \mathrm{de}(v)$*, if* $u \notin \mathrm{pa}(v)$*, then the population quantities for* P *satisfy*

$$\tau^{(K)}_{v.C \to u} = 0.$$

Condition (C2)**.** *The population covariance of* P *has minimum eigenvalue* $\lambda_{\min} > 0$.

Condition (C3)**.** *All population moments of* P *up to degree* K*,* $m_{V,\alpha}$ *for* $|\alpha| \leq K$*, are bounded by a constant* $\infty > M > \max(1, \lambda_{\min}/J)$ *for positive integer* J.

Condition (C4)**.** *All sample moments of* Y *up to degree* K*,* $\hat{m}_{V,\alpha}$ *for* $|\alpha| \leq K$*, are within* δ_1 *of the corresponding population values of* P *with* $\delta_1 < \lambda_{\min}/(2J)$.

The constraint in Condition (C3) that $M > \max(1, \lambda_{\min}/J)$ is only used to facilitate simplification of the error bounds, and not otherwise necessary. Condition (C1) is a faithfulness type assumption on P, and in Theorem 3.2 we make a further assumption on γ which ensures strong faithfulness. However, it is not strictly stronger or weaker than the typical strong faithfulness type assumption used in the Gaussian case. In particular the condition requires the linear coefficients and error moments to be jointly "sufficiently parentally faithful" and "sufficiently non-Gaussian," so even if the linear coefficients would not have otherwise satisfied strong faithfulness, sufficiently non-Gaussian moments could still ensure the strong faithfulness condition holds. Also, since we first make a deterministic statement, we do not actually make any assumption about the sample moments of Y being close to the true moments of $Y \sim P_Y$, but instead require that the sample moments are close to the true moments of some P, which may not be equivalent to P_Y.

Theorem 3.2. *For some p-variate distribution* P *and data* $Y = (Y_1, \ldots, Y_n)$*:*

(i) *Suppose (C1) holds. Then there exists a DAG G such that $P \in \mathcal{P}_{F_K}(G)$ and G is unique such that $P \notin \mathcal{P}_{F_K}(\tilde{G})$ for any other DAG $\tilde{G}$ with maximum in-degree $\leq J$.*

(ii) *Furthermore, suppose (C1)-(C4) hold for constants which satisfy*

$$\begin{aligned} \gamma/2 > \delta_3 := 4M\delta_1 \left[16(3^K)(J+K)^K K \frac{J^{(K+4)/2} M^{K+1}}{\lambda_{\min}^{K+1}} \right] \\ + 2\left(\delta_1 \left[16(3^K)(J+K)^K K \frac{J^{(K+4)/2} M^{K+1}}{\lambda_{\min}^{K+1}} \right] \right)^2. \end{aligned} \tag{3.11}$$

Then with pruning parameter $g = \gamma/2$, Algorithm 1 will output $\hat{G} = G$.

The main result of Theorem 3.2 is part (ii). The identifiability of a DAG from infinite data was previously shown by Shimizu et al. (2006) by appealing to results for independent component analysis; however, our direct analysis of rational functions of Y allows for an explicit tolerance for how sample moments of Y may deviate from the corresponding population moments of P. This implicitly allows for model misspecification, which is addressed more explicitly in Corollary 3.3. The proof requires Lemmas 3.1, 3.2 and 3.3 which we develop before presenting the proof of Theorem 3.2. Recall that β_{vC} are the population level coefficients when v is regressed onto C and let $\hat{\beta}_{vC}$ denote the coefficients estimated from Y.

Lemma 3.1. *Suppose that (C2), (C3), and (C4) hold. Then for any $v \in V$ and $C \subseteq V$ and $|C| \leq J$,*

$$\|\hat{\beta}_{vC} - \beta_{vC}\|_\infty < \delta_2 = 4\frac{J^{3/2} M \delta_1}{\lambda_{\min}^2}.$$

The proof of Lemma 3.1 uses well known results for matrix inversion and is given in the appendix. Recall, that $Y_{vi.C} = Y_{vi} - \sum_{c \in C} \beta_{vc.C} Y_{ci}$. Let $Z_{v.C}$ denote the analogous quantity for $Z \sim P$, and let $\hat{Y}_{vi.C} = Y_{vi} - \sum_{c \in C} \hat{\beta}_{vc.C} Y_{ci}$.

Lemma 3.2. *Suppose that (C2), (C3), and (C4) hold. Let s, r be non-negative integers such that $s + r \leq K$, and let $Z \sim P$. For any $v, u \in V$ and $C \subseteq V \setminus \{u, v\}$ such that $|C| \leq J$,*

$$\left| \frac{1}{n} \sum_i \hat{Y}_{vi.C}^s Y_{ui}^r - \mathbb{E}\left(Z_{v.C}^s Z_u^r \right) \right| < \delta_1 \Phi(J, K, M, \lambda_{\min})$$

where

$$\Phi(J,K,M,\lambda_{\min}) = \left[16(3^K)(J+K)^K K \frac{J^{(K+4)/2}M^{K+1}}{\lambda_{\min}^{K+1}}\right]. \tag{3.12}$$

Proof. For $a \in \mathbb{Z}_{\geq 0}^{|C|+1}$, let $\binom{s}{a} = \frac{s}{a_1!a_2!\ldots a_{|C|+1}!}$ be the multinomial coefficient. Define the map f as

$$\begin{aligned} f\left(\beta_{vC}, \{m_{V,\alpha}\}_{|\alpha|=s+r}\right) &= \mathbb{E}_P\left(Z_{v.C}^s Z_u^r\right) = \mathbb{E}_P\left(\left(Z_v - \sum_{c\in C}\beta_{vc.C}Z_c\right)^s Z_u^r\right) \\ &= \mathbb{E}_P\left(Z_u^r \sum_{|a|=s}\binom{s}{a}\prod_{c\in C}(-\beta_{vc.C}Z_c)^{a_c} Z_v^{a_v}\right) \\ &= \sum_{|a|=s}\left[\binom{s}{a} m_{(C,v,u),(a,r)}\prod_{c\in C}\left(-(\beta_{vc.C})^{a_c}\right)\right]. \end{aligned} \tag{3.13}$$

Since a is of length $|C|+1$, there are $\binom{|C|+1+s-1}{|C|+1-1}$ moments we consider. By Condition (C3) and (C4), each of the sample moments of Y is restricted to $(-M-\delta_1, M+\delta_1)$, and as shown in (A.17) each of the $\hat{\beta}_{vz.C}$ is restricted to $\left(-\frac{\sqrt{J}M}{\lambda_{\min}} - \delta_2, \frac{\sqrt{J}M}{\lambda_{\min}} + \delta_2\right)$. In this domain, the partial derivatives of f are bounded with

$$\left|\frac{\partial f}{m_{V,\alpha}}\right| \leq s!\left(\frac{\sqrt{J}M}{\lambda_{\min}} + \delta_2\right)^s, \text{ and} \tag{3.14}$$

$$\begin{aligned} \left|\frac{\partial f}{\partial \beta_{vz.C}}\right| &\leq \sum_{\substack{|a|=s\\ a_z>0}}\binom{s}{a}(a_z)\left|m_{(C,v,u),(a,r)}(-\beta_{vz.C})^{a_z-1}\prod_{c\in C\setminus z}(-\beta_{vc.C})^{a_c}\right| \\ &\leq \sum_{\substack{|a|=s\\ a_z>0}}\binom{s}{a} s\left[(M+\delta_1)\left(\frac{\sqrt{J}M}{\lambda_{\min}} + \delta_2\right)^{s-1}\right] \\ &\leq (|C|+1)^s s\left[(M+\delta_1)\left(\frac{\sqrt{J}M}{\lambda_{\min}} + \delta_2\right)^{s-1}\right]. \end{aligned}$$

Then by the mean value theorem for some $\left(\tilde{\beta}_{vC}, \{\tilde{m}_{V,\alpha}\}_{|\alpha|=s+r}\right)$, a convex combination of

$\left(\hat{\beta}_{vC}, \{\hat{m}_{V,\alpha}\}_{|\alpha|=s+r}\right)$ and $\left(\beta_{vC}, \{m_{V,\alpha}\}_{|\alpha|=s+r}\right)$,

$$
\begin{aligned}
&\left| f\left(\beta_{vC}, \{m_{V,\alpha}\}_{|\alpha|=s+r}\right) - f\left(\hat{\beta}_{vC}, \{\hat{m}_{V,\alpha}\}_{|\alpha|=s+r}\right)\right| \qquad (3.15)\\
&= \left|\left[\nabla f\left(\tilde{\beta}_{vC}, \{\tilde{m}_{V,\alpha}\}_{|\alpha|=s+r}\right)\right]^T \left[\left(\beta_{vC}, \{m_{V,\alpha}\}_{|\alpha|=s+r}\right) - \left(\hat{\beta}_{vC}, \{\hat{m}_{V,\alpha}\}_{|\alpha|=s+r}\right)\right]\right| \\
&\leq \left|\left[\nabla_\beta f\left(\tilde{\beta}_{vC}, \{\tilde{m}_{V,\alpha}\}_{|\alpha|=s+r}\right)\right]^T \left[\beta_{vC} - \hat{\beta}_{vC}\right]\right| \\
&\quad + \left|\left[\nabla_m f\left(\tilde{\beta}_{vC}, \{\tilde{m}_{V,\alpha}\}_{|\alpha|=s+r}\right)\right]^T \left[\{m_{V,\alpha}\}_{|\alpha|=s+r} - \{\hat{m}_{V,\alpha}\}_{|\alpha|=s+r}\right]\right| \\
&\leq |C|\delta_2 \max\left|\frac{\partial f}{\partial \beta_{vz.C}}\right| + \binom{|C|+s}{|C|}\delta_1 \max\left|\frac{\partial f}{m_{V,\alpha}}\right|
\end{aligned}
$$

where ∇_β and ∇_m indicate the gradient with respect to the linear coefficients and moments, respectively. The last inequality follows from Hölder's inequality. Plugging in δ_2 from Lemma 3.1 yields

$$
\begin{aligned}
&\binom{|C|+s}{|C|}\delta_1 \max\left|\frac{\partial f}{m_V, \alpha}\right| + |C|\delta_2 \max\left|\frac{\partial f}{\partial \beta_{vz.C}}\right| \qquad (3.16)\\
&\qquad \leq \binom{|C|+s}{|C|}\delta_1 s! \left(\frac{\sqrt{J}M}{\lambda_{\min}} + 4\frac{J^{3/2}M\delta_1}{\lambda_{\min}^2}\right)^s \\
&\qquad\quad + |C|4\frac{J^{3/2}M\delta_1}{\lambda_{\min}^2}(|C|+1)^s \left[(M+\delta_1)\, s \left(\frac{\sqrt{J}M}{\lambda_{\min}} + 4\frac{J^{3/2}M\delta_1}{\lambda_{\min}^2}\right)^{s-1}\right] \\
&\qquad \leq (|C|+s)^s\delta_1 \left(\frac{3\sqrt{J}M}{\lambda_{\min}}\right)^s \\
&\qquad\quad + |C|4\frac{J^{3/2}M\delta_1}{\lambda_{\min}^2}(|C|+1)^s \left[(M+\delta_1)\, s \left(\frac{3\sqrt{J}M}{\lambda_{\min}}\right)^{s-1}\right] \\
&\qquad \leq (|C|+s)^s\delta_1 \left(\frac{3\sqrt{J}M}{\lambda_{\min}}\right)^s \\
&\qquad\quad + |C|4\frac{J\delta_1}{\lambda_{\min}}(|C|+1)^s \left[2Ms \left(\frac{3\sqrt{J}M}{\lambda_{\min}}\right)^s\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \delta_1 \left[8(J+K)^K JK \left(\frac{3\sqrt{J}M}{\lambda_{\min}} \right)^K \left(1 + \frac{JM}{\lambda_{\min}} \right) \right] \\
&\leq \delta_1 \left[16(J+K)^K JK \left(\frac{3\sqrt{J}M}{\lambda_{\min}} \right)^K \left(\frac{JM}{\lambda_{\min}} \right) \right] \\
&= \delta_1 \left[16(3^K)(J+K)^K K \frac{J^{(K+4)/2} M^{K+1}}{\lambda_{\min}^{K+1}} \right].
\end{aligned}
$$

□

The second inequality holds because we assumed $J\delta_1/\lambda_{\min} < 1/2$; the third inequality holds because we assumed $\delta_1 < M$; the fourth inequality holds because we assumed $|C| \leq J$ and $s \leq K$; the fifth inequality holds because we assumed $JM/\lambda_{\min} > 1$.

We can now deduce that the distance between $\hat{\tau}_{v.C\to u}$ and $\tau_{v.C\to u}$ is bounded, which we formally state in the following lemma.

Lemma 3.3. *Suppose that (C2), (C3), and (C4) hold. Then*

$$
|\hat{\tau}_{v.C\to u} - \tau_{v.C\to u}| < 4M\delta_1 \Phi(J,K,M,\lambda_{\min}) + 2\left(\delta_1 \Phi(J,K,M,\lambda_{\min})\right)^2 = \delta_3
$$

for the function $\Phi(J,K,M,\lambda_{\min})$ given in Lemma 3.2.

The proof of Lemma 3.3 is given in the appendix. Using Lemmas 3.1-3.3, we now give a proof of Theorem 3.2.

Proof. We first show Theorem 3.2(ii) through induction. By Lemma 3.3 and assuming (3.11), each of the sample statistics $\hat{\tau}_{v.C\to u}$ is within $\delta_3 < \gamma/2$ of the corresponding population quantity. Thus, any statistic corresponding to a parameter with value 0 is less than $\gamma/2$ and by (C1) and the condition on γ in (3.11), all statistics corresponding to a non-zero parameter are greater than $\gamma/2$.

Recall that, $\Theta^{(z)}$, is a topological ordering of nodes, and assume for some step z, that $\Theta^{(z-1)}$ is consistent with a valid ordering of G. Let $R^{(z)} = \{v \in \Psi^{(z-1)} : \text{an}(r) \subseteq \Theta^{(z-1)}\}$ so

that any $r \in R^{(z)}$ is a root in the subgraph induced by $\Psi^{(z-1)}$ and $\Theta^{(z)} = (\Theta^{(z-1)}, \{r\})$ would still be consistent with G.

Setting $g = \gamma/2$ does not incorrectly prune any parents, so $\text{pa}(r) = \mathcal{C}_r^{(z)}$ which in turn implies $\hat{T}(r, \mathcal{C}_r^{(z)}, \Phi^{(z-1)}) < \gamma/2$ for any $r \in R^{(z)}$. Similarly, for any $v \in \Psi^{(z-1)} \setminus R^{(z)}$, there exists a $u \in \Psi^{(z-1)}$ such that $|\hat{\tau}_{v.C \to u}| > \gamma/2$ for all $C \subseteq \Theta^{(z-1)}$. Thus,

$$\hat{T}\left(r, \mathcal{C}_r^{(z)}, \Psi^{(z-1)}\right) < \hat{T}\left(v, \mathcal{C}_v^{(z)}, \Psi^{(z-1)}\right)$$

for every $r \in R^{(z)}$ and $v \in \Psi^{(z-1)} \setminus R^{(z)}$, so $\arg\min_{v \in \Psi^{(z-1)}} \hat{T}\left(v, \mathcal{C}_v^{(z)}, \Psi^{(z-1)}\right)$ must be in $R^{(z)}$ and $\Theta^{(z)}$ remains consistent with G. The base case for $z = 1$ is trivially satisfied since $\Theta^{(0)} = \emptyset$.

For Theorem 3.2(i), $P \in \mathcal{P}_{F_K}(G)$ follows directly from the definition. To show uniqueness, we use population quantities so that $\delta_1 = 0$ which in turn implies $\delta_3 = 0$. Then for any $\gamma > 0$, Algorithm 1 will return G. Thus, by 3.2(ii), G must be unique. □

Remark. *Under the conditions of Theorem 3.2, for $\alpha \leq 1$, the tuning parameter described in Section 3.3, the algorithm will return a topological ordering consistent with G, but $\hat{E}$ may be a superset of E. However, there exists some $\alpha \geq 1$ which will recover the exact graph.*

The statement in Theorem 3.2 concerned an explicit cut-off g; however, in practice we specify a tuning parameter α which is easier to interpret and tune. Letting $\alpha = 1$ ensures that under the specified conditions, the cut-off $g^{(z)} < \gamma/2$, so no parents will be incorrectly pruned and the estimated topological ordering will still be correct. This, however, may not remove all ancestors which are not parents so the estimated edge set may still a superset of the true edge set. Specifically, letting

$$\alpha = \frac{\min_v \min_{a \in \text{pa}(v)} \min_{C \cap \text{de}(v) = \emptyset} \hat{\tau}_{v.C \to a}}{\max_v \max_{a \in \text{an}(v) \setminus \text{pa}(v)} \min_{C \cap \text{de}(v) = \emptyset} \hat{\tau}_{v.C \to a}} \tag{3.17}$$

which under the assumptions is strictly greater than 1, will correctly prune ancestors and not parents. However, setting α too large may result in an incorrect estimate of the ordering since

a true parent may be errantly pruned. Thus, in practice we advocate a more conservative approach of setting $\alpha \leq 1$ which anecdotally is more robust to sampling error and violations of strong faithfulness.

Remark. *Suppose $P_Y \in \mathcal{P}(G)$ but is not necessarily parentally faithful with respect to G. If $\alpha = 0$ and $\beta_{vu} \neq 0$ for any $(u, v) \in E$, then for generic error moments a correct ordering will still be recovered consistently as $\delta_1 \to 0$.*

Note that Corollary 3.2(ii) holds even without parental faithfulness. So for generic error moments, $\gamma > 0$ such that $T(v, \mathcal{C}_v^{(z-1)}, \Phi^{(z-1)}) > \gamma$ for all $v \in \Phi^{(z-1)} \setminus R^{(z)}$ for all steps z. However, without parental faithfulness, a parent node may be errantly pruned if $\alpha > 0$. So to ensure Corollary 3.2(i) holds, we need $\text{pa}(r) \subseteq \mathcal{C}_v^{(z-1)}$ for all $r \in R^{(z)}$. This is satisfied by letting $\mathcal{C}_r^{(z-1)} = \Theta^{(z-1)}$. For fixed γ, since $\delta_3 \to 0$ as $\delta_1 \to 0$, then there exists some δ_1 sufficiently small so that $\gamma > 2\delta_3$.

3.3.2 High-dimensional consistency

We now consider a sequence of graphs, observations, and distributions indexed by the number of variables p: $G^{(p)}$, $Y^{(p)}$, $P_Y^{(p)}$, and $P^{(p)}$. For notational brevity, we do not explicitly include the index p in the notation. The following corollary states conditions sufficient for the deterministic conditions of Theorem 3.2 to hold with probability tending to 1. We make explicit assumptions on P_Y and let $m_{V,a}^\star$ denote the population moments of P_Y. Again, we allow for misspecification, but make assumptions about the L^∞ distance between population moments of P_Y and some $P \in \mathcal{P}_{F_K}(G)$.

Condition (C5)**.** *P_Y is a log-concave distribution.*

Condition (C6)**.** *All population moments of P_Y up to degree $2K$, $m_{V,\alpha}^\star$ for $|\alpha| \leq 2K$, are bounded by $M - \xi > \max(1, \lambda_{\min}/J)$.*

Condition (C7)**.** *Each population moment of Y up to degree $2K$, $m_{V,\alpha}^\star$ for $|\alpha| \leq 2K$, is within ξ of the corresponding population moment of P.*

When Y is actually generated from a recursive linear structural equation model, (C7) trivially holds with $\xi = 0$ and log-concave errors imply that Y is log-concave. The first condition in Corollary 3.3 shows how n must grow relative to p, and the second assumption shows how ξ may scale relative to the other quantities.

Corollary 3.3. *For a sequence of distributions P and data Y assume (C1), (C2), (C5), (C6), and (C7) hold. For pruning parameter $g = \gamma/2$, Algorithm 1 will return the graph $\hat{G} = G$ with probability tending to 1 if*

$$\frac{\log(p)}{n^{1/(2K)}} \frac{J^{5/2}K^{5/2}M^2}{\gamma^{1/2}\lambda_{\min}^{3/2}} \to 0, \text{ and} \tag{3.18}$$

$$\xi \frac{3^K K^{K+1} J^{(3K)/2+2} M^{K+2}}{\gamma \lambda_{\min}^{K+1}} \to 0 \tag{3.19}$$

when $p \to \infty$ and $\gamma, \lambda_{\min} < 1 < M$.

