PRAGMATIC CAUSAL INFERENCE

by
Rohit Bhattacharya

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
March 2021

Abstract

Data-driven causal inference from real-world multivariate systems can be biased for a number of reasons. These include unmeasured confounding, systematic censoring of observations, data dependence induced by a network of unit interactions, and misspecification of parametric models. This dissertation proposes statistical methods spanning three major steps of the causal inference workflow – *discovery* of a suitable causal model, which in our case, can be visualized via one of several classes of causal graphical models, *identification* of target causal parameters as functions of the observed data distribution, and *estimation* of these parameters from finite samples. The overarching goal of these methods is to augment the data scientist's toolkit to tackle the aforementioned challenges in real-world systems in theoretically sound yet practical ways. We provide a continuous optimization procedure for causal discovery in the presence of latent confounders, and a computationally efficient discrete search procedure for discovery and downstream estimation of causal effects in causal graphs encoding interactions between units in a network. For identification, we provide an algorithm that generalizes the state-of-the-art for recovery of target parameters in missing not at random distributions that can be represented graphically via directed acyclic graphs. Finally for estimation, we provide results on the tangent space of causal graphical models with latent variables which may be used to improve the efficiency of semiparametric estimators for any target parameter of interest. We also provide novel estimators, including influence-function based estimators, for the average causal effect of a point exposure on an outcome when there are latent variables in the system.

Committee Members

Dr. Ilya Shpitser (Primary Advisor)
John C. Malone Assistant Professor
Department of Computer Science
Johns Hopkins University

Dr. Elizabeth L. Ogburn
Associate Professor
Department of Biostatistics
Johns Hopkins University

Dr. Daniel O. Scharfstein
Professor
Department of Population Health Sciences
University of Utah

Dr. Ricardo Silva
Professor
Department of Statistical Science
University College London

Preface

As a computational genomics researcher, I learned that a principled causal analysis is often hindered by various "imperfections" in the data, ones which standard statistical tools and software are unable to correct for. For example, genomic sequencing is often performed on multiple tumors from an individual, or on multiple individuals from a single family or social network. This violates an assumption frequently employed in statistics and machine learning – that the data are independent and identically distributed. Further, the data may be censored or missing for a variety of reasons ranging from technical limitations of the instruments to complex patterns of dropout and re-enrollment in a longitudinal study, and non-response prompted by the social stigma attached to questions pertaining to drug-use or sexual orientation. Finally, it is unreasonable to expect that all sources of confounding in complex biological systems, such as those associated with cancer immunotherapy, can be accounted for. Each of these data imperfections pose a unique set of challenges to causal inference.

Data dependence, non-ignorable missingness, and unmeasured confounding, I realized, are by no means unique to the field of computational genomics. Indeed, "imperfect" data are pervasive and affect analyses across scientific disciplines, such as epidemiology, economics, and the social sciences. Since this realization, I have made it my goal to build theoretically sound, computationally feasible, and easy-to-use causal inference tools and software to analyze and correct for such understudied yet ubiquitous sources of bias. The contents of this dissertation are the start of my long journey in tackling what I consider to be the biggest challenges in data science today.

To my students – past, present, and future.

May I forever strive to be an instrument for positive change in your lives.

Acknowledgements

There are many folks that I must thank for bringing me to this moment. My advisor Ilya, for taking me on as mid-career graduate student with little to no background in causal inference, and believing in my ability to do the leg work to make novel contributions in the field (which I have now fallen in love with due in no small part to his contagious enthusiasm for it.) Members of my thesis committee Betsy, DanS, and Ricardo, for taking the time out of their busy schedules to read and lend their thoughts on my work. To Betsy and DanS especially, thank you for your mentorship, guidance, and support right from the start of my journey in causal inference to the end of my PhD and beginning of a new chapter in my life as an assistant professor. I am also grateful for all the wonderful mentors through the years who kindled and stoked my passion for teaching and research – Joanne, Sara, Peter, Tilak, Rachel, Victor, and Elsa. To Tilak, thank you for always staying in touch, even if it is to mock me regarding the latest Liverpool snafu.

To my partner Razieh, thank you for being a constant source of inspiration, support, and light in my life. The Terminator marathon, a real (half) marathon, slowly but surely working through Arnold's entire filmography, watching Iron Maiden live, watching all the Harry Potter movies, soon to have watched all the Star Wars movies, and making carrot jam (maybe some causal inference papers and faculty interviews inbetween?) – me thinks these are our greatest accomplishments over the last couple of years! There are no words to express the gratitude I feel knowing that I get to share all these moments with you – highs and lows, big and small, silly and

romantic – it's all a pleasure when it's done with you. Though our jobs may place us further apart for the moment, I am certain that exciting times lie ahead of us!

My family Bub, Mum, and Dada. Thank you for showering me with your love, and instilling in me humility and perseverance. My dad taught me maths, my brother taught me to program, and my mum taught me how to be an artist – all of which are vital to the work that I do today. I cannot thank you all enough for the lessons you have imparted, great and small, that have made me the person I am today.

My childhood friends Ajan, Shawn, Tanooj, Tushar, Varun, and Vijit. You are my ports in heavy weather, the ones who help me plug holes in this occasionally sinking ship, and whatever other maritime metaphors convey the importance of having all of you in my life. It gives me great comfort knowing that ours are bonds that shall never break, fade, nor crack. I look forward to the next "Boizgiving" that we are all together, so I can hug each and every one of you, and watch Tushar eat grapes during Korean BBQ. To my friends from undergrad, Bardia and Bedram, thank you for our shared memes, dreams, and hours playing Dota 2. And to my friends from high school, Mihir and Shreya, the brief overlap we had in Baltimore has been a lot of fun.

My House of Ayli lab mates Amir, Dan, Eli, Jaron, Noam, Numair, Ranjani, Razieh (again), and Zach. Thank you for welcoming me into the fold when I had no home, or for entertaining my shenanigans if you joined after/around the same time as me. To Dan, thank you for introducing me to structure learning, and to Jaron, thank you for co-habiting our tiny Malone space and meme-ing about the cold, the work, the Dota 2, and life in general. And to my lab mates from Karchin lab Ashok, Chris, Collin, Lily, Melody, Noushin, and Violeta, thank you for the wonderful memories!

My friends from soccer/football and excessive hours spent in the coffee room – Aaron, Alishah, Arka, Fabian, Jayant, Rachel, Teodor, and Yasamin. May our legacy live on in the CS department well past the time we have all graduated!

Finally, thank you to the city of Baltimore, which for the last 10 years of my life has been my home. I will carry a piece of you with me, wherever I may roam.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The study of qualitative and quantitative theories of causation in complex multivariate systems is a fundamental scientific endeavor. By virtue of the active nature of causal questions, e.g., "How does Y change when I *do* X?", causal reasoning also plays a key role in decision-making in the empirical sciences and public policy. Randomized controlled trials are often considered the gold standard for establishing cause-effect relations from data [2]. This is because in an ideal randomized controlled trial, probabilistic dependence between a "treatment" variable (potential cause) and its "outcome" (potential effect) due to factors other than causation are completely removed via randomization of the treatment assignment. However, for several causal questions, e.g., "Does smoking cause cancer?", running a randomized experiment is either unethical, infeasible, or too expensive. This has motivated the use of observational data to directly infer or narrow the space of feasible causal hypotheses [3, 4, 5, 6]. However, data from multivariate biological, healthcare, and socio-economic systems (even when derived from a randomized experiment) are "imperfect" in many ways. They exhibit systematically censored observations, data dependence, unmeasured confounding, and a myriad other issues that greatly complicate the task of inferring causal relationships from the data.

A data-driven causal inference workflow consists of three major tasks – *discovery* of the causal structure of the system, *identification* of target causal parameters as

functions of the observed data distribution, and *estimation* of these parameters from finite samples.[1] Data scientists have come to rely on a popular set of methods, such as imputation and covariate adjustment, as part of a standard toolkit used to accomplish various tasks in this workflow. However, issues that arise due to the complexity of real-world data result in serious violations of the underlying statistical assumptions used by these methods, which can lead to severely biased estimates. Publication and reliance on such estimates for informing public policy has contributed to the ongoing replication crisis [7, 8].

This dissertation proposes statistical methods spanning the three steps of the causal inference workflow outlined above. The overarching goal of these methods is to serve as theoretically sound and practical procedures that augment the data scientist's toolkit – enabling them to address challenges arising from systematically censored data, data dependence, unmeasured confounding, model misspecification, and finite sample estimation. Many of the methods described here have also been implemented as open-source software as part of a Python package for causal inference using graphical models called *Ananke* [9]. We hope that this serves to reduce the barrier-to-entry for the usage of these methods in standard data science workflows. The contributions of this dissertation and its organization are as follows.

- The rest of Chapter 1 introduces the necessary preliminaries and background information on causal modeling using graphical models.
- Chapters 2 and 3 describe methods for causal discovery in settings with "messy" observational data. Chapter 2 describes continuous optimization schemes for causal discovery in the presence of latent confounders based on work in [10], while Chapter 3 describes methods for causal discovery when units in the data are dependent due to an underlying network of unit interactions based on work

[1] A fourth task not covered in this dissertation is sensitivity analysis, which examines the relatonship between model assumptions and the robustness of estimates.

in [11].

- Chapter 4 describes methods for the identification of target distributions from missing not at random data when the target distribution and the missingness mechanism can be modeled via a directed acyclic graph based on work in [12].
- Chapter 5 describes methods for robust and efficient estimation of causal effects in the presence of latent confounders based on work in [13].
- Chapter 6 provides closing thoughts and concludes the dissertation.

The research for Chapters 4 and 5 was conducted jointly between the author of this dissertation and Razieh Nabi as co-first authors. Consequently, some text appearing in these chapters (and related introductory material in Chapter 1) may be similar across our dissertations. Longer proofs in each chapter are deferred to the Appendix of the dissertation.

1.1 Causal Inference Using Graphical Models

The cause-effect relationship between a treatment variable T and an outcome variable Y is typically quantified through the use of potential outcomes, a.k.a. counterfactuals. For example, the potential outcomes $Y(1)$ and $Y(0)$ may be used to represent a hypothetical randomized controlled trial where units are randomly assigned to the treatment arm (corresponding to $T = 1$), or the control arm (corresponding to $T = 0$). The *average causal effect* (ACE) is frequently used to compare the distribution of such counterfactual random variables on the mean difference scale. That is, $\text{ACE} \equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. More generally, one could define a random variable $Y(t)$[2] corresponding to the potential outcome had treatment T been *assigned* to some value t. This allows for the contrast of arbitrary treatment assignments t and t' as $\mathbb{E}[Y(t)] - \mathbb{E}[Y(t')]$.

[2]Or equivalently, $Y \mid \text{do}(t)$ in the do-calculus notation [5].

It is well understood that causal parameters (such as the ACE) cannot be expressed as functions of the observed data, or in other words are not *identified*, if no assumptions are made about the data generating process [5]. The use of graphical models to facilitate causal inference by encoding assumptions about the data generating process in a visual and intuitive fashion has gained traction across scientific disciplines [14, 15, 16]. The following subsections describe various graphical representations of causal models that appear in this dissertation as well as their intended use cases.

1.1.1 Causal Modeling with Directed Acyclic Graphs

Formally, a *directed acyclic graph* (DAG) $\mathcal{G}(V)$ is a set of nodes V connected by directed edges, such that there are no directed cycles. The *statistical model* of a DAG $\mathcal{G}(V)$, denoted by $\mathcal{M}^{\text{DAG}}(\mathcal{G})$, is the set of distributions that can be expressed as the following product of conditional densities,

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)), \qquad \textit{(DAG factorization)} \tag{1.1}$$

where $\text{pa}_{\mathcal{G}}(V_i)$ are the parents of V_i in $\mathcal{G}$.

Causal models are sets of distributions defined over counterfactual random variables. Causal models of a DAG $\mathcal{G}(V)$ are defined over counterfactual random variables $V_i(\text{pa}_i)$ for each $V_i \in V$, where pa_i is a set of values for $\text{pa}_{\mathcal{G}}(V_i)$. Alternatively, these counterfactuals can be viewed as being determined by a system of *structural equations* $f_i(\text{pa}_i, \epsilon_i)$ that map values pa_i as well as values of an exogenous noise term ϵ_i to values of V_i [5, 17]. Other counterfactuals may be defined from above via recursive substitution. Specifically, for any set $A \subseteq V$, and a variable V_i, we have:

$$V_i(a) \equiv V_i\Big(a \cap \text{pa}_{\mathcal{G}}(V_i), \{V_j(a) : V_j \in \text{pa}_{\mathcal{G}}(V_i) \setminus A\}\Big). \qquad \textit{(Recursive substitution)} \tag{1.2}$$

For any set $A \subset V$, we may define a joint distribution of potential outcomes $p(\{V \setminus A\}(a))$, or $p(V(a))$ for short, by applying recursive substitution as in Eq. 1.2.

Any such counterfactual distribution $p(V(a))$ is identified in causal models of a DAG $\mathcal{G}$ via the g-formula (a.k.a truncated factorization or manipulated distribution) [3, 4, 5],

$$p(V(a)) = \prod_{V_i \in V \setminus A} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)) \Big|_{A=a}. \qquad \textit{(g-formula)} \tag{1.3}$$

When A is the empty set, the above functional returns the DAG factorization in Eq. 1.1, meaning that the causal model of a DAG $\mathcal{G}$ also implies its statistical model.

Under the causal DAG framework, the presence of an edge $V_i \to V_j$ should be interpreted as saying "V_i is a *potential* direct cause of V_j." The absence of an edge between V_i and V_j not only implies the absence of a direct causal relation, but also conditional independences in the probability distribution $p(V)$. These independences can be read directly from $\mathcal{G}$ via the well-known d-separation criterion [18, 5]. That is, for disjoint sets X, Y, Z, the following *global Markov property* holds $(X \perp\!\!\!\perp_{\text{d-sep}} Y \mid Z)_{\mathcal{G}} \implies (X \perp\!\!\!\perp Y \mid Z)_{p(V)}$. When the context is clear, we drop the explicit reference to $p(V)$ and simply use $X \perp\!\!\!\perp Y \mid Z$ to denote conditional independence between X and Y given Z.

1.1.2 Causal Modeling in the Presence of Latent Confounders with Acyclic Directed Mixed Graphs

For many practical applications, causal models must account for the presence of latent variables. Reasoning directly with latent variable DAG models may lead to statistical issues as parameterizations of such models are generally not fully identifiable and may contain singularities [19]. Further, the use of latent variable models may require positing assumptions on the number of latent variables, their state-space, and relations between latent and observed variables. This can be difficult to do by the very nature of latent variables being unobserved, which increases the chance of model misspecification and biased estimation of target parameters.

Acyclic directed mixed graph (ADMGs) were proposed as an alternative graphical representation for modeling relations among just the observed variables in the problem,

while preserving much of the causal and statistical information from the original latent variable DAG model [20, 21, 22]. Formally, the latent projection of a hidden variable DAG $\mathcal{G}(V \cup H)$ onto observed variables V is an ADMG $\mathcal{G}(V)$ with directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges constructed as follows. The edge $V_i \rightarrow V_j$ exists in $\mathcal{G}(V)$ if there exists a directed path from V_i to V_j in $\mathcal{G}(V \cup H)$ with all intermediate vertices in H. An edge $V_i \leftrightarrow V_j$ exists in $\mathcal{G}(V)$ if there exists a collider-free path (i.e., there are no consecutive edges of the form $\rightarrow \circ \leftarrow$) from V_i to V_j in $\mathcal{G}(V \cup H)$ with all intermediate vertices in H, such that the first edge on the path is an incoming edge into V_i and the final edge is an incoming edge into V_j [23]. An example of applying the latent projection operator is provided in Figure 1-1.

It was shown by [24] that all non-parametric equality constraints in the observed data distribution $p(V)$ implied by Markov restrictions in the latent variable DAG model given by the factorization $p(V \cup H)$ with respect to $\mathcal{G}(V \cup H)$ are captured via *nested Markov models* of an ADMG. Further, identification of many causal parameters, such as the ACE, may be rephrased without loss of generality in terms of truncated functionals of the nested Markov factorization of the observed data distribution. [22, 12]. The following subsections describe the background necessary to describe the nested Markov factorization as well as some coarser factorizations of ADMG models used in Chapter 5 of this dissertation.

We adopt the following conventions in denoting standard genealogical relations in an ADMG $\mathcal{G}(V)$. The parents of a set of vertices S is defined as the set of parents of each vertex in S not contained in S, i.e., $\text{pa}_{\mathcal{G}}(S) \equiv \bigcup_{S_i \in S} \text{pa}_{\mathcal{G}}(S_i) \setminus S$. We follow the same convention for children of a set S, denoted $\text{ch}_{\mathcal{G}}(S)$. Other standard genealogical relations, such as ancestors $\text{an}_{\mathcal{G}}(V_i) \equiv \{V_j \in V \mid \exists\ V_j \rightarrow \cdots \rightarrow V_i \text{ in } \mathcal{G}\}$ and descendants $\text{de}_{\mathcal{G}}(V_i) \equiv \{V_j \in V \mid \exists\ V_i \rightarrow \cdots \rightarrow V_j \text{ in } \mathcal{G}\}$, include the vertex V_i itself by convention. The extension of these relations to a set S then uses the disjunctive definition which also includes the set itself. For example, $\text{an}_{\mathcal{G}}(S) = \bigcup_{S_i \in S} \text{an}_{\mathcal{G}}(S_i)$.

(a) $\mathcal{G}(V \cup H)$ (b) $\mathcal{G}(V)$

Figure 1-1. (a) A hidden variable DAG $\mathcal{G}(V \cup H)$; (b) The ADMG $\mathcal{G}(V)$ obtained via latent projection.

District and Topological Factorization of ADMGs

We first describe the district factorization of the observed distribution $p(V)$ with respect to an ADMG $\mathcal{G}(V)$ which bears resemblance to the more familiar DAG factorization (when all variables are observed) in the following sense. The district factorization of $p(V)$ with respect to an ADMG $\mathcal{G}(V)$ is a product of objects that resemble (but are not necessarily equal to) conditional densities defined over sets of variables given their parents. These objects are formally referred to as *kernels* and these sets of variables are known as *districts*. We also describe how terms that appear in the district factorization can be expressed as a product of ordinary conditional densities in $p(V)$ via the topological factorization of the joint distribution.

A *district* is defined as a bidirected connected component of an ADMG $\mathcal{G}(V)$, i.e., an induced subgraph of $\mathcal{G}$ in which any two vertices are connected to each other via at least one bidirected path. Districts of $\mathcal{G}(V)$ form a partitioning of its vertices. We use $\text{dis}_{\mathcal{G}}(V_i)$ to denote the district in $\mathcal{G}$ that contains V_i and $\mathcal{D}(\mathcal{G})$ to denote the set of all districts in $\mathcal{G}$.

For disjoint subsets X, Y, a *kernel* $q_X(X \mid Y)$ is defined as a mapping from values in Y to normalized densities over X [25]. That is, kernels behave like conditional densities in the sense that $\sum_X q_X(X \mid Y = y) = 1, \forall y$ in the state space of Y. For any $Z \subset V$, marginalization and conditioning in a kernel are defined as $q_{X \setminus Z}(V \setminus Z \mid$

$Y) \equiv \sum_Z q_X(X \mid Y)$ and $q_X(X \setminus Z \mid Z, Y) \equiv \frac{q_X(X|Y)}{q_Z(Z|Y)}$ respectively; more details and examples of such operations can be found in Appendix A.

Given a latent variable DAG $\mathcal{G}(V \cup H)$ and corresponding distribution $p(V \cup H)$, [26] showed that for each district D in the latent projection ADMG $\mathcal{G}(V)$, the kernel mapping values of the parents of D to normalized densities over D, denoted by $q_D(D \mid \text{pa}_\mathcal{G}(D))$, can be expressed as functionals of the observed distribution $p(V)$ as follows. Define the *Markov blanket* of a vertex V_i as the district of V_i and the parents of its district, excluding V_i itself, i.e., $\text{mb}_\mathcal{G}(V_i) = \text{dis}_\mathcal{G}(V_i) \cup \text{pa}_\mathcal{G}(\text{dis}_\mathcal{G}(V_i)) \setminus \{V_i\}$. Consider a valid topological order τ on all k vertices in V, that is a sequence $(V_1, \ldots, V_k)$ such that no vertex appearing later in the sequence is an ancestor of vertices earlier in the sequence. Let $\{\preceq_\tau V_i\}$ denote the set of vertices that precede V_i in this sequence, including V_i itself. Then for each $D \in \mathcal{D}(\mathcal{G})$,

$$q_D(D \mid \text{pa}_\mathcal{G}(D)) = \prod_{D_i \in D} p(D_i \mid \text{mp}_\mathcal{G}(D_i)), \qquad \textit{(Identification of district-level kernels)} \tag{1.4}$$

where $\text{mp}_\mathcal{G}(V_i)$, the *Markov pillow* of V_i, is defined as its Markov blanket in a subgraph restricted to V_i and its predecessors according to the topological ordering. More formally, $\text{mp}_\mathcal{G}(V_i) \equiv \text{mb}_{\mathcal{G}_S}(V_i)$ where $S = \{\preceq_\tau V_i\}$, and $\mathcal{G}_S$ is the subgraph of $\mathcal{G}$ that is restricted to vertices in S and the edges between these vertices.

The joint distribution $p(V)$ then satisfies the following district-level factorization with respect to the latent projection ADMG $\mathcal{G}(V)$ [26],

$$p(V) = \prod_{D \in \mathcal{D}(\mathcal{G})} q_D(D \mid \text{pa}_\mathcal{G}(D)), \qquad \textit{(District ADMG factorization)} \tag{1.5}$$

Equations 1.4 and 1.5 together lead to a factorization of the observed distribution as a product of simple conditional factors according to a valid topological order,

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{mp}_\mathcal{G}(V_i)). \qquad \textit{(Topological ADMG factorization)} \tag{1.6}$$

The above factorization (and the district factorization from which it is derived), does not always capture every equality restriction in $p(V)$ implied by the Markov property of the underlying hidden variable DAG $\mathcal{G}(V \cup H)$. However, it is particularly simple to work with, and under some conditions, which we derive in Chapter 5, is capable of capturing all such restrictions.

Nested Markov Models of Acyclic Directed Mixed Graphs

In addition to the district factorization in the previous subsection, the nested Markov factorization of the joint $p(V)$ with respect to an ADMG $\mathcal{G}(V)$ asserts that certain kernels associated with sets of vertices known as *reachable sets* may be expressed as products of a base set of kernels associated with sets of vertices known as *intrinsic sets*. This motivates the need for an intermediate graphical representation, known as *conditional ADMGs* (CADMGs), that can be used to visualize and operationalize the derivation of kernels associated with the reachable and intrinsic sets used to describe the nested Markov model of an ADMG.

Conditional ADMGs (CADMGs) $\mathcal{G}(V, W)$ are acyclic directed mixed graphs whose vertices can be partitioned into random variables V and *fixed* variables W with the restriction that variables in W may only have outgoing directed edges [22]. Fixed variables are also not considered to be part of any districts in $\mathcal{G}(V)$. However, the definitions of genealogic sets such as parents, descendants, and ancestors as well as other special sets such as the Markov blanket and Markov pillow of random variables $V_i \in V$ in a CADMG $\mathcal{G}(V, W)$, extend naturally by allowing for the inclusion of fixed variables into these sets as given by their standard definitions.

A vertex $V_i \in V$ is said to be *fixable* in a CADMG $\mathcal{G}(V, W)$ if $\mathrm{dis}_{\mathcal{G}}(V_i) \cap \mathrm{de}_{\mathcal{G}}(V_i) = \{V_i\}$. The graphical operation of fixing V_i denoted by $\phi_{V_i}(\mathcal{G})$, yields a new CADMG $\mathcal{G}(V \setminus V_i, W \cup V_i)$ where bidirected and directed edges into V_i are removed and V_i is fixed to a particular value v_i. Given a kernel $q_V(V \mid W)$ associated with the CADMG

$\mathcal{G}(V, W)$, the corresponding *probabilistic* operation of fixing, denoted by $\phi_{V_i}(q_V; \mathcal{G})$, yields a new kernel

$$\begin{aligned}\phi_{V_i}(q_V; \mathcal{G}) &\equiv q_{V \setminus V_i}(V \setminus V_i \mid W \cup V_i) \\ &\equiv \frac{q_V(V \mid W)}{q_V(V_i \mid \mathrm{mb}_{\mathcal{G}}(V_i), W)}. \qquad \textit{(Probabilistic fixing operator)} \end{aligned} \tag{1.7}$$

The notion of fixability can be extended to a set of vertices S by requiring that there exists an ordering $(S_1, \ldots, S_p)$ such that S_1 is fixable in $\mathcal{G}$, S_2 is fixable in $\phi_{S_1}(\mathcal{G})$ and so on. Such an ordering is said to be a *valid* fixing sequence and the set $V \setminus S$ is said to be *reachable*. It is known that any two valid fixing sequences on S yield the same CADMG, which we will denote by $\phi_S(\mathcal{G}(V, W))$. Fix a CADMG $\mathcal{G}(V, W)$ and a corresponding kernel $q(V \mid W)$. Given a valid fixing sequence σ_S on $S \subseteq V$ valid in $\mathcal{G}(V, W)$, define $\phi_{\sigma_S}(q_V; \mathcal{G})$ inductively to be $q(V \mid W)$ when S is empty, and $\phi_{\sigma_S \setminus S_1}(\phi_{S_1}(q_V; \mathcal{G}); \phi_{S_1}(\mathcal{G}))$ otherwise, where $\sigma_S \setminus S_1$ corresponds to the remainder of the sequence after S_1. A concrete example demonstrating sequential applications of the graphical and probabilistic operations of fixing can be found in Appendix A.

The nested Markov factorization of an ADMG $\mathcal{G}$ relies on the notion of *intrinsic* sets. A set $S \subseteq V$ is said to be intrinsic in $\mathcal{G}$ if S is reachable and $\phi_{V \setminus S}(\mathcal{G})$ contains a single district. The set of intrinsic sets of $\mathcal{G}$ is denoted by $\mathcal{I}(\mathcal{G})$. A distribution $p(V)$ is then said to obey the nested Markov factorization relative to an ADMG $\mathcal{G}(V)$ if for every fixable set S and every valid fixing sequence σ_S,

$$\phi_{\sigma_S}(p(V); \mathcal{G}) = \prod_{D \in \mathcal{D}(\phi_S(\mathcal{G}))} q_D(D \mid \mathrm{pa}_{\mathcal{G}}(D)), \qquad \textit{(Nested Markov factorization)} \tag{1.8}$$

where all kernels appearing in the product above can be constructed from the set of kernels corresponding to intrinsic sets in $\mathcal{G}$, i.e., $\{q_S(S \mid \mathrm{pa}_{\mathcal{G}}(S)) \mid S \in \mathcal{I}(\mathcal{G})\}$. If $p(V)$ obeys the nested Markov factorization, then for any fixable set S, applying any two distinct valid sequences σ_S^1, σ_S^2 to $p(V)$ and $\mathcal{G}(V)$ also yields the same kernel, which we define as $\phi_S(p(V); \mathcal{G}(V))$. Moreover, for every $D \in \mathcal{I}(\mathcal{G})$, $q_D(D \mid \mathrm{pa}_{\mathcal{G}}(D)) =$

$\phi_{V\setminus D}(p(V);\mathcal{G}(V))$. The nested Markov factorization above defines the *nested Markov model*, with associated Markov properties, described in [22].