Proof. (C6) and (C7) directly imply (C3). It remains to be shown that (C4) and (3.11) hold for the γ specified in (C1). Solving the inequality in Lemma 3.3 for δ_1 shows (3.11) will be satisfied if the sample moments of Y are within δ of the population moments of Y for some δ such that $\delta + \xi \leq \delta_1$ with δ_1 less than

$$\min\left(\frac{-8M\Phi + \sqrt{(8M\Phi)^2 + 16\Phi^2\gamma}}{8\Phi^2}, \frac{\lambda_{\min}}{2J}, M\right) = \min\left(\frac{\sqrt{M^2 + \gamma/4} - M}{\Phi}, \frac{\lambda_{\min}}{2J}\right) \tag{3.20}$$

for Φ defined in (3.12). Since $J, K, M > 1$, $\gamma, \lambda_{\min} < 1$ ensure that first term is the relevant term. We further simplify the expression since

$$\sqrt{M^2 + \gamma/4} \geq M + \gamma \min_{t \in (0,\gamma)} \left.\frac{\partial \sqrt{M^2 + \gamma/4}}{\partial \gamma}\right|_{\gamma = t} = M + \frac{\gamma}{8\sqrt{M^2 + \gamma/4}}.$$

Thus, the conditions of Theorem 3.2 will be satisfied if

$$\delta + \xi \leq \frac{\gamma}{8\sqrt{M^2 + \gamma/4}\Phi} =: \delta_4.$$

Specifically, we analyze the case when $\xi < \delta_4/2$ and $|\hat{m}_{V,a} - m_{V,a}| < \delta < \delta_4/2$ for all $|a| \leq K$. If Y_v follows a log-concave distribution, we can directly apply Lemma B.3 of Lin et al. (2016) which states for f, some K degree polynomial of log-concave random variables $Y = Y_1, \ldots Y_n$, and some absolute constant, L, if

$$\frac{2}{L}\left(\frac{\delta}{(e)\sqrt{\text{var}\left[f(Y)\right]}}\right)^{1/K} \geq 2 \tag{3.21}$$

then

$$\Pr\left(|f(Y) - \mathbb{E}\left[f(Y)\right]| > \delta\right) \leq \exp\left(\frac{-2}{L}\left(\frac{\delta}{\sqrt{\text{var}\left[f(Y)\right]}}\right)^{1/K}\right). \tag{3.22}$$

Letting $f(Y)$ be the sample moments of Y up to degree K, (C6) implies the variance is bounded by M/n. Since there are $\binom{p+K}{p}$ moments with degree at most K (which is less than p^K when $p > 2$), then by a union bound, when $0 < \xi < \delta_4/2$,

$$\begin{aligned}\Pr\left(\hat{G} = G\right) &\geq 1 - \Pr\left(|\hat{m}_{V,a} - m_{V,a}| > \delta_4/2 \text{ for any } |a| \leq K\right) \\ &\geq 1 - p^K \exp\left[\frac{-2}{L}\left\{\frac{\delta_4/2}{(M/n)^{1/2}}\right\}^{1/K}\right]\end{aligned} \tag{3.23}$$

when

$$\frac{2n^{1/(2K)}}{L}\left(\frac{\delta_4/2}{(e)M^{1/2}}\right)^{1/K} \geq 2. \tag{3.24}$$

In the asymptotic regime, where p, is increasing,

$$\frac{LM^{1/(2K)}K\log(p)}{(\delta_4/2)^{1/K}\, n^{1/(2K)}} \to 0 \tag{3.25}$$

implies that the inequality in (3.24) will be satisfied and

$$p^K \exp\left[\frac{-2}{L}\left\{\frac{\delta_4/2}{(M/n)^{1/2}}\right\}^{1/K}\right] \to 0.$$

Plugging in the expression for δ_4, we find

$$\frac{LM^{1/(2K)}K\log(p)}{(\delta_4/2)^{1/K}\,2n^{1/(2K)}} = \frac{LM^{1/(2K)}K\log(p)}{2n^{1/(2K)}} \times \left(\frac{16\sqrt{M^2+\gamma/4}\left(16(3^K)(J+K)^K K J^{(K+4)/2}M^{K+1}\right)}{\gamma\lambda_{\min}^{K+1}}\right)^{1/K}. \tag{3.26}$$

Assuming that $\gamma < M$, then this quantity is

$$\sim O\left(\frac{\log(p)}{n^{1/(2K)}} \times \frac{J^{5/2}K^{5/2}M^2}{\gamma^{1/2}\lambda_{\min}^{3/2}}\right).$$

In addition, $\xi < \delta_4/2$ will be satisfied if $\frac{2\xi}{\delta_4} \to 0$. This quantity

$$\frac{2\xi}{\delta_4} = 2\xi\left(\frac{16\sqrt{M^2+\gamma/4}\left(16(3^K)(J+K)^K K J^{(K+4)/2}M^{K+1}\right)}{\gamma\lambda_{\min}^{K+1}}\right); \tag{3.27}$$

and when again assuming that $\gamma < M$, this quantity is

$$O\left(\xi\frac{3^K K^{K+1}J^{(3K)/2+2}M^{K+2}}{\gamma\lambda_{\min}^{K+1}}\right).$$

□

3.4 Numerical results

3.4.1 Simulations: low dimensional performance

We first compare the proposed method using: (1) min-max $\hat{T}_1$ and (2) max-min $\hat{T}_2$ against (3) DirectLiNGAM (Shimizu et al., 2011) and (4) Pairwise LiNGAM (Hyvärinen and Smith, 2013). We randomly generate graphs and corresponding data with the following procedure. For each node $v = 2, \ldots, p$, select the number of parents d_v uniformly from $1, \ldots, \min(v, J)$. We always include edge $(v-1, v)$ to ensure that the ordering is unique and draw $\beta_{v,v-1}$ uniformly from $(-1, -.5) \cup (.5, 1)$. The remaining $d_v - 1$ parents are selected uniformly

from $[v-2]$ and the corresponding edges are set to $\pm 1/5$. The n error terms for variable v are generated by first drawing a standard deviation $\sigma_v \sim \text{unif}(.8, 1)$ and then drawing $\varepsilon_{vn} \sim \sigma_v \text{unif}(-\sqrt{3}, \sqrt{3})$.

We fix the max in-degree $J = 3$, let $p = 5, 10, 15, 20$, and let $n = 50p$ and $n = 10p$. Since $\gamma/2$ is not known, we set $\alpha = .8$ to ensure that at least the topological ordering can be recovered consistently. We compare the accuracy of the graph estimation algorithms by measuring Kendall's Tau between the returned ordering and the true ordering. The procedure is repeated 500 times for each setting of p and n.

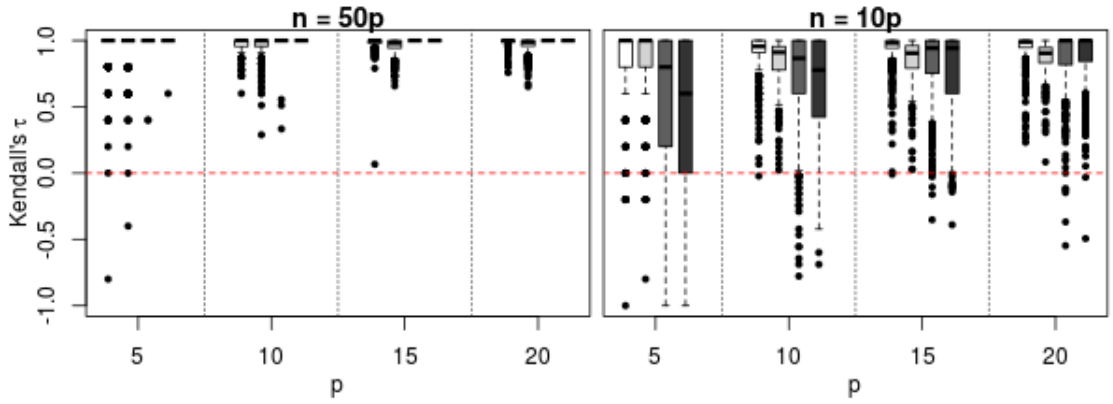


Figure 3.3: Each bar represents the results from 500 randomly drawn graphs and data. In each group, from left to right (lightest to darkest), the bars represent (1) min-max $\hat{T}_1$, (2) max-min $\hat{T}_2$, (3) DirectLiNGAM, and (4) Pairwise LiNGAM. In the left panel $n = 50p$ and the right panel $n = 10p$.

The simulation results are shown in Figure 3.3. In the low-dimensional case with $n = 50p$, the Pairwise LiNGAM and DirectLiNGAM methods outperform the proposed method (with either statistic). However, in the medium dimensional case of $n = 10p$, we see that the proposed method begins to out perform the low dimensional methods.

Comparing the T_1, the min-max, against T_2, the max-min, it appears that the min-max does outperform the max-min. However, in Figure 3.4, we observe a large difference in computational effort. We also include the cases where $p = 40, 80$ for further contrast. In the

remaining simulations, we use the max-min statistic, T_2. The proposed method compares

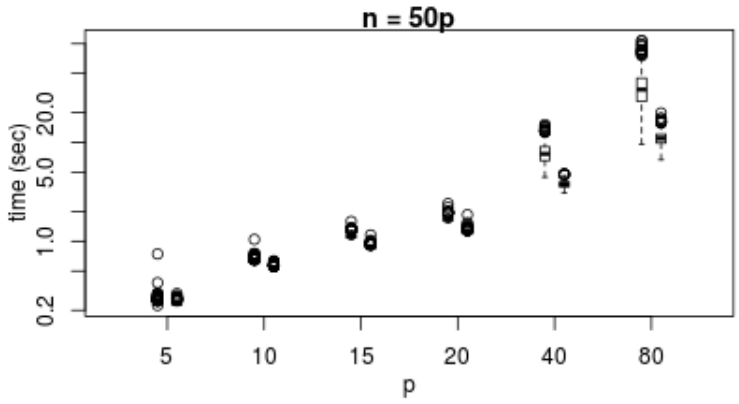


Figure 3.4: Each bar represents the results from 500 randomly drawn graphs and data with $n = 50p$. In each group from the left (light) represents min-max, $\hat{T}_1$ and the right (dark) max-min, $\hat{T}_2$. The y-axis is on a log scale

quite favorably to the DirectLiNGAM method in computational effort because of the expensive kernel mutual information calculation, and is comparable to the Pairwise LiNGAM. However, we do not otherwise present a direct timing comparison because DirectLiNGAM and Pairwise LiNGAM are both implemented in Matlab while our proposed method is implemented in R (R Core Team, 2017) and C++ using Rcpp (Eddelbuettel and François, 2011; Eddelbuettel and Sanderson, 2014).

3.4.2 Simulations: high-dimensional consistency

To show high-dimensional consistency for general graphs, we generate the graph and coefficients as described in Section 3.4.1. We let $p = 100, 200, 500, 1000, 1500$, $n = 3/4p$, and the maximum in-degree $J = 2$. We also consider the case where the graph may contain hub nodes, nodes with very large out-degree. Under our non-Gaussian setting, we only make assumptions about the in-degree, while the PC algorithm under Gaussian errors requires the maximum total degree (both in and out) to grow sub-linearly (Kalisch and Bühlmann, 2007). However, in gene regulatory networks, there are often hub nodes which have a very high out-

degree and regulate many downstream genes (Hao et al., 2012). We generate random graphs with hubs using the following procedure. We first include a directed edge from $v-1$ to v for all nodes $v = 2 \dots p$ and draw the edge weight uniformly from $(-1, -.65) \cup (.65, 1)$. The standard deviations for each of the error terms is drawn uniformly from $(.8, 1)$. We then set nodes $\{1, 2, 3\}$ as hubs and include an edge with weight $\pm 1/5$ to each non-hub node from a random hub. Thus, the degree for each of the hub nodes grows linearly with p, but the in-degree remains bounded by 2. Again, we let $p = 100, 200, 500, 1000, 1500$ and $n = 3/4p$. For both the general graph and hub graph cases, the results for 20 simulations at each value of $p = 100, 200, 500, 1000$ and 10 simulations for $p = 1500$ are shown in Figure 3.5.

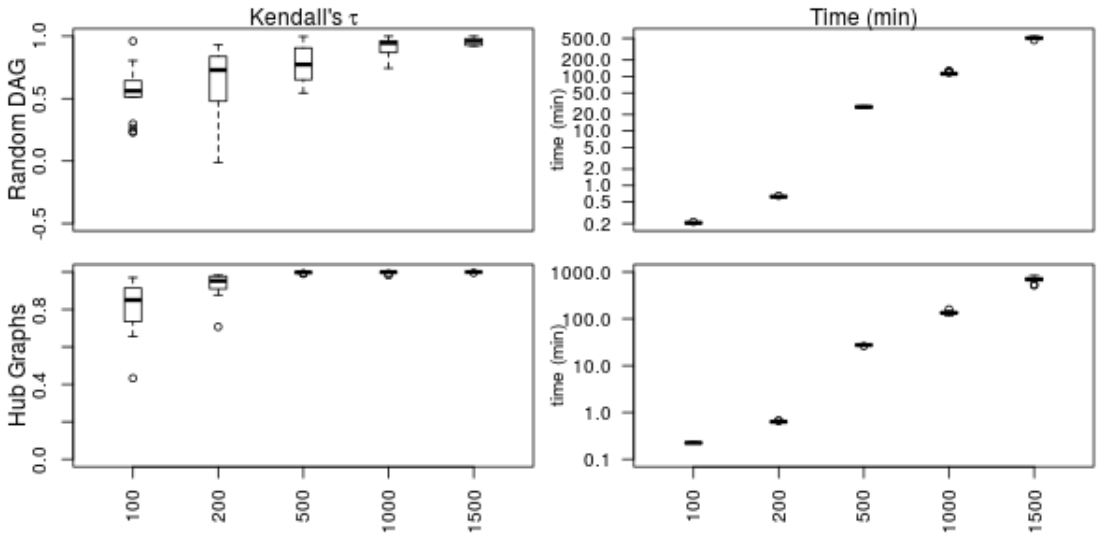


Figure 3.5: Each boxplot represents the results of 20 simulations for $p < 1500$ and 10 simulations for $p = 1500$. The top panels show results from randomly drawn DAGs while the bottom panel shows results from DAGs constructed to have hub structure.

3.4.3 Data example: high-dimensional performance

We apply the proposed high-dimensional LiNGAM method to estimate causal structure in the stocks which comprise the Standard and Poor's 500. Specifically, we consider the

percentage increase/decrease for each company's share price for each trading day between Jan 2007 to Sep 2017. After removing companies for which data is not available for the entire period, we are left with $p = 442$; we also scale and center the data so that each variable has mean 0 and variance 1. Because we believe the causal structure may vary over time and the observations may only be approximately identically distributed for short periods, we estimate causal structure for each of the following periods separately: 2007-2009, 2010-2011, 2012-2013, 2014-2015, 2016-2017 (ending in September). Across these periods, the sample size, n, ranges from 425 to 755; we let $J = 3$ for each time period. This is a setting where the only data we can gather is observational and high-dimensional.

It is worth noting that we only consider the daily percentage change of each company's stock price; thus, the causal structure we are modeling may not be a representation of longer term economic level causal effects, but rather short term trading structure. In addition, the underlying structure is unlikely to be causally sufficient or acyclic. Nonetheless, when using the sector of each individual company to assess the estimated causal structure, we still find that the method recovers interpretable structure. We first consider the most recent Jan 2016 - Sep 2017 period. In the estimated topological ordering, 160 of the companies directly follow another company in the same sector. When arranging the companies randomly, the probability of 160 or more matches is roughly 0. Figure 3.6 shows a box-plot for the estimated topological ordering of the companies within each sector, and the sectors are sorted top to bottom by median ordering (top is cause, bottom is effect). Near the top, we see utilities, energy, real estate, and finance. Since energy is an input for almost every other sector, intuitively price movements in energy should be causally upstream of other sectors. The estimated ordering of utilities might seem surprising since utility companies typically trade with very little volatility; however, utility stocks are typically thought of as a proxy for bond prices. Thus, the estimated ordering may reflect the fact that changes in utility stocks capture much of the causal effect of changing interest rates, which had stayed constant for much of 2011-2015 but began moving again in 2016. Real estate and finance are two other sectors which are highly impacted by interest rates and are estimated to be early in the

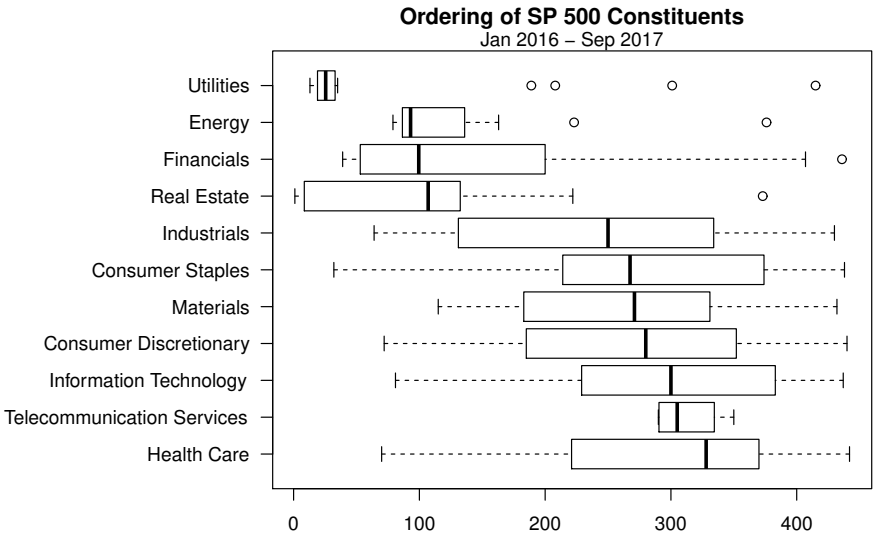


Figure 3.6: Estimated causal ordering by sector for Jan 2016 - Sep 2017.

causal ordering as well.

In Figure 3.7, we have ranked each sector by the median topological ordering for each period. We can see that the orderings are relatively stable over time, but there are a few notable changes. In particular, we see that in 2007, real estate was estimated to be the "root sector" while finance is in the middle. This aligns with the idea that the root of the 2008 financial crisis was actually caused by failing mortgage backed securities in real estate, which had a causal effect on finance. However, over time, real estate has moved more downstream.

3.5 Discussion

We have shown when the errors in a linear structural equation model are non-Gaussian, the underlying graph can be estimated consistently in the high-dimensional setting. The proofs for consistency involve our specific test statistic and log concave errors; however, a similar analysis could be used for other test statistics which are Lipschitz continuous in the sample moments over a bounded domain, can distinguish causal direction, and indicate the presence of confounding. This would include a normalized version of the proposed test statistics which

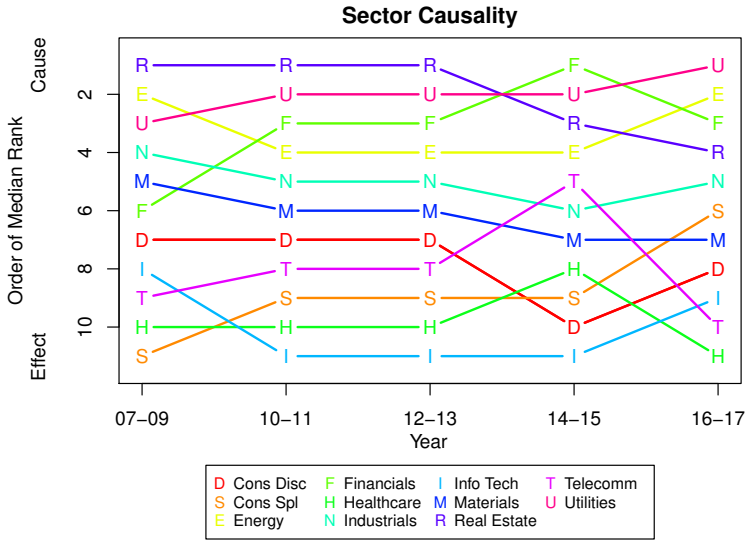


Figure 3.7: Sectors ranked across period by median estimated topological ordering.

accounts for the scaling of the data. In addition, the result would apply directly to other classes of distributions for which the sample moments concentrate at an exponential rate.

The proposed algorithm requires two inputs: a bound on the in-degree and a pruning parameter α. Typically a bound on the in-degree is unknown, but a reasonable upper bound may be used instead. The pruning parameter α plays a similar role to the nominal level for each conditional independence test in the PC algorithm. Both have a direct effect on the sparsity of the estimated graph and also regulate the maximum size of conditioning sets.

In (3.11), we have made a key restriction that the error moments must be adequately different from the moments of any Gaussian and the edge weights must be strongly parentally faithful. In practice, this is a difficult condition to satisfy, and Uhler et al. (2013) show strong faithfulness type restrictions can be problematic in practice. However, even if the distribution is not strongly parentally faithful, we can still consistently recover the correct ordering as long as each individual linear coefficient is non-zero and the errors are sufficiently non-Gaussian. Sokol et al. (2014) consider identifiability of independent component analysis for fixed p when the error terms are Gaussians contaminated with non-Gaussian noise. In particular, when

the effect of the non-Gaussian contamination decreases at an adequately slow rate, the entire mixing matrix is identifiable asymptotically. In our analysis, the measure of non-Gaussianity is treated by our assumptions on γ. Our results suggest that the results of Sokol et al. (2014) can also be extended to the asymptotic regime where the number of variables is increasing.

The modified procedure we propose retains the existing benefits of the DirectLiNGAM procedure. In particular, the output of algorithm is independent of the ordering of the variables in the input data. In the low-dimensional setting, the order of the variables is not typically a large issue, but in the high-dimensional case, the output of causal discovery methods may be highly dependent on ordering (Colombo and Maathuis, 2014). In addition, any edges (or non-edges) known in advance can be easily incorporated into the search procedure. Loh and Bühlmann (2014) show that the precision matrix recovers the moralized graph even with non-Gaussian errors. This could potentially be used as a pre-processing step to greatly decrease computational effort.

Chapter 4

CAUSAL DISCOVERY OF BOW-FREE ACYCLIC GRAPHS

4.1 Introduction

In this chapter, we also focus on causal discovery; however, instead of the high-dimensional DAG setting of Chapter 3, we relax the assumption of causal sufficiency and allow for latent confounding. In particular, we focus on discovering causal structure for SEMs which correspond to *bow-free acyclic path diagrams* (BAPs). Recall that a BAP is a graph $G = \{V, E_{\rightarrow}, E_{\leftrightarrow}\}$ where the directed portion $E_{\rightarrow}$ is acyclic and there is at most one edge between any pair of nodes $u, v \in V$.

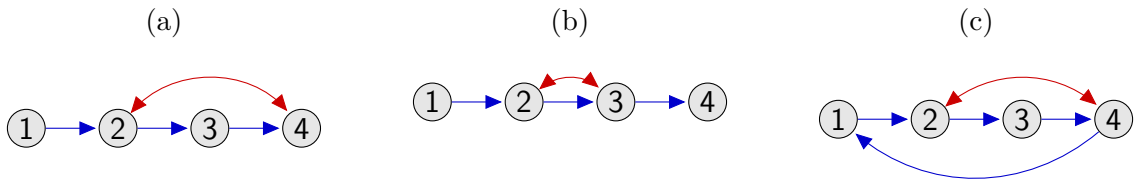


Figure 4.1: (a) is a bow-free acyclic path diagram (BAP), while (b) and (c) are not. In (b), there is a bow between nodes 1 and 3, while in (c), there is a directed cycle $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 1$.