In the previous subsection we saw that counterfactual distributions in the fully observed setting could be thought of as truncated functionals of the statistical DAG factorization given by the g-formula (Eq. 1.3). In an analogous fashion, identified counterfactual distributions over observed variables in a latent variable DAG can be given by truncated functionals of the nested Markov factorization. This strategy for identification counterfactual distributions in a latent variable DAG model is known to be sound and complete [27, 28, 22]. That is, identification of a counterfactual distribution $p(Y(a))$ in a hidden variable causal model associated with a DAG $\mathcal{G}(V\cup H)$ may be rephrased, without loss of generality, using its corresponding latent projection ADMG $\mathcal{G}(V)$. Specifically, for $Y^* \equiv \text{an}_{\mathcal{G}_{V\setminus T}}(Y)$,

$$p(Y(a)) = \sum_{Y^*\setminus Y} \prod_{D\in\mathcal{D}(\mathcal{G}_{Y^*})} \phi_{V\setminus D}(p(V);\mathcal{G}(V))\Big|_{T=t}, \quad \textit{(Truncated nested Markov factorization)} \tag{1.9}$$

provided every $D \in \mathcal{D}(\mathcal{G}_{Y^*})$ is intrinsic; otherwise, $p(Y(a))$ is not identifiable [22]. Recall from the definition of the nested Markov model that a set $S \subseteq V$ is said to be intrinsic in $\mathcal{G}$ if $V \setminus S$ is fixable, and $\phi_{V\setminus S}(\mathcal{G})$ contains a single district.

Implications of Presence and Absence of Edges in an ADMG

When viewing an ADMG $\mathcal{G}(V)$ as a graphical representation of the observed margin of a latent variable causal DAG $\mathcal{G}(V \cup H)$, an edge $V_i \to V_j$ in the ADMG may be interpreted as saying V_i is a potential direct cause of V_j, and an edge $V_i \leftrightarrow V_j$ may be interpreted as the presence of one or more latent confounders, e.g., $V_i \leftarrow H_k \to V_j$, between V_i and V_j in $\mathcal{G}(V \cup H)$. The absence of edges in ADMG $\mathcal{G}(V)$ imply the absence of a causal relation/unmeasured confounding, and *may* imply restrictions in the observed distribution $p(V)$. Conditional independences in $p(V)$ can be read off from the ADMG $\mathcal{G}(V)$ by a simple analogue of the d-separation criterion, known

as m-separation, that generalizes the notion of a collider to include mixed edges of the form $\rightarrow \circ \leftrightarrow$, $\leftrightarrow \circ \leftarrow$, and $\leftrightarrow \circ \leftrightarrow$, [29]. Sometimes, the absence of an edge may not imply any conditional independence statements on $p(V)$ but rather generalized equality restrictions, informally referred to as Verma constraints, that resemble ordinary conditional independences albeit in post-intervention distributions [3, 23]. At present, there exists no graphical criterion to read all non-parametric generalized equality restrictions implied in $p(V)$ directly from $\mathcal{G}(V)$. [30] provides a recursive algorithm that outputs a list of non-parametric equality constraints that comprise the nested Markov model and [22] rephrase this in terms of the fixing operator. It may also be the case that the absence of an edge implies no equality restrictions on the observed distribution $p(V)$. In Chapter 5 we provide a sound and complete algorithm to determine the existence of non-parametric equality restrictions in the nested Markov model of an ADMG $\mathcal{M}^{\text{nested}}(\mathcal{G})$. We use this algorithm to decide the statistical efficiency of estimators for causal effects we present in Chapter 5.

1.1.3 Causal Modeling of Network Data with Chain Graphs

Classical causal and statistical inference methods typically assume the observed data consists of independent realizations. However, in many applications this assumption is inappropriate due to a network of dependences between units in the data. For example, the COVID-19 global pandemic serves as a stark reminder that a disease may spread within a social network of friends, family, or neighbors by a causal process known as contagion, and that vaccinating a subset of the population may confer immunity to a larger subset than those who were vaccinated (a phenomenon known as herd immunity.) Further, the limited availability of beds in intensive care units can also induce data dependence when patients are triaged before admission. That is, the allocation of a bed in the intensive care unit to one individual not only improves their own chance of survival, but also has a harmful "spillover effect" on the outcomes of

other individuals to whom the bed was not given. Data dependence is not limited to the infectious disease setting and may appear, for example, in classroom studies on the efficacy of new pedagogical methods, and studies on the spread of misinformation on social media, in the form of peer-to-peer influence and homophily [31].

Recently, chain graphs (CGs) have been used to model causal phenomena that result in data dependence [32, 33, 34]. Formally, a chain graph is a mixed graph consisting of directed ($\rightarrow$) and undirected ($-$) edges, such that it is impossible to create a directed cycle by orienting any combination of the undirected edges [25]. A CG with no undirected edges is simply a DAG. The variables in a CG can be partitioned into subsets of undirected connected components known as *blocks*, which play a central role in defining causal and statistical models of CGs. The set of all blocks in a CG $\mathcal{G}$ is denoted by $\mathcal{B}(\mathcal{G})$.

In this dissertation we consider statistical and causal models of a chain graph under the Lauritzen-Wermuth-Frydenberg (LWF) interpretation. For a subset of vertices $S \subseteq V$, let $\mathcal{G}_S$ denote the induced subgraph of $\mathcal{G}$ with vertices S and edges in $\mathcal{G}$ whose endpoints are both in S. Given a CG $\mathcal{G}_S$, define the *augmented graph* $\mathcal{G}_S^a$ (sometimes referred to as the moralized graph) to be an undirected graph constructed from $\mathcal{G}_S$ by replacing all directed edges with undirected edges and connecting all vertices in $\mathrm{pa}_{\mathcal{G}_S}(B)$ for every block B in $\mathcal{G}_S$ by undirected edges [25]. Let $\mathcal{C}(\mathcal{G}_S^a)$ denote the set of all cliques in the augmented graph $\mathcal{G}_S^a$, where a clique C is defined as a maximal set of vertices that are pairwise connected by undirected edges. The statistical LWF model of a CG $\mathcal{G}(V)$, denoted $\mathcal{M}^{\mathrm{CG}}(\mathcal{G})$ is then the set of distributions that satisfy the following two-level factorization with respect to $\mathcal{G}$,

$$p(V) = \prod_{B \in \mathcal{B}(\mathcal{G})} p(B \mid \mathrm{pa}_{\mathcal{G}}(B)), \qquad \Big(\textit{CG factorization (i)}\Big) \tag{1.10}$$

and for each block

$$p(B \mid \mathrm{pa}_{\mathcal{G}}(B)) = \frac{\prod_{C \in \left\{ c\left(\mathcal{G}^{a}_{\mathrm{bd}_{\mathcal{G}}(B)}\right) \,:\, C \not\subseteq \mathrm{pa}_{\mathcal{G}}(B) \right\}} \kappa_C(C)}{Z\Big(\mathrm{pa}_{\mathcal{G}}(B)\Big)}, \qquad \Big(\textit{CG factorization (ii)}\Big) \tag{1.11}$$

where each $\kappa_C(C)$ is a non-negative clique potential function and $Z(\mathrm{pa}_{\mathcal{G}}(B))$ is a normalizing function.

Causal models associated with DAGs have been generalized to causal models associated with CGs as follows. The causal interpretation of LWF CGs may be understood as equilibria of dynamic models with feedback [35]. Under this interpretation, the distribution $p(B \mid \mathrm{pa}_{\mathcal{G}}(B))$ for each block $B \in \mathcal{B}(\mathcal{G})$ can be determined by a Gibbs sampler [36] on the variables $B_i \in B$. Here, each conditional distribution $p(B_i \mid B \setminus B_i, \mathrm{pa}_{\mathcal{G}}(B))$ is produced by structural equation of the form $f_{B_i}(B \setminus B_i, \mathrm{pa}_{\mathcal{G}}(B), \epsilon_{B_i})$. Interventions on elements of B are defined by replacing the appropriate line in the Gibbs sampler program. Under this interpretation, [35] showed that for any set $A \subset V$, the distribution of $p(V(a))$ is identified by a CG version of the DAG g-formula which can be interpreted as truncation of the two-level CG factorization as follows,

$$p(V(a)) = \prod_{B \in \mathcal{B}(\mathcal{G})} p(B \setminus A \mid \mathrm{pa}_{\mathcal{G}}(B), B \cap A) \Big|_{A=a}, \qquad \textit{(CG g-formula)} \tag{1.12}$$

and for each block

$$p(B \setminus A \mid \mathrm{pa}_{\mathcal{G}}(B), B \cap A) = \frac{\prod_{C \in \left\{ c\left(\mathcal{G}^{a}_{\mathrm{bd}_{\mathcal{G}}(B)}\right) \,:\, C \not\subseteq \mathrm{pa}_{\mathcal{G}}(B) \right\}} \kappa_C(C)}{Z\Big(\mathrm{pa}_{\mathcal{G}}(B), B \cap A\Big)} \tag{1.13}$$

where each $\kappa_C(C)$ is a clique potential function and $Z\Big(\mathrm{pa}_{\mathcal{G}}(B), B \cap A\Big)$ is a normalizing function as before. As was the case with causal and statistical models of a DAG, the causal model of a CG implies its statistical model by considering $A = \emptyset$.

Under the causal LWF CG framework, the presence of an edge $V_i \to V_j$ can be interpreted as encoding direct causation, while the presence of an edge $V_i - V_j$ can be interpreted as encoding symmetric causal relations induced by Gibbs equilibria as

discussed above. Similar to DAGs, the absence edges in a CG $\mathcal{G}$ imply conditional independences in the probability distribution $p(V)$. These independences can be read directly from $\mathcal{G}$ via the c-separation criterion or the equivalent augmentation criterion [37, 25]. That is, for disjoint sets $X, Y,$ and Z, the following *global Markov property* holds $(X \perp\!\!\!\perp_{\text{c-sep}} Y \mid Z)_{\mathcal{G}} \implies (X \perp\!\!\!\perp Y \mid Z)_{p(V)}$.

Finally, if only interventions on entire blocks are of interest, i.e., we only consider treatment assignments $A = a$ where elements of A consist of entire blocks in $\mathcal{B}(\mathcal{G})$, then there exists an alternative causal interpretation of a CG $\mathcal{G}$ that does not rely on the Gibbs sampler machinery of [35]. Specifically, in such a case we consider a causal DAG model where each block B corresponds to a supervariable V_B defined as a Cartesian product of variables in B, and a DAG causal model is defined on $V_B(a)$, where a are values assigned to the parents of V_B in $\mathcal{G}$. If for each block B in a CG $\mathcal{G}$, the graph $\mathcal{G}^a_{\text{bd}_{\mathcal{G}}(B)}$ has a single clique, then this yields a classical causal model of a DAG. If not, we can still view the model as a classical causal model of a DAG, but with an extra restriction that the observed data distribution factorizes as Equations 1.10 and 1.11 above; see also [34] for a perspective on interpreting chain graphs in the interference and dependent data setting. The methods we present in Chapter 3.3 for model selection of LWF CG models in the presence of data dependence induced by unit interactions in a network are agnostic to the choice of causal interpretation.

Chapter 2

Causal Discovery Under Unmeasured Confounding

Biological, economic, and social systems are often affected by unmeasured (latent) variables. In such scenarios, statistical and causal models of a directed acyclic graph (DAG) over the observed variables do not faithfully capture the underlying causal process. Chapter 1 introduced ADMG models as a principled alternative for modeling causal and statistical relations over the observed variables in confounded systems.

There are three popular classes of ADMGs that we discuss in this chapter – ancestral, arid, and bow-free ADMGs. Each class is suited to capturing information on the observed distribution at different levels of granularity while providing certain trade-offs in terms of statistical or computational benefits. We first briefly summarize the graphical criteria that characterize each of these classes and then describe their advantages/disadvantages.

An ADMG $\mathcal{G} = (V, E)$ is said to be *ancestral* if for any pair of vertices $V_i, V_j \in V$, a directed path $V_i \to \cdots \to V_j$ and bidirected edge $V_i \leftrightarrow V_j$ do not both appear in $\mathcal{G}$ [21]. An ADMG $\mathcal{G}$ is said to be *arid* if it does not contain any *c-trees* [38, 39]. A c-tree is a subgraph of $\mathcal{G}$ whose directed edges form an arborescence (the directed graph analogue of a tree) and bidirected edges form a single bidirected connected component within the subgraph. It is easy to confirm that the ADMG in Figure 2-1(b)

is ancestral while the one in Figure 2-1(c) is arid but not ancestral. An ADMG is called *bow-free* if for any pair of vertices, $V_i \to V_j$ and $V_i \leftrightarrow V_j$ do not both appear in $\mathcal{G}$ [40]. A graph that is bow-free but neither arid nor ancestral is displayed in Figure 2-1(e). The relation between these graph classes is the following:

$$\text{Ancestral} \subset \text{Arid} \subset \text{Bow-free}$$

Prior work in causal discovery in confounded systems has focused on discrete search procedures for selecting ancestral ADMGs via their associated *ordinary* Markov models [29] that encode ordinary conditional independence constraints among the observed variables of the system [4, 41, 42]. However, as we have seen in Chapter 1, confounded systems may also exhibit more general non-parametric equality restrictions, many of which are only captured via the nested Markov factorization with respect to more general classes of ADMGs [23, 30, 3]. While the general class of unrestricted ADMGs capture all such equality constraints [24], the associated parametric models (for the nested Markov models described in Chapter 1) are not guaranteed to form smooth curved exponential families with globally identifiable parameters – an important pre-condition for score-based model selection. A smooth parameterization for arbitrary ADMGs is known only when all observed variables are either binary or discrete [43]. For the common scenario when the data comes from a linear Gaussian system of structural equations, the statistical model of an ADMG is almost-everywhere identified if the ADMG is bow-free [40], and is globally identified and forms a smooth curved exponential family if and only if the ADMG is arid [38, 39]. From a causal perspective, arid and bow-free ADMGs, like ancestral ADMGs, have the desirable property of preserving ancestral relationships in the underlying latent variable DAG, while also capturing all equality restrictions on the observed margin [39]. While global identifability of arid ADMG models is highly appealing, the search space of bow-free ADMGs is computationally simpler to work with, and is often sufficient for accurate causal discovery despite their weaker guarantee of identifiability.

This chapter introduces a structure learning procedure for selecting arid, bow-free, or ancestral ADMGs from observational data. Our learning approach is based on reformulating the usual discrete combinatorial search problem into a more tractable constrained continuous optimization program. Such a reformulation was first proposed by [44] for the special case when the search space is restricted to DAGs. Subsequent extensions such as [45], [46], and [47] also restrict the search space in a similar fashion. In this work, we derive differentiable algebraic constraints on the adjacency matrices of the directed and bidirected portions of an ADMG that fully characterize the space of arid ADMGs. We also derive similar algebraic constraints that characterize the space of ancestral and bow-free ADMGs that are quite useful in practice and connect our work to prior methods. Having derived these differentiable constraints, we select the best fitting graph in the class by optimizing a penalized likelihood-based score. While the constraints we derive in this chapter are non-parametric, we focus our discovery methods on distributions that arise from linear Gaussian systems of equations.

Our structure learning procedure for arid and ancestral graphs is consistent in the following sense: asymptotically, convergence to the global optimum implies that the corresponding ADMG is either the true model or one that belongs to the same equivalence class. That is, if the optimization procedure succeeds in finding the global optimum, the resulting graph is either the true underlying structure or one that implies the same set of equality constraints on the observed data. While the L_0-regularized objective we propose is non-convex and so our optimization scheme may result in local optima, we show via experiments and application to protein expression data that our proposal works quite well in practice. It is important to note that optimizing non-convex objectives is unavoidable when pursuing score-based causal discovery for ADMGs as the likelihood of ADMG models is non-convex in general. We believe the proposed algebraic constraints are valuable for further research at the intersection of non-convex optimization techniques for L_0-regularization and causal discovery.

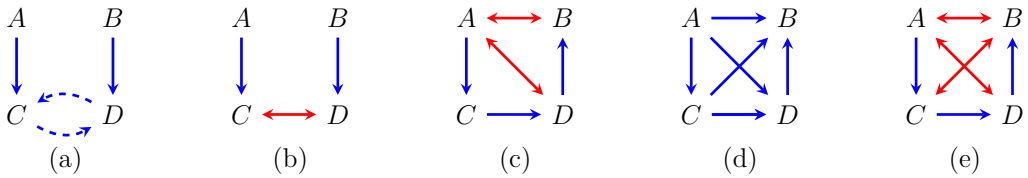


Figure 2-1. (a) A DAG if $C \to D$ or $D \to C$ exists but not both; (b) An ADMG that posits an unmeasured confounder between C and D; (c) An (arid) ADMG encoding a Verma constraint between C and B; (d) The ancestral version of (c); (e) A non-arid bow-free ADMG that is a super model of (c).

The rest of Chapter 2 is organized as follows. We begin with a motivating example and background on the structure learning problem for partially-observed systems. We then derive differentiable algebraic constraints that characterize arid, bow-free, and ancestral ADMGs. We then use these to formulate the first (to our knowledge) tractable method for learning arid ADMGs from observational data, by extending the continuous optimization scheme of causal discovery. Simply by modifying the constraint in the optimization program, the same procedure may also be leveraged to learn bow-free or ancestral graphs. Finally we evaluate the performance of our algorithms in simulation experiments and on protein expression data from [1].

2.1 Motivating Example

To motivate our work, we present an example of how our method may be used to reconstruct complex interactions in a network of genes, which is related to the data application we present towards the end of the chapter.

Consider a scenario in which an analyst has access to gene expression data on four genes: A, B, C, and D. Assume that the analyst is confident (due to prior analysis or background knowledge) about the structure corresponding to non-dashed edges shown in Figure 2-1(a), i.e., that A regulates C and B regulates D but A and B are independent. This leaves an important ambiguity regarding regulatory explanations of co-expression of genes C and D.

An observed correlation between C and D may be explained in different ways that provide very different mechanistic interpretations. If the hypothesis class is restricted to DAGs, the only explanations available to the analyst are that C is a cause of D or vice-versa as shown in Figure 2-1(a). If the analyst proceeds with either of these explanations and performs a gene-knockout experiment where C (or D) is removed but sees no change in D (respectively C), then the causal DAG fails to be a faithful representation of the true underlying mechanism. The correlation may instead be explained by an ADMG as in Figure 2-1(b) where $C \leftrightarrow D$ indicates that C and D are dependent due to the presence of at least one unmeasured confounding gene that regulates both of them. That is, if we had data on these unmeasured genes U the corresponding DAG would have contained a structure $C \leftarrow U \rightarrow D$. However, given observations only on A, B, C, D, Figure 2-1(b) provides a faithful representation of the underlying mechanism on the observed variables. It correctly encodes that intervention on C or D has no downstream effects on the other.

Importantly, each of these different explanations are not just different from a mechanistic point of view but also imply different independence restrictions on the observed data. The two DAGs in Figure 2-1(a) imply that $A \perp\!\!\!\perp D \mid C$ or $B \perp\!\!\!\perp C \mid D$ respectively, whereas Figure 2-1(b) implies $A \perp\!\!\!\perp D$ and $B \perp\!\!\!\perp C$. Hence, a causal discovery procedure that seeks the best fitting structure from the hypothesis class of ADMGs, will be able to distinguish between these different explanations and choose the correct one.

Some mechanisms, such as the one shown in Figure 2-1(c), are not distinguishable using ordinary conditional independence statements alone. In this graph, the only pair of genes with no edge between them is B and C. The absence of this edge implies that C does not directly regulate the expression of B and only does so through D. This missing edge does not correspond to any ordinary conditional independence (there are no independence constraints implied by the model at all), but does encode a Verma

constraint, namely that $B \perp\!\!\!\perp C \mid D$ in a re-weighed distribution derived from the joint, $p(A, B, C, D)/p(C \mid A)$.

Ancestral ADMGs can "hide" certain important information because they encode only ordinary conditional independence constraints. An ancestral ADMG that encodes the same ordinary independence constraints as the arid graph in Figure 2-1(c) is shown in Figure 2-1(d). It is a complete graph since there are no conditional independence constraints in Figure 2-1(c). That is, the absence of any $C \rightarrow B$ edge in Figure 2-1(c) is "masked" to preserve the ancestrality property. We can potentially learn a more informative structure if we do not limit our hypothesis class to the class of ancestral graphs.

2.2 Graphical Interpretation of Linear SEMs

In Chapter 1, we saw that causal models of a DAG may be interpreted as a system of structural equations that give rise to counterfactual random variables. In this section, we review linear structural equation models (linear SEMs) and their graphical representations. Such representations date back to the seminal work of geneticist Sewall Wright close to a century ago [48, 49]. We make use of the following standard matrix notation: A_{ij} refers to the element in the i^{th} row and j^{th} column of a matrix A, indexing $A_{-i,-j}$ refers to the sub matrix obtained by excluding the i^{th} row and j^{th} column of A, and $A_{:,i}$ refers to the i^{th} column of A.

2.2.1 Linear SEMs and DAGs

Consider a linear SEM on d variables parameterized by a weight matrix $\theta \in \mathbb{R}^{d \times d}$. For each variable $V_i \in V$, we have a structural equation

$$V_i \leftarrow \sum_{V_j \in V} \theta_{ji} V_j + \epsilon_i,$$

where the noise terms ϵ_i are mutually independent. That is, $\epsilon_i \perp\!\!\!\perp \epsilon_j$ for all $i \neq j$. Let $\mathcal{G}(\theta)$ and $D(\theta) \in \{0,1\}^{d \times d}$ be the induced directed graph and corresponding binary adjacency matrix obtained as follows: $V_i \to V_j$ exists in $\mathcal{G}(\theta)$ and $D(\theta)_{ij} = 1$ if and only if $\theta_{ij} \neq 0$. The induced graph $\mathcal{G}$ has no directed cycles if and only if θ can be made upper-triangular via a permutation of vertex labelings [50]. Such an SEM is said to be *recursive* or *acyclic* and the corresponding probability distribution $p(V)$ is said to be Markov with respect to the DAG $\mathcal{G}(\theta)$. This means that conditional independence statements in $p(V)$ can be read off from $\mathcal{G}$ via d-separation [5].

2.2.2 Systems with Unmeasured Confounding

A set of observed variables is called *causally insufficient* if there exist unobserved variables, commonly referred to as latent confounders, that cause two or more observed variables in the system. In the linear SEM setting, unmeasured variables manifest as correlated errors [5]. Such an SEM on d variables can be parameterized by two real-valued matrices $\delta, \beta \in \mathbb{R}^{d \times d}$ as follows. For each $V_i \in V$, we have a structural equation $V_i \leftarrow \sum_{V_j \in V} \delta_{ji} V_j + \epsilon_i$, and the dependence between the noise terms $\epsilon = (\epsilon_1, ..., \epsilon_d)$ is summarized via their covariance matrix $\beta = \mathbb{E}[\epsilon\epsilon^T]$. In the case when each noise term ϵ_i is normally distributed the induced distribution $p(V)$ is jointly normal with mean zero and covariance matrix $\Sigma = (I - \delta)^{-T} \beta (I - \delta)^{-1}$. The induced graph $\mathcal{G}$ is a mixed graph consisting of directed ($\to$) and bidirected ($\leftrightarrow$) edges and can be represented via two adjacency matrices D and B. $V_i \to V_j$ exists in $\mathcal{G}$ and $D_{ij} = 1$ if and only if $\delta_{ij} \neq 0$. $V_i \leftrightarrow V_j$ exists in $\mathcal{G}$ and $B_{ij} = B_{ji} = 1$ if and only if $\beta_{ij} \neq 0$. That is, the adjacency matrix B corresponding to bidirected edges in $\mathcal{G}$ is symmetric as the covariance matrix β itself is symmetric (and positive definite.)

We consider three classes of mixed graphs to represent causally insufficient linear SEMs: ancestral, arid, and bow-free ADMGs. All of these have no directed cycles and lack specific substructures as defined in the previous section. A distribution $p(V)$

induced by a linear Gaussian SEM is said to be Markov with respect to an ADMG $\mathcal{G}$ if absence of an edge between V_i and V_j implies $\delta_{ij} = \delta_{ji} = \beta_{ij} = \beta_{ji} = 0$ which in turn implies equality restrictions on the support of all possible covariance matrices $\Sigma(\mathcal{G})$ by forcing certain polynomial functions of entries in the covariance matrix to evaluate to 0 [51]. To facilitate causal discovery, we assume a generalized version of faithfulness, similar to the one in [52], stating that if a distribution $p(V)$ is induced by a linear Gaussian SEM where $\delta_{ij} = \delta_{ji} = \beta_{ij} = \beta_{ji} = 0$ then there is no edge present between V_i and V_j in $\mathcal{G}$. In other words, we define $p(V)$ to be Markov and faithful with respect to $\mathcal{G}$ if absence of edges in $\mathcal{G}$ occurs if and only if the corresponding entries in δ and β are 0.

As a concrete example, let Σ denote the covariance matrix of standardized normal random variables A, B, C, D drawn from a linear SEM that is Markov with respect to the ADMG in Figure 2-1(c), and let δ and β denote the corresponding normalized coefficient matrices. By standard rules of path analysis [48, 49], the Verma constraint due to the missing edge in Figure 2-1(c) corresponds to the equality constraint:

$$\Sigma_{BC} - \delta_{CD}\delta_{DB} - \delta_{AC}\beta_{AB} - \delta_{AC}\beta_{AD}\delta_{DB} = 0.$$

Since entries in the covariance matrix are rational functions of δ and β, the above constraint can be re-expressed solely in terms of entries in Σ. Our faithfulness assumption is used to ensure that such polynomial functions of the covariance matrix do not "accidentally" evaluate to zero, and only do so due to a missing edge in the underlying ADMG.

As mentioned earlier, ancestral ADMGs cannot encode such generalized equality restrictions but arid and bow-free ADMGs can. For any ADMG $\mathcal{G}$, an arid ADMG that shares all non-parametric equality constraints with $\mathcal{G}$ may be constructed by an operation called maximal arid projection [39]. We consider bow-free ADMGs because the algebraic constraint characterizing the bow-free property is simpler than the one

characterizing the arid property. Though the lack of global identifiability in bow-free ADMG models (only almost everywhere identifiable) can pose problems for model convergence, we confirm in our experiments that enforcing only the weaker bow-free property is often sufficient for accurate causal discovery in practice.