4.1.1 Previous work

Previous works on causal discovery with latent variables typically use constraint based methods which test conditional independence relationships. In particular, most of the work focuses on *maximal ancestral graphs* (MAGs). MAGs can be used to represent causal structure, but can also be used to represent Markov structure which may not necessarily correspond to directly to causality. MAGs are graphs with directed, bidirected, and undirected edges–which

encode the presence of latent selection variables. A graph is *ancestral* if an arrowhead on an edge from u to v implies that $v \notin \text{an}(u)$. Similar to BAPs, MAGs have at most one edge between any pair of nodes (i.e., the graph is also bow-free) and are acyclic. However, the ancestral restriction precludes a sibling of v from also being an ancestor of v. An ancestral graph is also *maximal* if for any two nodes u, v which are not adjacent in G, there exists some set Z in the set of observed variables such that $u \perp\!\!\!\perp v \mid Z$. A graph which is not maximal cannot be distinguished from a corresponding maximal graph via conditional independence relations. Figure 4.2 gives an example from Richardson and Spirtes (2002) of a graph which is ancestral but not maximal. MAGs are a useful class of graphs because they are closed (with respect to conditional independence relations) under marginalization and conditioning. However, when precluding selection variables (i.e., considering MAGs without undirected edges), as we do here, BAPs are a generalization of MAGs.

Spirtes et al. (2000) modify the PC-algorithm–originally designed for discovering DAG structure–to discover MAGs, and the proposed Fast Causal Inference (FCI) consistently discovers a *partial ancestral graph* (PAG). Like a CPDAG, the PAG represents an equivalence class of ancestral graphs which satisfy the same set of conditional independencies (Ali et al., 2009). Zhang (2008) added additional orientation rules such that the output of FCI is complete. Colombo et al. (2012) introduce Really Fast Causal Inference (RFCI) which requires a polynomial (of the variables considered) number of conditional independence tests in a sparse setting. Although it is not necessarily complete, it is sound, even in the high dimensional setting. Finally, Claassen et al. (2013) introduce FCI+, which is both complete and sound under a polynomial number of conditional independence tests if the degree of the graph is bounded.

A few greedy search approaches for causal discovery with latent variables have also been proposed which assume Gaussian errors. Triantafillou and Tsamardinos (2016) give a greedy algorithm for learning MAGs. In simulations, they show that the greedy search compares favorably to constraint based algorithms which can be more susceptible to propagated error.

In addition, Nowzohour et al. (2017) propose a greedy search procedure for discovering

Gaussian linear SEMs parameterized by BAPs by maximizing a BIC type score. Despite the relative dearth of work, BAPs are an important class of graphs which can give a more refined representation of causal structure which cannot be captured by conditional independence alone. Brito and Pearl (2002) show for linear structural equation models, when the graph is known and corresponds to a BAP, the directed and bidirected edge weights can be uniquely identified. However, complications can arise when a graph is not ancestral. For instance, there is no known graphical characterization of a Markov equivalence class of BAPs, which can complicate a greedy search (Nowzohour et al., 2017). Nowzohour et al. (2017) circumvent this problem by identifying graphs which are equivalent up to optimization error.

Although the previously proposed methods have enjoyed great success, when using conditional independencies or maximizing a Gaussian likelihood, only an equivalence class of graphs can be discovered. This can be unsatisfying to a practitioner since generally this set of graphs is quite large and may have conflicting causal interpretations. However, there are often additional non-parametric constraints which can also be tested (Verma and Pearl, 1990; Tian and Pearl, 2002; Shpitser et al., 2014; Evans, 2016).

When considering linear structural equation models where the errors are not Gaussian, additional structure can be determined (Shimizu et al., 2006). Explicitly assuming the errors to be non-Gaussian, Hoyer et al. (2008b) propose lvLiNGAM (latent variable LiNGAM). When the number of unobserved variables are known, lvLiNGAM uses overcomplete independent component analysis to identify the exact causal structure when the full model (including unobserved latents) correspond to a DAG. Although this makes no assumptions about bows in the observed variables, in practice, it is quite unlikely, with the exception of perhaps studies in psychometrics, that the number of latent variables would be known when one is interested in estimating the causal structure. However, even when the model is correctly specified, the overcomplete ICA procedure performs poorly in practice (Entner and Hoyer, 2010; Shimizu and Bollen, 2014). As an alternative, Entner and Hoyer (2010) propose Pairwise lvLiNGAM which discovers the pairwise ancestral relationships for the subset of variables which are not confounded by unobserved latents. In other words, the method certifies ancestral relation-

ships for unconfounded variables, but does not posit a relationship if there is unobserved confounding. Finally, Shimizu and Bollen (2014) present a Bayesian procedure which allows for latent confounding through individual level random effects. Although the method does not require inputting a specific number of confounders, it does require specifying a prior distribution for the edges and a likelihood for the errors. This can be quite difficult in practice and may lead to inconsistent estimation when misspecified.

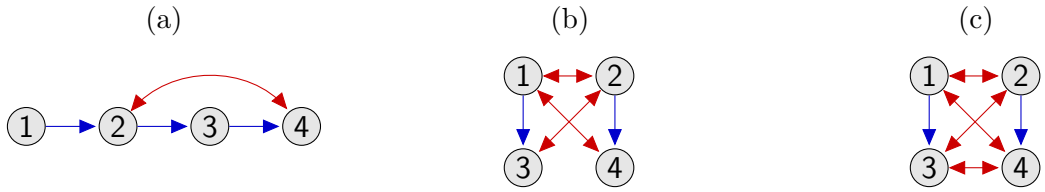


Figure 4.2: (a) is a BAP, but is not ancestral because 2 is both a sibling and ancestor of 4; (b) is an ancestral graph, but it is not maximal because there is no M-seperating set for 3 and 4 (c) is a maximal ancestral graph. Graphs (b) and (c) are from Richardson and Spirtes (2002).

4.1.2 Contribution

In this work, we focus on BAPs and let the graph represent honest causal structure–as opposed to simply a CI map. However, in contrast to Nowzohour et al. (2017) we assume non-Gaussian errors, similar to the LiNGAM setting (Shimizu et al., 2006). We show that when the data are generated by a linear structural equation model with non-Gaussian errors and corresponds to a BAP, the exact causal structure–not simply an equivalence class–is identifiable from observational data. The results do not require specifying a distributional form for the errors, but simply require some higher order moments which do not correspond to the moments of a Gaussian distribution. In addition, we do not require specifying the number of latent variables in advance.

Using these results, we propose the **B**ow-free **A**cyclic **n**on-**G**aussian (BANG) procedure, which consistently discovers the exact causal structure given observational data by testing

whether certain rational functions of the observed data are zero or non-zero. Similar to the approach in Chapter 3 and DirectLiNGAM Shimizu et al. (2011), we also use an iterative procedure to certify ancestral relationships. However, without causal sufficiency, we must consider sets larger than simple pairs. When the maximum in-degree (both directed and bidirected edges) is bounded, the total number of tests performed is bounded by a polynomial of p, the number of variables considered. In simulations, we show that the method reliably discovers exact causal structure when given a large sample.

4.2 Causal discovery setup

In order to determine causal direction, we again exploit the fact that in a linear model with no unobserved confounding and non-Gaussian errors, the residuals of v when regressed onto its parents $\text{pa}(v)$ are independent of $\text{pa}(v)$; however, the residuals of $\text{pa}(v)$ when regressed onto v are not generally independent of v (Shimizu et al., 2006; Shimizu and Kano, 2008). Using a least squares criterion, the first order conditions for the linear coefficients ensure that the correlation between regressors and residuals is 0 in either situation; however, dependence can still be detected by examining the higher order cross moments (i.e. $\mathbb{E}(x^k y) \neq 0$ for $k > 1$) of (incorrectly posited) parents and the residuals of the (incorrectly posited) child.

If the underlying graph is ancestral, one could directly apply this idea to discover a topological ordering of the variables by naively taking some set $C \subseteq V \setminus \{v\}$ and regressing v onto some set C to form residuals, $Y_{v.C}$. If $\text{pa}(v) \subseteq C \subseteq \text{an}(v)$ then $Y_c \perp\!\!\!\perp Y_{v.C}$ for each $c \in C$. However, if $C \cap \{\text{de}(v) \cup \text{sib}(v)\} \neq \emptyset$, then $Y_c \not\perp\!\!\!\perp Y_{v.C}$ for some $c \in C$. Thus, one could simply repeat this procedure to certify ancestral relationships and estimate a graph.

In the case of a bow-free acyclic graph, however, the naive strategy will no longer correctly certify ancestral relationships, since siblings of v can also be ancestors of v. In general, if a target node is regressed onto all of its parents, the resulting coefficients may not correspond to the true direct effects. In the graph presented in Figure 4.3, simple linear regression will

yield consistent estimates of β_{21}. However, regressing 3 onto 2 will consistently estimate

$$\mathbb{E}(Y_3|Y_2) = \frac{\beta_{32}(\beta_{21}^2\omega_{11} + \omega_{22}) + \beta_{21}\omega_{13}}{\beta_{12}^2\omega_{11} + \omega_{22}} \neq \beta_{32}.$$

Even if we were given the true direct effect, β_{32}, and could form $\varepsilon_3 = Y_2 - \beta_{32}Y_3$, in general, $Y_2 \not\perp\!\!\!\perp \varepsilon_3$ since 1 is an ancestor of 2 and $1 \in \text{sib}(3)$.

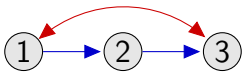


Figure 4.3: The graph is non-Ancestral because $1 \leftrightarrow 3$ and $1 \in \text{an}(3)$. However, it is acyclic in the directed portion and bow-free because there is at most one edge between each pair of nodes.

In the subsequent subsections, we discuss how the naive estimates of the direct effect can be corrected to calculate the true direct effect and how we can use these estimates to certify ancestral relationships. These results will be used to motivate the discovery algorithm presented in Section 4.3.

4.2.1 De-biased direct effect

For some node $v \in V$, let the estimated de-biased direct effect, $\delta_v(C, A, S, D)$, be a function of sets $C \subseteq A \subseteq V \setminus \{v\}$ and matrices $S, D \in \mathbb{R}^{p\times p}$:

$$\delta_v(C, A, S, D) = \left[(I - D)_{C,A}S_{A,A}(I - D_{A,A})^T((I - D_{A,A})^{-T})_{A,C}\right]^{-1}(I - D)_{C,A}S_{A,v} \quad (4.1)$$

Theorem 4.1. *Suppose Y is generated by a linear SEM with parameters B and Ω and corresponds to an acyclic mixed graph $G = \{V, E_{\rightarrow}, E_{\leftrightarrow}\}$. For node $v \in V$ and sets $C \subseteq A \subseteq V \setminus \{v\}$, suppose*

1. $pa(v) \subseteq C \subseteq \text{an}(v) \setminus \text{sib}(v)$,
2. $A = \text{An}(C)$,

3. $D_{A,A} = B_{A,A}$, *and*
4. $S_{\{A,v\},\{A,v\}} = \Sigma_{\{A,v\},\{A,v\}}$.

Then $\delta_v(C, A, S, D) = B_{v,C}$, *the true direct effect of* C *on* v.

Proof. The assumptions $pa(v) \subseteq C \subseteq \text{an}(v) \setminus \text{sib}(v)$ and $A = \text{An}(C)$ imply that $A = \text{an}(v)$. Recall that $\Pi = (I - B)^{-1}$, where π_{vu} represents the total effect of u onto v. Then we have the representation

$$Y_v = \varepsilon_v + \varepsilon_A(\Pi_{v,A})^T. \tag{4.2}$$

Note that since B is permutation similar to a lower triangular matrix (via the acyclic assumption), $(I - B)$ must be invertible. For convenience, let $\bar{A} = V \setminus A$. By well known block matrix inversion formulas,

$$\begin{aligned}\left[(I-B)^{-1}\right]_{A,A} &= \left[(I_{A,A} - B_{A,A}) - (I_{A,\bar{A}} - B_{A,\bar{A}})(I_{\bar{A},\bar{A}} - B_{\bar{A},V\setminus A})^{-1}(I_{\bar{A},A} - B_{\bar{A},A})\right]^{-1} \\ &= \left[I_{A,A} - B_{A,A}\right]^{-1}.\end{aligned} \tag{4.3}$$

The equality holds because $A = \text{An}(A)$ implies that

$$(I_{A,\bar{A}} - B_{A,\bar{A}}) = 0. \tag{4.4}$$

We also represent the residuals of C

$$\varepsilon_C = Y_A((I - B)_{C,A})^T. \tag{4.5}$$

Since $\Pi_{v,A}$ represents the total effect of A onto v, and $\text{pa}(v) \subseteq C$, then $\Pi_{v,A} = B_{v,C}\Pi_{C,A}$. We assume $\varepsilon_v \perp\!\!\!\perp \varepsilon_C$, so

$$\begin{aligned}(I - B)_{C,A}\Sigma_{A,v} &= (I - B)_{C,A}\mathbb{E}(Y_A^T Y_v) = \mathbb{E}((I - B)_{C,A}Y_A^T Y_v) \\ &= \mathbb{E}(\varepsilon_C^T Y_v) \\ &= \mathbb{E}(\varepsilon_C^T \varepsilon_v) + \mathbb{E}(\varepsilon_C^T \varepsilon_A (B_{v,C}\Pi_{C,A})^T)\end{aligned} \tag{4.6}$$

$$
\begin{aligned}
&= \Omega_{C,A}(B_{v,C}\Pi_{C,A})^T \\
&= \Omega_{C,A}(B_{v,C}[(I-B)^{-1}]_{C,A})^T.
\end{aligned}
$$

Again, by (4.4),

$$
\begin{aligned}
\Omega_{C,A} &= [(I-B)\Sigma(I-B)^T]_{C,A} = (I-B)_{C,V}\Sigma[(I-B)^T]_{V,A} \\
&= \begin{bmatrix}[(I-B)_{C,A} & (I-B)_{C,\bar{A}}\end{bmatrix} \begin{bmatrix}\Sigma_{A,A} & \Sigma_{A,\bar{A}} \\ \Sigma_{\bar{A},A} & \Sigma_{\bar{A},\bar{A}}\end{bmatrix} \begin{bmatrix}[(I-B)^T]_{A,A} \\ [(I-B)^T]_{\bar{A},A}\end{bmatrix} \\
&= (I-B)_{C,A}\Sigma_{A,A}[(I-B)^T]_{A,A}.
\end{aligned} \tag{4.7}
$$

Plugging (4.7) into (4.6) and solving yields

$$
\begin{aligned}
B_{v,C}^T &= \left[(I-B)_{C,A}\Sigma_{A,A}[(I-B)^T]_{A,A}[(I-B)^{-T}]_{A,C}\right]^{-1}(I-B)_{C,A}\Sigma_{A,v} \\
&= \delta_v(C, A, \Sigma, B) = \delta_v(C, A, S, D).
\end{aligned} \tag{4.8}
$$

The last equality holds by assumptions 3 and 4 since the calculation of δ only involves $\Sigma_{\{A,v\},\{A,v\}}$ and $D_{A,A}$. □

Theorem 4.1 implies that for a node $v \in V$, when we select correct sets C and A such that $\mathrm{pa}(v) \subseteq C \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$ and $A = \mathrm{An}(C)$, given the population covariance of the observed data, Σ, and the direct effects between A, we can recover the direct effect of C onto v if $C \cap \mathrm{sib}(v) = \emptyset$. We do not explicitly require that the entire graph G be bow-free, although we do require that $\mathrm{pa}(v) \cap \mathrm{sib}(v) = \emptyset$ in order for C to exist. Note that since δ_v only involves inversions and multiplication, it is a rational function of the elements of S and D. However, the specific form of the rational function is determined by the sets C and A.

Intuitively, one can think of calculating δ through the following procedure. We first form the errors ε_C and naively regress Y_v onto ε_C. This would give us a total effect of C on v, but may be biased by dependency between ε_C and ε_A. However, because we know the direct effects among A, we can compute the covariance of ε_A from Σ. Thus, we can

de-bias the naive regression coefficients to give the true direct effects. The assumption that $C \cap \text{sib}(v) = \emptyset$ ensures that we do not need to also correct for dependency between ε_C and ε_v, which we would not be able to calculate without already having the correct direct effects. In Section 4.3, we will iteratively apply this procedure in the discovery algorithm.

The following results will show how we can certify whether we have selected appropriate sets C and A. But first, we define and review some additional notions which will be helpful for the stated results.

Recall that for some directed path $l = v_1 \to \ldots \to v_s$, the path weight, $W(l)$ is the product of the edgeweights along the path, i.e $W(l) = \prod_{j=1}^{s-1} \beta_{v_{j+1},v_j}$. The *marginal direct effect* is the direct effect between a set of variables A in the sub-model defined by marginalizing away all $V \setminus A$. For convenience, let $\Lambda = I - B$ and $\bar{A} = V \setminus A$. Then the marginal direct effect for some set of variables A is

$$\begin{aligned}\tilde{B}(A) = I - \left[\left(\Lambda^{-1}\right)_{A,A}\right]^{-1} &= \left[\left(\Lambda_{A,A} - \Lambda_{A,\bar{A}}(\Lambda_{\bar{A},\bar{A}})^{-1}\Lambda_{\bar{A},A}\right)^{-1}\right]^{-1} \\ &= I - \Lambda_{A,A} - \Lambda_{A,\bar{A}}(\Lambda_{\bar{A},\bar{A}})^{-1}\Lambda_{\bar{A},A}.\end{aligned}$$

For $i \neq j$,

$$\tilde{B}(A)_{ij} = \beta_{ij} + \sum_{s\in\bar{A}} \beta_{is} \sum_{t\in\bar{A}} \bar{\pi}_{st}\beta_{tj} \tag{4.9}$$

where $\bar{\pi}_{st} = ((\Lambda_{\bar{A},\bar{A}})^{-1})_{st}$ is the total effect of t on s in the sub-graph of G induced by $V \setminus A$. In other words, this is the sum of the product of the edge weights over all directed paths from s to t which only contain edges between nodes in $\bar{A}$. This implies that for $i, j \in A$, $\tilde{B}(A)_{ij} \neq 0$ only if $j \in \text{an}(i)$.

For $D \in \mathbb{R}^{p\times p}$, define the *pseudo-parents* of c given D to be the set $\text{pa}_D(c) = \{v : D_{cv} \neq 0\}$, the *pseudo-ancestors* of c given D to be the set $\text{an}_D(c) = \{v : [(I - D)^{-1}]_{cv} \neq 0\}$. Note that $\text{an}_D(c)$ always includes c. Similar to previous notation, when an argument is a set we mean the union of the function applied to each element, i.e., for some set C, $\text{pa}_D(C) = \bigcup_{c\in C} \text{pa}_D(c)$.

In most cases, we will consider a matrix D such that $D_{ij} \neq 0$ only if $j \in \text{an}(i)$. This implies that $\text{an}_D(i) \subseteq \text{an}(i)$. However, it will sometimes be useful to place an additional restriction on D. Define a set of matrix valued functions, $\mathcal{D} = \{\mathbf{D}\}$ such that $\mathbf{D} : B \mapsto D$ for $B, D \in \mathbb{R}^{p \times p}$. Each $\mathbf{D}$ is defined by a set of sets $\{C_v\}_{v \in V}$ such that $D_{v,C_v} = \tilde{B}(\{C_v, v\})_{v,C_v}$ and $D_{v,q} = 0$ for any $q \notin C_v$. Note that this implies for all $\mathbf{D} \in \mathcal{D}$, if $D = \mathbf{D}(B)$, then D_{ij} is the sum of path weights for a (not necessarily strict) subset of the paths from j to i. Furthermore, for every $\mathbf{D}$, D_{ij}, is a polynomial of the elements of B, but the specific form of the polynomial varies across $\mathbf{D} \in \mathcal{D}$. Finally, note that $\mathcal{D}$ is a finite set.

For some matrix D and variable c, let $\gamma_c(D)$ denote the resulting residuals when positing D to be the direct effects (i.e., when D is an estimate of B). Furthermore, for given node v, set $C \subseteq A \subseteq V \setminus \{v\}$ and matrix S, let $\gamma_v(C, A, S, D)$ denote the residuals when positing parent set C with $A = \text{an}_D(C)$.

$$\begin{aligned} \gamma_c(D) &= Y_c - Y(D_{c,V})^T \\ \gamma_v(C, S, D) &= Y_v - Y_C \delta_v(C, \text{an}_D(C), S, D). \end{aligned} \tag{4.10}$$

When the arguments for γ_c and γ_v are clear from the context, we will suppress the additional notation.

Recall that $\mathcal{L}_{v,q}$ is the set of all paths from q to v. For some set C with $q \in C$, we can partition $\mathcal{L}_{v,q}$ into disjoint sets

$$\mathcal{L}_{v,q} = \bigcup_{c \in C} \mathcal{L}_{v,q}^{(c)}(C) \tag{4.11}$$

where $\mathcal{L}_{v,q}^{(c)}(C)$ is the set of paths l from q to v such that c is the last node in C to appear in l. Thus, $\mathcal{L}_{v,q}^{(q)}(C)$ are all paths which originate at q and do not pass through any other node in C which implies that $\sum_{l \in \mathcal{L}_{v,q}^{(q)}(C)} W(l) = \tilde{B}(\{C, v\})_{v,q}$. In addition, there is a bijection between the elements $l \in \mathcal{L}_{v,q}^{(c)}(C)$ and $(l_1, l_2) \in \mathcal{L}_{v,c}^{(c)}(C) \times \mathcal{L}_{c,q}$ and in particular for the corresponding elements, $W(l) = W(l_1)W(l_2)$.