2.3 Differentiable Algebraic Constraints

We now introduce differentiable algebraic constraints that precisely characterize when the parameters of a linear SEM induce a graph that belongs to any one of the ADMG classes described in the previous section. Our results are summarized in Table 2-I in terms of the binary adjacency matrices but as we explain below, the results extend in a straightforward manner to real-valued matrices that parameterize a linear SEM. In Table 2-I, $A \circ B$ denotes the Hadamard (elementwise) matrix product between A and B and e^A denotes the exponential of a square matrix A defined as the infinite Taylor series, $e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$. We formalize the properties of our constraints in the following theorem.

Theorem 1. *The constraints shown in Table 2-I are satisfied if and only if the adjacency matrices satisfy the relevant property of ancestrality, aridity, and bow-freeness respectively.*

We defer formal proofs to the Appendix but briefly provide intuition for our results. For a binary square matrix A, corresponding to a directed/bidirected adjacency matrix, the entry A^k_{ij} counts the number of directed/bidirected walks of length k from V_i to V_j; see for example [53]. For $k = 0$, D^k is the identity matrix by definition and for $k \geq 1$, each diagonal entry of the matrix D^k appearing in the infinite series e^D thus corresponds to the number of directed walks of length k from a vertex back to itself, i.e., the number of directed cycles of length k. The quantity $\text{trace}(e^D) - d$ is therefore a weighted count of the number of directed cycles in the induced graph and

Algorithm 1 GREENERY

Inputs: $d \times d$ matrices (D, B)
greenery $\leftarrow 0$ and $I \leftarrow d \times d$ identity matrix
for i in $(1, \ldots, d)$ **do**
    $D_f, B_f \leftarrow D, B$
    for j in $(1, \ldots, d-1)$ **do**
        $t \leftarrow$ row sums of $e^{B_f} \circ D_f$ ▷ $1 \times d$ vector
        $f \leftarrow \tanh(t + I_i)$ ▷ $1 \times d$ vector
        $F \leftarrow [f^T; \ldots; f^T]^T$ ▷ $d \times d$ matrix
        $D_f \leftarrow D_f \circ F$ and $B_f \leftarrow B_f \circ F \circ F^T$
    $C \leftarrow e^{D_f} \circ e^{B_f}$
    greenery $\leftarrow$ greenery $+ \operatorname{sum}(C_{:,i})$ ▷ sum of i^{th} column
return greenery $- d$

ADMG	Algebraic Constraint
Ancestral	$\operatorname{trace}(e^D) - d + \operatorname{sum}(e^D \circ B) = 0$
Arid	$\operatorname{trace}(e^D) - d + \text{GREENERY}(D, B) = 0$
Bow-free	$\operatorname{trace}(e^D) - d + \operatorname{sum}(D \circ B) = 0$

Table 2-I. Differentiable algebraic constraints that characterize the space of binary adjacency matrices that fall within each ADMG class. The GREENERY algorithm to penalize c-trees is described in Algorithm 1.

is zero precisely when no such cycles exist. Hence, this term appears in all algebraic constraints presented in Table 2-I as requiring $\operatorname{trace}(e^D) - d = 0$ enforces acylicity.

Similar reasoning can be used to show that requiring $\operatorname{sum}(e^D \circ B) = 0$ enforces ancestrality. An entry i, j of the matrix $D^k \circ B$ appearing in the infinite series counts the number of violations of ancestrality due to a directed path from V_i to V_j of length k and a bidirected edge $V_i \leftrightarrow V_j$. The sum of all such terms is then precisely zero when the induced graph is ancestral. The bow-free constraint $\operatorname{sum}(D \circ B) = 0$ is simply a special case of the ancestral constraint where directed paths of length ≥ 2 need not be considered.

C-trees are known to be linked to the identification of causal parameters, specifically,

the effect of each variable's parents on the variable itself [28]. The outer loop of Algorithm 1 iterates over each vertex V_i to determine if there is a V_i-rooted c-tree. The inner loop performs the following recursive simplification at most $d-1$ times. At each step, the sum of the i^{th} row of the matrix $e^{B_f} \circ D_f$ is zero if and only if there are no bidirected paths from V_j to any of its direct children. If this criterion – called primal fixability – is met, the effect of V_j on its children is identified and the post-intervention distribution can be summarized by a new graph with all incoming edges into V_j removed [13]. Lines 7-9 are the algebraic operations that correspond to deletion of incoming directed and bidirected edges into primal fixable vertices, except V_i itself as it is the root node of interest. The hyperbolic tangent function is used to ensure that recursive applications of the operation do not result in large values. At the end of the recursion, the co-existence of directed and bidirected paths to V_i imply the existence of a c-tree. Hence, the quantity $\text{sum}(C_{:,i})$ is non-negative and is zero if and only if there is no V_i-rooted c-tree. Concrete examples of applying Algorithm 1, and its connections to primal fixing are provided in Appendix B.1.

It is easy to see that the above results and intuitions can be applied to arbitrary non-negative real-valued matrices D and B. Theorem 1 then extends in a straightforward manner to parameters of a linear SEM by noting that for any real-valued matrix A, the matrix $A \circ A$ is real-valued and non-negative.

Corollary 1.1. *The result in Theorem 1 and the constraints in Table 2-I can be applied to linear SEMs by plugging in* $D \equiv \delta \circ \delta$ *and* $B \equiv \beta' \circ \beta'$, *where* $\beta'_{ij} = \beta_{ij}$ *for* $i \neq j$ *and* 0 *otherwise.*

Finally, while the matrix exponential makes theoretical arguments simple, the resulting constraints are not numerically stable as pointed out in [45]. The following corollary provides a more stable alternative that we use in our implementations.

Corollary 1.2. *The results in Theorem 1 and Corollary 1.1 hold if every occurrence*

of a matrix exponential e^A is replaced with the matrix power $(I + cA)^d$ for any $c > 0$, where I is the identity matrix.

2.4 Differentiable Causal Discovery for ADMGs

Let θ be the parameters of a linear SEM. We use θ here to refer to a generic parameter vector that can be reshaped into the appropriate parameter matrices δ, and β as discussed in Section 2.2. Let $\mathcal{G}(\theta)$ be the corresponding induced graph. Given a dataset $X \in \mathbb{R}^{n \times d}$ drawn from the linear SEM and a hypothesis class $\mathbb{G}$ that corresponds to one of ancestral, arid, or bow-free ADMGs, the combinatorial problem of finding an optimal set of parameters $\theta^* \in \Theta$ that minimizes some score $f(X; \theta)$ such that $\mathcal{G}(\theta) \in \mathbb{G}$ can be rephrased as a more tractable continuous program.

$$\begin{aligned} &\min_{\theta \in \Theta} \; f(X; \theta) \\ &\text{s.t.} \quad \mathcal{G}(\theta) \in \mathbb{G} \end{aligned} \quad \iff \quad \begin{aligned} &\min_{\theta \in \Theta} \; f(X; \theta) \\ &\text{s.t.} \quad h(\theta) = 0. \end{aligned} \tag{2.1}$$

The results in the previous section in Theorem 1, its Corollaries and Table 2-I tell us how to pick the appropriate function $h(\theta)$ for each hypothesis class $\mathbb{G}$. We now discuss choices of score function $f(X; \theta)$ and procedures to minimize it for different hypothesis classes.

2.4.1 Choice of Score Function

Given a dataset $X \in \mathbb{R}^{n \times d}$, the Bayesian Information Criterion (BIC) is given by $-2 \ln(\mathcal{L}(X; \theta)) + \ln(n) \sum_{i=1}^{\dim(\theta)} \mathbb{I}(\theta_i \neq 0)$, where $\mathcal{L}(\cdot)$ is the likelihood function and $\dim(\theta)$ is the dimensionality of θ. The BIC is consistent for model selection in curved exponential families [54, 55], i.e., as $n \to \infty$ the BIC attains its minimum at the true model (or one that is observationally equivalent to it). This results in the following desirable theoretical property when the BIC is used as our objective function.

Theorem 2. *Let $p(V; \theta^*)$ be a distribution in the curved exponential family that is Markov and faithful with respect to an arid ADMG $\mathcal{G}^*$. Finding the global optimum*

of the continuous program in display (2.1) with $f \equiv BIC$ yields an ADMG $\mathcal{G}(\theta)$ that implies the same equality restrictions as $\mathcal{G}^*$.

However, the presence of the indicator function makes the BIC non-differentiable and optimization of L_0 objectives like the BIC is known to be NP-hard [56]. While L_1 regularization is a popular alternative, it often leads to inconsistent model selection and overshrinkage of coefficients [57]. It is known that optimizing the L_1 regularized objective yields the true model when the data comes from a linear Gaussian SEM when all noise terms in the system have equal variance [58, 59]. However, recent work has shown that this is an untestable assumption that implies a known causal ordering on all variables in the system and that causal discovery procedures that rely on this information may be particularly prone to drops in performance when the data are re-scaled [60, 61]. Hence, we pursue approximations to the BIC score itself.

Several procedures have been devised in order to provide approximations of the BIC score; see [62] for an overview. In this work, we consider the approximate BIC (ABIC) obtained via replacement of the indicator function with the hyperbolic tangent function as outlined in [63] and [64]. That is, we seek to optimize $-2\ln(\mathcal{L}(X;\theta)) + \lambda\sum_{i=1}^{\dim(\theta)}\tanh(c|\theta_i|)$, where $c > 0$ is a constant that controls the sharpness of the approximation of the indicator function and λ controls the strength of regularization. As highlighted in [63], the ABIC is relatively insensitive to the choice of c. The main hyperparameter is the regularization strength λ. In our experiments we set $c = \ln(n)$ and report results for different choices of λ. In the next section we discuss our strategy to optimize the ABIC subject to the constraint that θ induces a valid ADMG within a hypothesis class $\mathbb{G}$.

2.4.2 Solving the Continuous Program

We formulate the optimization objective as minimizing the ABIC subject to one of the algebraic equality constraints in Table 2-I. We use the augmented Lagrangian

Algorithm 2 REGULARIZED RICF

Inputs: $(X, \text{tol}, \text{max iterations}, h, \rho, \alpha, \lambda)$
Initialize estimates δ^t and β^t and set $c = \ln(n)$
Define $\text{LS}(\theta)$ as $\frac{1}{2n}\sum_{i=1}^{d} ||X_{:,i} - X\delta_{:,i} - Z^{(i)}\beta_{:,i}||_2^2$
for t in $(1, \ldots, \text{max iterations})$ **do**
    $\forall i \in (1, \ldots, d)$ compute $\epsilon_i \leftarrow X_{:,i} - \delta_{:,i}^t X$
    $\forall i \in (1, \ldots, d)$ compute $Z^{(i)} \in \mathbb{R}^{n \times d}$ as $Z_{:,i}^{(i)} = 0$ and $Z_{:,-i}^{(i)} \leftarrow \epsilon_{-i}(\beta_{-i,-i}^t)^{-T}$
    $\delta^{t+1}, \beta^{t+1} \leftarrow \text{argmin}_{\theta \in \Theta} \left\{ \text{LS}(\theta) + \frac{\rho}{2}|h(\theta)|^2 + \alpha h(\theta) + \lambda \sum_{i=1}^{\dim(\theta)} \tanh(c|\theta_i|) \right\}$
    $\forall i \in (1, \ldots, d)$ compute $\epsilon_i \leftarrow X_{:,i} - \delta_{:,i}^{t+1} X$
    $\forall i \in (1, \ldots, d)$ set $\beta_{ii}^{t+1} \leftarrow \text{var}(\epsilon_i)$
    if $||\delta^{t+1} - \delta^t + \beta^{t+1} - \beta^t|| < \text{tol}$ **then**
        break
return δ^t, β^t

formulation [65] to convert the problem into an unconstrained optimization problem with a quadratic penalty term, which can be solved using a dual ascent approach. Specifically, in each iteration we first solve the primal equation:

$$\min_{\theta \in \Theta} \ \text{ABIC}_\lambda(X; \theta) + \frac{\rho}{2}|h(\theta)|^2 + \alpha h(\theta),$$

where ρ is the penalty weight and α is the Lagrange multiplier. Then we solve the dual equation $\alpha \leftarrow \alpha + \rho h(\theta^*)$. Intuitively, optimizing the primal objective with a large value of ρ would force $h(\theta)$ to be very close to zero thus satisfying the equality constraint.

However, unlike DAG models, maximum likelihood estimation of parameters under the restrictions of an ADMG does not correspond to a simple least squares regression that can be solved in one step. [66] proposed an iterative procedure known as Residual Iterative Conditional Fitting (RICF) that produces a sequence of maximum likelihood estimates for δ and β under the constraints implied by a fixed ADMG $\mathcal{G}$. Each RICF step is guaranteed to produce better estimates than the previous step and the overall procedure is guaranteed to converge to a local optimum or saddle point when $\mathcal{G}(\theta)$ is

Algorithm 3 DIFFERENTIABLE DISCOVERY

1: **Inputs**: $(X, \text{tol}, \text{max iterations}, s, h, \lambda, r \in (0, 1))$
2: Initialize $\theta^t, \alpha^t, m^t \leftarrow 1$
3: **while** $t <$ max iterations and $h(\theta^t) >$ tol **do**
4: $\quad \theta^{t+1} \leftarrow \theta^*$ from REGULARIZED RICF with inputs $(X, 10^{-4}, m^t, h, \rho, \alpha^t, \lambda)$ where ρ is such that $h(\theta^*) < rh(\theta^t)$
5: $\quad \alpha^{t+1} \leftarrow \alpha^t + \rho h(\theta^{t+1})$
6: $\quad m^{t+1} \leftarrow m^t + s$
7: **return** $\mathcal{G}(\theta^t)$

arid/ancestral, i.e., globally identified [38].

In Algorithm 2 we describe a modification of RICF that directly inherits the aforementioned properties with respect to the regularized maximum likelihood objective, and can be used to solve the primal equation of our procedure. Briefly, for Gaussian ADMG models, maximization of the likelihood corresponds to minimization of a least squares regression problem where each variable i is regressed on its direct parents $V_j \to V_i$ and pseudo-variables Z formed from the residual noise terms and bidirected coefficients of its siblings $V_j \leftrightarrow V_i$. At each RICF step, we compute Z with respect to the current parameter estimates, and then solve the primal equation in line 7 of the algorithm. We repeat this until convergence or a pre-specified maximum number of iterations. As RICF is not expected to converge during initial iterations of the augmented Lagrangian procedure when the penalty applied to $h(\theta)$ is quite small (resulting in non-arid graphs), we start with a small number of maximum RICF iterations and at each dual step increment this number. Our simulations show this works quite well in practice with convergence of the algorithm obtained typically within 10-15 steps of the augmented Lagrangian procedure.

We summarize our structure learning algorithm in Algorithm 3. Though optimization of the objective in display (2.1) is non-convex, standard properties of dual ascent procedures as well as the RICF algorithm guarantee that at each step in the process

we recover parameter estimates that do not increase the objective we are trying to minimize. Further, per Theorem 2, if optimization of the ABIC objective for a given level of λ provides a good enough approximation of the BIC, the global minimizer (if found by our optimization procedure) yields a graph that implies the same equality restrictions as the true graph.

2.4.3 Reporting Equivalent Structures

Our procedure only reports a single ADMG but there may exist multiple ADMGs that imply the same equality restrictions on the observed data. In the linear Gaussian setting, exact recovery of the skeleton of the ADMG (i.e., adjacencies without any orientations) is possible, but complete determination of all edge orientations is not. Reporting the uncertainty in edge orientations is important for downstream causal inference tasks. When limiting our hypothesis class $\mathbb{G}$ to ancestral ADMGs, the non-parametric equivalence class can be represented via a Partial Ancestral Graph (PAG). After obtaining a single ADMG using our procedure, we can easily reconstruct its equivalence class using rules in [67] to create the summary PAG. For arid and bow-free ADMGs, a full theory of equivalence that captures Verma constraints is still an open problem. Thus, while we are able to recover the exact skeleton, we coarsen reporting of edge orientations by converting the estimated ADMG into an ancestral ADMG and reporting the PAG. Connections in this PAG may be pruned using sound rules from [68] and [69] though we do not pursue this approach in the present work. Deriving a summary structure that captures the class of all ADMGs that are equivalent up to equality restrictions is an important open problem. The authors in [70] made progress on equivalence theory for 4-variable ADMGs by enumerating all possible 4-variable ADMGs and evaluating the BIC score for each one, grouping graphs with equal scores to form an "empirical equivalence class." We believe something similar could be done for larger graphs using our proposed causal discovery procedure. If

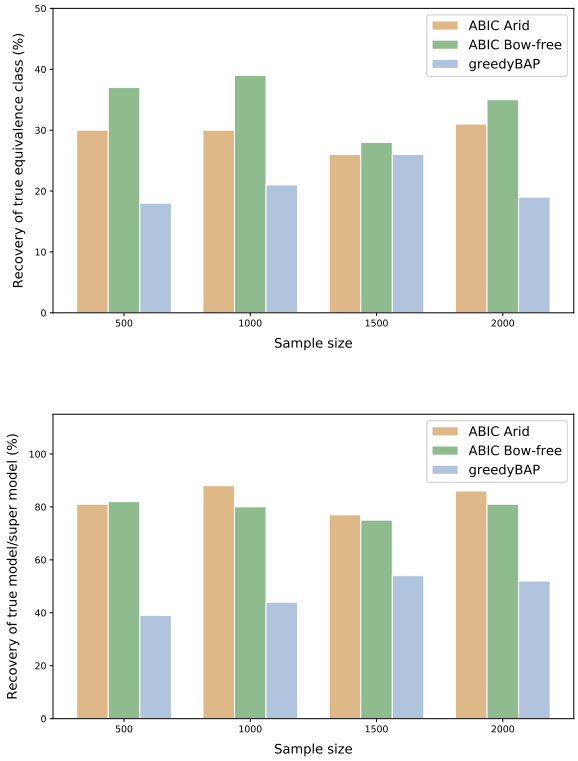


Figure 2-2. Top panel: plots showing rate of recovery of the true equivalence class of ADMGs with a Verma constraint as a function of sample size. Bottom panel: plots showing rate of recovery of the true equivalence class *or* a super model of the true equivalence class of ADMGs with a Verma constraint as a function of sample size.

relevant patterns in larger empirical equivalence classes become apparent, this may result in significant progress towards characterizing nested Markov equivalence.

2.5 Experiments

For a given ADMG, we generate data as follows. For each $V_i \to V_j$ we uniformly sample δ_{ij} from $\pm[0.5, 2.0]$, for $V_i \leftrightarrow V_j$, we sample $\beta_{ij} = \beta_{ji}$ from $\pm[0.4, 0.7]$, and for each β_{ii} we sample from $\pm[0.7, 1.2]$ and add $\text{sum}(|\beta_{i,-i}|)$ to ensure positive definiteness of β.

Since randomly generated ADMGs are unlikely to exhibit Verma constraints, we first consider recovery of the ADMG shown in Figure 2-1(c) and two other ADMGs

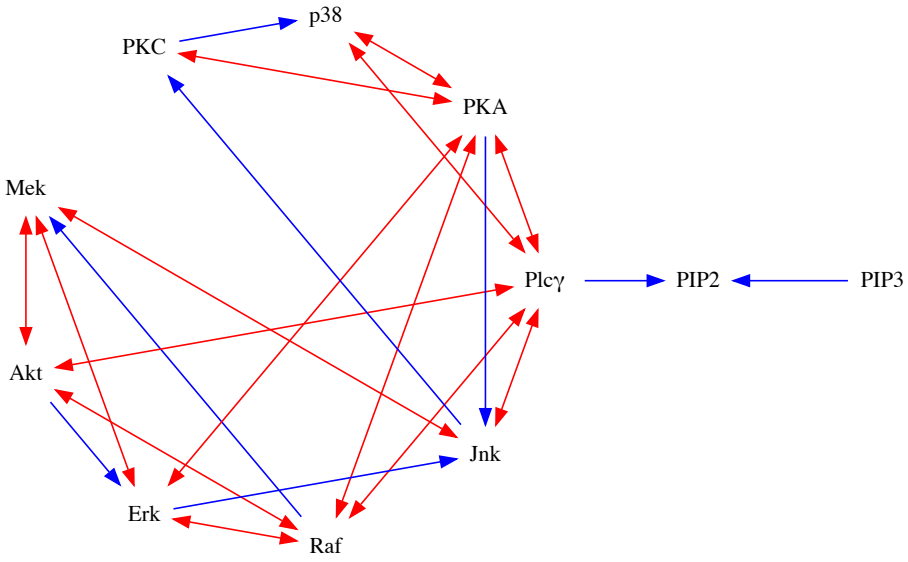


Figure 2-3. Application of the ABIC bow-free method to protein expression data from [1] .

$A \to B \to C \to D, B \leftrightarrow D$ and a Markov equivalent ADMG obtained by replacing $A \to B$ with $A \leftrightarrow B$ which have Verma constraints established in the prior literature. Exact recovery of Figure 2-1(c) is possible while the latter ADMGs can be recovered up to ambiguity in the adjacency between A and B as $A \to B$ or $A \leftrightarrow B$. We compare our arid and bow-free algorithms to the greedyBAP method proposed in [68] (the only other method available for recovering such constraints). Since greedyBAP is designed to perform random restarts, we allow all methods 5 uniformly random restarts and pick the final best fitting ADMG. As mentioned earlier, our main hyperparameter is the regularization strength λ, which we set to 0.05 for all experiments. Choice of other hyperparameters and additional experiments with varying λ are provided in Appendix B. We generate 100 datasets for each sample size of $[500, 1000, 1500, 2000]$ from a uniform sample of the 3 aforementioned ADMGs. The results are summarized via barplots in Figure 2-2.

The ABIC arid and bow-free procedures both outperform the greedyBAP procedure in recovering the true equivalence class; see top panel of Figure 2-2. The highest recovery rate is shown by the bow-free procedure with 39% at $n = 1000$. Though this seems low, these results are quite promising in light of geometric arguments in [71] that show reliable recovery of Verma constraints may require very large sample sizes. In examining the modes of failure of each algorithm, our ABIC procedures often fail to recover the true ADMG by returning a super model of the true equivalence class while the greedyBAP procedure often returns an incorrect independence model; see bottom panel of Figure 2-2. The former kind of mistake does not yield bias in downstream inference tasks while the latter does. Our bow-free procedure yields more accurate results than the arid one most likely due to posing an easier optimization problem. In the 400 runs used to generate plots in Figure 2-2, the bow-free procedure failed to converge only 3 times and the arid one never failed to converge, which is consistent with established theoretical results on almost-everywhere and global identifiability of these models.

For larger randomly generated arid ADMGs, to save computation time, we only compare our bow-free procedure with greedyBAP, and for ancestral ADMGs, we compare our ancestral procedure with FCI [4] and greedySPo [42]. We also obtained results for GFCI which were slightly worse than FCI and greedySPo so we only report the latter. Random arid/ancestral ADMGs on 10 variables were generated by first producing a random bow-free ADMG with directed and bidirected edge probabilities of 0.4 and 0.3 respectively, and then applying the arid/ancestral projection. We report true positive and false discovery rates for exact skeleton recovery of the true ADMG as well as recovery of tails and arrowheads in the true PAG for 100 datasets of 1000 samples each. For FCI, we used a significance level of 0.15 which gave the most competitive results. Our method performs favorably in recovery of both arid and ancestral ADMGs. Results for 10 variable arid and ancestral ADMGs, which roughly

	Skeleton		Arrowhead		Tail	
Method	tpr ↑	fdr ↓	tpr ↑	fdr↓	tpr ↑	fdr ↓
greedyBAP [68]	0.80	0.30	0.41	0.58	0.11	0.65
ABIC (bow-free)	**0.89**	**0.17**	**0.72**	**0.29**	**0.30**	**0.45**

Table 2-II. Comparison of our method to greedyBAP for recovering 10 variable arid ADMGs. We report true positive rate (tpr) and false discovery rate (fdr) — the fraction of predicted edges that are actually present in the target structure or the fraction that are absent from the target structure, respectively — for skeleton, arrowhead and tail recovery. (↑/↓ indicates higher/lower is better.)

	Skeleton		Arrowhead		Tail	
Method	tpr ↑	fdr ↓	tpr ↑	fdr↓	tpr ↑	fdr ↓
FCI [4]	0.51	0.12	0.41	0.53	0.10	0.73
greedySPO [42]	**0.88**	0.27	0.45	0.59	0.32	0.81
ABIC (ancestral)	0.85	**0.11**	**0.72**	**0.23**	**0.66**	**0.47**

Table 2-III. Comparison of our method to FCI and gSPo for recovering 10 variable ancestral ADMGs. We report true positive rate (tpr) and false discovery rate (fdr) — the fraction of predicted edges that are actually present in the target structure or the fraction that are absent from the target structure, respectively — for skeleton, arrowhead and tail recovery. (↑/↓ indicates higher/lower is better.)

matches the dimensionality of our data application, are summarized in Tables 2-II and 2-III respectively.

Finally we apply our ABIC bow-free method to a cleaned version of the protein expression dataset in [1] from [72]. The result is shown in Figure 2-3. The precision and recall of our procedure with respect to the true adjacencies provided in [72] are 0.77 and 0.61 respectively. We do not provide evaluation of orientations as there is no consensus regarding many of them. However, we briefly highlight the importance of a Verma restriction in producing a model that is consistent with an intervention experiment performed by [1]. The authors found that manipulation of Erk produced no downstream effect on PKA though they are correlated. The ADMG in Figure 2-3 has an edge Erk $\leftrightarrow$ PKA that is consistent with this finding. Moreover, this edge cannot be oriented in either direction without producing different independence models

than the one implied by Figure 2-3. This is due to a Verma restriction between Akt and PKC; we provide more details in Appendix B. We confirm that orienting the edge as Erk $\leftarrow$ PKA or Erk $\rightarrow$ PKA leads to an increase in the BIC score, indicating that the Verma restriction capturing the ground truth is preferred over these other explanations.

2.6 Related and Future Work

Causal discovery methods for learning ancestral ADMGs from data are well developed [4, 73, 41], but procedures for more general ADMGs are understudied. [74] propose a constraint-based satisfiability solver approach for mixed graphs with cycles. However, their proposal relies on an independence oracle that does not address how to perform valid statistical tests for arbitrarily complex equality restrictions and their procedure may lead to models where the corresponding statistical parameters are not identified (so goodness-of-fit cannot be evaluated). A score-based approach to discovery for linear Gaussian bow-free ADMGs was proposed in [68]. Their method relies on heuristics that may lead to local optima and is not guaranteed to be consistent. Similar issues are faced by the method in [75], which makes a linear non-Gaussian assumption. Currently, there does not exist any consistent fully score-based procedure for learning general ADMGs (besides exhaustive enumeration which is intractable); there are greedy algorithms [42] and hybrid greedy algorithms [41] for ancestral ADMGs, but these are computationally intensive due to the large discrete search space and extending these to arid or bow-free ADMGs would be non-trivial. The procedure we have proposed has the benefit of being easy to adapt to either ancestral, arid, or bow-free ADMGs while avoiding the need to solve a complicated discrete search problem, instead exploiting state-of-the-art advances in continuous optimization. Approaches to learning latent variable DAGs, e.g., [76, 77], rather than ADMGs over the observed variables are also related to our problem. However, these approaches impose assumptions, such as the

purity assumption which requires the absence of edges between observed variables, which may be considered restrictive in many applied settings.

We have extended the continuous optimization scheme of causal discovery to include models that capture all equality constraints on the observed margin of hidden variable linear SEMs with Gaussian errors. Extensions to other parametric settings (including non-linear models) rely on the development of general curved exponential ADMG models and methods for maximum likelihood estimation which, outside of the discrete and linear Gaussian case, is still an open problem. However, the differentiable algebraic constraints we have provided are non-parametric and may thus also enable future development of non-parametric causal discovery methods. We conjecture that using the ABIC as the objective function may also give our causal discovery procedure nice theoretical properties, such as robustness to data re-scaling. However, these claims require further investigation (both theoretical and empirical) and careful study. Joint optimization of parameters and causal structure may also lend itself to ideas for simultaneous causal discovery and estimation as pursued by [78] for the case of covariate adjustment functionals. There also exists room for improving computational efficiency of the procedure to make its run time more competitive with existing greedy procedures, for e.g., by implementing a faster acyclicity constraint proposed in [79] and using the Sherman-Morrison formula [80] to perform efficient matrix inversions in the RICF algorithm. Studying the properties of hybrid procedures that use the methods proposed here as a first step before applying a constraint-based method may also be of interest. Finally, we propose that the methods developed in this work may also help explore questions regarding distributional equivalence and Markov equivalence with respect to all equality restrictions in ADMG models.

2.7 Acknowledgements

Chapter 2 is, in part, a reprint of material from [10].

Chapter 3

Causal Inference Under Interference and Network Uncertainty

In many scientific and policy settings, research subjects do not exist in isolation but in interacting networks. For instance, data drawn from an online social network will exhibit *homophily* (friends are similar, because they are friends) and *contagion* (friends may causally influence each other) [81, 31, 82, 32]. A related phenomenon that is well-documented in the infectious disease epidemiology literature is that of *herd immunity* – vaccinating some subset of a population may confer immunity to the entire population. Resource constraints in allocation problems may also induce data dependence – a phenomenon known as *allocational interference* [83, 84, 85]. The above phenomena imply that the treatment given to one individual or *unit* may affect the outcomes for others within their (social) network.

In the context of causal inference, methods for dealing with data dependence are developed under the heading of *interference* [86, 84, 87, 88, 89, 31, 85, 33, 90]. Most such work assumes the structure of the dependence (which units depend on which others, and how) is known precisely. For example, [89] assumes units in the data may be organized into equal sized neighborhoods, where units within a neighborhood are pairwise dependent and units across neighborhoods are not. Some work makes

alternative assumptions, e.g., [33] assumes that neighborhoods are drawn from a known Markov random field.

In many applications, the network inducing dependence between units may not be known exactly. For instance, in vulnerable, stigmatized, or isolated communities (such as groups of individuals with intravenous substance use disorders, or remote rural communities), we may have no way of reconstructing the precise social ties between individuals. Often, online databases of social media users may be anonymized, with friendship ties deliberately omitted. There has been some work in such settings that involves adapting the data collection method itself in order to discover the underlying networks: e.g., snowball sampling in [91] and [92]. Unfortunately, such study designs are not always possible to arrange in advance, and most data available on networks of interacting units is not collected under such designs.

While there is a rich literature on model selection from observational data in the context of causal inference (e.g., [4, 93, 94, 58]), to our knowledge all previous work has assumed the absence of interference. This chapter explores learning the dependence structure using graphical model selection methods. Techniques for structure learning from probabilistic relational models and physical module networks are also related to this work [95, 96, 97]. However, these models are not particularly well suited to modeling the phenomena that arise in the infectious disease setting or settings involving collective-decision making processes, i.e., biological and social contagion respectively. As described in Chapter 1, chain graph models under the LWF interpretation have been used to model such phenomena and form the focus of this chapter.

The contributions of this chapter may be viewed in one of two ways. From the point of view of causal inference under interference, it provides methods for estimating causal effects when there is substantial uncertainty about network structure. From the point of view of causal discovery, it introduces novel algorithms for model selection when units are dependent due to a network, the structure of which is unknown.

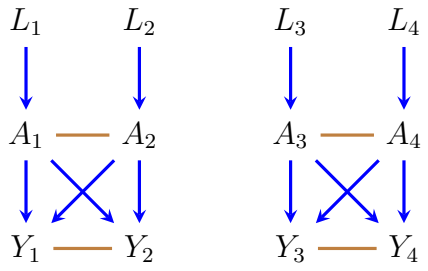


Figure 3-1. A chain graph over three variables (L, A, and Y) on 4 individuals, representing possible relationships between disposable needle use and risk of blood-borne disease among heroin-users.

3.1 Motivating Example and Background Assumptions

To motivate the work in this chapter, we discuss an example application. Consider a public health program aimed at lowering the incidence of blood-borne diseases such as HIV in at-risk individuals who are addicted to heroin and share needles when injecting intravenously. An example of such a program is described in [98]. The program creates pop-up clinics around the city where disposable needles are distributed for free to individuals in need, but due to limited resources only a limited number of individuals will actually receive these needles. We would like to know, in this restricted resource setting, if the use of disposable needles spreads amongst the rest of the population. Additionally, we would like to detect the phenomenon of herd immunity – whether some members of the population being protected due to taking advantage of the clean needles confer this protection to others who do not.

Data on heroin users was collected via such program, with users arranged by neighborhood or municipality. Users in different neighborhoods are assumed independent, but users within the same neighborhood are likely dependent. This setting is known as *partial interference* [88]. For each individual i, data is collected on their use of disposable needles A_i, their subsequent health outcome Y_i (risk of obtaining blood-born disease), along with a vector of pre-treatment covariates $L_i = (L_{1,i}, ..., L_{p,i})$.

We may be interested in quantifying the causal effects of A_i on Y_j, for arbitrary i and j within a neighborhood, or network-averaged versions of such effects [85]. We may assume that background knowledge or study design implies a "known" individual-level causal structure for each i, namely that $A_i \rightarrow Y_i$ and $A_i \leftarrow L_i \rightarrow Y_i$, but that we are uncertain about network ties among users. One approach is to assume the least restrictive model, where all users in a neighborhood are arbitrarily dependent. This would correspond to a complete network, where every pair of vertices is directly connected. However, assuming a complete network when the true network is sparse ignores useful structure in the problem and leads to inefficient estimates of target quantities. In addition, complete networks often lead to likelihoods that are intractable to evaluate. An alternative is to a select a sparse network supported by the data. In addition to enabling tractable and statistically efficient inference, such an approach may also rule out the presence of certain causal effects without explicitly estimating them, if corresponding pathways are absent in the selected network.

As an example, if neighborhoods have 4 units, we may aim to learn a chain graph (CG) model such as the one shown in Figure 3-1. Recall that statistical and causal models of chain graphs were described in Chapter 1. In the network setting, directed edges in the model represent direct causal influences while undirected edges representing symmetric network ties. The CG model in Figure 3-1 tells us that we should expect some spread of disposable needle use from one unit to another. However, it also tells us that users in neighborhoods are split into two non-interacting groups: $\{1, 2\}$ and $\{3, 4\}$. This implies the absence of contagion from one group to another. In addition, the conditional independences among units implied by this split suggests that contagion effects within groups may be estimated more efficiently as compared to a statistically saturated model, with a complete network across units.

The algorithms proposed in this chapter are consistent (in the sense that they asymptotically converge on the true model) under a set of assumptions which we

now informally summarize. We assume the true data-generating process corresponds perfectly (satisfying Markov and faithfulness conditions) to some unknown chain graph, with two restrictions: (1) the unit-level graph is known, reflecting the aforementioned causal ordering between pre-treatment covariates, treatment variables, and outcomes; and (2) the graph respects what we later call *tier symmetry*, which restricts connections between variables at the same "tier" in the causal ordering to be symmetric. We assume the data is distributed with some (known) likelihood in the exponential family, as well as some weak statistical regularity conditions. We also present algorithms that make an additional simplifying assumption on the graphical structure – namely that influence between units is the same for all unit pairs – but such an assumption is not strictly necessary for consistency.

The rest of this chapter is organized as follows. We first describe an example causal estimand that one may be interested in estimating using a CG network model, such as the one shown in Figure 3-1. We then briefly describe a taxonomy of problems that may be pursued with regards to CG network model selection and define the subset that we choose to tackle in this chapter. We then present algorithms to learn the structure of CG network models under varying degrees of uncertainty (corresponding to various problems in our defined subset) with respect to edges between units in the network. Finally we present experiments demonstrating the efficacy of our methods for network recovery and downstream estimation of network causal effects.

3.2 The Conditionally Ignorable Network Model and Network Causal Effects

We consider CGs decomposed into three disjoint sets of variables similar to the one shown in Figure 3-1: L, represents vectors of baseline (pre-treatment) factors; A, represents treatments; and Y, represents outcomes. For each unit i in a neighborhood, we assume $L_i \subseteq \text{pa}_{\mathcal{G}}(A_i)$, and $L_i \cup A_i \subseteq \text{pa}_{\mathcal{G}}(Y_i)$. This represents a common assumption

(which we call *causal ordering*) in causal inference that for each unit both baseline factors and treatment potentially affect the outcome, and that the baseline factors also affect treatment assignment. Here each unit has one treatment variable A_i, one outcome variable Y_i, and possibly many baseline variables L_i. In interference settings, it is standard to allow that variables for another unit j may influence variables for unit i. In our case, there is a further complication: the precise nature of this influence is unknown.

This model implies, for positive $p(V)$, the following standard assumptions from the interference literature: $Y(a) \perp\!\!\!\perp A \mid L$ (network ignorability); $p(A = a \mid L) > 0 \; \forall a$ (positivity); and $Y(a) = Y$ if $A = a$ (network consistency). Under these assumptions, the joint counterfactual outcome is identified, regardless of the underlying network structure, as the following special case of the chain graph g-formula in Eq. 1.12 and Eq. 1.13, $p(Y(a)) = \sum_L p(Y \mid A = a, L) \times p(L)$.

Given a particular treatment assignment probability $\pi(A)$, a number of causal effects of interest may be defined; see [89] for an extensive discussion. In this chapter, we focus on a single effect, the *population average overall effect (PAOE)*, as an exemplar network effect though our results generalize to any identified causal effect of interest in network settings (for example, spillover effects.) Consider identical and independent neighborhoods of size m as in a partial interference setting and two fixed treatment assignment probabilities π_1 and π_2. Then the PAOE is defined as:

$$\frac{1}{m} \sum_{i=1}^{m} \sum_{A} \mathbb{E}[Y_i(A)] \times \{\pi_1(A) - \pi_2(A)\}. \tag{3.1}$$

Under the aforementioned assumptions, this effect is identified by the following functional [89]:

$$\frac{1}{m} \sum_{i=1}^{m} \sum_{L,A} \mathbb{E}[Y_i \mid A, L] \times p(L) \times \{\pi_1(A) - \pi_2(A)\}. \tag{3.2}$$

A number of estimation strategies for Eq. 3.2 are possible under various assumptions on network structure. For example, [89] considered an inverse probability weighted

estimator. In this chapter, we use the auto-g-computation algorithm from [33] to estimate the PAOE, which makes use of the Gibbs sampler interpretation of CGs and allows for arbitrary network structure; details of this estimator are in [33].

3.3 Taxonomy of Problems in Network Model Selection

We are interested in estimating causal effects like the PAOE under the assumptions of network ignorability, positivity, and consistency when there is uncertainty about the network structure. We first provide a taxonomy of problems of this type, having different levels of difficulty depending on the degree of uncertainty present. We will use chain graphs to represent both causal relationships and network dependence among units that form a ("social") network. We define an undirected network $\mathcal{N}$ as a graph (distinct from our CG of interest) where the vertices correspond to units (e.g. individuals $i, j, ...$), not random variables. Units may be adjacent or non-adjacent in $\mathcal{N}$ based on whether they are "friends" or otherwise directly dependent. For each unit i, we denote the unit-level variables for i in the CG $\mathcal{G}$ (e.g., L_i, A_i, and Y_i in Figure 3-1) by V_i, and edges among those variables by $\mathcal{E}_i$. Similarly, for a pair of units i, j which are adjacent in $\mathcal{N}$, we represent the set of edges from V_i to V_j (and vice versa) by $\mathcal{E}_{ij}$. It is the presence of these edges that induces data dependence between i and j in the analysis of dependent data. The set of $\mathcal{E}_{ij}$ for all pairs i, j adjacent in $\mathcal{N}$ (i.e., the set of all cross-unit edges) will be denoted by $\mathcal{E}_{\mathcal{N}}$.

The most general version of the network selection problem occurs when neither the causal structure of each unit, nor the network structure inducing dependence between units, is known. In this case the problem reduces to a structure learning problem for arbitrary chain graphs, as considered in [99, 100, 101, 102]. We do not pursue this version of the problem here for two reasons. First, the causal structure for each unit is often known due to background knowledge on temporal ordering and study design,

as is the case for our needle-dispensary motivating example. Second, model selection of arbitrary CGs is known to be a very challenging problem which (in the worst case) may require very large sample sizes [71].

In many settings, the causal structure for each individual unit is known and is typically assumed to be the same for every unit, i.e., $\mathcal{E}_i = \mathcal{E}_j$ for all i, j. The problem of model selection then amounts to learning the structure of the connections between units i.e., $\mathcal{E}_{ij}$ for all i, j. The search space for such a problem, while much smaller than the general problem, is still exponential. For a block that contains m units, there are $\binom{m}{2}$ possible pairings of units, leading to $2^{\binom{m}{2}}$ possible networks. The number of possible valid chain graphs is even larger, since units i, j adjacent in a network could be connected in a variety of ways via (undirected or directed) edges in $\mathcal{E}_{ij}$. Learning these connections requires a search through all possible combinations of edges that form $\mathcal{E}_{ij}$ such that the overall graph is a CG.

We may restrict the problem further by requiring that the connections between any two units, if present, are *homogenous*, meaning that dependence between any two units, if it exists, arises in the same way. Formally, we define homogeneity such that, for all pairs $(i, j), (k, l) \in \mathcal{N}$, $\mathcal{E}_{ij} = \mathcal{E}_{kl}$. Notice that the space of homogenous networks is still fairly large. The problem may be made more tractable by one of the following two assumptions. We may assume the existence of network connections is known, but that their types are unknown, i.e., we know $\mathcal{N}$ and would like to learn $\mathcal{E}_{ij}$. Alternatively, we may assume we know how two adjacent units are connected, but not which pairs are adjacent, i.e., we know $\mathcal{E}_{ij}$ and would like to learn $\mathcal{N}$. We may also have no such background knowledge. In the following, we present algorithms for both homogenous and heterogenous settings.

Throughout, we make an assumption which we call *tier symmetry*, which is commonly made implicitly or explicitly in the interference literature [89, 33]. That is, we require connections between variables in the same "tier" of causal ordering

to represent symmetric relations between the variables. This restricts edges $L_i - L_j$, $A_i - A_j$, and $Y_i - Y_j$ to always be undirected. Also it is natural to extend the known causal ordering of variables to connections between units: while we allow for e.g., $A_i \to Y_j$, the reverse, $Y_j \to A_i$ is ruled out. Finally, we rule out the existence of undirected edges connecting variables across tiers, e.g, edges of the form $A_i - Y_j$, since the existence of such edges, coupled with our causal ordering assumption, leads to graphs which are not CGs.

3.4 Greedy Network Search

In this section, we will describe a greedy score-based procedure (readers may contrast this with the continuous optimization based procedure for ADMGs described in Chapter 2) to learn the true underlying chain graph network model. We begin by describing an assumption used to facilitate learning graphical structures in the network based on independences in the data. In Chapter 1, we introduced the global Markov property of CGs which stated that the absence of edges in a CG $\mathcal{G}$ imply conditional independences in the joint distribution $p(V)$, which can be obtained via c-separation. In what follows, we make the *faithfulness* assumption, which is the converse of the global Markov property: if $X \perp\!\!\!\perp Y \mid Z$ in $p(V)$, then X is c-separated from Y given Z in $\mathcal{G}$. This is directly analogous to the faithfulness assumption made when selecting DAG models from data via constraint-based or score-based methods [4, 93].

3.4.1 Model Scores and the Pseudolikelihood

We will learn the structure of the network using a score-based approach to model selection. Score-based methods proceed by choosing the graph (from among some space of candidates) that optimizes a model score. Exhaustive model search is typically infeasible, so it is popular to employ greedy methods that optimize only "locally," that is, they traverse the space of candidate graphs considering only single-edge additions

and deletions. Under some conditions, such greedy procedures can be shown to asymptotically converge to the globally optimal model [93]. Scores used for greedy search typically satisfy three properties that are sufficient for finding the globally optimal model: decomposability, score-equivalence, and consistency.

A score is said to be decomposable if it can be written as a sum of local contributions, each a function of one vertex and its boundary. A score is said to be score-equivalent if two Markov equivalent graphs (i.e., graphs that imply the same set of conditional independences by the global Markov property) yield the same score. A score is said to be consistent if, as the sample size goes to infinity, the following two conditions hold. First, when two models both contain the true generating model, the model of lower dimension will have a better score. Second, when one model contains the true model and another does not, the former will have a better score.

A popular score satisfying these properties for model selection among DAG models is the Bayesian Information Criterion (BIC) [54]. Given a d-dimensional data set X of size n and model likelihood $\mathcal{L}(X;\mathcal{G}) \equiv \prod_{i=1}^{n} p(x_{1,i},\ldots,x_{d,i};\mathcal{G})$, the BIC is given by $2\ln\mathcal{L}(X;\mathcal{G}) - k\ln(n)$ where k is model dimension.[1] For CG models, the BIC is only decomposable for blocks, not for variables within the block. In addition, the score is not easy to evaluate. Both of these issues arise due to the presence of normalizing functions in the likelihood; refer to the CG factorization in Eq. 1.10 and Eq. 1.11. We present an alternative score which avoids some of these problems, based on the *pseudolikelihood function* [103]:

$$\mathcal{PL}(X;\mathcal{G}) \equiv \prod_{i=1}^{n}\prod_{j=1}^{d} p(x_{j,i} \mid x_{-j,i};\mathcal{G}),$$

where x_{-j} is the vector $(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_d)$. We define a score based on the pseudolikelihood called Pseudo-BIC (PBIC): $2\ln\mathcal{PL}(X;\mathcal{G}) - k\ln(n)$.

We propose a greedy score-based model selection procedure based on the PBIC

[1] In Chapter 2 we used the BIC in the context of minimizing the score. The definition of the BIC here with flipped signs is equivalent and is often used in the context of maximizing the score.

score, which is consistent and obeys a weaker notion of decomposability for exponential families, as we show below.

Lemma 1. *With dimension fixed and sample size increasing to infinity, the PBIC is a consistent score for curved exponential families whose natural parameter space* Θ *forms a compact set.*

Decomposability of a scoring criterion makes greedy search a practical procedure, by limiting the number of terms in the overall score that need to be recomputed for each considered edge modification. While the BIC score for DAG models is decomposable, the PBIC score for CG models is not. Nevertheless, a weaker notion of decomposability holds, which implies that two CG models that differ by a single edge differ by a subset of components of the score, which we now describe.

Consider a candidate edge between V_i and V_j in a CG $\mathcal{G}$. Let B_{loc} denote the block to which V_j belongs when the edge is directed $V_i \rightarrow V_j$, or to which V_i and V_j belong when the edge is undirected $V_i - V_j$. We use $\text{loc}(V_i, V_j; \mathcal{G})$ to denote a set of vertices called the *local set*, defined as:

$$\bigcup_{C} \Big\{ C \in \mathcal{C}\Big((\mathcal{G}_{\text{bd}_{\mathcal{G}}(B_{\text{loc}})})^a\Big) : V_i, V_j \in C \not\subseteq \text{pa}_{\mathcal{G}}(B_{\text{loc}}) \Big\}.$$

As we show, the score difference for graphs $\mathcal{G}$ and $\mathcal{G}'$ which differ by a single edge can be written as the difference between terms that involve only variables in the local set of $\mathcal{G}$. The next result, and much subsequent discussion in the chapter, is stated for conditional Markov random fields (MRFs). This is because statistical CG models can be equivalently described as sets of conditional MRF models. We elaborate on this relationship in Appendix C.

Lemma 2. *Let* $\mathcal{G}$ *and* $\mathcal{G}'$ *be graphs which differ by a single edge between* V_i *and* V_j*. For conditional MRFs in the exponential family, the local score difference between* $\mathcal{G}$ *and* $\mathcal{G}'$ *is given by:* $\sum_{V_k \in \text{loc}(V_i, V_j; \mathcal{G}) \cap B_{\text{loc}}} \{ s_{V_k}\Big(X; \mathcal{G}\Big) - s_{V_k}\Big(X; \mathcal{G}'\Big) \}$*, where* $s_{V_k}(.)$ *denotes the component of the score for* V_k*.*

Note that the above definition of the local set may simplify further in certain special cases of MRF models in the exponential family. In particular, if we consider an MRF that is multivariate normal, or a log linear discrete model with only main effects and pairwise interactions, then the sum in Lemma 2 reduces to either a sum over elements V_i and V_j (for an undirected edge $V_i - V_j$) or only V_j (for a directed edge $V_i \rightarrow V_j$). We omit these straightforward proofs. As we aim for our method to be fairly general, we do not consider these special instances of the exponential family in the remainder of this chapter, and provide an informal discussion of the incurred computational costs for exponential families in general in Appendix C.

Having described the PBIC score and its properties, we are now ready to propose our greedy search procedures. We begin by describing a greedy search procedure that learns network ties $\mathcal{E}_{\mathcal{N}}$ without imposing homogeneity. Model selection proceeds by solving 3 independent sub problems: learning a Markov random field (MRF) over the baseline covariates L, learning a conditional MRF on the treatments A, and learning a conditional MRF on the outcomes Y. The resulting network ties learned from each of these, are combined to produce the final result (Algorithm 5). Each of the above subproblems is solved by a greedy search procedure (Algorithm 4) that starts with the complete conditional MRF (or MRF), and deletes the edge that yields the greatest improvement to the PBIC score on each iteration.

We now describe procedures for learning network ties in the homogenous setting, after defining some preliminaries. The *homologs* of an edge $E_{ij} \in \mathcal{E}_{\mathcal{N}}$ with endpoints $U_i, W_j \in V$, are defined as: $h(E_{ij}) \equiv \{E_{kl} \in \mathcal{E}_{\mathcal{N}} : \text{endpoints}(E_{kl}) = U_k, W_l\}$. The network tie prototypes in a homogenous graph $\mathcal{G}$ are defined as: $\mathcal{E}_{\text{proto }\mathcal{N}} \equiv \{E_{ij} \in \mathcal{E}_{ij}\text{for any } (i,j) \in \mathcal{N}\}$. $h(\mathcal{E}_{\text{proto }\mathcal{N}})$ can then be defined as: $\{h(E) : E \in \mathcal{E}_{\text{proto }\mathcal{N}}\}$.

When the types of connections $\mathcal{E}_{\text{proto }\mathcal{N}}$ between any two connected units is known, we start with a CG that is fully connected as $\mathcal{E}_{\text{proto }\mathcal{N}}$ for every pairwise combination of units. Search proceeds by deleting $\mathcal{E}_{ij}$ between two units i and j that yields the best

Algorithm 4 GREEDY NETWORK SEARCH

Inputs: $\mathcal{G}^{\text{init}}, X$
Initialize $\mathcal{G}^* \leftarrow \mathcal{G}^{\text{init}}$, score_change $\leftarrow$ True
while score_change **do**
    score_change $\leftarrow$ False
    $\mathcal{E}^*_{\mathcal{N}} \leftarrow$ network ties in $\mathcal{G}^*$
    $E_{max} \leftarrow \text{argmax}_{E \in \mathcal{E}_{\mathcal{N}^*}} \text{PBIC}(X; \mathcal{G}^* \setminus E)$
    if $\text{PBIC}(X; \mathcal{G}^* \setminus E_{max}) > \text{PBIC}(X; \mathcal{G}^*)$ **then**
        $\mathcal{G}^* \leftarrow \mathcal{G}^* \setminus E_{max}$ ▷ delete edge E_{max}
        score_change $\leftarrow$ True
return $\mathcal{E}^*_{\mathcal{N}}$

Algorithm 5 HETEROGENOUS

Inputs: $\mathcal{G}^{\text{complete}}, X$
$\mathcal{G}^L, \mathcal{G}^A, \mathcal{G}^Y \leftarrow$ conditional MRFs on L, A, and Y formed from $\mathcal{G}^{\text{complete}}$
$\mathcal{E}^*_{\mathcal{N}_L} \leftarrow$ GREEDY NETWORK SEARCH$(\mathcal{G}^L, X)$
$\mathcal{E}^*_{\mathcal{N}_A} \leftarrow$ GREEDY NETWORK SEARCH$(\mathcal{G}^A, X)$
$\mathcal{E}^*_{\mathcal{N}_Y} \leftarrow$ GREEDY NETWORK SEARCH$(\mathcal{G}^Y, X)$
return $\mathcal{E}^*_{\mathcal{N}_L} \cup \mathcal{E}^*_{\mathcal{N}_A} \cup \mathcal{E}^*_{\mathcal{N}_Y}$

improvement in the PBIC on each iteration (Algorithm 6). When the social network $\mathcal{N}$ is known, we start with a CG where pairs of units in $\mathcal{N}$ are fully connected in network ties. Search proceeds by deleting all homologs of the type of edge in $\mathcal{E}_{\text{proto}\,\mathcal{N}}$ that yields the best improvement in the PBIC on each iteration (Algorithm 7). Finally, when there is no background knowledge, homogenous search (Algorithm 8) can be performed by chaining the operations of Algorithm 6 and Algorithm 7 (or vice versa) on the CG complete in network ties for every pairwise combination of units.

Clearly we could use the heterogenous procedure even if the true underlying network ties are homogenous since it is more general. However, intuitively we expect the homogenous procedures to fare better in a finite data setting, because the homogeneity assumption allows pooling data from samples across units for each edge deletion test. This intuition is confirmed in our simulations.

Algorithm 6 HOMOGENOUS

Inputs: $\mathcal{G}^{\text{complete}}, X, \mathcal{E}_{\text{proto }\mathcal{N}}$
$\mathcal{G}^* \leftarrow$ graph obtained by removing all edges between units i, j in $\mathcal{G}^{\text{complete}}$ when $E_{ij} \notin h(\mathcal{E}_{\text{proto }\mathcal{N}})$
score change $\leftarrow$ True
while score change **do**
    score change $\leftarrow$ False
    $\mathcal{N}^* \leftarrow$ network in $\mathcal{G}^*$
    $(i,j)_{max} \leftarrow \text{argmax}_{(i,j)\in\mathcal{N}^*} \text{PBIC}(X; \mathcal{G}^* \setminus \mathcal{E}_{ij})$
    if $\text{PBIC}(X; \mathcal{G}^* \setminus \mathcal{E}_{ij_{max}}) > \text{PBIC}(X; \mathcal{G}^*)$ **then**
        $\mathcal{G}^* \leftarrow \mathcal{G}^* \setminus \mathcal{E}_{ij_{max}}$
        score change $\leftarrow$ True
return $\mathcal{N}^*$

3.4.2 Size of the Search Space

In the heterogenous case, the search space grows as $O(|\mathcal{E}_{\text{proto }\mathcal{N}}|\binom{m}{2})$ i.e., as a function of the number of possible edges between two units i and j multiplied by the number of possible pairings on m units. Under homogeneity when $\mathcal{E}_{\text{proto }\mathcal{N}}$ is known, this reduces to $O(\binom{m}{2})$; when $\mathcal{N}$ is known, it reduces to $O(|\mathcal{E}_{\text{proto }\mathcal{N}}|)$; and under homogeneity where neither is available, it is $O(\binom{m}{2}) + O(|\mathcal{E}_{\text{proto }\mathcal{N}}|)$.