4.2.2 Testing ancestral relationships

The following corollary gives a testable direct implication of when the conditions in Theorem 4.1 are satisfied.

Corollary 4.1. *Suppose the conditions in Theorem 4.1 hold, then* $\gamma_c \perp\!\!\!\perp \gamma_v$ *for every* $c \in C$. *Thus,* $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ *for all* $c \in C$.

Proof. Since, $D_{C,A} = B_{C,A}$ so $\gamma_C = \varepsilon_C$. Theorem 4.1 shows that we can recover the direct effect $B_{v,C}$, so $\gamma_v = \varepsilon_v$. Thus, by assumption, since $C \cap \text{sib}(v) = \emptyset$, so $\gamma_c \perp\!\!\!\perp \gamma_v$ for all $c \in C$. This implies and $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = \mathbb{E}(\gamma_c^{K-1})\mathbb{E}(\gamma_v) = 0$. □

For some node v and set $C \subset V \setminus \{v\}$, the algorithm proposed in Section 4.3 will require a certification that $C \subseteq \{\text{an}(v) \setminus \text{sib}(v)\}$. We will do this by testing whether $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ for all $c \in C$. Theorem 4.1 gives sufficient conditions, but in general they are not necessary. The necessary condition given in Lemma 4.1 will be useful in deriving further results.

Lemma 4.1. *Consider fixed* v *and sets* $C \subseteq A \subseteq V \setminus \{v\}$. *Let* $D \in \mathbb{R}^{p\times p}$ *such that* $D_{ij} \neq 0$ *only if* $j \in \text{an}(i)$ *Then, for any* B *and generic error moments,* $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ *for all* $c \in C$ *only if* $\delta_v(C, A, S, D) = \tilde{B}(\{C, v\})_{v,C}$, *the marginal direct effect of* C *onto* v.

Proof. By Okamoto (1973), since $\mathbb{E}\left(\gamma_{c_i}^{K-1}\gamma_v\right)$ is a rational function of the error moments, showing that the quantity is non-zero for some point is sufficient for showing that it vanishes only over a set of measure zero. We prove the statement by showing the contrapositive holds. Without loss of generality, let $C = \{c_1, \ldots, c_{|C|}\}$ where c_i is not a descendant of c_j for any $j < i$. Note that

$$\gamma_v = \varepsilon_v + \sum_{a\in\text{an}(v)} \pi_{va}\varepsilon_a - \sum_{c\in C} \delta_{vc}(\varepsilon_c + \sum_{a\in\text{an}(c)} \pi_{ca}\varepsilon_a).$$

Then suppose i is the minimum index for which $\delta_{c_i} \neq \tilde{B}_{v,c_i}$ such that $\delta_{c_j} = \tilde{B}_{v,c_j}$ for all $j < i$. Then, the coefficient of ε_{c_i} in $Y_v - \sum_{j<i} \delta_{v,c_j} Y_{c_j}$ is

$$\begin{aligned}\pi_{v,c_i} - \sum_{j<i} \delta_{v,c_j}\pi_{c_j,c_i} &= \pi_{v,c_i} - \sum_{j<i} \tilde{\beta}_{v,c_j}\pi_{c_j,c_i} \\ &= \sum_{l\in\mathcal{L}_{v,c_i}} W(l) - \sum_{j<i}\left[\left(\sum_{l\in\mathcal{L}^{(c_j)}_{v,c_j}(C)} W(l)\right)\left(\sum_{l\in\mathcal{L}_{c_j,c_i}} W(l)\right)\right] \\ &= \sum_{l\in\mathcal{L}_{v,c_i}} W(l) - \sum_{j<i}\left[\sum_{l\in\mathcal{L}^{(c_j)}_{v,c_i}(C)} W(l)\right] \\ &= \sum_{l\in\mathcal{L}^{(c_i)}_{v,c_i}} W(l) = \tilde{B}(C)_{v,c_i}. \end{aligned} \tag{4.12}$$

Since $\delta_{c_i} \neq \tilde{B}_{v,c_i}$ by assumption, then let $\delta_{c_i} = \tilde{B}_{v,c_i} - \alpha$ for $\alpha \neq 0$. Since no other Y_{c_j} for $j > i$ includes any terms of ε_{c_i}, then

$$\gamma_v = \alpha\varepsilon_{c_i} + \eta \qquad \text{and} \qquad \gamma_{c_i} = \varepsilon_{c_i} + \zeta \tag{4.13}$$

where η and ζ do not contain ε_{c_i}. Then,

$$\begin{aligned}\mathbb{E}\left(\gamma_{c_i}^{K-1}\gamma_v\right) &= \mathbb{E}\left(\left[\varepsilon_{c_i} + \zeta\right]^{K-1}\left[\alpha\varepsilon_{c_i} + \eta\right]\right) \\ &= \mathbb{E}\left(\left[\varepsilon_{c_i}^{K-1} + \sum_{k=0}^{K-2}\varepsilon_{c_i}^{k}\zeta^{K-1-k}\right]\left[\alpha\varepsilon_{c_i} + \eta\right]\right) \\ &= \alpha\mathbb{E}\left(\varepsilon_{c_i}^{K}\right) + \mathbb{E}\left(\varepsilon_{c_i}^{K-1}\eta\right) + \mathbb{E}\left(\left[\sum_{k=0}^{K-2}\varepsilon_{c_i}^{k}\zeta^{K-1-k}\right]\left[\alpha\varepsilon_{c_i} + \eta\right]\right).\end{aligned}$$

Note that the last two terms do not involve $\mathbb{E}(\varepsilon_{c_i}^K)$. Since $\alpha \neq 0$, selecting

$$\mathbb{E}\left(\varepsilon_{c_i}^{K}\right) \neq -\frac{\left(\mathbb{E}\left(\varepsilon_{c_i}^{K-1}\eta\right) + \mathbb{E}\left(\left[\sum_{k=0}^{K-2}\varepsilon_{c_i}^{k}\zeta^{K-1-k}\right]\left[\alpha\varepsilon_{c_i} + \eta\right]\right)\right)}{\alpha}$$

ensures that $\mathbb{E}\left(\gamma_{c_i}^{K-1}\gamma_v\right) \neq 0$. □

Note that Lemma 4.1 is true regardless of whether each node c_i is a descendant, sibling, parent or ancestor of v. If $c_i \notin \text{an}(v)$, then the marginal direct effect must be 0, so δ_{c_i} must be zero if $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ for all $c \in C$. Lemma 4.1 implies if we only update $D_{v,C} = \delta_v(C, A, S, D)$ when $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ for all $c \in C$ and set all other elements in $D_{v,C}$ to 0, then $D = \mathbf{D}(B)$ for some $\mathbf{D} \in \mathcal{D}$.

The following lemmas show that when we use $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ for all $c \in C$ to certify if $C \subseteq \text{an}(v) \setminus \text{sib}(v)$, we will not mistakenly certify sets for generic B and error moments. Lemma 4.2 shows that for generic B and error moments if $C \cap \text{sib}(v) \neq \emptyset$, then there exists some $c \in C$, such that $\mathbb{E}(\gamma_c^{K-1}\gamma_v) \neq 0$. Furthermore, Corollary 4.2 shows that we can restrict the sets C which we attempt to certify, such that we avoid testing any set C where $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ for all $c \in C$, but $C \not\subseteq \text{an}(v)$. Finally, Corollary 4.3 shows that by restricting the sets C which we attempt to certify, we will not mistakenly remove a parent or sibling of v.

Lemma 4.2. *Consider $v \in V$, sets A, C such that $C \subseteq A \subseteq V \setminus \{v\}$, and generic B and error moments. Suppose $C \cap \text{sib}(v) \neq \emptyset$, $D = \mathbf{D}(B)$ for any $\mathbf{D} \in \mathcal{D}$, and $S = \Sigma$. Then there exists some $q \in C$ such that $\mathbb{E}\left(\gamma_q^{K-1}\gamma_v\right) \neq 0$.*

Proof. We again appeal to Okamoto (1973), and show that the quantity is non-zero for generic B and the error moments, by first showing that it is a rational function of B and the error moments. Then, constructing a single point at which the quantity of interest is non-zero is sufficient to show that the function vanishes only over a null set with respect to Lebesgue measure.

In particular, select $q \in C$ such that $q \in \text{sib}(v)$. We then represent γ_v as

$$\begin{aligned} \gamma_v &= \varepsilon_v + \sum_{a \in \text{an}(v)} \pi_{va}\varepsilon_a - \sum_{c \in C} \delta_{vc} \sum_{z \in \text{An}(c)} \pi_{cz}\varepsilon_z \\ &= \alpha\varepsilon_q + \eta, \end{aligned} \tag{4.14}$$

where

$$\alpha = \pi_{vq} + \sum_{c \in C} \delta_{vc} \pi_{cq}$$
$$\eta = (1 - \sum_{c \in C} \delta_{vc} \pi_{cv}) \epsilon_v + \sum_{a \in \text{an}(v) \setminus q} \pi_{va} \varepsilon_a - \sum_{c \in C} \delta_{vc} \sum_{z \in \text{An}(c) \setminus q} \pi_{cz} \varepsilon_z$$

and δ_{vc} is the c-th element of δ_v from (4.1). Similarly, we represent γ_q

$$\begin{aligned} \gamma_q &= \varepsilon_q + \sum_{a \in \text{an}(q)} \pi_{va} \varepsilon_a - \sum_{s \in \text{pa}_D(q)} d_{qs} \sum_{t \in \text{An}(s)} \pi_{st} \varepsilon_t \\ &= \varepsilon_q + \zeta \end{aligned} \tag{4.15}$$

where ζ does not involve ε_q. The coefficient on ε_q is 1 since we assume $d_{qs} \neq 0$ only if $s \in \text{an}(q)$. Note that for fixed $\mathbf{D}$, α is a rational function of B and Ω because both Π and δ only involve matrix inversions and multiplications. We now show that for some point B and Ω, $\alpha \neq 0$. In particular, let $B = 0$ and $\omega_{qv} \neq 0$, but $\omega_{ij} = 0$ for all other $i \neq j$. At this point, $\pi_{vq} = \pi_{cq} = 0$ for all $c \in C \setminus q$ so that

$$\alpha = \delta_{vq}. \tag{4.16}$$

Also since $B = 0$ implies $D = 0$, then $S_{C,C} = \Omega_{C,C}$, which is diagonal by construction, and $S_{C \setminus q, v} = 0$ since all path weights for treks among nodes in C and between $C \setminus \{q\}$ and v are zero. However, there is a single trek between q and v, namely the bidirected edge, so $S_{qv} = \omega_{qv}$. Then,

$$\alpha = \delta_{vC} = [S_{C,C}]^{-1} S_{C,v} = \frac{\omega_{qv}}{\omega_{qq}} \neq 0. \tag{4.17}$$

Thus, for generic choice of B and Ω, $\alpha \neq 0$. Now, we finally examine the quantity of interest, which is a rational function of the error moments and B, and play the same game as before, where

$$\begin{aligned}\mathbb{E}\left(\gamma_q^{K-1}\gamma_v\right) &= \mathbb{E}\left(\left[\varepsilon_q+\zeta\right]^{K-1}\left[\alpha\varepsilon_q+\eta\right]\right)\\ &= \mathbb{E}\left(\left[\varepsilon_q^{K-1}+\sum_{k=0}^{K-2}\varepsilon_q^k\zeta^{K-1-k}\right]\left[\alpha\varepsilon_q+\eta\right]\right)\\ &= \alpha\mathbb{E}\left(\varepsilon_q^K\right)+\mathbb{E}\left(\varepsilon_q^{K-1}\eta\right)+\mathbb{E}\left(\left[\sum_{k=0}^{K-2}\varepsilon_q^k\zeta^{K-1-k}\right]\left[\alpha\varepsilon_q+\eta\right]\right).\end{aligned}$$

Note that the last two terms do not involve $\mathbb{E}(\varepsilon_q^K)$. So when $\alpha \neq 0$, selecting

$$\mathbb{E}\left(\varepsilon_q^K\right) \neq -\frac{\left(\mathbb{E}\left(\varepsilon_q^{K-1}\eta\right)+\mathbb{E}\left(\left[\sum_{k=0}^{K-2}\varepsilon_q^k\zeta^{K-1-k}\right]\left[\alpha\varepsilon_q+\eta\right]\right)\right)}{\alpha} \tag{4.18}$$

ensures that $\mathbb{E}\left(\gamma_q^{K-1}\gamma_v\right) \neq 0$. Thus, there exists some point such that $\mathbb{E}\left(\gamma_q^{K-1}\gamma_v\right) \neq 0$. This implies there is a null set of B and error moments which we must avoid for each $\mathbf{D} \in \mathcal{D}$, but since $|\mathcal{D}|$ is finite, then the union of these null sets is again a null set. □

In practice, when using $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \mathrm{an}_D(C), S, D)) = 0$ to certify ancestral relationships, we may miscertify (as an ancestor) a variable which is neither a descendant or an ancestor of v or we may miscertify a descendant s if we have removed the effect of v on s by adjusting for descendants of v which are ancestors of s, but have not yet otherwise identified that s is a descendant of v.

However, Corollary 4.2 together with Lemma 4.1 imply that if $C \not\subseteq \mathrm{an}(v)$, then either $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \mathrm{an}_D(C), S, D)) \neq 0$ for some $c \in C$, or if $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \mathrm{an}_D(C), S, D)) = 0$ for all $c \in C$, we can pre-screen the set to avoid miscertification. In particular, let $C = C_1 \cup C_2$, where we are certain that $C_1 \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$ and we are interested in testing whether $C_2 \subseteq \mathrm{an}(v)\setminus\mathrm{sib}(v)$. If $C_2 \not\subseteq \mathrm{an}(v)\setminus\mathrm{sib}(v)$, but $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \mathrm{an}_D(C), S, D)) = 0$– which would result in a miscertification–then for $C' = C_1 \cup \{C_2 \cap \mathrm{an}(v)\}$, we could have certified that $C' \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$ because $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C', \mathrm{an}_D(C'), S, D)) = 0$ for all $c \in C'$.

Furthermore, $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C'), \mathrm{an}_D(C'), S, D)) = 0$ for all $c \in C_2 \setminus \mathrm{an}(v)$. This implies we can pre-screen the non-ancestors of v, which we would have otherwise miscertified, as long as we have already tried to certify all subsets of C_2. This is implemented in line 11 of Algorithm 2.

Corollary 4.2. *Consider $v \in V$ and set $C \subseteq V \setminus \{v\}$. Let $D \in \mathbb{R}^{p\times p}$ such that $D_{ij} \neq 0$ only if $j \in \mathrm{an}(i)$. Suppose $C \not\subseteq \mathrm{an}(v)$, but $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \mathrm{an}_D(C), S, D)) = 0$ for all $c \in C$ for generic B and error moments. Then for $C_1 = C \cap \mathrm{an}(v) \setminus \mathrm{sib}(v)$*

$$\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C_1, \mathrm{an}_D(C_1), S, D)) = 0$$

for all $c \in C$.

Proof. For convenience, let $A = \mathrm{an}_D(C)$, $A_1 = \mathrm{an}_D(C_1)$, $A_2 = A \setminus A_1$, and $\Lambda = I - D$ and $\Pi = (I - D)^{-1}$. This implies $D_{A_1,A_2} = 0$ and $[(I - D_{A,A})^{-1}]_{A_1,A_2} = 0$. So that

$$\begin{aligned}
&(I - D)_{C_1,A} S_{A,A} (I - D_{A,A})^T ((I - D_{A,A})^{-T}_{A,C_1}) \\
&\qquad = \begin{bmatrix} \Lambda_{C_1,A_1} & \Lambda_{C_1,A_2} \end{bmatrix} \begin{bmatrix} S_{A_1,A_1} & S_{A_1,A_2} \\ S_{A_2,A_1} & S_{A_2,A_2} \end{bmatrix} \begin{bmatrix} (\Lambda_{A_1,A_1})^T & (\Lambda_{A_2,A_1})^T \\ (\Lambda_{A_1,A_2})^T & (\Lambda_{A_2,A_2})^T \end{bmatrix} \begin{bmatrix} (\Pi_{C_1,A_1})^T \\ (\Pi_{C_1,A_2})^T \end{bmatrix} \\
&\qquad = \begin{bmatrix} \Lambda_{C_1,A_1} & 0 \end{bmatrix} \begin{bmatrix} S_{A_1,A_1} & S_{A_1,A_2} \\ S_{A_2,A_1} & S_{A_2,A_2} \end{bmatrix} \begin{bmatrix} (\Lambda_{A_1,A_1})^T & (\Lambda_{A_2,A_1})^T \\ 0 & (\Lambda_{A_2,A_2})^T \end{bmatrix} \begin{bmatrix} (\Pi_{C_1,A_1})^T \\ 0 \end{bmatrix} \\
&\qquad = \Lambda_{C_1,A_1} S_{A_1,A_1} (\Lambda_{A_1,A_1})^T (\Pi_{C_1,A_1})^T ,
\end{aligned}$$

and

$$(I - D)_{C_1,A}\Sigma_{A,v} = \begin{bmatrix} \Lambda_{C_1,A_1} & \Lambda_{C_1,A_2} \end{bmatrix} \begin{bmatrix} \Sigma_{A_1,v} \\ \Sigma_{A_2,v} \end{bmatrix} = \begin{bmatrix} \Lambda_{C_1,A_1} & 0 \end{bmatrix} \begin{bmatrix} \Sigma_{A_1,v} \\ \Sigma_{A_2,v} \end{bmatrix} = (I - D)_{C_1,A_1}\Sigma_{A_1,v}.$$

Thus,

$$\begin{aligned}\delta_v(C_1, A, S, D) &= \left[(I-D)_{C_1,A} S_{A,A} (I-D_{A,A})^T ((I-D_{A,A})^{-T}_{A,C})\right]^{-1} (I-D)_{C_1,A} \Sigma_{A,v} \\ &= \left[(I-D)_{C_1,A_1} S_{A_1,A_1} (I-D_{A_1,A_1})^T ((I-D_{A_1,A_1})^{-T}_{A_1,C_1})\right]^{-1} (I-D)_{C_1,A_1} \Sigma_{A_1,v} \\ &= \delta_v(C_1, A_1, S, D).\end{aligned}$$

By Lemma 4.1, for generic B and error moments, if $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \mathrm{an}_D(C), S, D)) = 0$, then for every $q \notin C_1$, $\delta_{vq}(C, A, S, D) = 0$.

$$\begin{aligned}\gamma_v(C, A, S, D) &= Y_v - Y_C \delta_v(C, A, S, D) \\ &= Y_v - Y_{C_1}(C_1, A_1, S, D) \\ &= \gamma_v(C_1, A_1, S, D).\end{aligned}$$

So if for all $c \in C$,

$$\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(C, A, S, D)\right) = 0, \tag{4.19}$$

then for all $c \in C$

$$\mathbb{E}\left(\gamma_c(D)^{K-1}\gamma_v(C_1, A_1, S, D)\right) = 0. \tag{4.20}$$

□

We now show that for generic B and error moments, that this pre-screening procedure will not mistakenly remove a parent (which has not already been identified) or sibling of v.

Corollary 4.3. *Consider generic B and error moments. Suppose $D = \mathbf{D}(B)$ and for some $v \in V$, $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(\mathrm{pa}_D(v), \mathrm{an}_D(v), S, D)) = 0$ for all $c \in \mathrm{pa}_D(v)$. If $q \in \{\mathrm{pa}(v) \backslash \mathrm{pa}_D(v)\} \cup \mathrm{sib}(v)$, then $\mathbb{E}\left(\gamma_q(D)^{K-1}\gamma_v(D)\right) \neq 0$.*

Proof. First consider $q \in \mathrm{pa}(v) \setminus \mathrm{pa}_D(v)$. $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \mathrm{an}_D(C), S, D)) = 0$ for all $c \in$

$\text{pa}_D(v)$ implies that

$$\begin{aligned}\gamma_v(D) &= Y_v - Y_{\text{pa}_D(v)}(D_{v,\text{pa}_D(v)})^T \\ &= \left(\pi_{v,q} - \sum_{c\in C} \tilde{B}(\{C,v\})_{v,c}\pi_{c,q}\right)\epsilon_q + \eta \\ &= \left(\pi_{v,q} - \sum_{c\in C\cap \text{de}(q)} \tilde{B}(\{C,v\})_{v,c}\pi_{c,q}\right)\epsilon_q + \eta \\ &= \alpha\epsilon_q + \eta\end{aligned} \tag{4.21}$$

where η does not involve ϵ_q. For any $c \in \text{de}(q)$, $\tilde{B}(\{C,v,q\})_{v,C} = \tilde{B}(\{C,v\})_{v,C}$ because there are no paths from c to v which pass through q, so marginalizing q does not change the marginal direct effect. Thus, as shown in Lemma 4.1,

$$\begin{aligned}\alpha &= \pi_{q,v} - \sum_{c\in C\cap \text{de}(q)} \tilde{B}(\{C,q,v\})_{v,C}\pi_{c,q} \\ &= \tilde{B}(\{C,q,v\})_{v,q}.\end{aligned} \tag{4.22}$$

The points, B such that $q \in \text{pa}(v)$, but the marginal direct effect $\tilde{B}(\{C,q,v\})_{vq} = 0$ have Lebesgue measure 0, so by the same argument as Lemma 4.1 when $\alpha \neq 0$, for generic error moments, $\mathbb{E}(\gamma_q^{K-1}\gamma_v) \neq 0$.

Now consider $q \in \text{sib}(v)$. Since $\text{pa}_D(v) \subseteq \text{an}(v)$ for all $v \in V$, then $\gamma_v = \epsilon_v + \eta$ where η does not involve ε_v and $\gamma_q = \epsilon_q + \zeta$ where ζ does not involve ϵ_q. Then, using the same argument as the previous lemmas, selecting

$$\mathbb{E}(\epsilon_q^{K-1}\epsilon_v) \neq -\mathbb{E}\left(\sum_{t=0}^{K-2}\binom{K-1}{t}\varepsilon_q^t\zeta^{K-1-t}(\epsilon_v+\eta) + \epsilon_q^{K-1}\eta\right) \tag{4.23}$$

ensures that $\mathbb{E}(\gamma_q^{K-1}\gamma_v) \neq 0$ □

So far we have been concerned with discovering the ancestors of some node v. The final corollary shows that when we have identified a superset of the parents of v, we can prune

away ancestors which are not parents. The corollary motivates the pruning procedure which is used in line 38 of Algorithm 2.

Corollary 4.4. *Consider v and generic B and error moments and let $D = B$. Suppose $\mathrm{pa}(v) \subseteq C \subseteq \mathrm{an}(v) \setminus \mathrm{sib}(v)$ and $\mathbb{E}(\gamma_c(D)^{K-1}\gamma_v(C, \mathrm{an}_D(C), S, D) = 0$ for all $c \in C$. Then for any $q \in C \setminus \mathrm{pa}(v)$, $\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C \setminus \{q\}, \mathrm{an}_D(C \setminus \{q\}), S, D)) = 0$, but for any $q \in \mathrm{pa}(v)$, $\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C \setminus \{q\}, \mathrm{an}_D(C \setminus \{q\}), S, D)) \neq 0$.*

Proof. Lemma 4.1 implies for any $q \in C \setminus \mathrm{pa}(v)$,

$$\delta_v(C, \mathrm{an}_D(C), S, D) = B_{v,(C\setminus\{q\},q)} = \begin{bmatrix} B_{v,(C\setminus\{q\})} & 0 \end{bmatrix} = \begin{bmatrix} \delta_v(C \setminus \{q\}, \mathrm{an}_D(C \setminus \{q\}), S, D) & 0 \end{bmatrix}$$

so that

$$\mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C \setminus \{q\}, \mathrm{an}_D(C \setminus \{q\}), S, D)) = \mathbb{E}(\gamma_q(D)^{K-1}\gamma_v(C, \mathrm{an}_D(C), S, D)) = 0.$$

The second statement follows directly from Corollary 4.3. □

4.3 Graph estimation algorithm

Using the claims established above, we present in Algorithm 2 the **B**ow-free **A**cyclic **n**on-**G**aussian (BANG) algorithm which will consistently discover the underlying causal structural of the linear structural equation model when it corresponds to a BAP with non-Gaussian errors if the sample moments of Y consistently estimate the population moments of Y.

Roughly speaking, the algorithm starts with a fully connected bidirected graph and, for each variable, iteratively certifies ancestors which are not siblings by considering progressively larger sets. When a set is certified as containing ancestors but not siblings, it is added to $\widehat{\mathrm{pa}}$ and removed from $\widehat{\mathrm{sib}}$ and D is updated. This procedure is repeated until no additional ancestral relationships can be certified, and any remaining dependency which cannot be accounted for via linear adjustments is assumed to be from correlated errors. In the algorithm, whenever we test for $x \perp\!\!\!\perp y$, in practice, we mean testing $\mathbb{E}(x^{K-1}y) = 0$.

Algorithm 2 BANG procedure

$S = \frac{1}{n} Y^T Y$
$l = 1$
$\widehat{\text{pa}}(v) = \emptyset \qquad \forall v$
$\widehat{\text{sib}}(v) = V \setminus \{v\} \qquad \forall v$
$D_{uv} = 0 \qquad \forall u, v$
$\gamma = Y$

Find Ancestral Relationships
while $\max |\widehat{\text{sib}}(v)| \geq l$ **do**

  Cycle through all variables
  for $v \in [p]$ **do**

    Check for independencies
    for $u \in \widehat{\text{sib}}(v)$ **do**
      if $\gamma_u \perp\!\!\!\perp \gamma_v$ **then**
        $\widehat{\text{sib}}(v) = \widehat{\text{sib}}(v) \setminus \{u\}$
        $\widehat{\text{sib}}(u) = \widehat{\text{sib}}(u) \setminus \{v\}$
      end if
    end for

    Certify pseudo-parents
    $C^\star = \emptyset$
    for $C \in \binom{\widehat{\text{sib}}(v)}{l}$ **do**
      if $\gamma_C \perp\!\!\!\perp \gamma_v(C \cup \widehat{\text{pa}}(v), \text{an}_D(C \cup \widehat{\text{pa}}(v)), S, D)$ **then**
        $C^\star = C^\star \cup C$
      end if
    end for
    $\widehat{\text{pa}}(v) = \widehat{\text{pa}}(v) \cup C^\star$
    $D_{v,\widehat{\text{pa}}(v)} = \delta_v(\widehat{\text{pa}}(v), \text{an}_D(\widehat{\text{pa}}(v)), S, D)$
    $\widehat{\text{sib}}(v) = \widehat{\text{sib}}(v) \setminus \widehat{\text{pa}}(v)$
    $\widehat{\text{sib}}(s) = \widehat{\text{sib}}(s) \setminus \{v\} \qquad \forall s \in \widehat{\text{pa}}(v)$
    $\gamma_v = Y_v - Y(D_{v,V})^T$
  end for

  If no updates, consider larger sets
  if D was updated **then**
    $l = 1$
  else
    $l = l + 1$
  end if
end while

Pruning Phase
Form topological ordering σ
for $v \in \sigma^{-1}([p])$ **do**
  for $s \in \widehat{\text{pa}}(v)$ **do**
    if $\gamma_s \perp \gamma_v(\widehat{\text{pa}(v)} \setminus \{s\}, \text{an}_D(\widehat{\text{pa}}(v) \setminus \{s\}), S, D)$ **then**
      $\widehat{\text{pa}}(v) = \widehat{\text{pa}}(v) \setminus \{s\}$
      $D_{v,s} = 0$
    end if
  end for
end for

Return
Let $\hat{E}_{\rightarrow} = \{(u, v) : u \in \widehat{\text{pa}}(v)\}$
Let $\hat{E}_{\leftrightarrow} = \{\{u, v\} : u \in \widehat{\text{sib}}(v)\}$
Return $\hat{G} = \{V, \hat{E}_{\rightarrow}, \hat{E}_{\leftrightarrow}\}$; $\hat{B} = D$; $\hat{\Omega} = \text{var}(\gamma)$

4.3.1 Graph identification

Theorem 4.2. *Suppose Y is generated by a linear SEM which corresponds to some bow-free acyclic path diagram $G = \{V, E_{\rightarrow}, E_{\leftrightarrow}\}$. Then for generic choices of B and error moments, when given population level moments of Y, Algorithm 2 will output $\hat{G} = G$.*

Proof. First, we note that the statements from Lemmas 4.1 and 4.2 and Corollaries 4.3 and 4.4 about certain quantities being non-zero for generic B and error moments only explicitly pertained to a single quantity. However, since we only only consider a finite set of these quantities, the union of the null sets to be avoided is also a null set. Thus, for the remainder of the proof, we assume that quantities which are generically non-zero are actually non-zero.

Assume that σ is some topological ordering of the directed portion of underlying graph $G = \{V, E_{\rightarrow}, E_{\leftrightarrow}\}$. Let a single step z be an entire iteration through all $v \in V$ for which all relevant sets of up to size $z-1$ are tested. We will now show by induction that the algorithm recovers the correct graph.

As the induction hypothesis, suppose for $z \in [p]$, $v = \sigma^{-1}(z)$, (1) $D_{A,A} = B_{A,A}$ for $A = \sigma^{-1}([z-1])$, $\widehat{\text{pa}}(a) \supseteq \text{pa}(a)$ for all $a \in A$; (2) $D = \mathbf{D}(B)$ for some $\mathbf{D} \in \mathcal{D}$; and (3) $\widehat{\text{sib}}(j) \supseteq \text{sib}(j)$, $\text{pa}(j) \subseteq \{\widehat{\text{sib}}(j) \cup \widehat{pa}(j)\}$ and $\widehat{\text{pa}}(j) \subseteq \{\text{an}(j) \setminus \text{sib}(j)\}$ for all $j \in V$.

By assumption $D_{A,A} = B_{A,A}$ and $D = \mathbf{D}(B)$ for some $\mathbf{D} \in \mathcal{D}$, so $\text{an}_D(\text{pa}(v)) \subseteq \text{an}(\text{pa}(v))$. By Corollary 4.2, for all $s \in V$ the pruning procedure on Line 11 removes any $q \notin \text{an}(s)$ which may have otherwise been mistakenly certified into $\widehat{\text{pa}}(s)$. However, Corollary 4.3 implies that this procedure does not remove any siblings of s or remaining parents of s which have not yet been certified in $\widehat{\text{pa}}(s)$, so $\text{pa}(s) \subseteq \{\widehat{\text{sib}}(s) \cup \widehat{\text{pa}}(s)\}$ and $\widehat{\text{sib}}(s) \supseteq \text{sib}(s)$.

Together with Lemma 4.2, this ensures that for all $s \in V$ no set C will be certified such that $C \not\subseteq \text{an}(s) \setminus \text{sib}(s)$. Thus, $\widehat{\text{pa}}(s) \subseteq \text{an}(s) \setminus \text{sib}(s)$ is preserved for all $s \in V$ and any updates to D continue to satisfy $D = \mathbf{D}(B)$.

By the acyclic assumption, $|\text{pa}(v)| \leq z-1$ and since we test all sets C of size less than z so we will eventually consider some C such that $\text{pa}(v) \subseteq C \subseteq \text{an}(v) \setminus \text{sib}(v)$. By Lemma 4.1 and Corollary 4.1, $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ for all $c \in C$ so that C will be certified into $\widehat{\text{pa}}(v)$. In

addition, the resulting update which sets $D_{v,\widehat{\text{pa}}(v)} = \delta_{v,\widehat{\text{pa}}(v)}$ will result in $D_{v,V} = B_{v,V}$. This implies $D_{\{A,v\},\{A,v\}} = B_{\{A,v\},\{A,v\}}$.

Thus, for $z+1$, $v' = \sigma^{-1}(z+1)$ and $A' = A \cup \{v\}$, the induction condition is still satisfied. That is, (1) $D_{A',A'} = B_{A',A'}$ and $\widehat{\text{pa}}(a) \supseteq \text{pa}(a)$ for all $a \in A'$; (2) $D = \mathbf{D}(B)$ for some $\mathbf{D} \in \mathcal{D}$; and (3) $\widehat{\text{sib}}(j) \supseteq \text{sib}(j)$, $\text{pa}(j) \subseteq \{\widehat{\text{sib}}(j) \cup \widehat{pa}(j)\}$ and $\widehat{\text{pa}}(j) \subseteq \{\text{an}(j) \setminus \text{sib}(j)\}$ for all $j \in V$.

The base case is trivially satisfied since for step 1, $v = \sigma^{-1}(1)$ has no ancestors so $D = 0$ trivially $D_{A,A} = B_{A,A}$. Furthermore, we initialize $\widehat{\text{sib}}(s) = V \setminus s$, and $\widehat{\text{pa}}(s) = \emptyset$ so the induction conditions are satisfied.

Thus, after p steps, $D = B$, so $\mathbb{E}(\gamma_i(D)^{K-1}\gamma_j(D)) \neq 0$ if and only if $j \in \text{sib}(i)$ so $\widehat{\text{sib}}(v) = \text{sib}(v)$ for all $v \in V$. Furthermore, $\text{pa}(v) \subseteq \widehat{\text{pa}}(v) \subseteq \text{an}(v) \setminus \text{sib}(v)$. By Corollary 4.4, the pruning phase removes any ancestors from $\widehat{\text{pa}}(v)$ which are not parents, but does not remove any parents. So that the final $\widehat{\text{pa}}(v) = \text{pa}(v)$. □

Since all the tests involve rational functions of the moments of Y, it immediately follows that the method is consistent if the sample moments of Y are consistent for the population moments of Y.

Corollary 4.5. *Suppose $Y_1, \ldots, Y_n$ is generated by a linear structural equation model which corresponds to some BAP $G = \{V, E_{\rightarrow}, E_{\leftrightarrow}\}$. Then, for generic choices of B and error moments, there exists some $\eta > 0$ such that if the sample moments are within an η-ball of the population moments Y, Algorithm 2 will output $\hat{G} = G$ when the independence tests are appropriately tuned.*

Proof. First, we note that when $D = \mathcal{D}(B)$ for some $\mathcal{D}$, the maps which take moments of Y to $E(\gamma_c(D)^{K-1}\gamma_v(C, A, S, D))$ are rational functions and are thus Lipschitz within a bounded domain around the population moments of Y. So, when using the sample moments of Y to form estimates of $\widehat{E}(\gamma_c(D)^{K-1}\gamma_v(C, A, S, D)$, for any $\zeta > 0$, there exists some $\eta > 0$ such that if the sample moments are within a η-ball of the population quantities, the estimates of $\mathbb{E}(\gamma_c(D)^{k-1}\gamma_v(C, A, S, D)$ are within $\zeta/2$ of the true quantities.

For generic B and error moments, let $\zeta = \min_{\mathbb{E}(\gamma_c^{K-1}\gamma_v)\neq 0} |\mathbb{E}(\gamma_c^{k-1}\gamma_v)|$. When all estimates of the tested quantities are within $\zeta/2$ of the population quantities, all estimates which correspond to quantities which are 0 are less than $\zeta/2$ in absolute value, and all estimates which correspond to quantities which are generically non-zero are greater than $\zeta/2$ in absolute value. Thus, we could construct a correct test of $\mathbb{E}(\gamma_c^{k-1}\gamma_v) = 0$ vs $\mathbb{E}(\gamma_c^{K-1}\gamma_v) \neq 0$ by simply using a cut-off of $\zeta/2$ for the sample moments. Thus, we will always certify or reject the correct sets and by Theorem 4.2, $\hat{G} = G$. □

4.3.2 Practical concerns

For any BAP, the results stated hold for all but a null set of B and error moments. In particular, as shown in Corollary 4.3, any point where the direct marginal effect of q on v vanishes for some $q \in pa(v)$ lies in the null set. This trivially includes any points where $B_{vq} = 0$ for some $q \in \text{pa}(v)$. We also know that error moments which correspond to a Gaussian distribution ought to be avoided. In general, for finite samples, accurate estimation would require strong faithfulness assumptions similar to the conditions required in Chapter 3 such that the errors are sufficiently non-Gaussian and the linear coefficients are bounded away from 0. For BAPs, we also need the error covariances to be bounded from 0 and the higher order cross moments to be sufficiently non-Gaussian.

Any consistent test for whether $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ could be used with the algorithm. We use empirical likelihood to jointly test whether $\mathbb{E}(\gamma_c^{K-1}\gamma_v) = 0$ for all $c \in C$. Empirical likelihood is useful in this context because it does not require explicit estimation of the variances of $\gamma_c^{K-1}\gamma_v$ in order to form a well-calibrated test statistic and the empirical likelihood ratio statistic converges to a known reference distribution under very mild conditions. In addition, pooling together all the tests into one omnibus test helps limit multiple testing. Unfortunately, the computational burden of empirical likelihood is highly dependent on the sample size.

In practice, we also need to select some K for the testing procedure. This should correspond to a moment of the errors which is not consistent with the Gaussian distribution.

In theory, one could test $\mathbb{E}(\gamma_c^{k-1}\gamma_v) = 0$ for all $k = \{3, \ldots, K\}$ for some arbitrarily large K when given infinite data; in practice, letting $K = 3, 4$ should suffice.

Finally we note that when given an oracle for conducting independence tests, if the in-degree of each node (counting both directed and bidirected edges) is bounded by some constant J, then the number of tests required is bounded by a polynomial of the number of variables. Without loss of generality assume that $1, \ldots, p$ is a topological ordering of the nodes. For any $z \in [p]$, assuming that we have discovered the parents of nodes $[z-1]$, then testing all subsets up to size J of possible parents is sufficient for discovering the parents of z. So D will be updated and l, the counter for set size, will be reset so that l will never exceed J. Thus, for each of the p nodes, in between updates to D, there will be at most $\sum_{k=0}^{J} \binom{p}{k} \leq p^{J+1}$ independence tests. We include $k = 0$ to count the independence screening procedure before any of the regressions are performed. By the acyclic assumption, there are at most $p(p-1)/2$ ancestral relationships to discover, which would cause an update to D. Once D is fully updated so that $D = B$, then $\widehat{\text{sib}}(v) = \text{sib}(v)$ for all $v \in V$. This implies there will be at most an additional round of p^{J+1} tests for each v before the counter surpasses $\max_v |\widehat{\text{sib}}(v)|$ and the algorithm terminates. Then, there are at most $p(p-1)/2$ discovered ancestral relationships which much be checked and pruned. Thus, there are $O(p \times p^{J+1} \times p^2)$ total independence tests.

4.4 Numerical results

We compare BANG to two existing methods for Gaussian data–FCI+ (Claassen et al., 2013) with Gaussian conditional independence tests and Greedy BAP Search (GBS) (Nowzohour et al., 2017). For FCI+, we use the implementation in the `R` package `pcalg` (Kalisch et al., 2012) and for GBS we use the implementation in the `R` package `greedyBaps` (Nowzohour, 2017). We consider structure learning in two settings: ancestral graphs and BAPs. Finally, we show that when applied to ecology data with a ground truth model, the BANG method performs well.

4.4.1 Comparison with FCI+ and Greedy BAP search

In the first setting, we generate random ancestral graphs with $p = 5$ by selecting 4 directed edges uniformly from the set $\{(i, j) : i < j\}$ and then selecting 2 bidirected edges from the set $\{\{i, j\} : i \notin \text{an}(j) \text{ and } j \notin \text{an}(i)\}$. Note that these graphs are not necessarily maximal. We then draw the directed edges uniformly from (.6, 1). For the errors, we first draw $\delta_i \sim N(0, \Delta)$ where $\Delta_{vv} = 1$ and $\Delta_{uv} = .2$ if $(u, v) \in E_{\leftrightarrow}$. We then let $\varepsilon_i = \exp(\delta_i) - \sqrt{e}$ so ε_i are centered log-normal variables with variances $e \times (e - 1) = 4.67$ and covariances $e \times (e^{\cdot 2} - 1) = 0.60$. We let $Y_i = (I - B)^{-1}\varepsilon_i$ and randomly permute the ordering of the columns of Y. We repeat this for 1000 random realizations of a graph and data.

Although there is no MAG restriction on the graphs returned by BANG and GBS, for a direct comparison, we take the BAP estimated from BANG and GBS and project it to a PAG by using FCI+ with population quantities. For GBS, we allow 100 random restarts, the same number used in the simulations by (Nowzohour, 2017). The proportion of times BANG, GBS, and FCI+ identify the correct PAG given non-Gaussian data is shown in Figure 4.4 with the solid lines. We also include the proportion of times that BANG identifies the exact graph for comparison. We see that when the nominal level for the independence tests is properly tuned, BANG outperforms existing methods in identifying the correct PAG, and is even able to get the exact graph a higher proportion of the times than the other methods identify the correct PAG. Note that because GBS has no user specified nominal test level, the lines in each of the panels are similar.

For further comparison, we also simulate 1000 random realizations with Gaussian data using the same procedure to generate edge weights and error covariances. The proportion of times which FCI+ and GBS identify the true PAG for the Gaussian data is shown with dotted lines in Figure 4.4. We note that there does not seem to be a substantial difference for either method when the data are Gaussian or non-Gaussian.

When considering BAPs, we consider three settings and compare the performance of BANG and GBS. First, we generate random BAPs with 4 directed edges and 2 bidirected

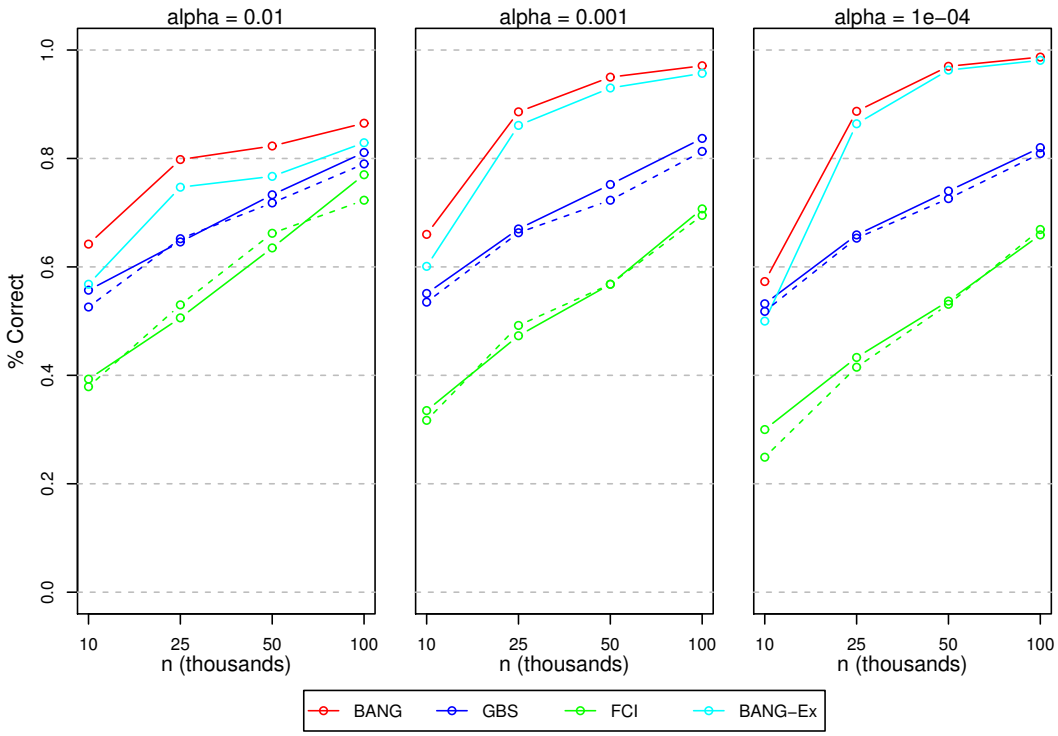


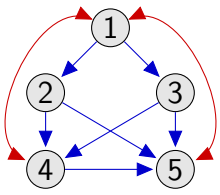
Figure 4.4: Proportion of times that BANG, GBS and FCI+ discover the correct PAG (solid lines are for non-Gaussian errors; dotted lines show FCI+ and GBS performance for Gaussian errors). We also include the proportion of times that BANG discovers the exact ancestral graph. The panels show the results when using a nominal test level of $\alpha = .01, .001, .0001$. GBS has no nominal test level parameter, but we still show 1000 different simulations in each panel.

edges with the same procedure for the ancestral graphs except we randomly draw bidirected edges from the set $\{\{i, j\} : i \notin \text{pa}(j) \text{ and } j \notin \text{pa}(i)\}$ which enforces bow-freeness but not the ancestral condition. Second, we consider denser graphs with 8 total edges by selecting the number of directed edges, d uniformly from $\{3, \ldots, 7\}$ and then let the number of bidirected edges be $8-d$. Finally, we also consider a setting where we fix the graph to the two adversarial BAPs shown in Figure 4.5, but draw random edgeweights and data. Again, we allow GBS to use 100 random restarts. Because GBS can only identify graphs up to an equivalence class, we consider the estimated BAP correct if it is in the empirical Markov equivalence class of the

true BAP; following Nowzohour (2017), since there is no known graphical characterization for the equivalence class of BAPs, we say that the estimated graph is correct if the score of the estimated structure is within 10^{-10} of the score of the true BAP structure. We hold BANG to a a higher standard and only consider the estimate correct if it identifies the exact BAP.