3.4.3 Consistency of Greedy Network Search

Lemma 3. *If the generating distribution is Markov to a CG satisfying tier symmetry and the causal ordering assumption, then the search space of* GREEDY NETWORK SEARCH *consists of graphs belonging to their own equivalence classes of size 1.*

Theorem 3. *If the generating distribution is in the exponential family (with compact natural parameter space Θ) and is Markov and faithful to a CG satisfying tier symmetry and causal ordering then* GREEDY NETWORK SEARCH *is consistent.*

Under the same assumptions in the theorem above, we have the following corollary results.

Algorithm 7 HOMOGENOUS

Inputs: $\mathcal{G}^{\text{complete}}, X, \mathcal{N}$

$\mathcal{G}^* \leftarrow$ graph obtained by removing all edges between units i, j in $\mathcal{G}^{\text{complete}}$ when $(i,j) \notin \mathcal{N}$

score change $\leftarrow$ True

while score change **do**

    score change $\leftarrow$ False

    $\mathcal{E}^*_{\text{proto}\,\mathcal{N}} \leftarrow$ prototypes of network ties in $\mathcal{G}^*$

    $E_{max} \leftarrow \text{argmax}_{E \in \mathcal{E}^*_{\text{proto}\,\mathcal{N}}} \text{PBIC}(X; \mathcal{G}^* \setminus h(E))$

    if $\text{PBIC}(X; \mathcal{G}^* \setminus h(E_{max}) > \text{PBIC}(X; \mathcal{G}^*)$ **then**

        $\mathcal{G}^* \leftarrow \mathcal{G}^* \setminus h(E_{max})$

        score change $\leftarrow$ True

return $\mathcal{E}^*_{\text{proto}\,\mathcal{N}}$

Algorithm 8 HOMOGENOUS

Inputs: $\mathcal{G}^{\text{complete}}, X$

$\mathcal{E}_{\text{proto}\,\mathcal{N}} \leftarrow$ prototypes of network ties in $\mathcal{G}^{\text{complete}}$

$\mathcal{N}^* \leftarrow$ HOMOGENOUS$(\mathcal{G}^{\text{complete}}, X, \mathcal{E}_{\text{proto}\,\mathcal{N}})$

$\mathcal{E}^*_{\text{proto}\,\mathcal{N}} \leftarrow$ HOMOGENOUS$(\mathcal{G}^{\text{complete}}, X, \mathcal{N}^*)$

return $\mathcal{N}^*$, $\mathcal{E}^*_{\text{proto}\,\mathcal{N}}$

Corollary 3.1. *The* HETEROGENOUS *procedure is consistent.*

Corollary 3.2. *When the true network ties are homogenous, the* HOMOGENOUS *procedure is consistent.*

3.5 Experiments

We evaluate the performance of our proposed algorithms on networks of varying size, for various block sizes, and for different regularity settings. (Regularity refers to the number of neighbors for each unit i in the dependency network $\mathcal{N}$. This setting thus controls the density of the graph.) We consider neighborhoods/blocks of size 4, 8, 16, and 32, with regularity 2 or 3. The ground truth models are homogenous and of the

form shown in Figures 3-2 and 3-3, where we display the case of block size 4. Data is generated from each network via a Gibbs sampler with a burn-in period of 1000 iterations and thinning every 100 iterations using the following equations:

$$p(L_i = 1) = \text{expit}(\tau_1),$$
$$p(A_i = 1 | L_i, \{A_j : j \in \text{nb}_{\mathcal{N}}(i)\}) = \text{expit}(\beta_1 L_i + \beta_2 \sum_{j \in \text{nb}_{\mathcal{N}}(i)} A_j),$$
$$p(Y_i = 1 | L_i, A_i, \{A_j : j \in \text{nb}_{\mathcal{N}}(i)\}) = \text{expit}(\nu_1 L_i + \nu_2 A_i + \nu_3 \sum_{j \in \text{nb}_{\mathcal{N}}(i)} A_j),$$

where $\text{expit}(x) = (1 + \exp(-x))^{-1}$. We emphasize that some of these networks are quite large; for example, the network with block size 32 and 2000 iid blocks has an effective size of 64,000 individuals. For each network setting we run 100 bootstraps (bootstrapping blocks rather than individuals) of structure learning to get an average estimate of precision and recall as shown in Figures 3-4 and 3-5. To spare computation time, we use only Algorithm 6 on the latter two block settings. An interesting feature of the results in Figures 3-4 and 3-5, which matches our earlier intuition, is the faster convergence of the homogenous procedures to the true model – which we attribute to the parameter sharing (effectively using of more data when testing each edge deletion.)

In order to demonstrate the utility of learning the structure in dealing with network uncertainty, we consider the population average overall effect in Eq. 3.1. We first execute structure learning, and then estimate the PAOE, contrasting a treatment assignment determined with probability 0.7 with the naturally observed probability (on the mean difference scale.) We do this for 2-regular networks with 2000 realizations of iid blocks of varying size. We use the heterogenous procedure and one of the homogenous procedures (Alg. 6) to learn the structure of the networks. Estimation of the causal effect is done by the auto-g-computation algorithm described in [33]. We perform 1000 bootstraps of both structure learning and effect estimation to compare the bias and variance of the estimates from the learned graphs to the estimates provided by utilizing the maximally uninformative complete graph. Unfortunately the

auto-g-computation procedure is computationally intensive because it requires Gibbs sampling. To spare computation time we do not run the heterogenous procedure on the larger graphs with block sizes 16 and 32 (networks with 32,000 and 64,000 individuals). We also only perform 8 bootstraps for these larger networks. In order to emphasize the need to deal with interference and network uncertainty appropriately, we estimated the bias for 200 bootstraps of the network with blocks of size 8 using the empty graph (a complete iid assumption), and an incorrect graph where $\mathcal{N}$ is shuffled randomly to have incorrect adjacencies. In both cases the bias turned out to be approximately .06, an order of magnitude higher than the bias from using the complete or learned graphs.

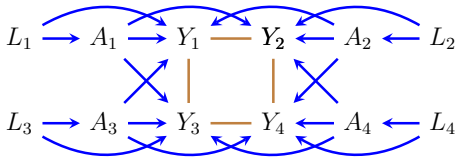


Figure 3-2. The 2-regular CG for a block/neighborhood of size 4

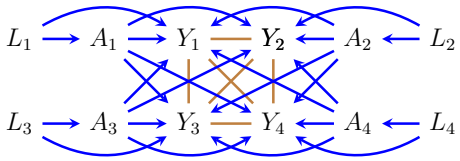


Figure 3-3. The 3-regular CG for a block/neighborhood of size 4

Block Size	Complete	Homogenous	Heterogenous
4	.009, 9.2e-5	.008, 8.1e-5	.009, 9.7e-5
8	.007, 6.6e-5	.006, 4.1e-5	.006, 4.5e-5
16	.006, 3.8e-5	.005, 1.9e-5	x
32	.007, 6.1e-5	.002, 7.6e-6	x

Table 3-I. Bias and variance for estimating the PAOE.

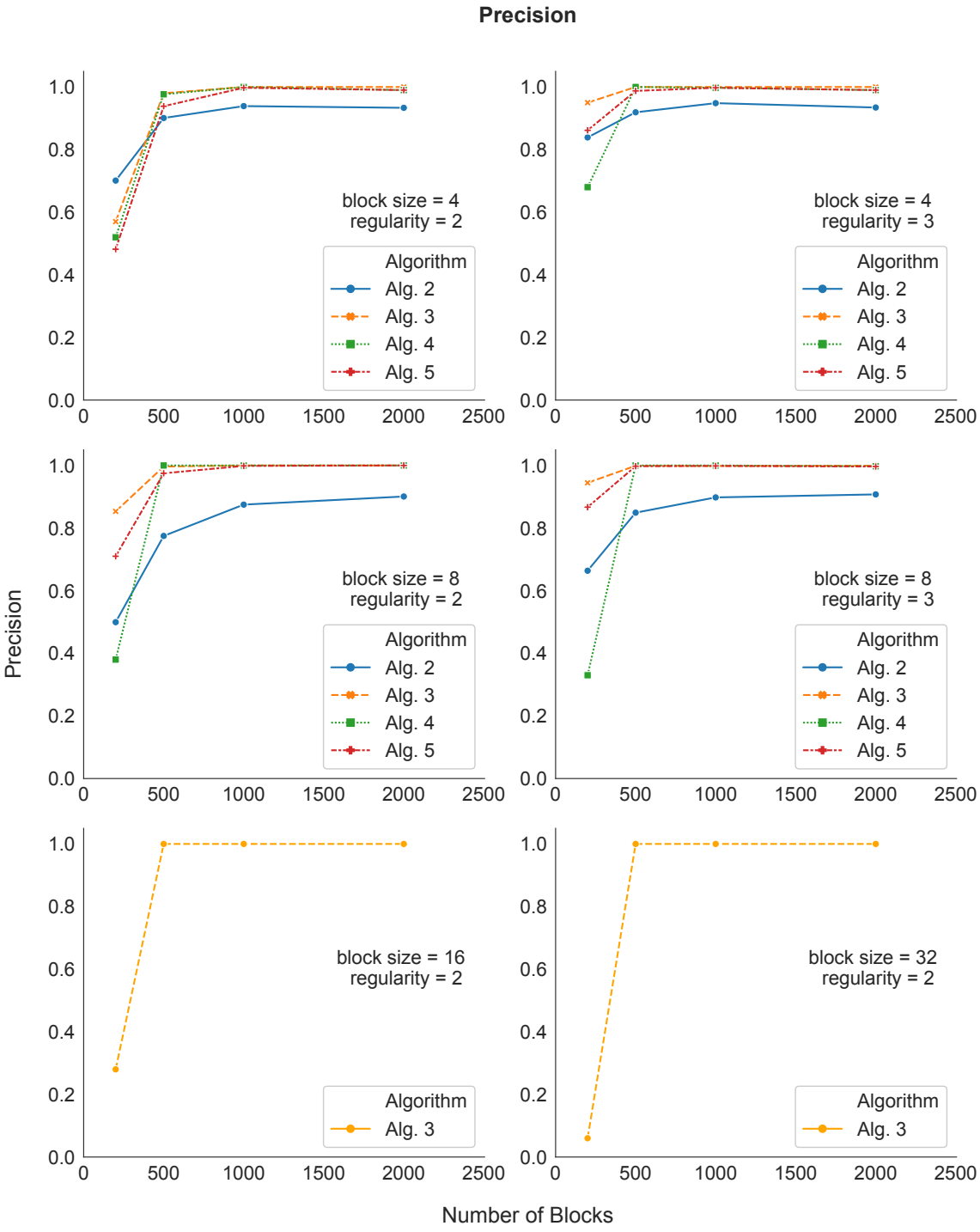


Figure 3-4. Performance of structure learning algorithms as measured by precision.

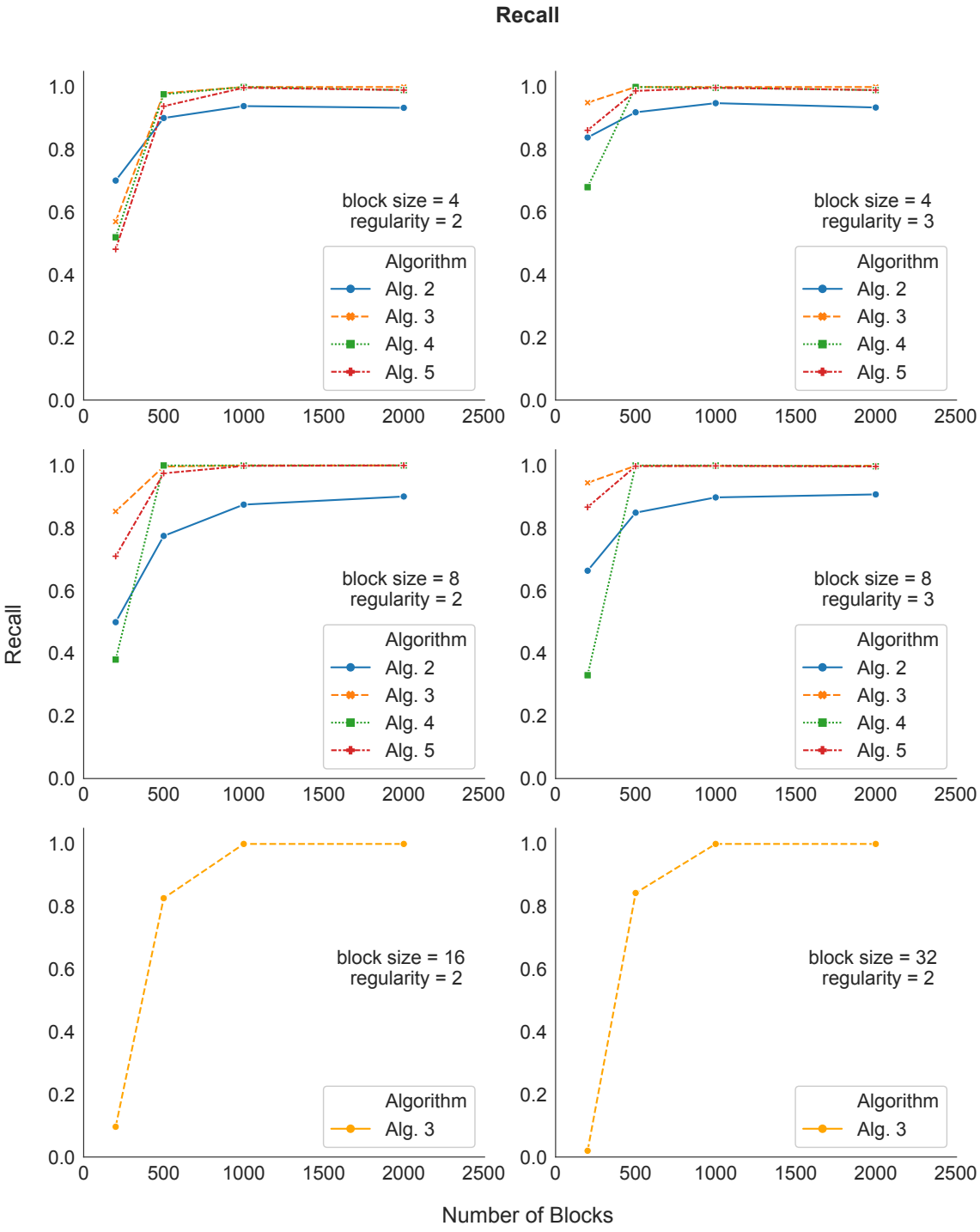


Figure 3-5. Performance of structure learning algorithms as measured by recall.

From Table 3-I we see that causal effect estimates based on learned structure have the same or lower bias as compared with using the complete graph. Furthermore, the

sparsity of the learned graph reduces variance of the estimates in most cases. This reduction in bias and variance is more easily achieved when we are able to exploit homogeneity in the network structure. In experiments with lower sample sizes, we see that the bias of effect estimates may increase (because the learning procedure may fail to recover the true graph) but that the variance of the estimates remains comparable to or lower than the estimates based on the complete graph.

3.6 Related and Future Work

We have developed a method for estimating causal effects under unit dependence induced by a network represented by a chain graph (CG) model [25], when there is uncertainty about network structure. Instead of estimating causal effects given a completely uninformative network where each pair of units is connected, as is typically done in the interference literature [89], we estimated causal effects given a sparser network learned via a score-based model selection method based on the pseudolikelihood function [103]. We showed that this strategy can yield lower variance in estimates without sacrificing bias, if the underlying true network structure is recovered accurately. It is worth noting that similar results may be achieved if one were able to develop regularization techniques for causal parameters in the interference setting, such as the PAOE. However, regularization for causal inference must be given careful treatment as the intended purpose of regularization is typically to improve out-of-sample prediction rather than perform inference. This may lead to biased estimates of the desired causal parameters; see [104, 105] for details.

Our model selection method relied on weak parametric assumptions, specifically that all Markov factors in the CG model corresponded to conditional Markov random fields in the exponential family. The approach here is a generalization of local score-based search algorithms for directed acyclic graph (DAG) models [93] to CG models. As a price of this generalization, our local search algorithms recompute a potentially

larger part of the model score with every move through the model space. An alternative to the pseudolikelihood approach we have taken here is Monte Carlo methods for efficient computation of the normalizing functions as in [106]. However, this technique is restricted to Gaussian graphical models where recomputation of the PBIC is fairly straightforward.

Though the discussion in this chapter was focused on interacting units due to an underlying social network, the methods proposed are readily adaptable to settings where units (possibly even non-human units, such as individual power plants on a grid) may be connected based on spatial proximity as in [107] and [108].

Finally, our approach only works for settings with *partial interference*, where units within a block exhibit dependence, but data on blocks is iid. The restriction to blocks of identical size may be relaxed by combining our heterogeneous procedure with a scheme of parameter sharing and hierarchical modeling across blocks that are of different sizes. In future work, we aim to extend our methods to *full interference* settings and settings with unmeasured confounders modeled by segregated graphs (mixed graphs with directed, undirected, and bidirected edges) [109]. In such settings, model selection techniques will have to rely on alternatives to the BIC, such as the PBIC, as the normalizing function is intractable [33]. Additionally, quantifying uncertainty post model selection for certain effects that necessarily rely on the presence/absence of edges when defining the estimand (such as the spillover effect) may be quite challenging.

3.7 Acknowledgements

Chapter 3 is, in part, a reprint of material from [11].

Chapter 4

Identification in the Presence of Missing Data

Missing data is ubiquitous in applied data analyses resulting in target distributions that are systematically censored by a missingness process. A common modeling approach assumes data entries are censored in a way that does not depend on the underlying missing data, known as the missing completely at random (MCAR) model, or only depends on observed values in the data, known as the missing at random (MAR) model [110]. However, these simple models are insufficient for problems where missingness status may depend on underlying values that are themselves censored. This type of missingness is known as missing not at random (MNAR).

Similar to causal inference, recovery of parameters from censored distributions requires the analyst to pose restrictions on the full data distribution which consists of the target distribution and its missingness process. While there exist MNAR models whose restrictions cannot be represented graphically [111], the restrictions posed in several popular MNAR models such as the permutation model [112], the block-sequential MAR model [113], the itemwise conditionally independent nonresponse (ICIN) model or no self-censoring model [114, 115], and those posed in [116, 117, 118, 119, 120, 121] are either explicitly graphical or can be interpreted as such.

The problem of identification of the target distribution from the observed data

distribution in missing data DAG models bears many similarities to the problem of identification of counterfactual distributions from the observed distribution in causal DAG models with hidden variables. This observation prompted recent work [119, 120, 122] on adapting identification methods from causal inference to identifying target distributions in graphical models of missing data and leads to a natural interpretation of missing data problems as causal inference problems.

In this chapter we show that the most general methods known for identification in missing data DAG models retain a significant gap in the sense that they fail to identify the target distribution in many models where it is identified. We show that methods used to obtain a complete characterization (ensuring successful recovery of identifying functionals whenever possible) of identification of counterfactual distributions in hidden variable DAG models, e.g., via truncated nested Markov factorization as described in Chapter 1, and simple generalizations of them, are insufficient for obtaining a complete characterization for missing data problems. We describe, via a set of examples, that in order for a missing data identification algorithm to be complete, the algorithm must recursively simplify the problem by removing sets of variables rather than single variables, and these must be removed according to a *partial order* rather than a total order. Furthermore, the algorithm must be able to handle subproblems where selection bias, hidden variables, or both, are present even if these complications are missing in the original problem. We pose a new identification algorithm that exploits these observations and significantly narrows the identifiability gap in existing methods. Finally we show that in certain classes of missing data DAGs, our algorithm takes on a particularly simple formulation to identify the underlying target distribution.

Chapter 4 is organized as follows. We begin with the necessary preliminaries on missing data models, their graphical representations, and identification of the target and full data distributions. We then present examples that demonstrate a gap in missing data identification theory. We then present our algorithm, simplifications of

it, and some results on graphical structures that prevent non-parametric identifiability of the full data distribution of missing data DAG models. We conclude with a short discussion of related work and new missing data identification theory that draw on results presented in this chapter.

4.1 Missing Data Models

A missing data model is a set of distributions defined over the following sets of random variables.

$X^{(1)}$: where $X_i^{(1)} \in X^{(1)}$ is a variable that is potentially missing.

R : where $R_i \in R$ is a missingness indicator for $X_i^{(1)}$.

X : where X_i is an observed proxy for $X_i^{(1)}$.

O : where $O_i \in O$ is a variable that is always observed.

While the state space of variables in $X^{(1)}$ and O are unrestricted, missingness indicators are binary random variables by definition. The state space of each observed proxy X_i is also restricted to be the same as $X_i^{(1)}$ along with a special symbol, such as "?", which can be used to represent that the true value of the variable $X_i^{(1)}$ is unobserved. More formally, given $X_i^{(1)} \in X^{(1)}$ and its corresponding missingness indicator $R_i \in R$, the observed proxy X_i is defined as $X_i \equiv X_i^{(1)}$ if $R_i = 1$, and $X_i = ?$ if $R_i = 0$. Hence, $p(X \mid R, X^{(1)})$ is deterministically defined. The non-deterministic part of a missing data distribution, i.e, $p(O, X^{(1)}, R)$, is known as the *full law*, and can be partitioned into two pieces: the *target law* $p(O, X^{(1)})$ and the *missingness mechanism* $p(R \mid X^{(1)}, O)$. The censored version of the full law $p(O, R, X)$, that the analyst actually has access to is known as the *observed data distribution*.

Missing data DAG models are defined as follows. Let $\mathcal{G}(V)$ be a DAG, where $V = O \cup X^{(1)} \cup R \cup X$. Following the convention in [119], edges in $\mathcal{G}$ are subject to the following restrictions in addition to acyclicity: there are no outgoing edges from

indicators in R to variables in $X^{(1)}$ and/or O, each observed proxy $X_i \in X$ has only two parents R_i and $X_i^{(1)}$ which encodes the deterministic nature of X_i (these edges are shown in gray in all the figures below), and there are no outgoing edges from X_i (i.e., the proxy X_i does not cause any variable on the DAG, however the corresponding full data variable $X_i^{(1)}$ may cause other variables.) The statistical model of a missing data DAG $\mathcal{G}(V)$ is defined as the set of distributions $p(O, X^{(1)}, R, X)$ that factorize as,

$$\prod_{V_i \in O \cup X^{(1)} \cup R} p(V_i \mid \mathrm{pa}_{\mathcal{G}}(V_i)) \prod_{X_i \in X} p(X_i \mid X_i^{(1)}, R_i). \qquad \textit{(Missing data DAG factorization)} \tag{4.1}$$

By standard results on DAG models, conditional independences in $p(X^{(1)}, O, R)$ can still be read off from $\mathcal{G}$ by the d-separation criterion [5]. For convenience, we will drop the deterministic terms of the form $p(X_i \mid X_i^{(1)}, R_i)$ from the identification analyses in the following sections since these terms are always identified by construction.

We also consider hidden variable DAGs $\mathcal{G}(V \cup H)$, where $V = O \cup X^{(1)} \cup R \cup X$ and variables in H are unobserved, to encode missing data models in the presence of unmeasured confounders. In such cases, the full law satisfies the nested Markov factorization described in Chapter 1 with respect to a missing data ADMG $\mathcal{G}(V)$, obtained by applying the latent projection operator to the hidden variable DAG $\mathcal{G}(V \cup H)$. Similar to when there was no missingness in the problem, marginalization of latents H may give rise to bidirected edges encoding hidden common causes between variables in V. It is straightforward to see that the missing data ADMG obtained via latent projection of a hidden variable missing data DAG satisfies all the restrictions mentioned in the previous paragraph. In particular, $\mathcal{G}(V)$ has no directed cycles, $\mathrm{pa}_{\mathcal{G}}(X_i) = \{X_i^{(1)}, R_i\}$, every $X_i \in X$ is childless, and there are no outgoing edges from R_i to any variables in $X^{(1)} \cup O$. Ordinary conditional independences in a missing data ADMG model can be read off from $\mathcal{G}(V)$ via the m-separation criterion as before.

4.1.1 Identification in Missing Data Models

An analyst may pursue a few different tasks related to non-parametric identification in missing data models. These include identification of the target law $p(O, X^{(1)})$, identification of specific functions of the target law $f(p(O, X^{(1)}))$, and identification of the full law $p(O, X^{(1)}, R)$. Though the focus of this chapter is on identification of the target law of missing data DAG models, some of our results naturally extend to identification of the full law as well as specific functionals of the target law. By the chain rule of probability, the target law $p(O, X^{(1)})$ is identified *if and only if* $p(R = 1 \mid O, X^{(1)})$ is identified. The identifying functional is given by,

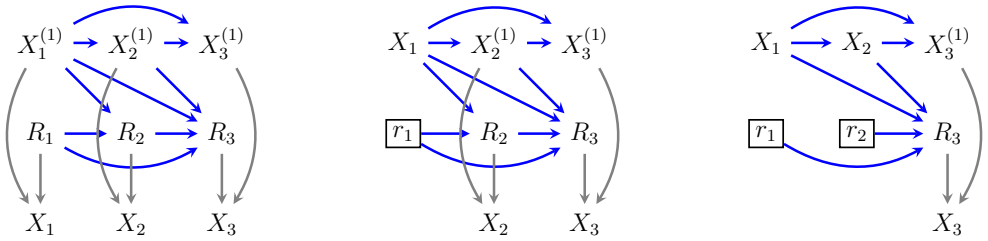
$$p(O, X^{(1)}) = \frac{p(O, X^{(1)}, R = 1)}{p(R = 1 \mid O, X^{(1)})}, \qquad \textit{(Identification of target law)} \tag{4.2}$$

where in the numerator of the right hand side $X^{(1)} = X$, and is observed when $R = 1$ by definition. The full law $p(O, X^{(1)}, R)$ is identified *if and only if* $p(R \mid O, X^{(1)})$ is identified. According to Eq. 4.2, the identifying functional is,

$$p(O, X^{(1)}, R) = \frac{p(O, X^{(1)}, R = 1)}{p(R = 1 \mid O, X^{(1)})} \times p(R \mid O, X^{(1)}). \quad \textit{(Identification of full law)} \tag{4.3}$$

4.2 Gaps in Current Identification Theory

In this section, we describe a set of examples of missing data models that factorize as in Eq. 4.1 for different DAGs, where the target law is identified. We start with simpler examples where sequential fixing techniques from causal inference suffice to obtain identification, then move on to describe more complex examples where existing algorithms in the literature suffice, and finally proceed to examples where no published method known to us obtains identification, illustrating an identifiability gap in existing methods. In these examples, we show how identification may be obtained by appropriately generalizing existing techniques. In these discussions, we focus on obtaining identification of the missingness mechanism $p(R \mid X^{(1)}, O)$ evaluated at

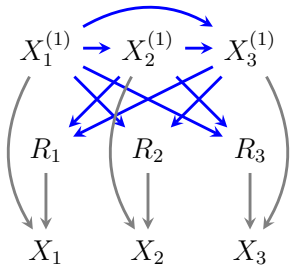


(a) $\mathcal{G}$ (b) $\mathcal{G}_1 \equiv \phi_{R_1}(\mathcal{G})$ (c) $\mathcal{G}_2 \equiv \phi_{R_2}(\mathcal{G}_1)$

Figure 4-1. (a), (b), (c) are intermediate graphs obtained in identification of a block-sequential model by fixing $\{R_1, R_2, R_3\}$ in sequence.