The results of 1000 random realizations for the random BAPs are shown in Figure 4.6. The results of the non-Gaussian realizations are shown with solid lines and the results for GBS with Gaussian data are shown with the dotted lines. Again, we see in the 6 edge and 8 edge case, when properly tuned, BANG, with non-Gaussian data, outperforms GBS when the data are Gaussian or non-Gaussian. In almost every case where GBS does not identify the correct equivalence class, the estimated graph has a higher score than the true graph; this indicates that the greedy procedure is not the main shortcoming, but that the true structure is not actually a global maximum. Finally, we see that GBS tends to perform worse when the graph is more dense (8 edges vs 6). This phenomenon is further accentuated in results of the two adversarial graphs shown in Figure 4.7 which have 9 and 10 edges. Here we see that GBS almost never identifies the correct equivalence class while BANG still performs well given a large sample size. We posit that this occurs because GBS uses a BIC type score, but doubles the penalty for the number of edges. Thus, in finite samples, this tends to favor sparser graphs.

(a) House Graph



(b) Complete Graph

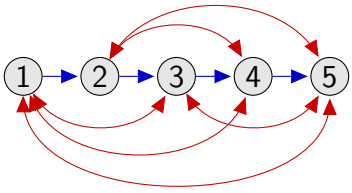


Figure 4.5: Adversarial graphs used in simulations

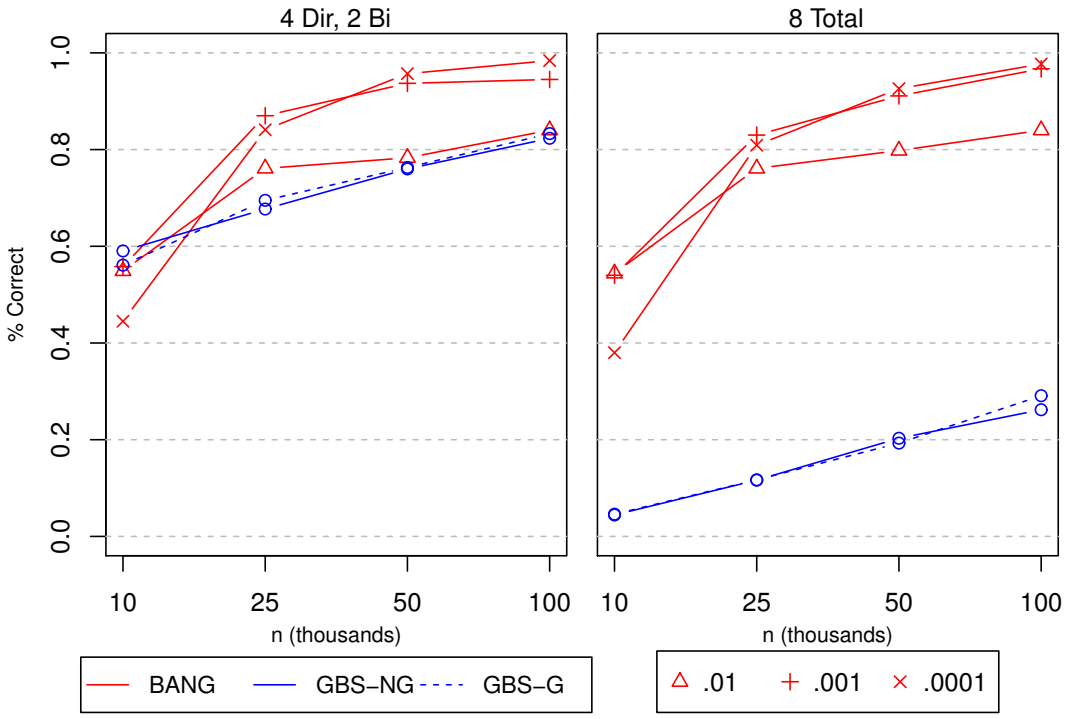


Figure 4.6: Proportion of times GBS discovers a BAP in the empirical Markov equivalence class and BANG discovers the exactly correct graph. BANG is shown in red and plot symbols indicate nominal test level. GBS is shown in blue and the dotted line indicates performance on Gaussian data.

4.4.2 Data example

Grace et al. (2016) use a structural equation model to examine the relationships between land productivity and the richness of plant diversity. They consider measurements taken at 39 different sites and 1126 specific plots, which are locations within the 39 sites. They fit two models: (1) a site level model ($n = 39$) which only includes the site level measures (shown in light gray in Figure 4.8), and (2) a plot level model which includes the plot specific measurements as well as any posited parents of the plot specific variables. We only consider the variables that they include in the plot level model, and omit the variables which were only used in the site level model. We then "project" the causal model considered by Grace et al.

Figure 4.7: Proportion of times GBS discovers a BAP in the empirical Markov equivalence class and BANG discovers the exactly correct graph. BANG is shown in red and plot symbols indicate nominal test level. GBS is shown in blue and the dotted line indicates performance on Gaussian data.

(2016) into a BAP through the procedure described below. When marginalizing variables, we roughly follow the procedure of Koster (2002); however, because we want to arrive at a bow-free graph, some slight modifications are required.

We first remove any edges which they had posited, but found were not significant (denoted by NS in Figure 4.8). Note that this removes the cycle in the plot specific measurements, but there is still a cycle between site productivity, biomass and richness. The nodes for climate, disturbance and suitability, actually represent multiple variables which are used in the SEMs. For climate and disturbance, the separate measures are both highly correlated, so it seems reasonable to use bidirected edges between site productivity, biomass and richness

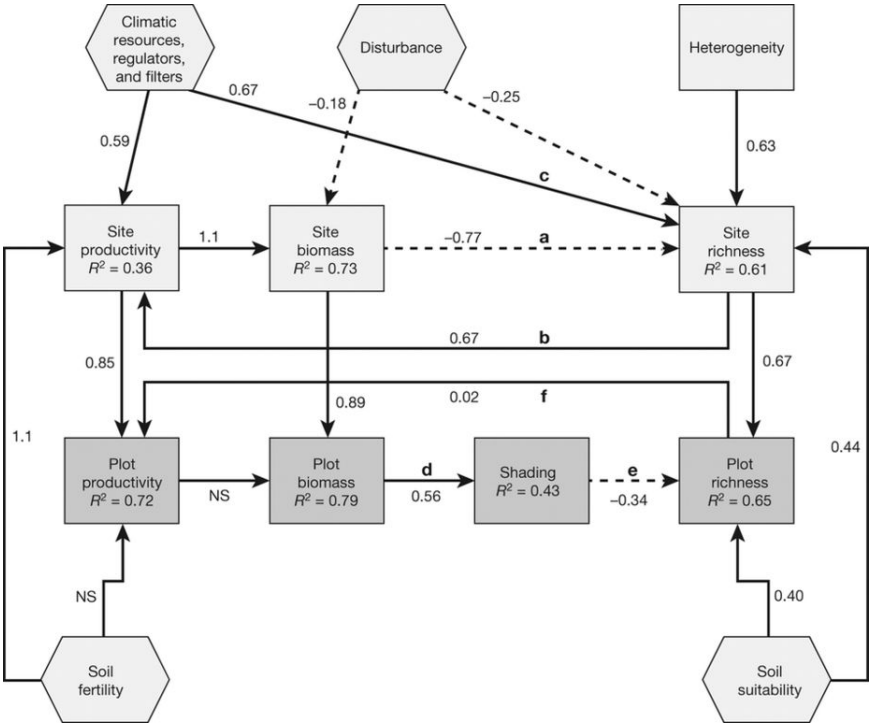


Figure 4.8: Full model from Grace et al. (2016)

when marginalizing out those variables, despite the fact that they are actually separate measures. To keep the bow-free assumption, we do not include the directed edges between site productivity, site biomass and site richness. This results in ancestral relationships in the full model which are not otherwise captured in the marginalized model. Thus, we add directed edges from site productivity to plot biomass and plot richness; from site biomass to plot productivity and plot richness; from site richness to plot productivity and plot biomass. For suitability, there is both a site suitability, which is a parent of site richness, and a plot suitability which is a parent of plot richness. Although there is no explicit specification in their SEM of how site suitability relates into plot suitability, it seems reasonable to assume that site suitability has a direct effect on plot suitability, as is the case for all other site vs plot measures. Thus, we include a bidirected edge between plot suitability and site richness. This results in the BAP shown in Figure 4.9. We consider this model the ground truth.

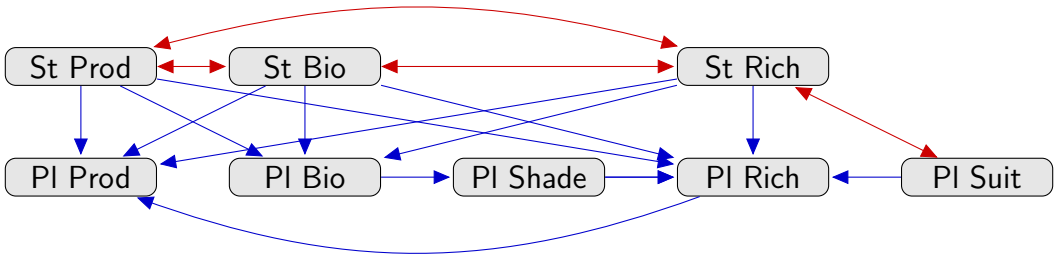


Figure 4.9: BAP representation of plot specific model from Grace et al. (2016).

For BANG, we selected the nominal test level, .01, so that there are roughly the same number of directed edges in the estimated and ground truth graphs, 11 and 13 respectively. The discovered graph is shown in Figure 4.10. Of the 28 pairs of nodes, BANG correctly identifies the correct relation ($\rightarrow, \leftarrow, \leftrightarrow$ or no edge) for 16 of the pairs. Naively, letting the probability of guessing each relationship to be 1/4, this results in a binomial probability of $P(X \geq 16) = .00029$. This probability does not account for the dependency between edges since there is an acyclic restriction, but it suggests that the method is doing much better than random guessing. There are 7 bidirected edges in the estimated graph compared to 4 in the ground truth model. This behavior is somewhat expected since there is still likely to be uncontrolled confounding which is either not actually fully accounted for in the ground truth model or direct causes which cannot be fully explained by a linear relationship. We compare the discovered model to the BAP selected using GBS with 100 random restarts. The graph posited by GBS only matches 7 out of the possible 28 pairs. It is quite possible that another BAP in the same equivalence class matches the ground truth much better than the BAP that was produced by the greedy search; however, this illustrates the point that only considering equivalence classes can lead to unsatisfactory scientific conclusions.

4.5 Discussion

In this chapter, we have discussed the identification of causal structure when the errors in a linear structural equation model are correlated. In particular, we borrow intuition from

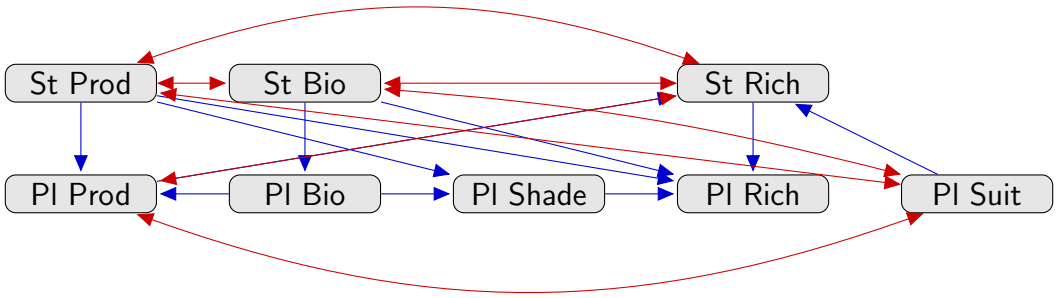


Figure 4.10: Discovered model (BANG)

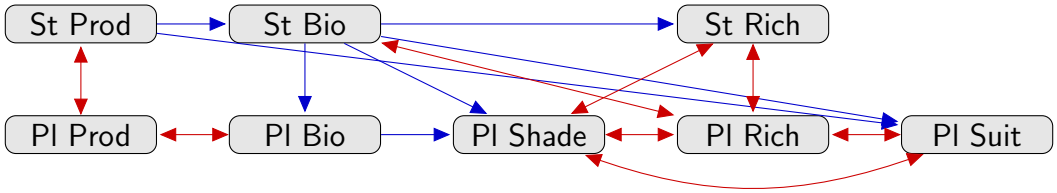


Figure 4.11: Discovered model (GBS)

the LiNGAM line of work (Shimizu et al., 2006), to show that when a SEM corresponds to a BAP and the errors are non-Gaussian, one can identify the exact causal structure from observational data. We propose the BANG algorithm and show that it consistently identifies the graph. This extends previous work on BAPs by Nowzohour et al. (2017) by identifying an exact graph rather than a larger equivalence class. In addition, this extends the work on non-Gaussian SEMs with errors by Hoyer et al. (2008b) and Shimizu and Bollen (2014) by consistently identifying an exact graph without pre-specifying the number of latent variables or a distribution of the errors.

Since the number of independence tests considered is a polynomial of the number of variables, under additional assumptions, we might be able to consistently estimate the graph in a sparse high dimensional setting where the maximum in-degree of the graph is bounded, but the number of variables p grows at a faster rate than the number of samples n. If the errors are assumed to be log-concave, but non-Gaussian, one could then again directly apply

the concentration results in Lin et al. (2016) which allows for arbitrary dependence. For computational reasons, using empirical likelihood would be intractable, so one might use a set of statistics or tests which are easier to compute as in Chapter 3.

Other possible extensions of the current work include discovery of graphs with bows by incorporating possible instrumental variables into the search procedure. In particular, Silva and Shimizu (2017) combine Tetrad constraints with LiNGAM ideas to test for valid instrumental variables. This could be incorporated into our proposed discovery procedure to consistently estimate direct effects even in the presence of bows.

Finally, as discussed in Chapter 3, Loh and Bühlmann (2014) show for DAGs, even with non-Gaussian errors, the precision matrix encodes causal structure. A similar statement can be made for BAPs, where a non-zero entry in the precision implies that two nodes are in the same Tian component. Thus, using a consistent procedure for estimating the precision could also be a fruitful pre-processing step in the BAP setting.

Chapter 5

DISCUSSION

In this work, we studied three related problems involving linear structural equation models where the stochastic error terms do not correspond to a Gaussian distribution.

We first study the setting where the underlying structure is known and corresponds to a mixed graph which may contain bows and cycles. We show that when the data is non-Gaussian, empirical likelihood can be an attractive alternative to existing methods when several modifications are applied to a naive formulation.

Next, we show that under certain conditions, when the underlying structure is unknown, but corresponds to a DAG, the graph can be consistently recovered in the high dimensional setting where the number of variables may exceed the number of observations. In particular, we require the maximum in-degree of the graph to be bounded and the errors to correspond to some log-concave distribution.

Finally, we show that the underlying graph can be consistently recovered in the presence of unobserved latent variables if the errors are non-Gaussian. In particular, we assume the graph corresponds to a bow-free acyclic path diagram (BAP), and show that the graph can be recovered with a polynomial number of independence tests when the maximum in-degree of the graph is bounded.

While the proposed causal discovery algorithms are theoretically sound and work well in simulations, we do acknowledge a few drawbacks of our approach. Although the assumption of non-Gaussian errors is almost always true, the assumption of linearity typically does not hold. When this assumption fails, it is not clear if the methods consistently estimate a single graph, and if so what that graph represents. In the previous conditional independence based constraint testing methods, if a graph is consistently estimated, it can still be interpreted as

a conditional independence map. However, a corresponding statement is not straightforward in the setting we consider. Thus, although we can make a much stronger statement when our assumptions hold, when they do not hold, the output can be quite hard to interpret. In addition, the methods require a type of strong faithfulness for estimation with finite samples. This is exacerbated by the constraint based approach which is already sensitive to error propagation.

We close by briefly describing a few possible avenues of future work.

5.0.1 Beyond constraint based methods

Although the constraint based methods are theoretically sound, greedy algorithms have been shown to be quite successful in practice because they are less susceptible to error propagation. The empirical likelihood with estimating equations that also constrain higher order moments may prove a useful criterion to greedily optimize because it is agnostic to error distribution. This would require a penalty on the number of parameters so that we consistently identify the exact graph and not a super-model of the truth. If the empirical likelihood could be decomposed in a sound way, this could be applied with dynamic programming ideas similar to Silander and Myllymäki (2006) to extend computational tractability to larger graphs. Similar to the graphical lasso, we have also explored using penalized empirical likelihood techniques for graph selection, but further analysis and refinement is needed.

5.0.2 Exploration of latent structure

We have only considered specification of latent confounding on a pairwise level. Preliminary work shows that one might be able to use non-Guassianity to also detect the number of latent variables. Combined with ideas about mDAGs (Evans, 2016), this may yield a finer grain approach to discovering causal structure.

5.0.3 Non-linear relationships

Peters et al. (2014) show that when the underlying structure corresponds to a DAG, causal discovery of the exact graph is also possible when the functional relationships implied by the structural equation model are non-linear and unknown. Combined with ideas from Chapter 4, in which the estimated covariance structure of previously certified ancestors is explicitly used, one may be able to extend their results to settings with latent confounding.

BIBLIOGRAPHY

Ali, R. A., Richardson, T. S., and Spirtes, P. (2009). Markov equivalence for ancestral graphs. *The Annals of Statistics*, pages 2808–2837.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

Brito, C. and Pearl, J. (2002). A new identification condition for recursive models with correlated errors. *Struct. Equ. Model.*, 9(4):459–474.

Burghy, C. A., Stodola, D. E., Ruttle, P. L., Molloy, E. K., Armstrong, J. M., Oler, J. A., Fox, M. E., Hayes, A. S., Kalin, N. H., Essex, M. J., et al. (2012). Developmental pathways to amygdala-prefrontal function and internalizing symptoms in adolescence. *Nature neuroscience*, 15(12):1736–1741.

Calis, J. C., Phiri, K. S., Faragher, E. B., Brabin, B. J., Bates, I., Cuevas, L. E., de Haan, R. J., Phiri, A. I., Malange, P., Khoka, M., Hulshof, P. J., van Lieshout, L., Beld, M. G., Teo, Y. Y., Rockett, K. A., Richardson, A., Kwiatkowski, D. P., Molyneux, M. E., and van Hensbroek, M. B. (2008). Severe anemia in Malawian children. *New England Journal of Medicine*, 358(9):888–899. PMID: 18305266.

Chaudhuri, S., Drton, M., and Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216.

Chaudhuri, S., Mondal, D., and Yin, T. (2017). Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(1):293–320.

Chen, B. (2016). Identification and overidentification of linear structural equation models. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 1587–1595. Curran Associates, Inc.

Chen, J., Variyath, A. M., and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *J. Comput. Graph. Statist.*, 17(2):426–443.

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

Claassen, T., Mooij, J. M., and Heskes, T. (2013). Learning sparse causal models is not np-hard. In Nicholson, A. and Smyth, P., editors, *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*. AUAI Press.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15:3741–3782.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.*, 40(1):294–321.

Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314.

Drton, M., Eichler, M., and Richardson, T. S. (2009). Computing maximum likelihood estimates in recursive linear models with correlated errors. *J. Mach. Learn. Res.*, 10:2329–2348.

Drton, M., Fox, C., and Wang, Y. S. (2017). Computation of maximum likelihood estimates in cyclic structural equation models. *Ann. Statist., to appear.* arXiv:1610.03434.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4(1):365–393.

Drton, M. and Weihs, L. (2016). Generic identifiability of linear structural equation models by ancestor decomposition. *Scand. J. Stat.*, 43(4):1035–1045.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063.

Entner, D. and Hoyer, P. O. (2010). Discovering unconfounded causal relationships using linear non-gaussian models. In Onada, T., Bekki, D., and McCready, E., editors, *New Frontiers in Artificial Intelligence - JSAI-isAI 2010 Workshops, LENLS, JURISIN, AMBN, ISS, Tokyo, Japan, November 18-19, 2010, Revised Selected Papers*, volume 6797 of *Lecture Notes in Computer Science*, pages 181–195. Springer.