$R = 1$, as this suffices to identify the target law $p(X^{(1)}, O)$ by Eq. 4.2. In the course of describing these examples, we will obtain intermediate graphs and kernels. In these graphs, lower case letters (e.g. v) indicates the variable V is evaluated at v (for $R_i, r_i = 1$). A square vertex indicates V had been fixed. Drawing the vertex normally with lower case indicates V was conditioned on (creating selection bias in the subproblem.) For brevity, we use 1_{R_i} to denote $\{R_i = 1\}$.

We first consider the block-sequential MAR model [113], shown in Figure 4-1 for three variables. The target law is identified by applying the (valid) fixing sequence (R_1, R_2, R_3) via the fixing operator ϕ to $\mathcal{G}$ and $p(R, X)$ as follows. $p(R_1 \mid \mathrm{mb}_{\mathcal{G}}(R_1)) = p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) = p(R_1)$ is identified immediately. Applying the fixing operator ϕ_{R_1} yields the graph $\mathcal{G}_1 \equiv \phi_{R_1}(\mathcal{G})$ shown in Figure 4-1(b), and a corresponding kernel $q_1(X_1^{(1)}, X_2, X_3, R_2, R_3 \mid 1_{R_1}) \equiv p(X_1, X_2, X_3, R_2, R_3, 1_{R_1})/p(1_{R_1})$ where $X_1^{(1)}$ is now observed. Thus, in the new subproblem represented by $\mathcal{G}_1$ and q_1, $p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))|_{R=1} = q_1(R_2 \mid X_1^{(1)}, 1_{R_1})$ is identified. Applying the fixing operator ϕ_{R_2} to $\mathcal{G}_1$ and q_1 yields $\mathcal{G}_2 \equiv \phi_{R_2}(\mathcal{G}_1)$ shown in Figure 4-1(c), and $q_2(X_1^{(1)}, X_2^{(1)}, X_3, R_3 \mid 1_{R_1,R_2}) = q_1(X_1^{(1)}, X_2, X_3, R_2, R_3 \mid 1_{R_1})/q_1(R_2 \mid X_1^{(1)}, 1_{R_1})$. Finally, in the new subproblem represented by $\mathcal{G}_2$ and q_2, $p(R_3 \mid \mathrm{pa}_{\mathcal{G}}(R_3))|_{R=1} = q_2(R_3 \mid X_1^{(1)}, X_2^{(1)}, 1_{R_1,R_2})$ is identified. Applying the fixing operator ϕ_{R_3} to $\mathcal{G}_2$ and q_2 yields $q_3(X_1^{(1)}, X_2^{(1)}, X_3^{(1)} \mid 1_{R_1,R_2,R_3}) = p(X_1^{(1)}, X_2^{(1)}, X_3^{(1)})$. The identifying functional for the target law only involves monotone



(d)

Figure 4-2. An MNAR model that is identifiable by fixing all Rs in parallel.

cases (cases where $R_i = 0$ implies $R_{i+1} = 0$) just as would be the case under the monotone MAR model, although this model does not assume monotonicity and is not MAR. In this simple example, identification may be achieved purely by causal inference methods, by treating the missingness indicators R as treatments and finding a valid fixing sequence on them. In this example, each R_i in the sequence is fixable given that the previous variables are fixable, since all parents of R_i become observed at the time it is fixed.

Following a total order to fix is not always sufficient to identify the target law, as noted in [119, 120, 122]. Consider the model represented by the DAG in Figure 4-2. For any R_i in this model, say R_1, we have, by d-separation, that $p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) = p(R_1 \mid X_2^{(1)}, X_3^{(1)}, 1_{R_2,R_3})$, which is identified. However, if we were to fix R_1 in $p(X, R)$, we would obtain a kernel $q_1(X_1^{(1)}, X_2, X_3, 1_{R_2,R_3} \mid 1_{R_1})$ where selection bias on R_2 and R_3 is introduced. The fact that q_1 is not available at all levels of R_2 and R_3 prevents us from sequentially obtaining $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))$, for $R_i = R_2, R_3$, due to our inability to sum out those variables from q_1.

However, the model in Figure 4-2 allows identification of a target law in another way. This follows from the fact that $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))$ is identified for each R_i by exploiting conditional independences in $p(X, R)$ displayed by Figure 4-2. Since $p(R \mid X^{(1)}) = \prod_{i=1}^{3} p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))$, the missingness mechanism is identified, which means

the target law is also identified, as long as we fix R_1, R_2, R_3 *in parallel* (as in Eq. 4.2) rather than sequentially. In other words, the model is identified, but no total order on fixing operations suffices for identification. A general algorithm that aimed to fix indicators in R in parallel, while potentially exploiting causal inference fixing operations to identify each $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))$ was proposed in [122]. Our subsequent examples show that this algorithm is insufficient to obtain identification of the target law in general, and thus is incomplete.

Consider the DAG in Figure 4-3. Since R_2 is a child of R_3 and $X_2^{(1)}$ is a parent of R_3, we cannot obtain $p(R_3 \mid \mathrm{pa}_{\mathcal{G}}(R_3)) = p(R_3 \mid X_2^{(1)})$ by d-separation in any kernel (including the original distribution) where R_2 is not fixed. Thus, any total order on fixing operations of elements in R must start with R_1 or R_2. Fixing either of these variables entails dividing $p(X, R)$ by some factor $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))$, which is identified as either $p(R_1 \mid X_3^{(1)}, 1_{R_3})$ or $p(R_2 \mid X_1^{(1)}, 1_{R_1})$. This division entails inducing selection bias on the subsequent kernel q_1 for a variable not yet fixed (either R_1 or R_3). Thus, no total order on fixing operations works to identify the target law in this model. At the same time, attempting to fix all R variables in parallel would fail as well, since we cannot identify $p(R_3 \mid X_2^{(1)})$ either in the original distribution or any kernel obtained by standard causal inference operations described in [122]. In particular, in any such kernel or distribution R_3 remains dependent on R_2 given $X_2^{(1)}$.

However, the target law in this model is identified by following a *partial order* of fixing operations. In this partial order, R_1 is incompatble with R_2, and $R_2 \prec R_3$. This results in an identification strategy where we fix each variable *only* given that variables earlier than it in the *partial* order are fixed. That is, distributions $p(R_1 \mid X_3^{(1)}) = p(R_1 \mid X_3, 1_{R_3})$ and $p(R_2 \mid X_1^{(1)}, R_3) = p(R_2 \mid X_1, 1_{R_1}, R_3)$ are obtained directly in the original distribution without fixing anything. The distribution $p(R_3 \mid \mathrm{pa}_{\mathcal{G}}(R_3))$, on the other hand, is obtained in the kernel $q_1(X_1, X_2^{(1)}, X_3, 1_{R_1}, R_3 \mid 1_{R_2}) = p(X, R)/p(R_2 \mid X_1, 1_{R_1}, R_3)$ after R_2 (the variable earlier than R_3 in the partial

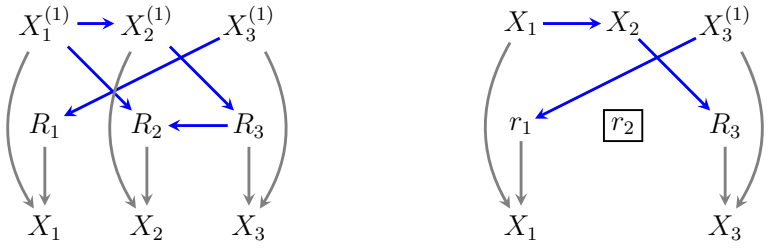


Figure 4-3. (a) A DAG where Rs are fixed according to a partial order. (b) The CADMG obtained by fixing R_2.

order) is fixed. The graph corresponding to this kernel is shown in Figure 4-3 (b). Note that in this graph $X_2^{(1)}$ is observed, and there is selection bias on R_1. However, it easily follows by d-separation that R_3 is independent of R_1. It can thus be shown that $p(R_3 \mid X_2^{(1)}) = q_1(R_3 \mid X_2^{(1)}, 1_{R_2})$ even if q_1 is only available at value $R_1 = 1$. Since all $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))$ are identified, so is the target law in this model, by Eq. 4.2.

Next, we consider the model in Figure 4-4. Here, $p(R_2 \mid X_1^{(1)}, X_3^{(1)}, R_1) = p(R_2 \mid X_1, X_3, 1_{R_1,R_3})$ and $p(R_3 \mid X_2^{(1)}, R_1) = p(R_3 \mid X_2, 1_{R_2}, R_1)$ are identified immediately. However, $p(R_1 \mid X_2^{(1)})$ poses a problem. In order to identify this distribution, we either require that R_1 is conditionally independent of R_2, possibly after some fixing operations, or we are able to render $X_2^{(1)}$ observable by fixing R_2 in some way. Neither seems to be possible in the problem as stated. In particular, fixing R_2 via dividing by $p(R_2 \mid X_1^{(1)}, X_3^{(1)}, R_1)$ will necessarily induce selection bias on R_1, which will prevent identification of $p(R_1 \mid X_2^{(1)})$ in the resulting kernel.

However, we can circumvent the difficulty by treating $X_1^{(1)}$ as an *unobserved variable* U_1, and attempting the problem in the resulting (hidden variable) DAG shown in Figure 4-4 (b), and its latent projection ADMG $\tilde{\mathcal{G}}$ shown in Figure 4-4 (c), where U_1 is "projected out." In the resulting problem, we can fix variables according to a partial order $\prec$ where R_2 and R_3 are incompatible, $R_2 \prec R_1$, and $R_3 \prec R_1$. Thus, we are able to fix R_2 and R_3 in parallel by dividing by $p(R_2 \mid \mathrm{mb}_{\tilde{\mathcal{G}}}(R_2)) = p(R_2 \mid X_1, R_1, X_3^{(1)}, 1_{R_3})$

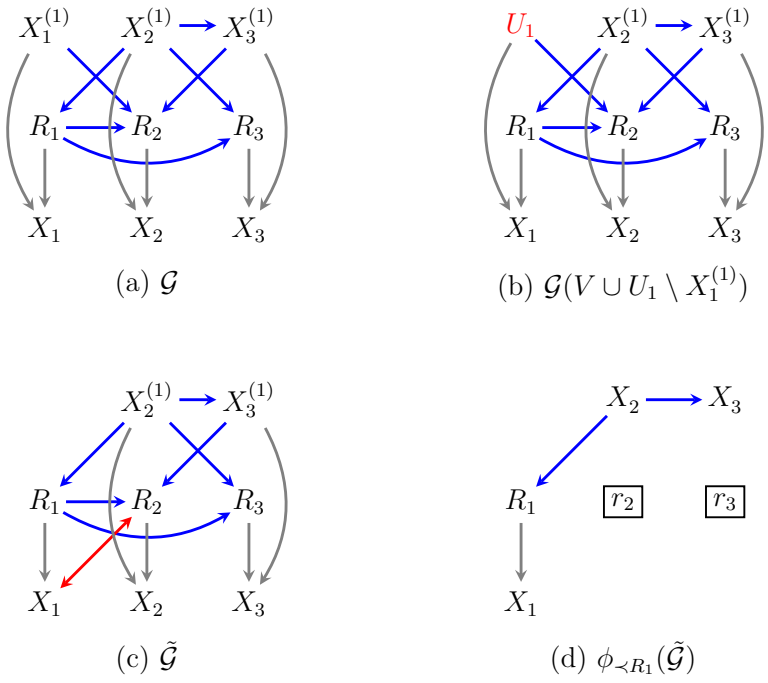


Figure 4-4. A DAG where selection bias on R_1 is avoidable by following a partial order fixing schedule on an ADMG induced by latent projecting out $X_1^{(1)}$.

and $p(R_3 \mid R_1, X_2^{(1)}) = p(R_3 \mid R_1, X_2, 1_{R_2})$, leading to a kernel $\tilde{q}_1(X_1, X_2^{(1)}, X_3^{(1)}, R_1 \mid 1_{R_2,R_3})$, and the graph $\phi_{\prec R_1}(\tilde{\mathcal{G}})$ shown in Figure 4-4(d), where notation $\phi_{\prec R_1}$ means "fix all necessary elements that occur earlier than R_1 in the partial order, in a way consistent with that partial order." In this example, this means fixing R_2 and R_3 in parallel. We will describe how fixing operates given general *fixing schedules* given by a partial order later in the paper. In the kernel $\tilde{q}_1$ the parent of R_1 is observed data, meaning that $p(R_2 \mid X_2^{(1)})$ is identified as $\tilde{q}_1(R_1 \mid X_2, 1_{R_2,R_3})$. This implies the target law is identified in this model.

In general, to identify $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))$, we may need to use separate partial fixing orders on different sets of variables for different $R_i \in R$. In addition, the fact that fixing introduces selection bias sometimes results in having to divide by a kernel where a *set* of variables are random, something that was never necessary in causal inference problems. In general, for a given R_i, the goal of a fixing schedule is to arrive at a kernel

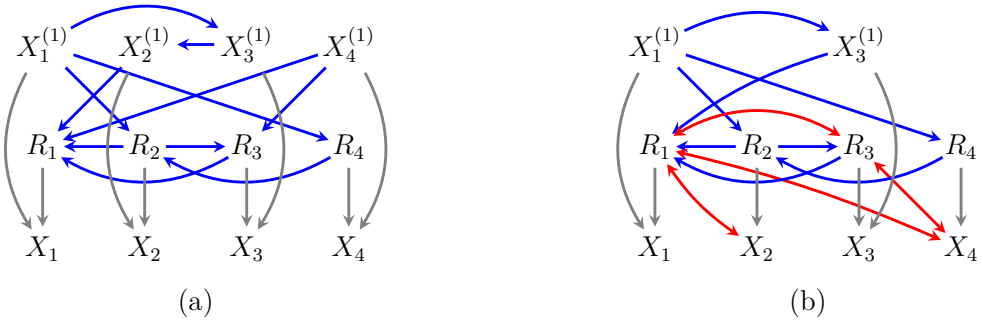


Figure 4-5. (a) A DAG where the fixing operator must be performed on a set of vertices. (b) A latent projection of a subproblem used for identifiation of $p(R_4 \mid X_4^{(1)})$.

where an independence exists allowing us to identify $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))$, even if some elements of $\text{pa}_{\mathcal{G}}(R_i)$ are in $X^{(1)}$ in the original problem. This fixing must be given by a partial order, and sometimes on sets of variables. In addition, some elements of $X^{(1)}$ must be treated as hidden variables. These complications are necessary in general to avoid creating selection bias in subproblems, and ultimately to identify the missingness mechanism. The following example is a good illustration.

Consider the graph in Figure 4-5(a). For R_1 and R_3, the fixing schedules are empty, and we immediately obtain their distributions as $p(R_1 \mid X_2^{(1)}, X_4^{(1)}, R_2, R_3) = p(R_1 \mid X_2, X_4, R_3, 1_{R_2,R_4})$ and $p(R_3 \mid X_4^{(1)}, R_2) = p(R_3 \mid X_4, 1_{R_4}, R_2)$. For R_2, the partial order is $R_3 \prec R_1$ in a graph where we treat $X_2^{(1)}$ as a hidden variable U_2. This yields $p(R_2 \mid X_1^{(1)}, R_4) = q_2(R_2 \mid X_1^{(1)}, R_4, 1_{R_1,R_3})$, where

$$q_2(X_1^{(1)}, X_2, X_3^{(1)}, X_4, R_2, 1_{R_4} \mid 1_{R_1,R_3}) = \frac{q_1}{q_1(1_{R_1} \mid 1_{R_3}, R_2, X_2)}, \text{ and}$$

$$q_1(X_1, X_2, X_3^{(1)}, X_4, R_1, R_2, 1_{R_4} \mid 1_{R_3}) = \frac{p(X, R_1, R_2, 1_{R_3,R_4})}{p(1_{R_3} \mid R_2, X_4, 1_{R_4})}.$$

In order to obtain the propensity score for R_4 we must either render $X_1^{(1)}$ observable through fixing R_1 or perform valid fixing operations until we obtain a kernel in which R_4 is conditionally independent of R_1 given its parent $X_1^{(1)}$. There exists no partial order on elements of R, however, all partial orders on elements in R induce selection bias on variables higher in the order, preventing the identification of the required

distribution for R_4. For example, choosing a partial fixing order of $R_1 \prec R_3$, where we treat $X_2^{(1)}, X_4^{(1)}$ as hidden variables results in selection bias on R_3 as soon as we fix R_1. Other partial orders fail similarly. However, the following approach is possible in the graph in which we treat $X_2^{(1)}, X_4^{(1)}$ as hidden variables.

R_1, R_3 lie in the same district in the resulting latent projection ADMG, shown in Figure 4-5(b). Moreover, $\{R_1, R_3\}$ is closed under descendants in the district in Figure 4-5(b). As a result, R_1 and R_3 can essentially be viewed as a single vertex from the point of view of fixing. Indeed we may choose a partial order $\{R_1, R_3\} \prec R_2$, where we fix R_1 and R_3 as a set. The fixing operation on the set is possible since $p(R_1, R_3 \mid \mathrm{mb}_{\mathcal{G}}(R_1, R_3))$ is a function of $p(X, R)$. Specifically it is equal to

$$p(1_{R_1,R_3} \mid X_3^{(1)}, R_2, R_4, X_2, X_4) = p(1_{R_3} \mid R_2, R_4, X_2, X_4) \\ \times p(1_{R_1} \mid R_2, R_4, X_2, X_3, X_4, 1_{R_3}),$$

where the above equality holds by d-separation $(R_3 \perp\!\!\!\perp X_3^{(1)} \mid R_2, R_4, X_4, X_2)$. We then obtain $p(R_4 \mid X_1^{(1)}) = q_2(R_4 \mid X_1^{(1)}, 1_{R_1,R_2,R_3})$, where

$$q_2(X_1^{(1)}, X_3^{(1)}, X_4, R_4 \mid 1_{R_1,R_2,R_3}) = \frac{q_1}{q_1(R_2 \mid X_1^{(1)}, R_4, 1_{R_1,R_3})},$$

$$q_1(X_1^{(1)}, X_2, X_3^{(1)}, X_4, R_2, R_4 \mid 1_{R_1,R_3}) = \frac{p(X, R_2, R_4, 1_{R_1,R_3})}{p(1_{R_3} \mid R_2, R_4) \times p(1_{R_1} \mid R_2, R_4, X_3, 1_{R_3})}.$$

Our final example demonstrates that in order to identify the target law, we may potentially need to fix variables outside R, including variables in $X^{(1)}$ that become observed after fixing or conditioning on some elements of R. Figure 4-6 (a) contains a generalization of the model considered in [122], where O_3 is fully observed. In this model, distributions for R_4 and R_1 are identified immediately, while identification of R_2 requires a partial order $R_4 \prec X_4^{(1)} \prec O_3 \prec R_1$ in the graph where we treat $X_1^{(1)}, X_2^{(1)}, X_4^{(1)}$ as latent variables (with the latent projection ADMG shown in Figure 4-6 (b)) until they are rendered observed by fixing the corresponding

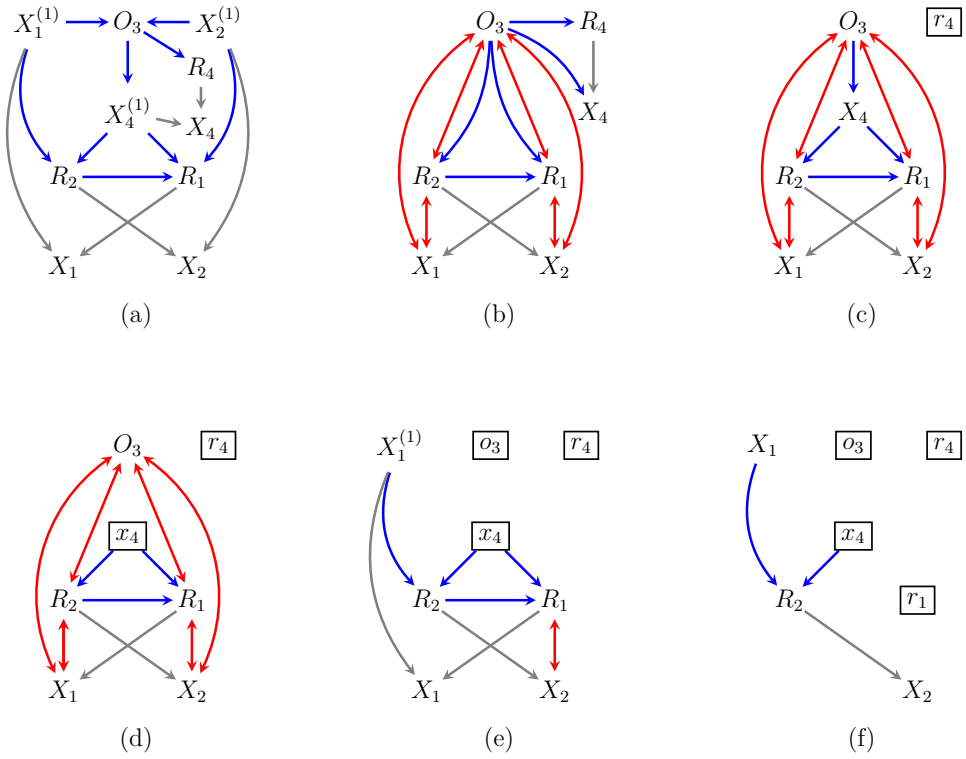


Figure 4-6. A DAG where variables besides Rs are required to be fixed.

missingness indicators. To illustrate fixing operations according to this order, the intermediate graphs that arise are shown in Figures 4-6 (c),(d),(e),(f).

4.3 A New Identification Algorithm

In order to identify the target law in examples discussed in the previous section, we had to consider situations where some variables were viewed as hidden, and marginalized out, and others were conditioned on, introducing selection bias. In addition, fixing operations were performed according to a partial order rather than a total order as was the case in causal inference problems. Finally, we sometimes fixed sets of variables jointly, rather than individual variables. We now introduce relevant definitions that allow us to formulate a general identification algorithm that takes advantage of all these techniques.

Let V be a set of random variables (and corresponding vertices) consisting of observed variables O, R, X, missing variables $X^{(1)}$, and selected variables S. Let W be a set of fixed observed variables. The following definitions apply to a latent projection $\mathcal{G}(V \setminus X^{(1)}, W)$, for some $X_U^{(1)} \subseteq X^{(1)}$, and a corresponding kernel $q(V \setminus X^{(1)} \mid W) \equiv \sum_{X^{(1)}} q(V, W)$. $\mathcal{G}$ can be viewed as a latent variable CADMG for q where $X_U^{(1)}$ are latent. Such CADMGs represent intermediate subproblems in our identification algorithm.

For $Z \subseteq D_Z \in \mathcal{D}(\mathcal{G})$, let $R_Z = \{R_j \mid X_j^{(1)} \in Z \cup \mathrm{mb}_{\mathcal{G}}(Z), R_j \notin Z\}$, and $\mathrm{mb}_{\mathcal{G}}(Z) \equiv (D_Z \cup \mathrm{pa}_{\mathcal{G}}(D_Z)) \setminus Z$. We say Z is *fixable* in $\mathcal{G}(V \setminus X_U^{(1)}, W)$ if

(i) $\mathrm{de}_{\mathcal{G}}(Z) \cap D_Z \subseteq Z$,

(ii) $S \cap Z = \emptyset$,

(iii) $Z \perp\!\!\!\perp (S \cup R_Z) \setminus \mathrm{mb}_{\mathcal{G}}(Z) \mid \mathrm{mb}_{\mathcal{G}}(Z)$.

In words, these conditions apply to some Z that is a subset of its own district (which is trivial when the set Z is a singleton). The conditions, in the listed order, require that Z is closed under descendants in the district, should not contain any selected variables, and should be independent of both selected variables S and the missingness indicators R_Z of the corresponding counterfactal parents given the Markov blanket of Z, respectively. Consider the graph in Figure 4-5(b) where $S = \emptyset$ and let $Z = \{R_1, R_3\}$. Z is fixable since $Z \subseteq D_Z = \{R_1, R_3, X_2, X_4\}, \mathrm{de}_{\mathcal{G}}(Z) = \{R_1, R_3, X_1, X_3\} \cap D_Z = \{R_1, R_3\}$ is closed, and S, R_Z are empty sets.

A set $\tilde{Z}$ spanning multiple elements in $\mathcal{D}(\mathcal{G})$ is said to be fixable if it can be partitioned into a set $\mathcal{Z}$ of elements $Z \equiv \tilde{Z} \cap D \in \mathcal{D}(\mathcal{G})$ such that each such Z is fixable.

Given an ordering $\prec$ on vertices $V \cup W$ topological in $\mathcal{G}$, and $\tilde{Z}$ fixable in $\mathcal{G}$, define

$\phi_{\tilde{Z}}(q;\mathcal{G})$ as

$$\frac{q(V \setminus (X_U^{(1)} \cup R_Z), R_Z = 1 \mid W)}{\prod_{Z \in \mathcal{Z}} \prod_{Z \in Z} q(Z \mid \mathrm{mb}_{\mathcal{G}}(Z; \mathrm{an}_{\mathcal{G}}(D_Z) \cap \{\preceq Z\})), R_Z)|_{(R \cap Z) \cup R_Z = 1}}, \qquad \textit{(Fixing a set)} \quad (4.4)$$

where $\mathrm{mb}_{\mathcal{G}}(V;S) \equiv \mathrm{mb}_{\mathcal{G}_S}(V)$, and $\{\preceq Z\}$ is the set of all elements earlier than Z in the order $\prec$ (this includes Z itself).

Given a set $Z \subseteq R \cup O \cup X^{(1)}$, and an equivalence relation $\sim$, let ${}^{Z}\!/_{\sim}$ be the partition of Z into equivalence classes according to $\sim$. Define a *fixing schedule* for ${}^{Z}\!/_{\sim}$ to be a partial order $\lhd$ on ${}^{Z}\!/_{\sim}$. For each $\tilde{Z} \in {}^{Z}\!/_{\sim}$, define $\{\trianglelefteq \tilde{Z}\}$ to be the set of elements in ${}^{Z}\!/_{\sim}$ earlier than $\tilde{Z}$ in the order $\lhd$, and $\{\lhd \tilde{Z}\} \equiv \{\trianglelefteq \tilde{Z}\} \setminus \tilde{Z}$. Define $\trianglelefteq_{\tilde{Z}}$ and $\lhd_{\tilde{Z}}$ to be restrictions of $\lhd$ to $\{\trianglelefteq \tilde{Z}\}$ and $\{\lhd \tilde{Z}\}$, respectively. Both restrictions, $\trianglelefteq_{\tilde{Z}}$ and $\lhd_{\tilde{Z}}$, are also partial orders.

We inductively define a *valid* fixing schedule (a schedule where fixing operations can be successfully implemented), along with the fixing operator on valid schedules. The fixing operator will implement fixing as in (4.4) on $\tilde{Z}$ within an intermediate problem represented by a CADMG where some $X_{\tilde{Z}}^{(1)} \subseteq X^{(1)}$ will become observed after fixing $\tilde{Z}$, with $X^{(1)} \setminus X_{\tilde{Z}}^{(1)}$ treated as latent variables, and a kernel associated with this CADMG defined on the observed subset of variables. We also define $X_{\{\trianglelefteq \tilde{Z}\}}^{(1)} \equiv \bigcup_{Z \in \{\trianglelefteq \tilde{Z}\}} X_Z^{(1)}$.

We say $\lhd_{\tilde{Z}}$ is valid for $\{\lhd \tilde{Z}\}$ in $\mathcal{G}$ if for every $\lhd$-largest element $\tilde{Y}$ of $\{\lhd \tilde{Z}\}$, $\trianglelefteq_{\tilde{Y}}$ is valid for $\{\trianglelefteq \tilde{Y}\}$. If $\lhd_{\tilde{Z}}$ is valid for $\{\lhd \tilde{Z}\}$, we define $\phi_{\lhd_{\tilde{Z}}}(\mathcal{G})$ to be a new CADMG $\mathcal{G}(V \setminus \bigcup_{Z \in \{\lhd \tilde{Z}\}} Z, W \cup \bigcup_{Z \in \{\lhd \tilde{Z}\}} Z)$ obtained from $\mathcal{G}(V,W)$ by:

(i) Removing all edges with arrowheads into $\bigcup_{Z \in \{\lhd \tilde{Z}\}} Z$,

(ii) Marking any $\{X_j^{(1)} \mid X_j^{(1)} \in Z \cup \mathrm{mb}_{\phi_{\lhd_Z}(\mathcal{G})}(Z), Z \in \{\lhd \tilde{Z}\}\}$ as observed,

(iii) Marking any $\{R_Z \cap V \mid Z \in \{\lhd \tilde{Z}\}\} \setminus \bigcup_{Z \in \{\lhd \tilde{Z}\}} Z$ as selected to value 1, where R_Z is defined with respect to $\phi_{\lhd_Z}(\mathcal{G})$

(iv) Treating elements of $X^{(1)} \setminus X_{\tilde{Z}}^{(1)}$ as hidden variables.

We say $\trianglelefteq_{\tilde{Z}}$ is valid for $\{\trianglelefteq\tilde{Z}\}$, if $\triangleleft_{\tilde{Z}}$ is valid for $\{\triangleleft\tilde{Z}\}$, and $\tilde{Z}$ is fixable in $\phi_{\triangleleft_{\tilde{Z}}}(\mathcal{G})$. If $\trianglelefteq_{\tilde{Z}}$ is valid, we define

$$\phi_{\trianglelefteq_{\tilde{Z}}}(q;\mathcal{G}) \equiv \phi_{\tilde{Z}}\left(\phi_{\triangleleft_{\tilde{Z}}}(q;\mathcal{G});\phi_{\triangleleft_{\tilde{Z}}}(\mathcal{G})\right), \tag{4.5}$$

where $\phi_{\triangleleft_{\tilde{Z}}}(q;\mathcal{G}) \equiv \frac{q(V|W)}{\prod_{\tilde{Y}\in\triangleleft\tilde{Z}} q_{\tilde{Y}}}$, and $q_{\tilde{Y}}$ are defined inductively as the denominator of (4.4) for $\tilde{Y}$, $\phi_{\triangleleft_{\tilde{Y}}}(\mathcal{G})$ and $\phi_{\triangleleft_{\tilde{Y}}}(q;\mathcal{G})$. This leads to the following identification results.

Theorem 4. *Given a DAG $\mathcal{G}(X^{(1)},R,O,X)$, the distribution $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))|_{\mathrm{pa}_{\mathcal{G}}(R_i)\cap R=1}$ is identifiable from $p(R,O,X)$ if there exists*

(i) $Z \subseteq X^{(1)} \cup R \cup O$,

(ii) *an equivalence relation* $\sim$ *on* Z *such that* $\{R_i\} \in {}^{Z}\!/_{\sim}$,

(iii) *a set of elements* $X^{(1)}_{\tilde{Z}}$ *such that* $X^{(1)}_{\{\triangleleft\tilde{Z}\}} \subseteq X^{(1)}_{\tilde{Z}} \subseteq X^{(1)}$ *for each* $\tilde{Z} \in {}^{Z}\!/_{\sim}$,

(iv) $X^{(1)} \cap \mathrm{pa}_{\mathcal{G}}(R_i) \subseteq (Z \setminus \{R_i\}) \cup X^{(1)}_{\{R_i\}}$,

(v) *and a valid fixing schedule* $\triangleleft$ *for* ${}^{Z}\!/_{\sim}$ *in* $\mathcal{G}$ *such that for each* $\tilde{Z} \in {}^{Z}\!/_{\sim}$, $\tilde{Z} \triangleleft \{R_i\}$.

Moreover, $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))|_{\mathrm{pa}_{\mathcal{G}}(R_i)\cap R=1}$ is equal to $q_{\{R_i\}}$, defined inductively as the denominator of Eq. 4.4 for $\{R_i\}$, $\phi_{\triangleleft_{\{R_i\}}}(\mathcal{G})$ and $\phi_{\triangleleft_{\{R_i\}}}(p;\mathcal{G})$, and evaluated at $\mathrm{pa}_{\mathcal{G}}(R_i)\cap R = 1$.

Theorem 4 implies that $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))$ is identified if we can find a set of variables that can be fixed according to a partial order (possibly through set fixing) within subproblems where certain variables are hidden. At the end of the fixing schedule, we require that R_i itself is fixable given its Markov blanket in the original DAG. We encourage the reader to view the example provided in Appendix D, for a demonstration of valid fixing schedules that may be chosen by Theorem 4.

Corollary 4.1. *Given a DAG $\mathcal{G}(X^{(1)},R,O,X)$, the target law $p(X^{(1)},O)$ is identified if $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))$ is identified via Theorem 4 for every $R_i \in R$.*

Proof. Follows from Theorem 4 and Eq. 4.2. □

Further, in special classes of models, in addition to the target law, the full law is also identified as follows.

Theorem 5. *Given a DAG $\mathcal{G}(X^{(1)}, R, O, X)$, the full law $p(R, X^{(1)}, O)$ is identifiable from $p(R, O, X)$ if for every $R_i \in R$, all conditions in Theorem 4 (i-v) are met, and also for each $\tilde{Z} \in {}^{Z}/_{\sim}$, $X^{(1)}_{\tilde{Z}}$ does not contain any elements in $\{X^{(1)}_j \mid R_j \in \text{pa}_{\mathcal{G}}(R_i)\}$. Moreover, $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))$ is equal to $q_{\{R_i\}}$, defined inductively as the denominator of Eq. 4.4 for $\{R_i\}$, $\phi_{\lhd_{\{R_i\}}}(\mathcal{G})$ and $\phi_{\lhd_{\{R_i\}}}(p; \mathcal{G})$, and*

$$p(R, X^{(1)}, O) = \Big(\prod_{R_i \in R} q_{R_i} \Big) \times \frac{p(R = 1, O, X)}{\big(\prod_{R_i \in R} q_{R_i} \big) \,|_{R=1}}$$

Proof. Under conditions (i-v) in Theorem 4, we are guaranteed to identify the target law and obtain $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))$ where some $R_j \in \text{pa}_{\mathcal{G}}(R_i)$ may be evaluated at $R_j = 1$. Under the additional restriction stated above, all $R_j \in \text{pa}_{\mathcal{G}}(R_i)$ can be evaluated at all levels. □

Theorem 5 always fails if a special collider structure $X^{(1)}_j \to R_i \leftarrow R_j$, which we call a *colluder*, exists in $\mathcal{G}$. The following theorem establishes that the presence of a colluder in the missing data DAG $\mathcal{G}$ is a sufficient condition for non-identifiability of the full law.

Theorem 6. *In a DAG $\mathcal{G}(X^{(1)}, R, O, X)$, if there exists $R_i, R_j \in R$ such that $\{R_j, X^{(1)}_j\} \in \text{pa}_{\mathcal{G}}(R_i)$, then $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))|_{R_j=0}$ is not identified. Hence, the full law $p(X^{(1)}, R)$ is not identified.*

Proof. Follows by providing two different full laws that agree on the observed law (see Appendix D) on a DAG with 2 counterfactual random variables. This result holds for any arbitrary missing DAG containing the colluder structure defined above. □

Theorems 4 and 5 do not address a computationally efficient search procedure for a valid fixing schedule $\lhd$ that permit identification of $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))$ for a particular $R_i \in R$. Nevertheless, the following theorem shows how to easily obtain identification of the target law in a restricted class of missing data DAGs.

Theorem 7. *Consider a DAG $\mathcal{G}(X^{(1)}, R, O, X)$ such that for every $R_i \in R$, $\{R_j \mid X_j^{(1)} \in \mathrm{pa}_{\mathcal{G}}(R_i)\} \cap \mathrm{an}_{\mathcal{G}}(R_i) = \emptyset$. Then for every $R_i \in R$, a fixing schedule $\lhd$ for $\{\{R_j\} \mid R_j \in \mathcal{G}_{R \cap \mathrm{de}_{\mathcal{G}}(R_i)}\}$ given by the partial order induced by the ancestrality relation on $\mathcal{G}_{R \cap \mathrm{de}_{\mathcal{G}}(R_i)}$ is valid in $\mathcal{G}(X^{(1)}, R, O, X)$, by taking each $X^{(1)}_{\tilde{Z}} = \bigcup_{Z \in \{\trianglelefteq \tilde{Z}\}} X^{(1)}_Z$, for every $\tilde{Z} \in \{\trianglelefteq \{R_i\}\}$. Thus the target law is identified.*

Theorem 7 is particularly helpful for identification in missing data DAG models that may be used in longitudinal studies. In such studies, there may exist a natural temporal ordering on the variables $X^{(1)}$ and the missingness indicators R, where patient attendance at a future time point j, denoted by R_j, may be determined by their past attendance at a point in time i, denoted by R_i, which in turn may be affected by the patient's view of their future outcome $X_j^{(1)}$. This model bears many similarities to a popular missing data model known as the permutation model proposed in [112].

4.4 Related and Future Work

This chapter addressed a significant gap present in identification theory for missing data models representable as DAGs. We showed, by examples, that straightforward application of identification machinery in causal inference with hidden variables do not suffice for identification in missing data, and discussed the generalizations required to make it suitable for this task. These generalizations included fixing (possibly sets of) variables on a partial order and avoiding selection bias by introducing hidden variables into the problem though they were not present in the initial problem statement. Theorem 4 gives a characterization of how to utilize these generalized procedures to

obtain identification of the target law, while Theorem 5 gives a similar characterization for the full law. While neither of these results alluded to a computationally efficient algorithm to obtain identification in general, Theorem 7 provides such a procedure for a special class of missing data models where the partial order of fixing operations required for each R is easy to determine. Further, Theorem 6 provided sufficient graphical conditions, in the form of colluders, for non-parametric non-identifiability of the full data law in missing data DAG models.

In follow-up work done in [123], we went on to show that in fact, the absence of colluders are neccessary and sufficient for identifiability of the full data law in missing data DAG models. The notion of a colluder was also extended to colluding paths in missing data ADMG models. Using this generalization of the work on colluders done in this chapter, [123] provides a sound and complete characterization for identification of the full law in missing data models in the presence of unmeasured confounders. It was further shown in [123] that all such identified missing data models are sub models of a non-parametric saturated missing data CG model known as the no self-censoring model [115, 114]. This enables adaptation of semiparametric estimation theory from [124] for the no self-censoring model to these identified DAG/ADMG sub models. However, designing estimators for gain in statistical efficiency and/or robustness for specific identifying functionals (e.g., covariate adjustment in missing data DAG models for which a sound and complete characeterization was provided by [121]) that exploit restrictions in such missing data models is still an open problem. As is a complete characterization for identification of the *target law* in missing data DAG models.

The work done in this chapter also relies heavily on the missing data DAG being known. This can, in part, be alleviated by applying a structure learning algorithm as in [125], [126], or [127] to first learn the missing data DAG or a set of Markov equivalent missing data DAGs to use for identification. However, the improved understanding of identification theory in missing data DAG models from this chapter (and its extensions

in [123]) create the possibility of structure learning algorithms that allow for more complicated forms of missingness than the ones considered in prior work.

4.5 Acknowledgements

Chapter 4 is, in part, a reprint of material from [12] and [123].

Chapter 5

Estimation of Causal Effects in the Presence of Unmeasured Confounders

In Chapter 1 we saw that truncated nested Markov factorization (Eq. 1.9) is a sound and complete procedure to identify counterfactual distributions in a latent variable DAG model. Despite the sophistication of causal identification theory, estimators based on simple covariate adjustment remain the most common strategy for evaluating the ACE from data. Estimates obtained in this way are often biased due to the presence of unmeasured confounding and/or model misspecification. A popular approach for addressing the latter issue has been to use semiparametric estimators developed using the theory of influence functions [128, 129, 110]. The most popular of these estimators is known as the *augmented inverse probability weighted (AIPW)* estimator and is *doubly robust* in that it gives the analyst two chances to obtain a valid estimate for the ACE – either by specifying the correct model for the treatment assignment given observed covariates that render the treatment assignment *ignorable*, or by specifying the correct model for the dependence of the outcome on the treatment and these covariates. Recent work by [130] and [131] yields methods for constructing statistically efficient versions of AIPW that take advantage of Markov restrictions implied on the obseved data by a fully observed causal model associated with a DAG.

If a causal model contains hidden variables, a.k.a. unmeasured confounders, causal inference becomes considerably more complicated. In this chapter, we study estimation strategies for the average causal effect of a single treatment variable on a single outcome variable in scenarios where unmeasured confounding prevents us from finding a valid covariate adjustment set. Our contributions can be summarized as follows.

We first study equality restrictions on the tangent space implied by a hidden variable DAG model. Such restrictions are important as they play a role in deriving the most efficient influence function based estimator (one that attains the lowest asymptotic variance) for any given parameter of interest. In the special case where the model is nonparametric saturated, no restrictions are imposed on the tangent space, and the influence function is unique (and thus efficient). We provide Algorithm 9 as a *sound* and *complete* procedure for checking whether a hidden variable causal model that factorizes as a DAG imposes equality restrictions on the observed data tangent space, provided the hidden variables in the model are unrestricted. We then define a class of hidden variable causal models, termed *mb-shielded acyclic directed mixed graphs* (ADMGs), for which the restrictions on the tangent space resemble those of a DAG model with no hidden variables, which makes derivations of the efficient influence function (from a given nonparametric influence function) exceptionally simple.

For estimation of the ACE in a large class of hidden variable causal models characterized by a simple graphical criterion which we term *primal fixability*, we propose two new *inverse probability weighted* (IPW) estimators, *primal IPW* and *dual IPW*. We show that these estimators use variationally independent components of the joint likelihood on the observed margin of the hidden variable DAG. This leads to an influence function based semiparametric estimator, derived in [13], that can be viewed as augmentation of primal IPW. This semiparametric estimator, known as *augmented primal IPW (APIPW)*, was shown to be doubly robust in the models involved in the primal and dual IPW estimators. In this chapter, we study the efficiency of the

APIPW estimator using our results on nonparametric saturation of ADMG models. We then derive the efficient influence function for APIPW in mb-shielded ADMGs.

Finally, we propose the *nested IPW* estimator that generalizes IPW to all hidden variable causal models where the target parameter is identified. We propose a *sound* and *complete* algorithm (Algorithm 10) that derives the corresponding nested IPW estimator when possible. We show that the nested IPW estimator can help alleviate issues related to model misspecification by requiring the analyst fit parametric models for only a subset of the observed data likelihood related to the treatment assignment.

5.1 Overview of Semiparametric Estimation Theory

In Chapter 1 we saw that the average causal effect may be expressed in terms of a contrast between two counterfactual means $\mathbb{E}[Y(t)]$ and $\mathbb{E}[Y(t')]$ – the expected value of the outcome Y had treatment T been assigned to some value t or t' respectively. Since the results in this chapter hold for any value of treatment assignment, we will set our target of inference $\psi(t)$ to be $\mathbb{E}[Y(t)]$ without loss of generality. That is,

$$\psi(t) \equiv \mathbb{E}[Y(t)] \qquad \textit{(Target parameter)}$$

We now briefly review semiparametric estimation theory and its application to the standard covariate adjustment functional for the counterfactual mean. Given a statistical model $\mathcal{M} = \{p_\eta(Z) : \eta \in \Gamma\}$ where Γ is the parameter space and η is the parameter indexing a specific model. Let P_{η_0} and ψ_0 denote the true model and the true value of our target parameter ψ_0 respectively. Then, an estimator $\widehat{\psi}_n$ of the (scalar)[1] parameter ψ based on n i.i.d copies $Z_1, \ldots, Z_n$ drawn from $p_\eta(Z)$, is *asymptotically linear* if there exists a measurable random function $U_\psi(Z)$ with mean

[1] $\mathbb{E}[Y(t)]$ is a scalar parameter. For an extension to vector valued functionals in $\mathbb{R}^q, q > 1$, refer to [110, 132].

zero and finite variance such that

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) = \frac{1}{\sqrt{n}} \times \sum_{i=1}^{n} U_\psi(Z_i) + o_p(1), \tag{5.1}$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity. The random variable $U_\psi(Z)$ is called the *influence function* of the estimator $\widehat{\psi}_n$. The term influence function is derived from the robustness literature [133]. Given a collection of probability laws $\mathcal{M}$, an estimator $\widehat{\psi}$ of $\psi(P)$ is said to be *regular* in $\mathcal{M}$ at P if its convergence to $\psi(P)$ is locally uniform [128]. Our focus here will be on estimators that are both regular and asymptotically linear, or RAL for short. For a review and justification for restricting to such estimators, we refer the reader to [110] and [134].

Via application of the central limit theorem and Slutsky's theorem, it can be shown that the RAL estimator $\widehat{\psi}_n$ is *consistent* and *asymptotically normal* (CAN), with asymptotic variance equal to the variance of its influence function U_ψ,

$$\sqrt{n} \times (\widehat{\psi}_n - \psi) \xrightarrow{d} N\Big(0,\ \mathrm{var}(U_\psi)\Big). \tag{5.2}$$

The first step in dealing with a semiparametric model, is to consider a simpler finite-dimensional parametric submodel that is contained within the semiparametric model and contains the truth. Consider a (regular) parametric submodel $\mathcal{M}_{\text{sub}} = \{P_{\eta_\kappa} : \kappa \in [0, 1) \text{ where } P_{\eta_{\kappa=0}} = P_{\eta_0}\}$ of the model $\mathcal{M}$. Given P_{η_0}, define the corresponding score to be $S_{\eta_0}(Z) = \dfrac{d}{d\kappa} \log p_{\eta_\kappa}(Z)\Big|_{\kappa=0}$. It is known that

$$\frac{d}{d\kappa}\psi(\eta_\kappa)\Big|_{\kappa=0} = \mathbb{E}\Big[U_\psi(Z) \times S_{\eta_0}(Z)\Big], \tag{5.3}$$

where $\psi(\eta_\kappa)$ is the target parameter in the parametric submodel, $U_\psi(Z)$ is the corresponding influence function evaluated at law P_{η_0}, $S_{\eta_0}(Z)$ is the score of the law P_{η_0}, and the expectation is taken with respect to P_{η_0}. Thus, Eq. 5.3 provides a method for deriving an influence function for a given target parameter ψ.

We now discuss how this relates to estimation of counterfactual quantities under the assumption of no unmeasured confounding. Often, data analysts will assume (as a

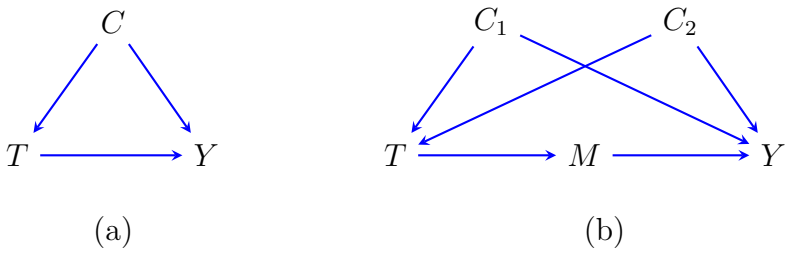


Figure 5-1. (a) DAG representing conditional ignorability; (b) A DAG where missing edges impose restrictions on the observed data distribution.

matter of convenience rather than substantive reasoning) that the treatment assignment is ignorable conditional on a set of baseline covariates. That is, $Y(t) \perp\!\!\!\perp T \mid C$. Graphically speaking, the conditionally ignorable model is often represented by the DAG shown in Fig. 5-1(a). Under the assumptions of this model, the counterfactual mean is identified via the adjustment functional as,

$$\psi(t) = \mathbb{E}\Big[\mathbb{E}[Y \mid T = t, C]\Big] \qquad \textit{(Adjustment functional)} \tag{5.4}$$

If a correct parametric model, say $\mu_t(C;\eta_1)$ can be specified for the outcome regression $\mathbb{E}[Y \mid T = t, C]$, the target parameter $\psi(t)$ can be estimated via the plug-in principle. That is, $\psi(t) = \mathbb{P}_n[\mu_t(C;\widehat{\eta_1})]$, where $\mathbb{P}_n[.] \equiv \frac{1}{n}\sum_{i=1}^{n}(.)$ and $\widehat{\eta_1}$ are the maximum likelihood values of η_1. If such a parametric specification is not possible, an alternative is to pursue an inverse probability weighted (IPW) estimator that relies on specification of a parametric model, say $\pi_t(C;\eta_2)$, for the treatment assignment probability $p(T = t \mid C)$. The IPW estimator takes the form $\mathbb{P}_n[\frac{\mathbb{I}(T=t)}{\pi_t(C;\widehat{\eta_2})} \times Y]$, where $\mathbb{I}(.)$ is the indicator function and $\widehat{\eta_2}$ are the maximum likelihood estimates of η_2. The estimators based on outcome regression and IPW are $\sqrt{n}$-consistent and asymptotically normal under correct specification of the parametric models they rely on. However, correct specification of such models is often not possible. Further, even if correct specification is possible for the IPW estimator or if the treatment assignment probability is known via experimental design, the resulting estimates are inefficient,

i.e., have high variance.

A principled alternative is to consider influence functions and RAL estimators. In the nonparametric saturated model, corresponding to the complete DAG shown in Fig. 5-1(a), the unique influence function for $\psi(t)$ derived from the adjustment functional in Eq. 5.4 using the method suggested by Eq. 5.3 is given by $U_{\psi_t} = \frac{\mathbb{I}(T=t)}{\pi_t(C)} \times \{Y - \mu_t(C)\} + \mu_t(C) - \psi(t)$ [110]. This yields the *AIPW* estimator: $\mathbb{P}_n\Big[\frac{\mathbb{I}(T=t)}{\pi_t(C;\widehat{\eta_2})} \times \{Y - \mu_t(C;\widehat{\eta_1})\} + \mu_t(C;\widehat{\eta_1})\Big]$. Given the standard factorization of the complete DAG as $p(Y \mid A, C) \times p(A \mid C) \times p(C)$, the propensity score model $\pi_t(C)$ and the outcome regression model $\mu_t(C)$ are variationally independent. Further, the bias of this estimator is a product of the biases of its nuisance functions $\pi_t(C)$ and $\mu_t(C)$. As a result, the AIPW estimator exhibits the *double robustness property*, where it remains consistent if *either* of the two nuisance models $\pi_t(C)$ or $\mu_t(C)$ is specified correctly, even if the other is arbitrarily misspecified.

Influence functions provide a geometric view of the behavior of RAL estimators. Consider a Hilbert space[2] H of all mean-zero scalar functions, equipped with an inner product defined as $\mathbb{E}[h_1 \times h_2], h_1, h_2 \in \mathrm{H}$. The *tangent space* in the model $\mathcal{M}$, denoted by Λ, is defined to be the mean-square closure of parametric submodel tangent spaces, where a parametric submodel tangent space is the set of elements $\Lambda_{\eta_\kappa} = \{\alpha S_{\eta_\kappa}(Z)\}, \alpha$ is a constant and S_{η_κ} is the score for the parameter ψ_{η_κ} for some parametric submodel. In mathematical form, $\Lambda = \overline{[\Lambda_{\eta_\kappa}]}$.

The tangent space Λ is a closed linear subspace of the Hilbert space H ($\Lambda \subseteq \mathrm{H}$). The orthogonal complement of the tangent space, denoted by $\Lambda^\perp$, is defined as $\Lambda^\perp = \{h \in \mathrm{H} \mid \mathbb{E}[h \times h'] = 0, \forall h' \in \Lambda\}$. Note that $\mathrm{H} = \Lambda \oplus \Lambda^\perp$, where $\oplus$ is the direct sum, and $\Lambda \cap \Lambda^\perp = \{0\}$. Given an arbitrary element $h \in \Lambda^\perp$, it holds that for any submodel $\mathcal{M}_{\text{sub}}$, with score S_{η_0} corresponding to P_{η_0}, $\mathbb{E}[h \times S_{\eta_0}] = 0$. Consequently,

[2]The Hilbert space of all mean-zero scalar functions is the L^2 space. For a precise definition of Hilbert spaces see [135].

using Eq. 5.3, $h + U_\psi(Z)$ is also an influence function. The vector space $\Lambda^\perp$ is then of particular importance because we can now construct the class of all influence functions, denoted by $\mathcal{U}$, as $\mathcal{U} = U_\psi(Z) + \Lambda^\perp$. Upon knowing a single IF $U_\psi(Z)$ and the tangent space orthogonal complement $\Lambda^\perp$, we can obtain the class of all possible RAL estimators that admit the CAN property.

Out of all the influence functions in $\mathcal{U}$ there exists a unique one which lies in the tangent space Λ, and which yields the most efficient RAL estimator by recovering the *semiparametric efficiency bound*. This efficient influence function can be obtained by projecting any influence function, call it U^*_ψ, onto the tangent space Λ. This operation is denoted by $U^{\text{eff}_\psi} = \pi[U^*_\psi \mid \Lambda]$, where U^{eff}_ψ denotes the efficient IF.