Evans, R. J. (2016). Graphs for margins of Bayesian networks. *Scand. J. Stat.*, 43(3):625–648.

Fox, J., Nie, Z., and Byrnes, J. (2017). *sem: Structural Equation Models*. R package version 3.1-9.

Foygel, R., Draisma, J., and Drton, M. (2012). Half-trek criterion for generic identifiability of linear structural equation models. *Ann. Statist.*, 40(3):1682–1713.

Grace, J. B., Anderson, T. M., Seabloom, E. W., Borer, E. T., Adler, P. B., Harpole, W. S., Hautier, Y., Hillebrand, H., Lind, E. M., Pärtel, M., et al. (2016). Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature*, 529(7586):390–393.

Grendár, M. and Judge, G. (2009). Empty set problem of maximum empirical likelihood methods. *Electron. J. Stat.*, 3:1542–1555.

Hao, D., Ren, C., and Li, C. (2012). Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. *BMC systems biology*, 6(1):34.

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14:3365–3383.

He, Y., Jia, J., and Yu, B. (2015). Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.*, 16:2589–2609.

Horn, R. A. and Johnson, C. R. (2013). *Matrix analysis*. Cambridge University Press, Cambridge, second edition.

Hoyer, P. O., Hyvärinen, A., Scheines, R., Spirtes, P., Ramsey, J., Lacerda, G., and Shimizu, S. (2008a). Causal discovery of linear acyclic models with arbitrary distributions. In *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008*, pages 282–289.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. (2008b). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362 – 378. Special Section on Probabilistic Rough Sets and Special Section on PGM06.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *J. Mach. Learn. Res.*, 14:111–152.

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

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal

inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Kolenikov, S. and Yuan, Y. (2009). Empirical likelihood estimation and testing in covariance structure models. available at http://staskolenikov.net.

Koster, J. T. A. (2002). Marginalizing and conditioning in graphical models. *Bernoulli*, 8(6):817–840.

Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications.

Lin, L., Drton, M., and Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806–854.

Liu, B., de la Fuente, A., and Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, 178(3):1763–1776.

Loh, P.-L. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *J. Mach. Learn. Res.*, 15:3065–3105.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, 37(6A):3133–3164.

Matsueda, R. L. and Heimer, K. (1987). Race, family structure, and delinquency: A test of differential association and social control theories. *American Sociological Review*, pages 826–840.

Muthen, B. and Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, 45(1):19–30.

Nandy, P., Hauser, A., and Maathuis, M. H. (2015). High-dimensional consistency in score-based and hybrid structure learning. *arXiv preprint arXiv:1507.02608*.

Nowzohour, C. (2017). *greedyBAPs: Greedy BAP Learning Using Penalised Maximum Likelihood Score*. R package version 0.0.0.9000.

Nowzohour, C., Maathuis, M. H., Evans, R. J., and Bühlmann, P. (2017). Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electron. J. Stat.*, 11(2):5342–5374.

Okamoto, M. (1973). Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.*, 1:763–765.

Olsson, U. H., Foss, T., Troye, S. V., and Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, 7(4):557–595.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A. B. (2001). *Empirical likelihood*. CRC press.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, second edition. Models, reasoning, and inference.

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

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15:2009–2053.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, 22(1):300–325.

R Core Team (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *Ann. Statist.*, 30(4):962–1030.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Shimizu, S. and Bollen, K. (2014). Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-Gaussian distributions. *Journal of Machine Learning Research*, 15(1):2629–2652.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. *J. Mach. Learn. Res.*, 12:1225–1248.

Shimizu, S. and Kano, Y. (2008). Use of non-normality in structural equation modeling: Application to direction of causation. *Journal of Statistical Planning and Inference*, 138(11):3483–3491.

Shpitser, I., Evans, R. J., Richardson, T. S., and Robins, J. M. (2014). Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39.

Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 1219–1226.

Silander, T. and Myllymäki, P. (2006). A simple approach for finding the globally optimal bayesian network structure. In *UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*. AUAI Press.

Silva, R. and Shimizu, S. (2017). Learning instrumental variables with structural and non-gaussianity assumptions. *Journal of Machine Learning Research*, 18(120):1–49.

Sokol, A., Maathuis, M. H., and Falkeborg, B. (2014). Quantifying identifiability in independent component analysis. *Electron. J. Stat.*, 8(1):1438–1459.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book.

Steinsky, B. (2013). Enumeration of labelled essential graphs. *Ars Combin.*, 111:485–494.

Thrien, J.-P. and Nol, A. (2000). Political parties and foreign aid. *American Political Science Review*, 94(1):151162.

Tian, J. and Pearl, J. (2002). On the testable implications of causal models with hidden variables. In Darwiche, A. and Friedman, N., editors, *UAI '02, Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence, University of Alberta, Edmonton, Alberta, Canada, August 1-4, 2002*, pages 519–527. Morgan Kaufmann.

Triantafillou, S. and Tsamardinos, I. (2016). Score-based vs constraint-based causal learning in the presence of confounders. In Eberhardt, F., Bareinboim, E., Maathuis, M. H., Mooij, J. M., and Silva, R., editors, *Proceedings of the UAI 2016 Workshop on Causation: Foundation to Application co-located with the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), Jersey City, USA, June 29, 2016.*, volume 1792 of *CEUR Workshop Proceedings*, pages 59–67. CEUR-WS.org.

Tsao, M. and Wu, F. (2014). Extended empirical likelihood for estimating equations. *Biometrika*, 101(3):703–710.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *Ann. Statist.*, 41(2):436–463.

van der Linden, S. (2018). Warm glow is associated with low- but not high-cost sustainable behaviour. *Nature Sustainability*, 1(1):28–30.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *UAI '90: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, MIT, Cambridge, MA, USA, July 27-29, 1990*, pages 255–270.

Wermuth, N. (2011). Probability distributions with summary graph structure. *Bernoulli*, 17(3):845–879.

Wright, S. (1921). Correlation and causation. *J. Agric. Res.*, 20:557–585.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896.

Appendix A

APPENDIX

A.1 Proof of Theorem 3.1

Theorem 3.1. *Let $P \in \mathcal{P}(G)$ be a distribution in the model given by an acyclic digraph G, and let $Y_i \sim P$. For $K > 2$, two distinct nodes u and v, and any set $C \subseteq V \setminus \{u, v\}$, define*

$$\tau^{(K)}_{v.C\to u} := \mathbb{E}_P(Y^{K-1}_{vi.C}Y_{ui})\mathbb{E}_P(Y^2_{vi.C}) - \mathbb{E}_P(Y^K_{vi.C})\mathbb{E}_P(Y_{vi.C}Y_{ui}). \tag{3.4}$$

(i) *If $u \notin \mathrm{pa}(v)$ and $\mathrm{pa}(v) \subseteq C \subseteq V \setminus \{\mathrm{de}(v), v, u\}$, then*

$$\tau^{(K)}_{v.C\to u} = 0.$$

(ii) *Suppose P is parentally faithful with respect to G. If $u \in \mathrm{pa}(v)$ and $C \subseteq V \setminus \{\mathrm{de}(v), v, u\}$, then for generic error moments up to order $K > 2$, we have*

$$\tau^{(K)}_{v.C\to u} \neq 0.$$

Proof. **Statement (i):** Consider any set C such that $\mathrm{pa}(v) \subseteq C$ and $C \bigcap \mathrm{de}(v) = \emptyset$. Since we condition on all parents, $\beta_{vc.C} = 0$ for any $c \in C$ which is not a parent of v. Then,

$$\begin{aligned} Y_{v.C} &= Y_v - \sum_{c\in\mathrm{pa}(v)} \beta_{vc.C}Y_c - \sum_{k\in C\setminus\mathrm{pa}(v)} \beta_{vc.C}Y_c \\ &= Y_v - \sum_{c\in\mathrm{pa}(v)} \beta_{vc.C}Y_c \\ &= \varepsilon_v. \end{aligned} \tag{A.1}$$

We then directly calculate the parameter for the set C.

$$\begin{aligned}
\tau^{(K)}_{v.C\to u} &= \mathbb{E}(Y^{K-1}_{v.C}Y_u)\mathbb{E}(Y^2_{v.C}) - \mathbb{E}(Y^K_{v.C})\mathbb{E}(Y_{v.C}Y_u) \\
&= \mathbb{E}\left(\varepsilon_v^{K-1}\left[\varepsilon_u + \pi_{uv}\varepsilon_v + \sum_{z\in \mathrm{an}(u)\setminus\{v\}} \pi_{uz}\varepsilon_z\right]\right)\mathbb{E}\left(\varepsilon_v^2\right) \\
&\quad - \mathbb{E}\left(\varepsilon_v^K\right)\mathbb{E}\left(\varepsilon_v\left[\varepsilon_u + \pi_{uv}\varepsilon_v + \sum_{z\in \mathrm{an}(u)\setminus\{v\}} \pi_{uz}\varepsilon_z\right]\right) \\
&= \pi_{uv}\mathbb{E}\left(\varepsilon_v^K\right)\mathbb{E}\left(\varepsilon_v^2\right) - \pi_{uv}\mathbb{E}\left(\varepsilon_v^K\right)\mathbb{E}\left(\varepsilon_v^2\right) \\
&= 0.
\end{aligned} \tag{A.2}$$

The penultimate equality follows from the assumption of independent errors. If there is no directed path from v to u, then $\pi_{uv} = 0$ and the statement trivially holds. □

Statement (ii): For fixed C, $u, v \in V$, and parentally faithful linear coefficients and variances, $\tau^{(K)}_{v.C\to u}$ is a polynomial of the error moments of degree $k = 3, \ldots, K$. Thus, selecting a single point (of error moments) where the quantity $\tau_{v.C\to u}$ is non-zero is sufficient for showing that the quantity is non-zero for generic error moments of degree $k = 3, \ldots, K$ (Okamoto, 1973). Specifically, we select that point by letting all error moments for $k < K$ be consistent with the Gaussian moments implied by σ^2_v, the variance of ε_v, but select the Kth degree moment to be inconsistent with the corresponding Gaussian moment. Since there are a finite number of sets $C \subseteq V$ such that $C \bigcap \mathrm{de}(v) = \emptyset$, then the set of error moments of degree $k = 3, \ldots, K$ which yield $\tau_{v.C\to u} = 0$ for any $C \subseteq V$ also has Lebesgue measure zero.

Recall that the total residual effect of u on v given C is

$$\pi_{vu.C} = \pi_{vu} - \sum_{c\in C} \beta_{vc.C}\pi_{cu},$$

where π_{vu} is the total effect of u on v and $\pi_{uu} = 1$. For any set C, where $C \bigcap \mathrm{de}(v) = \emptyset$,

$$\begin{aligned} Y_{v.C} &= Y_v - \sum_{c \in C} \beta_{vc.C} Y_c \\ &= \varepsilon_v + \sum_{k \in \mathrm{an}(v)} \pi_{vk} \varepsilon_k - \sum_{c \in C} \beta_{vc.C} \sum_{d \in \mathrm{An}(c)} \pi_{cd} \varepsilon_d \\ &= \varepsilon_v + \sum_{k \in \{\mathrm{an}(v) \cup \mathrm{An}(C)\}} \pi_{vk.C} \varepsilon_k. \end{aligned} \tag{A.3}$$

By the parental faithfulness assumption, since $u \in \mathrm{pa}(v)$, $\pi_{vu.C} \neq 0$. We partition $\{\mathrm{An}(v) \cup \mathrm{An}(C)\}$ into three disjoint sets

$$\begin{aligned} Z_1 &= \{\mathrm{An}(v) \cup \mathrm{An}(C)\} \setminus \mathrm{An}(u) \\ Z_2 &= \{z \in \mathrm{An}(u) : \text{such that } \pi_{vz.C} = \pi_{vu.C} \pi_{uz}\} \\ Z_3 &= \mathrm{An}(u) \setminus Z_2. \end{aligned} \tag{A.4}$$

Note that $u \in Z_2$ and graphically, $Z_2 \setminus \{u\}$ corresponds to ancestors of u which only have directed paths to v through $C \bigcup \{u\}$, while Z_3 corresponds to ancestors of u which have directed paths to v that do not pass through $C \bigcup \{u\}$. If there is no confounding between $Y_{v.C}$ and Y_u, then Z_3 is empty. For notational convenience, we define

$$\begin{aligned} \varepsilon_{Z_1} &= \varepsilon_v + \sum_{z \in Z_1} \pi_{vz.C} \varepsilon_z \\ \varepsilon_{Z_2} &= \sum_{z \in Z_2} \pi_{uz} \varepsilon_z \\ \varepsilon_{Z_3} &= \sum_{z \in Z_3} \left(\pi_{vz} - \sum_{c \in C} \beta_{vc.C} \pi_{cz} \right) \varepsilon_z = \sum_{z \in Z_3} \pi_{vz.C} \varepsilon_z \end{aligned} \tag{A.5}$$

so that

$$Y_{v.C} = \varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2} + \varepsilon_{Z_3}$$

and

$$Y_u = \varepsilon_{Z_2} + \sum_{z \in Z_3} \pi_{uz}\varepsilon_z.$$

For $\boldsymbol{w} = (w_1, w_2, w_3)$ with $|w| = w_0$, let $\binom{w_0}{\boldsymbol{w}} = \frac{w_0!}{w_1!w_2!w_3!}$, the multinomial coefficient. Then

$$\begin{aligned} Y_{v.C}^{w_0} &= \left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2} + \varepsilon_{Z_3}\right)^{w_0} = \sum_{|\boldsymbol{w}|=w_0} \binom{w_0}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1} (\pi_{vu.C}\varepsilon_{Z_2})^{w_2} \varepsilon_{Z_3}^{w_3} \\ &= \sum_{\substack{|\boldsymbol{w}|=w_0 \\ w_3=0}} \binom{w_0}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1} (\pi_{vu.C}\varepsilon_{Z_2})^{w_2} \varepsilon_{Z_3}^{w_3} + \sum_{\substack{|\boldsymbol{w}|=w_0 \\ w_3>0}} \binom{w_0}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1} (\pi_{vu.C}\varepsilon_{Z_2})^{w_2} \varepsilon_{Z_3}^{w_3} \\ &= \left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^{w_0} + \sum_{\substack{|\boldsymbol{w}|=w_0 \\ w_3>0}} \binom{w_0}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1} (\pi_{vu.C}\varepsilon_{Z_2})^{w_2} \varepsilon_{Z_3}^{w_3}, \end{aligned} \tag{A.6}$$

so that

$$\begin{aligned} \tau_{v.C\to u}^{(K)} &= \mathbb{E}(Y_{v.C}^{K-1}Y_u)\mathbb{E}(Y_{v.C}^2) - \mathbb{E}(Y_{v.C}^K)\mathbb{E}(Y_{v.C}Y_u) \\ &= \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2} + \varepsilon_{Z_3}\right)^{K-1} Y_u\right) \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2} + \varepsilon_{Z_3}\right)^2\right) \\ &\quad - \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2} + \varepsilon_{Z_3}\right)^K\right) \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2} + \varepsilon_{Z_3}\right) Y_u\right) \\ &= \left[\mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^{K-1} Y_u\right) + \mathbb{E}\left(Y_u \sum_{\substack{|\boldsymbol{w}|=K-1 \\ w_3>0}} \binom{K-1}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1} (\pi_{vu.C}\varepsilon_{Z_2})^{w_2} \varepsilon_{Z_3}^{w_3}\right)\right] \\ &\quad \times \left[\sigma_{Z_1}^2 + \pi_{vu.C}\sigma_{Z_2}^2 + \sigma_{Z_3}^2\right] \\ &\quad - \left[\mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^K\right) + \mathbb{E}\left(\sum_{\substack{|\boldsymbol{w}|=K \\ w_3>0}} \binom{K}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1} (\pi_{vu.C}\varepsilon_{Z_2})^{w_2} \varepsilon_{Z_3}^{w_3}\right)\right] \end{aligned} \tag{A.7}$$

$$\times \left[\pi_{vu.C}\sigma^2_{Z_2} + \mathbb{E}\left(\varepsilon_{Z_3} Y_u\right)\right].$$

We first consider the case where Z_3 is empty so that the expansion above reduces to

$$\begin{aligned}\tau_{v.C\to u} = &\left[\mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^{K-1} Y_u\right)\right] \times \left[\sigma^2_{Z_1} + \sigma^2_{Z_2}\right] \\ &- \left[\mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^{k}\right)\right] \times \left[\pi_{vu.C}\sigma^2_{Z_2}\right].\end{aligned} \tag{A.8}$$

Let the moments of degree k for $2 < k < K$ of all the error terms be consistent with some Gaussian distribution. This implies that the error moments for $z = Z_1, Z_2$ are also consistent with some Gaussian distribution since the sum of Gaussians is also Gaussian. So for $\mathbb{E}(\varepsilon_z^2) = \sigma_z^2$ and for $k < K$

$$E(\varepsilon_z^k) = \begin{cases} 0 & \text{if k is odd} \\ (k-1)!!\sigma_z^k & \text{if k is even,} \end{cases}$$

where $k!!$ denotes the double factorial of k. However, let the Kth degree error moments be inconsistent with the specified Gaussian distribution so that for $z = Z_1, Z_2$ and $\eta_z > 0$,

$$E(\varepsilon_z^K) = \begin{cases} \eta_z & \text{if K is odd} \\ (K-1)!!\sigma_z^K + \eta_z & \text{if K is even.} \end{cases}$$

By direct calculation we see

$$\begin{aligned}\tau^{(K)}_{v.C\to u} &= \mathbb{E}(Y_{v.C}^{K-1} Y_u)\mathbb{E}(Y_{v.C}^2) - \mathbb{E}(Y_{v.C}^K)\mathbb{E}(Y_{v.C} Y_u) \\ &= \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^{K-1} Y_u\right) \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^2\right) \\ &\quad - \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^K\right) \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)(Y_u)\right)\end{aligned} \tag{A.9}$$

$$
\begin{aligned}
&= \left(\sum_{a=0}^{K-1}\binom{K-1}{a}\mathbb{E}\left(\varepsilon_{Z_1}^a\right)\mathbb{E}\left(\pi_{vu.C}^{K-1-a}\varepsilon_{Z_2}^{K-1-a}Y_u\right)\right)\left(\mathbb{E}(\varepsilon_{Z_1}^2)+\mathbb{E}(\pi_{vu.C}\varepsilon_{Z_2}Y_u)\right)\\
&\quad-\left(\sum_{a=0}^{K}\binom{K}{a}\mathbb{E}\left(\varepsilon_{Z_1}^a\right)\mathbb{E}\left(\pi_{vu.C}^{K-1-a}\varepsilon_{Z_2}^{K-1-a}Y_u\right)\right)\left(\mathbb{E}\left(\pi_{vu.C}\varepsilon_{Z_2}Y_u\right)\right)\\
&= \left(\sum_{a=1}^{K-1}\binom{K-1}{a}\mathbb{E}\left(\varepsilon_{Z_1}^a\right)\pi_{vu.C}^{K-1-a}\mathbb{E}\left(\varepsilon_{Z_2}^{K-a}\right)\right)\left(\mathbb{E}(\varepsilon_{Z_1}^2)+\pi_{vu.C}^2\mathbb{E}(\varepsilon_{Z_2}^2)\right)\\
&\quad-\left(\sum_{a=1}^{K-1}\binom{K}{a}\mathbb{E}\left(\varepsilon_{Z_1}^a\right)\pi_{vu.C}^{K-a}\mathbb{E}\left(\varepsilon_{Z_2}^{K-a}\right)\right)\left(\pi_{vu.C}\mathbb{E}\left(\varepsilon_{Z_2}^2\right)\right)\\
&\quad+\pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\left[\mathbb{E}(\varepsilon_{Z_1}^2)+\pi_{vu.C}^2\mathbb{E}(\varepsilon_{Z_2}^2)\right]-\left(\pi_{vu.C}^K\mathbb{E}(\varepsilon_{Z_2}^K)+\mathbb{E}(\varepsilon_{Z_1}^K)\right)\left(\pi_{vu.C}\mathbb{E}(\varepsilon_{Z_2}^2)\right)\\
&= \left(\sum_{a=1}^{K-1}\binom{K-1}{a}\pi_{vu.C}^{K-1-a}\mathbb{E}\left(\varepsilon_{Z_1}^a\right)\mathbb{E}\left(\varepsilon_{Z_2}^{K-a}\right)\right)\left(\mathbb{E}(\varepsilon_{Z_1}^2)+\pi_{vu.C}^2\mathbb{E}(\varepsilon_{Z_2}^2)\right)\\
&\quad-\left(\sum_{a=1}^{K-1}\binom{K}{a}\mathbb{E}\left(\varepsilon_{Z_1}^a\right)\pi_{vu.C}^{K-a}\mathbb{E}\left(\varepsilon_{Z_2}^{K-a}\right)\right)\left(\pi_{vu.C}\mathbb{E}\left(\varepsilon_{Z_2}^2\right)\right)\\
&\quad+\pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\mathbb{E}(\varepsilon_{Z_1}^2)-\pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\mathbb{E}(\varepsilon_{Z_2}^2).
\end{aligned}
$$

When K is odd, $\mathbb{E}(\varepsilon_{Z_1}^a)\mathbb{E}(\varepsilon_{Z_2}^{K-a})=0$ for all $a=1,\ldots,K-1$, so we are left with

$$
\begin{aligned}
\tau_{v.C\to u} &= \pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\sigma_{Z_1}^2-\pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\sigma_{Z_2}^2\\
&= \pi_{vu.C}^{K-1}\eta_{Z_2}\sigma_{Z_1}^2-\pi_{vu.C}\eta_{Z_1}\sigma_{Z_2}^2.
\end{aligned}
\tag{A.10}
$$

When K is even, then $\mathbb{E}(\varepsilon_{Z_1}^a)\mathbb{E}(Y_u^{K-a})=0$ when a is odd, so we are left with

$$
\begin{aligned}
\tau_{v.C\to u} &= \left(\sum_{a=2,4,\ldots,K-2}\binom{K-1}{a}\pi_{vu.C}^{K-1-a}\mathbb{E}\left(\varepsilon_{Z_1}^a\right)\mathbb{E}\left(\varepsilon_{Z_2}^{K-a}\right)\right)\left(\mathbb{E}(\varepsilon_{Z_1}^2)+\pi_{vu.C}^2\mathbb{E}(\varepsilon_{Z_2}^2)\right)\\
&\quad-\left(\sum_{a=2,4,\ldots,K-2}\binom{K}{a}\mathbb{E}\left(\varepsilon_{Z_1}^a\right)\pi_{vu.C}^{K-a}\mathbb{E}\left(\varepsilon_{Z_2}^{K-a}\right)\right)\left(\pi_{vu.C}\mathbb{E}\left(\varepsilon_{Z_2}^2\right)\right)
\end{aligned}
\tag{A.11}
$$

$$+ \pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\mathbb{E}(\varepsilon_{Z_1}^2) - \pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\mathbb{E}(\varepsilon_{Z_2}^2).$$

Evaluating the moments yields

$$\begin{aligned}
\tau_{v.C\to u} &= \left(\sum_{a=2,\ldots,K-2}\binom{K-1}{a}\pi_{vu.C}^{K-1-a}(a-1)!!\sigma_{Z_1}^a(K-a-1)!!\sigma_{Z_2}^{K-a}\right)\left(\sigma_{Z_1}^2+\pi_{vu.C}^2\sigma_{Z_2}^2\right) \\
&\quad - \left(\sum_{a=2,\ldots,K-2}\binom{K}{a}(a-1)!!\sigma_v^a\pi_{vu.C}^{K-a}(K-a-1)!!\sigma_{Z_2}^{K-a}\right)\left(\pi_{vu.C}\sigma_{Z_2}^2\right) \\
&\quad + \pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\sigma_{Z_1}^2 - \pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\sigma_{Z_2}^2 \\
&= \left(\sum_{a=2,\ldots,K-2}\binom{K-1}{a}\pi_{vu.C}^{K-1-a}(a-1)!!\sigma_{Z_1}^{a+2}(K-a-1)!!\sigma_{Z_2}^{K-a}\right) \\
&\quad + \pi_{vu.C}\sigma_{Z_2}^2\left(\sum_{a=2,\ldots,K-2}\binom{K-1}{a}\pi_{vu.C}^{K-1-a}(a-1)!!\sigma_{Z_1}^a(K-a-1)!!\sigma_{Z_2}^{K-a}\right) \\
&\quad - \left(\sum_{a=2,\ldots,K-2}\binom{K-1}{a}\frac{K}{K-a}(a-1)!!\sigma_{Z_1}^a\pi_{vu.C}^{K-a}(K-a-1)!!\sigma_{Z_2}^{K-a}\right)\left(\pi_{vu.C}\sigma_{Z_2}^2\right) \\
&\quad + \pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\sigma_{Z_1}^2 - \pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\sigma_{Z_2}^2 \\
&= \left(\sum_{a=2,\ldots,K-2}\binom{K-1}{a}\pi_{vu.C}^{K-1-a}(a-1)!!\sigma_{Z_1}^{a+2}(K-a-1)!!\sigma_{Z_2}^{K-a}\right) \\
&\quad + \left(\sum_{a=2,\ldots,K-2}\binom{K-1}{a}\left(1-\frac{K}{K-a}\right)(a-1)!!\sigma_{Z_1}^a\pi_{vu.C}^{K-a}(K-a-1)!!\sigma_{Z_2}^{K-a}\right)\left(\pi_{vu.C}\sigma_{Z_2}^2\right) \\
&\quad + \pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\sigma_{Z_1}^2 - \pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\sigma_{Z_2}^2 \\
&= \left(\sum_{a=2,\ldots,K-4}\binom{K-1}{a}\pi_{vu.C}^{K-1-a}(a-1)!!\sigma_{Z_1}^{a+2}(K-a-1)!!\sigma_{Z_2}^{K-a}\right) \\
&\quad - \left(\sum_{a=4,\ldots,K-2}\binom{K-1}{a}\frac{a}{K-a}(a-1)!!\sigma_{Z_1}^a\pi_{vu.C}^{K-a}(K-a-1)!!\sigma_{Z_2}^{K-a}\right)\left(\pi_{vu.C}\sigma_{Z_2}^2\right)
\end{aligned} \tag{A.12}$$

$$+ \pi_{vu.C}(K-1)!!\sigma_{Z_1}^K\sigma_{Z_2}^2 - \pi_{vu.C}^{K-1}(K-1)!!\sigma_{Z_2}^K\sigma_{Z_1}^2$$
$$+ \pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\sigma_{Z_1}^2 - \pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\sigma_{Z_2}^2.$$

Rewriting terms and a change of variables show that the first two lines cancel leaving

$$= \left[\left(\sum_{a=2,\ldots,K-4}\binom{K-1}{a+2}\frac{(a+1)(a+2)}{(K-(a+1))(K-(a+2))}\pi_{vu.C}^{K-(a+2)}(a-1)!!\sigma_{Z_1}^{a+2}(K-a-1)!!\sigma_{Z_2}^{K-(a+2)}\right)\right. \tag{A.13}$$
$$\left. - \left(\sum_{a=4,\ldots,K-2}\binom{K-1}{a}\left(\frac{a}{K-a}\right)(a-1)!!\sigma_{Z_1}^a\pi_{vu.C}^{K-a}(K-a-1)!!\sigma_{Z_2}^{K-a}\right)\right]\pi_{vu.C}\sigma_{Z_2}^2$$
$$+ \pi_{vu.C}(K-1)!!\sigma_{Z_1}^K\sigma_{Z_2}^2 - \pi_{vu.C}^{K-1}(K-1)!!\sigma_{Z_2}^K\sigma_{Z_1}^2$$
$$+ \pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\sigma_{Z_1}^2 - \pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\sigma_{Z_2}^2$$
$$= \pi_{vu.C}\sigma_{Z_2}^2\left[\left(\sum_{a=2,\ldots,K-4}\binom{K-1}{a+2}\frac{a+2}{K-(a+2)}\pi_{vu.C}^{K-(a+2)}((a+2)-1)!!\sigma_{Z_1}^{a+2}(K-(a+2)-1)!!\sigma_{Z_2}^{K-(a+2)}\right)\right.$$
$$\left. - \left(\sum_{a=4,\ldots,K-2}\binom{K-1}{a}\left(\frac{a}{K-a}\right)(a-1)!!\sigma_{Z_1}^a\pi_{vu.C}^{K-a}(K-a-1)!!\sigma_{Z_2}^{K-a}\right)\right]$$
$$+ \pi_{vu.C}(K-1)!!\sigma_{Z_1}^K\sigma_{Z_2}^2 - \pi_{vu.C}^{K-1}(K-1)!!\sigma_{Z_2}^K\sigma_{Z_1}^2$$
$$+ \pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\sigma_{Z_1}^2 - \pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\sigma_{Z_2}^2$$
$$= \pi_{vu.C}(K-1)!!\sigma_{Z_1}^K\sigma_{Z_2}^2 - \pi_{vu.C}^{K-1}(K-1)!!\sigma_{Z_2}^K\sigma_{Z_1}^2 + \pi_{vu.C}^{K-1}\mathbb{E}(\varepsilon_{Z_2}^K)\sigma_{Z_1}^2 - \pi_{vu.C}\mathbb{E}(\varepsilon_{Z_1}^K)\sigma_{Z_2}^2$$
$$= \pi_{vu.C}^{K-1}\eta_{Z_2}\sigma_{Z_1}^2 - \pi_{vu.C}\eta_{Z_1}\sigma_{Z_2}^2.$$

This is same expression as when K is odd. So for any $K > 2$,

$$\eta_{Z_1} \neq \frac{\pi_{vu.C}^{K-2}\sigma_{Z_1}^2\eta_2}{\sigma_{Z_2}^2} \qquad \text{iff} \qquad \tau_{v.C\to u} \neq 0. \tag{A.14}$$

Since Z_1 and Z_2 are disjoint, we can always select the Kth moments of the individual error

moments so that this holds.

Now consider the case when Z_3 is not empty. From Equations (A.7), (A.10), and (A.13),

$$
\begin{aligned}
\tau_{v.C\to u} &= \pi_{vu.C}^{K-1}\eta_{Z_2}\sigma_{Z_1}^2 - \pi_{vu.C}\eta_{Z_1}\sigma_{Z_2}^2 \\
&\quad + \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^{K-1} Y_u\right)\sigma_{Z_3}^2 \\
&\quad + \mathbb{E}\left(Y_u \sum_{\substack{|\boldsymbol{w}|=K-1 \\ w_3>0}} \binom{K-1}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1}(\pi_{vu.C}\varepsilon_{Z_2})^{w_2}\varepsilon_{Z_3}^{w_3}\right) \times \left[\sigma_{Z_1}^2 + \pi_{vu.C}\sigma_{Z_2}^2 + \sigma_{Z_3}^2\right] \\
&\quad - \mathbb{E}\left(\left(\varepsilon_{Z_1} + \pi_{vu.C}\varepsilon_{Z_2}\right)^{K}\right)\mathbb{E}\left(\varepsilon_{Z_3}Y_u\right) \\
&\quad - \mathbb{E}\left(Y_u \sum_{\substack{|\boldsymbol{w}|=K \\ w_3>0}} \binom{K}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1}(\pi_{vu.C}\varepsilon_{Z_2})^{w_2}\varepsilon_{Z_3}^{w_3}\right) \\
&\quad \times \left[\pi_{vu.C}\sigma_{Z_2}^2 + \mathbb{E}\left(\varepsilon_{Z_3}Y_u\right)\right] \\
&= \pi_{vu.C}^{K-1}\eta_{Z_2}\sigma_{Z_1}^2 - \pi_{vu.C}\eta_{Z_1}\sigma_{Z_2}^2 \\
&\quad + \left(\pi_{vu.C}^{K-1}(K-1)!!\sigma_{Z_2}^K + \eta_{Z_2}\right)\sigma_{Z_3}^2 + \sigma_{Z_3}^2 \sum_{a=1}^{K-1} \mathbb{E}(\varepsilon_{Z_1}^a)\pi_{vu.C}^{K-1-a}\mathbb{E}(\varepsilon_{Z_2}^{K-a}) \\
&\quad + \mathbb{E}\left(Y_u \sum_{\substack{|\boldsymbol{w}|=K-1 \\ w_3>0}} \binom{K-1}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1}(\pi_{vu.C}\varepsilon_{Z_2})^{w_2}\varepsilon_{Z_3}^{w_3}\right) \times \left[\sigma_{Z_1}^2 + \pi_{vu.C}\sigma_{Z_2}^2 + \sigma_{Z_3}^2\right] \\
&\quad - \left((K-1)!!\sigma_{Z_1}^K + \eta_{Z_1} + \pi_{vu.C}^K\left((K-1)!!\sigma_{Z_2}^K + \eta_{Z_2}\right) + \sum_{a=1}^{K-1}\mathbb{E}(\varepsilon_{Z_1}^a)\mathbb{E}(\varepsilon_{Z_2}^K - a)\right)\mathbb{E}\left(\varepsilon_{Z_3}Y_u\right) \\
&\quad - \mathbb{E}\left(Y_u \sum_{\substack{|\boldsymbol{w}|=K \\ w_3>0}} \binom{K}{\boldsymbol{w}} \varepsilon_{Z_1}^{w_1}(\pi_{vu.C}\varepsilon_{Z_2})^{w_2}\varepsilon_{Z_3}^{w_3}\right)\left[\pi_{vu.C}\sigma_{Z_2}^2 + \mathbb{E}\left(\varepsilon_{Z_3}Y_u\right)\right].
\end{aligned}
\tag{A.15}
$$

Since the unevaluated terms are fixed with respect to η_{Z_1} and η_{Z_2}, selecting values such that

$$\begin{aligned}\eta_{Z_1} \neq \Bigg\{ & \pi_{vu.C}^{K-1}\eta_{Z_2}\sigma_{Z_1}^2 \\ & + \left(\pi_{vu.C}^{K-1}(K-1)!!\sigma_{Z_2}^K + \eta_{Z_2}\right)\sigma_{Z_3}^2 + \sigma_{Z_3}^2\sum_{a=1}^{K-1}\mathbb{E}(\varepsilon_{Z_1}^a)\pi_{vu.C}^{K-1-a}\mathbb{E}(\varepsilon_{Z_2}^{K-a}) \\ & + \mathbb{E}\left(Y_u\sum_{\substack{|w|=K-1\\ w_3>0}}\binom{K-1}{w}\varepsilon_{Z_1}^{w_1}(\pi_{vu.C}\varepsilon_{Z_2})^{w_2}\varepsilon_{Z_3}^{w_3}\right)\times\left[\sigma_{Z_1}^2 + \pi_{vu.C}\sigma_{Z_2}^2 + \sigma_{Z_3}^2\right] \\ & - \left((K-1)!!\sigma_{Z_1}^K + \pi_{vu.C}^K\left((K-1)!!\sigma_{Z_2}^K + \eta_{Z_2}\right) + \sum_{a=1}^{K-1}\mathbb{E}(\varepsilon_{Z_1}^a)\mathbb{E}(\varepsilon_{Z_2}^K - a)\right)\mathbb{E}\left(\varepsilon_{Z_3}Y_u\right) \\ & - \mathbb{E}\left(Y_u\sum_{\substack{|w|=K\\ w_3>0}}\binom{K}{w}\varepsilon_{Z_1}^{w_1}(\pi_{vu.C}\varepsilon_{Z_2})^{w_2}\varepsilon_{Z_3}^{w_3}\right)\left[\pi_{vu.C}\sigma_{Z_2}^2 + \mathbb{E}\left(\varepsilon_{Z_3}Y_u\right)\right]\Bigg\} \Big/ \left[\pi_{vu.C}\sigma_{Z_2}^2 + \mathbb{E}(\varepsilon_{Z_3}Y_u)\right]\end{aligned} \tag{A.16}$$

implies that $\tau_{v.C\to u} \neq 0$.

□

A.2 Proof of Lemma 3.1

Lemma 3.1. *Suppose that (C2), (C3), and (C4) hold. Then for any $v \in V$ and $C \subseteq V$ and $|C| \leq J$,*

$$\|\hat{\beta}_{vC} - \beta_{vC}\|_\infty < \delta_2 = 4\frac{J^{3/2}M\delta_1}{\lambda_{\min}^2}.$$

Proof. We use $\|\cdot\|$ to denote the vector norm and $|||\cdot|||$ to denote the matrix norm. Conditions (C2) and (C3) imply that

$$\begin{aligned}||\beta_{vC}||_\infty \leq ||\beta_{vC}||_2 &\leq \left|\left|\left|\Sigma_{cc}^{-1}\right|\right|\right|_2 ||\Sigma_{Cv}||_2 \\ &\leq \left|\left|\left|\Sigma_{cc}^{-1}\right|\right|\right|_2\sqrt{J}||\Sigma_{Cv}||_\infty \leq \frac{\sqrt{J}M}{\lambda_{\min}}.\end{aligned} \tag{A.17}$$

Condition (C4) implies $\hat{\Sigma}_{CC} = \Sigma_{CC} + E$ and $\hat{\Sigma}_{Cv} = \Sigma_{Cv} + e$ with $\|E\|_\infty < \delta_1$ and $\|e\|_\infty < \delta_1$. Using results from Horn and Johnson (2013, Equation 5.8.7) for the third and fourth inequalities below yields

$$\begin{aligned}
\|\hat{\beta}_{vC} - \beta_{vC}\|_\infty &\leq \|\hat{\beta}_{vC} - \beta_{vC}\|_2 = \|(\Sigma_{CC} + E)^{-1}(\Sigma_{Cv} + e) - \Sigma_{CC}^{-1}\Sigma_{Cv}\|_2 \\
&\leq \|(\Sigma_{CC} + E)^{-1}\Sigma_{Cv} - \Sigma_{CC}^{-1}\Sigma_{Cv}\|_2 + \|(\Sigma_{CC} + E)^{-1}e\|_2 \\
&\leq \frac{\left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right|_2}{1 - \left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right|_2}\|\beta_{vC}\|_2 + \left|\left|\left|(\Sigma_{CC} + E)^{-1}\right|\right|\right|_2\|e\|_2 \\
&\leq \frac{\left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right|_2}{1 - \left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right|_2}\|\beta_{vC}\|_2 + \frac{1/\lambda_{\min}}{1 - \left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right|_2}\|e\|_2 \\
&\leq \frac{\left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right|_2}{1 - \left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right|_2}\|\beta_{vC}\|_2 + \frac{\sqrt{J}\delta_1/\lambda_{\min}}{1 - \left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right|_2}.
\end{aligned} \tag{A.18}$$

The term $\left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right|_2 \leq \left|\left|\left|\Sigma_{CC}^{-1}\right|\right|\right|_2 |||E|||_2 \leq \frac{J\|E\|_\infty}{\lambda_{\min}} \leq \frac{J\delta_1}{\lambda_{\min}} < 1/2$. Since the bound in (A.18) is increasing in each of its arguments for $\left|\left|\left|\Sigma_{CC}^{-1}E\right|\right|\right| < 1$,

$$\begin{aligned}
\|\hat{\beta}_{vC} - \beta_{vC}\|_\infty &\leq \frac{\frac{J\delta_1}{\lambda_{\min}}}{1 - \frac{J\delta_1}{\lambda_{\min}}}\frac{\sqrt{J}M}{\lambda_{\min}} + \frac{\sqrt{J}\delta_1/\lambda_{\min}}{1 - \frac{J\delta_1}{\lambda_{\min}}} \\
&= \frac{\sqrt{J}\delta_1/\lambda_{\min}}{1 - \frac{J\delta_1}{\lambda_{\min}}}\left(JM/\lambda_{\min} + 1\right) \\
&\leq 2\frac{\sqrt{J}\delta_1}{\lambda_{\min}}\left(JM/\lambda_{\min} + 1\right) \\
&\leq 4\frac{J^{3/2}M\delta_1}{\lambda_{\min}^2} = \delta_2.
\end{aligned} \tag{A.19}$$

The penultimate inequality holds because by assumption $\frac{J\delta_1}{\lambda_{\min}} < 1/2$ and the last inequality holds because by assumption $M > \frac{\lambda_{\min}}{J}$. □

A.3 Proof of Lemma 3.3

Lemma 3.3. *Suppose that (C2), (C3), and (C4) hold. Then*

$$|\hat{\tau}_{v.C\to u} - \tau_{v.C\to u}| < 4M\delta_1\Phi(J,K,M,\lambda_{\min}) + 2\left(\delta_1\Phi(J,K,M,\lambda_{\min})\right)^2 = \delta_3$$

for the function $\Phi(J,K,M,\lambda_{\min})$ given in Lemma 3.2.

Proof. Similar the previous notation where $m_{H,\alpha} = \mathbb{E}(\prod_{h\in H} Z_h^{\alpha_h})$, we also allow for $v.C \in H$ indicating the population moment involving $Z_{v.C}$. By the triangle inequality,

$$\begin{aligned}
|\hat{\tau}_{v.C\to u} - \tau_{v\to u}| &= \big|\hat{m}_{(v.C,u),(K-1,1)}\hat{m}_{(v.C),(2)} - \hat{m}_{(v.C),(K)}\hat{m}_{(v.C,u),(1,1)} \\
&\quad -(m_{(v.C,u),(K-1,1)}m_{v.C(2)} - m_{(v.C),(K)}m_{(v.C,u),(1,1)})\big| \\
&\le |\hat{m}_{(v.C,u)(K-1,1)}\hat{m}_{(v.C),(2)} - m_{(v.C,u),(K-1,1)}m_{(v.C),(2)}| \\
&\quad + |\hat{m}_{(v.C),(K)}\hat{m}_{(v.C),u),(1,1)} - m_{(v),(K)}m_{(v,u),(1,1)})|.
\end{aligned} \tag{A.20}$$

Consider each of the two terms separately. For some $0 < \eta_1 < \delta_1\Phi(J,K,M,\lambda_{\min})$ and $0 < \eta_2 < \delta_1\Phi(J,K,M,\lambda_{\min})$ we have

$$\begin{aligned}
|\hat{m}_{(v.C,u),(K-1,1)}\hat{m}_{(v.C),(2)} - m_{(v.C,u),(K-1,1)}m_{(v.C),(2)}| &= \big|(m_{(v.C,u),(K-1,1)} + \eta_1)(m_{(v.C),(2)} + \eta_2) \\
&\quad - m_{(v.C,u),(K-1,1)}m_{(v.C),(2)}\big| \\
&= |(m_{(v.C,u),(K-1,1)}\eta_2 + m_{(v.C),(2)}\eta_1) + \eta_1\eta_2| \\
&\le M\eta_2 + M\eta_1 + \eta_1\eta_2 \\
&= 2M\delta_1\Phi(J,K,M,\lambda_{\min}) + \left(\delta_1\Phi(J,K,M,\lambda_{\min})\right)^2.
\end{aligned} \tag{A.21}$$

Using the analogous argument for the second term, we can bound the entire quantity as

$$|\hat{\tau}_{v.C\to u} - \tau_{v.C\to u}| < \delta_3 = 4M\delta_1\Phi(J,K,M,\lambda_{\min}) + 2\left(\delta_1\Phi(J,K,M,\lambda_{\min})\right)^2.$$

□

VITA

Yu-hsuan Samuel Wang was born in Holmdel, NJ and lived in Dallas, TX for most of his childhood. At Rice University, he was a member of Sid Richardson College and majored in Applied Math, Economics, and Policy Studies. He worked in management consulting before beginning his PhD at the University of Washington. When not thinking about structural equation models, he enjoys cycling, soccer, and attempting to cook.