In a semiparametric model of a DAG with missing edges, such as the one shown in Fig. 5-1(b) (this model also implies conditional ignorability $Y(t) \perp\!\!\!\perp T \mid C$, where $C = \{C_1, C_2\}$), defined by conditional independence restrictions on the tangent space implied by the DAG factorization, the AIPW influence function can be projected onto the tangent space of the model to improve efficiency; see [131] for details. On the other hand, if the tangent space contains the entire Hilbert space, i.e., $\Lambda = \text{H}$, then the statistical model $\mathcal{M}$ is called a *nonparametric* model. In a nonparametric model, we only have one influence function since $\Lambda^\perp = \{0\}$. This unique influence function can be obtained via Eq. 5.3 and corresponds to the efficient influence function U^{eff}_ψ (the unique element in the tangent space Λ) in the nonparametric model $\mathcal{M}$. For a more detailed description of the concepts outlined here refer to [110, 132].

As we are interested in estimation of our target parameter $\psi(t)$ in causal models with unmeasured confounders, we first derive new results for restrictions on the tangent space of ADMG models that we will use to study the efficiency of semiparametric estimators derived in [13].

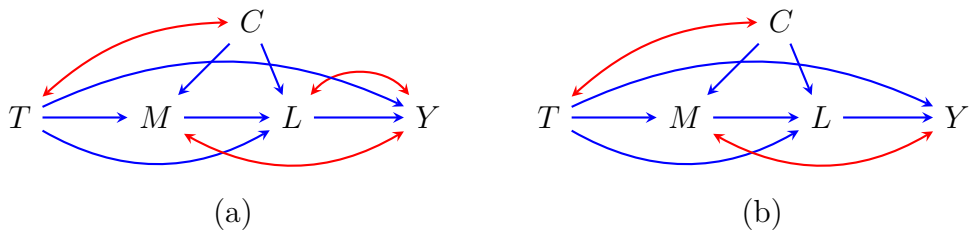


Figure 5-2. (a) Example of an ADMG whose underlying nested Markov model is NPS even though there is a missing edge between C and Y. (b) Absence of the bidirected edge $L \leftrightarrow Y$ in (a) introduces a Verma constraint $C \perp\!\!\!\perp Y \mid T$ in $p(V)/p(L \mid T, M, C)$.

5.2 Restrictions on the Tangent Space of ADMG Models

In this section, we present results regarding restrictions on the tangent space of ADMG models. Though we use these results in order to reason about the efficiency of estimators proposed for our target of interest $\psi(t)$, they are in fact applicable to *any* target parameter of interest, not just the one defined in the present work.

Recall from Chapter 1, an ADMG $\mathcal{G}(V)$ may encode two types of equality constraints: ordinary conditional independence statements such as $V_i \perp\!\!\!\perp V_j \mid V_k$, and more general equality constraints, known as *Verma* constraints, that resemble conditional independences albeit in post-intervention distributions [23]. Grouped together, these are known as *equality constraints.* We first describe an algorithm that characterizes when the statistical model of an ADMG $\mathcal{G}(V)$, i.e., $\mathcal{M}(\mathcal{G})$, is *nonparametric saturated;* meaning $\mathcal{M}(\mathcal{G})$ imposes no equality restrictions on $p(V)$. As mentioned earlier, when $\mathcal{M}(\mathcal{G}) = \mathcal{M}_{\text{nps}}$, then the tangent space of the corresponding ADMG model consists of the entire Hilbert space.

5.2.1 Algorithm to Detect Nonparametric Saturation

The easiest way to confirm whether the model implied by an ADMG is nonparametric saturated (NPS) is to simply check that all vertices are pairwise connected by a

directed or bidirected edge. However, the absence of edges between two vertices in an ADMG do not necessarily correspond to a conditional independence or even generalized conditional independence (Verma) constraint [30]; see Fig. 5-2(a) for example where the missing edge between C and Y does not correspond to any constraint. The missing edge between C and Y in Fig. 5-2(b) on the other hand, does correspond to the Verma constraint $C \perp\!\!\!\perp Y \mid T$ in the distribution $p(V)/p(L \mid T, M, C)$; see [23, 70] for more details.

We propose a procedure to check if the model implied by an ADMG $\mathcal{G}(V)$ with missing edges is nonparametric saturated by checking if it is equivalent to the model implied by another ADMG where there are no missing edges. In order to do this, we use the *maximal arid projection* of an ADMG as described in [39]. Such projections yield another ADMG that implies the same set of equality restrictions as the original, albeit one in which the absence of edges facilitates easier study of the constraints in the model. The algorithm that we now describe closely relates to this projection in that it declares a model to be NPS when the input ADMG's maximal arid projection is a complete graph, and not NPS otherwise. For more details see proof of Theorem 8 and [39].

The concept of a *reachable closure,* plays a key role in constructing the maximal arid projection. The reachable closure of a set of vertices S, denoted by $\langle S \rangle_{\mathcal{G}}$ is the unique minimal superset of S such that $V \setminus \langle S \rangle_{\mathcal{G}}$ is fixable [39]. We provide our procedure for checking if a model is nonparametric saturated in Algorithm 9. We show that our algorithm is sound and complete for this purpose in the following theorem. Further, an informal complexity analysis of Algorithm 9 shows that it is computationally tractable as it runs in polynomial time with respect to the number of vertices and edges in the graph $\mathcal{G}$. The complexity of the outer loop is $\mathcal{O}(|V|^2)$ as it requires the selection of all possible pairs of random vertices. Further, naive implementations for computing reachable closures of sets are $\mathcal{O}(|V|^2 + |V| \times |E|)$ as it

Algorithm 9 Check Nonparametric Saturation

1: **Inputs**: $\mathcal{G}$
2: Define the disjunctive definition of parents of a set S as $\text{pa}^{\text{d}}_{\mathcal{G}}(S) \equiv \bigcup_{S_i \in S} \text{pa}_{\mathcal{G}}(S_i)$
3: **for** all V_i, V_j pairs in $\mathcal{G}$ such that $i \neq j$ **do**
4: **if not** $\Big\{ V_i \in \text{pa}^{\text{d}}_{\mathcal{G}}(\langle V_j \rangle_{\mathcal{G}})$ **or** $V_j \in \text{pa}^{\text{d}}_{\mathcal{G}}(\langle V_i \rangle_{\mathcal{G}})$ **or**
5: $\langle V_i, V_j \rangle_{\mathcal{G}}$ is bidirected connected in $\mathcal{G} \Big\}$ **then**
6: **return** Not NPS
7: **return** NPS

involves repeated applications of depth first search (popular algorithms for which are linear in complexity $\mathcal{O}(|V| + |E|)$ [136]) in order to determine the fixability of a set of vertices.

Theorem 8. *Algorithm 9 is sound and complete for deciding the nonparametric saturation status of the model implied by an ADMG $\mathcal{G}(V)$ by determining the absence of equality constraints.*

Example: Application of Algorithm 9

As an example of the application of Algorithm 9, we return to the ADMGs in Fig. 5-2. As all pairs of vertices besides C and Y are adjacent in these ADMGs, the negation of the condition in lines 4 and 5 trivially evaluates to False for these pairs. We now focus on steps executed by the algorithm when examining the pair (C, Y). In the case of Fig. 5-2(a), the algorithm computes $\langle Y \rangle_{\mathcal{G}} = \{L, M, Y\}$. Therefore, C is indeed a parent of the reachable closure of Y, i.e., $C \in \text{pa}^{\text{d}}_{\mathcal{G}}(\langle Y \rangle_{\mathcal{G}})$ (note the use of the disjunctive definition of parents as defined in Algorithm 9), and the algorithm completes execution by confirming that the model is NPS. In the case of Fig. 5-2(b), the reachable closure of Y is $\langle Y \rangle_{\mathcal{G}} = \{Y\}$ and therefore, $C \notin \text{pa}^{\text{d}}_{\mathcal{G}}(\langle Y \rangle_{\mathcal{G}})$. It is also easy to confirm that $Y \notin \text{pa}^{\text{d}}_{\mathcal{G}}(\langle C \rangle_{\mathcal{G}})$, and that $\langle C, Y \rangle_{\mathcal{G}} = \{C, Y\}$ is not bidirected connected. Thus, with all these conditions evaluating to False, the negation is True, resulting in the algorithm correctly identifying Fig. 5-2(b) as not NPS.

5.2.2 mb-shielded ADMGs

Both ordinary conditional independences and Verma constraints restrict the tangent space of a given ADMG model. Hence, both sets of equality constraints play a role in formulating estimators that achieve the semiparametric efficiency bound. Deriving restrictions on the tangent space implied by Verma constraints may be quite difficult in general as these restrictions hold in kernels obtained after recursive fixing operations. In what follows, we identify a large class of ADMGs where all Verma constraints are implied by ordinary conditional independences and derive the tangent space of such ADMG models.

Assume the existence of a class of ADMGs where, given a topological order τ, all equality constraints implied by the ADMG $\mathcal{G}(V)$ can be written as ordinary conditional independence statements of the form,

$$V_i \perp\!\!\!\perp \{\prec_\tau V_i\} \setminus \mathrm{mp}_{\mathcal{G}}(V_i) \mid \mathrm{mp}_{\mathcal{G}}(V_i). \tag{5.5}$$

Such a property immediately implies that the topological factorization of the observed data distribution $p(V)$ shown in Eq. 1.6 captures *all* equality constraints implied by the ADMG $\mathcal{G}(V)$. A sound criterion for identifying ADMGs that satisfy this property is to check that an edge between two vertices V_i and V_j in $\mathcal{G}$ is absent only if $V_i \notin \mathrm{mb}_{\mathcal{G}}(V_j)$ and $V_j \notin \mathrm{mb}_{\mathcal{G}}(V_i)$. We call this class of ADMGs *mb-shielded ADMGs*, as pairs of vertices are always adjacent if either one is in the Markov blanket of the other. We formalize this criterion in the following theorem, and show that mb-shielded ADMGs possess the desired property.

Theorem 9. *Consider a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where an edge between two vertices is absent only if $V_i \notin \mathrm{mb}_{\mathcal{G}}(V_j)$ and $V_j \notin \mathrm{mb}_{\mathcal{G}}(V_i)$. Then, given any valid topological order on V, all equality constraints in $p(V)$ are implied by the set of restrictions: $V_i \perp\!\!\!\perp \{\prec V_i\} \setminus \mathrm{mp}_{\mathcal{G}}(V_i) \mid \mathrm{mp}_{\mathcal{G}}(V_i)$, $\forall V_i \in V$.*

Since the factorization given in Eq. 1.6 captures all equality constraints, the tangent space of the statistical model corresponding to an mb-shielded ADMG will be the same as that of a Markov equivalent DAG obtained by orienting all bidirected edges according to a valid topological order. This fact follows directly from Lemma 1.6 in [137] and Theorem 4.5 in [110]. For the sake of completeness, we reiterate these results and provide the tangent space of mb-shielded ADMGs and its orthogonal complement in the following lemma.

Lemma 4. *Consider the statistical model $\mathcal{M}(\mathcal{G})$ where $\mathcal{G}(V)$ is an mb-shielded ADMG. The tangent space of $\mathcal{M}(\mathcal{G})$ is given by a direct sum of mutually orthogonal spaces: $\Lambda = \oplus_{V_i \in V} \Lambda_i$, where*

$$\begin{aligned}\Lambda_i &= \Big\{\alpha_i(V_i, \mathrm{mp}_{\mathcal{G}}(V_i)) \in \mathbb{H} \ s.t.\ \mathbb{E}[\alpha_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)] = 0\Big\} \\ &= \Big\{\alpha_i(V_i, \mathrm{mp}_{\mathcal{G}}(V_i)) - \mathbb{E}[\alpha_i \mid \mathrm{mp}_{\mathcal{G}}(V_i)],\ \forall \alpha_i(V_i, \mathrm{mp}_{\mathcal{G}}(V_i)) \in \mathbb{H}\Big\}.\end{aligned}$$

In addition, the projection of an element $h(V) \in \mathbb{H}$ onto Λ_i, denoted by h_i, is given by $h_i \equiv \Pi[h(V) \mid \Lambda_i] = \mathbb{E}[h(V) \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)] - \mathbb{E}[h(V) \mid \mathrm{mp}_{\mathcal{G}}(V_i)]$. Consequently, the orthogonal complement of the tangent space $\Lambda^{\perp}$ is given as follows,

$$\Lambda^{\perp} = \left\{ \sum_{V_i \in V} \alpha_i(V_1, \ldots, V_i) - \mathbb{E}\Big[\alpha_i(V_1, \ldots, V_i) \ \Big|\ V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\Big] \right\},$$

where $\alpha_i(V_1, \ldots, V_i)$ is any function of V_1 through V_i in $\mathbb{H}$, such that $\mathbb{E}[\alpha_i \mid V_1, \ldots, V_{i-1}] = 0$.

The following section studies estimators in a wide class of hidden variable causal models characterized by a simple graphical criterion that we term *primal fixability*.

5.3 Estimating the ACE Under Primal Fixability

Consider the ADMGs shown in Fig. 5-3. It is easy to check that in either case there exists no valid adjustment set to identify the causal effect of T on Y. However, such an

effect is indeed identified in both graphs. The defining characteristic of these ADMGs that permits identification of the target $\psi(t)$, is that the district of T does not intersect with its children.

More formally, we will discuss ADMGs where $\text{dis}_{\mathcal{G}}(T) \cap \text{ch}_{\mathcal{G}}(T) = \emptyset$. This criterion encompasses many popular models in the literature, including those that satisfy the back-door and front-door criteria [5, 138], as special cases. We name this criterion primal fixability or *p-fixability* for short, due to its generalization of the fixing criterion introduced in the definition of the nested Markov model. For scenarios when the treatment is p-fixable, we introduce two new estimators (primal and dual IPW) for $\psi(t)$ that use variationally independent pieces of the observed data likelihood. Since these new estimators offer different perspectives on estimating the same target, we draw inspiration from the optimization literature [139, 140] in naming them primal and dual IPW.

5.3.1 Primal and Dual IPW Estimators

Primal fixability is known to be a necessary and sufficient condition for identification of the causal effect of T on all other variables $V \setminus T$ [26]. In observed data distributions $p(V)$ that district factorize according to an ADMG $\mathcal{G}(V)$ where T is primal fixable, the resulting identifying functional for the target is as follows.

$$\psi(t) = \sum_{V \setminus T} Y \times \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_{T} \prod_{V_j \in D_T} p(V_j \mid \text{mp}_{\mathcal{G}}(V_j)) \Big|_{T=t}, \tag{5.6}$$

where D_T denotes the district of T [26]. We provide this special notation for the district of T as D_T due to its frequent occurrence in subsequent results.

For the remainder of the chapter, we assume a fixed valid topological ordering τ where the treatment T appears later than all of its non-descendants i.e., $T \succ_\tau V \setminus \text{de}_{\mathcal{G}}(T)$ and the outcome Y appears earlier than all of its non-descendant non-ancestors i.e., $Y \prec_\tau V \setminus (\text{de}_{\mathcal{G}}(Y) \cup \text{an}_{\mathcal{G}}(Y))$. This allows for easier exposition by fixing the definition

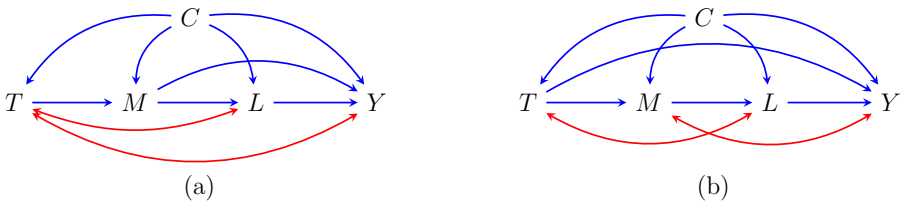


Figure 5-3. Examples of acyclic directed mixed graphs where T is primal fixable.

of pre-treatment covariates as being any variable that appears earlier than T under the ordering τ. In introducing primal IPW below, we use $\{\succeq T\}$ (dropping subscript τ for readability) to mean the set of vertices (including T) that succeed T under the topological order τ.

Lemma 5. *Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, $\psi(t) = \psi(t)_{primal} \equiv \mathbb{E}[\beta(t)_{primal}]$ where*

$$\begin{aligned}\beta(t)_{primal} &\equiv \frac{\mathbb{I}(T=t)}{q_{D_T}(T \mid \mathrm{mb}_{\mathcal{G}}(T))} \times Y \\ &= \mathbb{I}(T=t) \times \frac{\sum_T \prod_{V_i \in D_T \cap \{\succeq T\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))}{\prod_{V_i \in D_T \cap \{\succeq T\}} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))} \times Y. \end{aligned} \tag{5.7}$$

The kernel $q_{D_T}(T \mid \mathrm{mb}_{\mathcal{G}}(T))$ in Lemma 5 may be viewed as a *nested* propensity score derived from the post-intervention distribution $q_{D_T}(D_T \mid \mathrm{pa}_{\mathcal{G}}(D_T))$ where all variables outside of D_T are intervened on and held fixed to some constant value. Recall that the kernel $q_{D_T}(D_T \mid \mathrm{pa}_{\mathcal{G}}(D_T))$ is identified as $\prod_{V_i \in D_T} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))$ as in Eq. 1.4. Consequently, $q_{D_T}(T \mid \mathrm{mb}_{\mathcal{G}}(T))$ is identified by the definition of conditioning on all elements in D_T outside of T in the kernel $q_{D_T}(D_T \mid \mathrm{pa}_{\mathcal{G}}(D_T))$ as,

$$\begin{aligned} q_{D_T}(T \mid \mathrm{mb}_{\mathcal{G}}(T)) = q_{D_T}(T \mid D_T \cup \mathrm{pa}_{\mathcal{G}}(D_T) \setminus T) &= \frac{q_{D_T}(D_T \mid \mathrm{pa}_{\mathcal{G}}(D_T))}{q_{D_T}(D_T \setminus T \mid \mathrm{pa}_{\mathcal{G}}(D_T))} \\ = \frac{q_{D_T}(D_T \mid \mathrm{pa}_{\mathcal{G}}(D_T))}{\sum_T q_{D_T}(D_T \mid \mathrm{pa}_{\mathcal{G}}(D_T))} &= \frac{\prod_{V_i \in D_T} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))}{\sum_T \prod_{V_i \in D_T} p(V_i \mid \mathrm{mp}_{\mathcal{G}}(V_i))}. \end{aligned}$$

The final expression simplifies further by noticing that all vertices appearing prior to T under the topological order τ, do not contain T in their Markov pillows. Consequently, $p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$ is not a function of T if $V_i \prec T$. Thus, these terms may be pulled out of the summation in the denominator, and cancel with the corresponding term in the numerator. This gives us the resulting primal IPW formulation in Eq. 5.7.

We now introduce the dual formulation. Define the *inverse Markov pillow* of a vertex V_i to be all other vertices V_j outside of the district of V_i, such that V_i is a member of the Markov pillow of V_j. More formally, $\text{mp}_{\mathcal{G}}^{-1}(V_i) = \{V_j \in V \mid V_j \notin \text{dis}_{\mathcal{G}}(V_i), V_i \in \text{mp}_{\mathcal{G}}(V_j)\}$.

Lemma 6. *Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, $\psi(t) = \psi(t)_{dual} \equiv \mathbb{E}[\beta(t)_{dual}]$ where*

$$\beta(t)_{dual} = \frac{\prod_{V_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} \; p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \mid_{T=t}}{\prod_{V_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} \; p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))} \times Y. \tag{5.8}$$

The representation of $\psi(t)$ as $\beta(t)_{\text{primal}}$ and $\beta(t)_{\text{dual}}$ in Lemmas 5 and 6 immediately yields the corresponding primal and dual IPW estimators. In what follows, we occasionally assume dependence on t to be implicit and for simplicity of notation, write $\psi(t)_{\text{primal}}$ as simply ψ_{primal} and $\beta_{\text{primal}}(t)$ as β_{primal} for example. Assume a finite set of parameters η_{primal} used to parameterize the nuisance models $\Big\{p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)\Big), \forall V_i \in D_T \cap \{\succeq T\}\}$ that appear in β_{primal}. Similarly, assume a finite set of parameters η_{dual} used to parameterize the nuisance models $\{p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)), \forall V_i \in \text{mp}_{\mathcal{G}}^{-1}(T)\}$ that appear in β_{dual}. Let $\widehat{\eta}_{\text{primal}}$ and $\widehat{\eta}_{\text{dual}}$ denote the respective maximum likelihood estimates. The primal IPW estimator $\widehat{\psi}_{\text{primal}}$ and dual IPW estimator $\widehat{\psi}_{\text{dual}}$ are obtained by evaluating empirical versions of the estimating equations $\mathbb{E}\Big[U(\psi(t), \widehat{\eta}_{\text{primal}})\Big] = 0$ and $\mathbb{E}\Big[U(\psi(t), \widehat{\eta}_{\text{dual}})\Big] = 0$, where $U(\psi(t), \eta_{\text{primal}}) = \beta_{\text{primal}} - \psi(t)$, and $U(\psi(t), \eta_{\text{dual}}) = \beta_{\text{dual}} - \psi(t)$. That is,

$$\widehat{\psi}_{\text{primal}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(T_i = t) \times \frac{\sum_{T} \prod_{V_j \in D_T \cap \{\succeq T\}} p(V_{j,i} \mid \text{mp}_{\mathcal{G}}(V_{j,i}); \widehat{\eta}_{\text{primal}})}{\prod_{V_j \in D_T \cap \{\succeq T\}} p(V_{j,i} \mid \text{mp}_{\mathcal{G}}(V_{j,i}); \widehat{\eta}_{\text{primal}})} \times Y_i,$$

$$\widehat{\psi}_{\text{dual}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\prod_{V_j \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(V_{j,i} \mid \text{mp}_{\mathcal{G}}(V_{j,i}); \widehat{\eta}_{\text{dual}}) \mid_{T=t}}{\prod_{V_j \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(V_{j,i} \mid \text{mp}_{\mathcal{G}}(V_{j,i}); \widehat{\eta}_{\text{dual}})} \times Y_i.$$

We now show that the sets of nuisance models in the primal and dual IPW estimators form variationally independent components of the observed data distribution $p(V)$. That is, correct specification of the nuisance models in the primal IPW estimator do not rely in any way on the correct specification of nuisance models in the dual IPW estimator.

Theorem 10. *Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, the IPW estimators ψ_{primal} and ψ_{dual} proposed in Lemmas 5 and 6 respectively, use variationally independent components of the observed distribution $p(V)$.*

We now briefly discuss some intuition regarding the primal and dual IPW estimators. In the regular conditionally ignorable model, the primal and dual IPW estimators correspond to the standard IPW and outcome regression plug-in estimators respectively. More generally, primal IPW can be viewed as a generalization of the g-formula to kernel factorizations that arise in ADMGs. The ordinary g-formula for a DAG model involves truncation of the DAG factorization, namely dropping a simple conditional factor of the treatment given its parents, i.e., $p(V(t)) = \{p(V)/p(T = t \mid \text{pa}_{\mathcal{G}}(T))\}|_{T=t}$. On the other hand, the primal formulation, or the *nested* g-formula, can be viewed as truncation of the district factorization in Eq. 5.6, where the nested conditional factor for the treatment given its Markov blanket is dropped from the observed joint distribution, i.e., $p(V(t)) = \{p(V)/q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T))\}|_{T=t}$. Intuition for the dual IPW

can be gained by viewing it as a probabilistic formalization of the node splitting operation in single world intervention graphs (SWIGs) described in [141]. To provide more concrete intuition on the primal and dual IPW estimators, we discuss their application to the ADMGs shown in Fig. 5-3.

Examples: Primal and Dual IPW Estimators

Consider the ADMG in Fig. 5-3(a). T is primal fixable as there is no bidirected path from T to any of its children, namely M. The inverse Markov pillow of T in Fig. 5-3(a) is just M. Per Lemmas 5 and 6, the primal and dual IPW estimators for the target parameter $\psi(t)$ in Fig. 5-3(a) are given by,

$$\text{(Fig. 5-3a)} \qquad \psi_{\text{primal}} = \mathbb{E}\Big[\ \mathbb{I}(T=t) \times \frac{\sum_T\ p(T \mid C) \times p(L \mid T, M, C) \times p(Y \mid T, M, L, C)}{p(T \mid C) \times p(L \mid T, M, C) \times p(Y \mid T, M, L, C)} \times Y\Big],$$
$$\psi_{\text{dual}} = \mathbb{E}\Big[\frac{p(M \mid T=t, C)}{p(M \mid T, C)} \times Y\Big].$$

In order to estimate $\psi(t)$ using finite samples, we proceed as follows. In case of the primal IPW, we can fit parametric models (generalized linear models for instance) for the conditional densities $p(T \mid C), p(L \mid T, M, C)$, and $p(Y \mid T, M, L, C)$,. The target parameter is then obtained by empirically evaluating the outer expectation using the fitted models. Note that we can also avoid modeling the conditional density of Y, as the outcome regression $\mathbb{E}[Y \mid T, M, L, C]$ suffices to estimate $\psi(t)$, i.e., ψ_{primal} can be expressed equivalently as

$$\mathbb{E}\Big[\ \mathbb{I}(T=t) \times \frac{\sum_T\ p(T \mid C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T, M, L, C]}{p(T \mid C) \times p(L \mid T, M, C)}\ \Big].$$

A simple procedure to estimate the dual IPW involves modeling the conditional density $p(M \mid T, C)$. However, a more sophisticated procedure may take advantage of modeling the density ratio directly as suggested by [142].

We now turn our attention to the ADMG in Fig. 5-3(b). The inverse Markov pillow

of T in Fig. 5-3(b) is $\{M, Y\}$. The corresponding primal and dual IPW estimators are given by,

$$\text{(Fig. 5-3b)} \qquad \psi_{\text{primal}} = \mathbb{E}\Big[\mathbb{I}(T=t) \times \frac{\sum_T p(T \mid C) \times p(L \mid T, M, C)}{p(T \mid C) \times p(L \mid T, M, C)} \times Y\Big].$$
$$\psi_{\text{dual}} = \mathbb{E}\Big[\frac{p(M \mid T=t, C)}{p(M \mid T, C)} \times \frac{p(Y \mid T=t, M, L, C)}{p(Y \mid T, M, L, C)} \times Y\Big].$$

Similar strategies can be used to estimate $\psi(t)$ as in the previous example. Also, note that the conditional density of Y in ψ_{dual} can be replaced by the outcome regression $\mathbb{E}[Y \mid T=t, M, L, C]$, i.e., ψ_{dual} can be expressed equivalently as $\mathbb{E}\Big[\frac{p(M|T=t,C)}{p(M|T,C)} \times \mathbb{E}[Y \mid T=t, M, L, C]\Big]$.

5.3.2 Augmented Primal IPW Estimators

In the previous subsection we have shown the existence of two estimators for the target $\psi(t)$ that use variationally independent portions of the likelihood when T is p-fixable. The question naturally arises if it is possible to combine these estimators to yield a single estimator that exhibits double robustness in the sets of models used in each one. In [13], we showed that in fact, the nonparametric influence function obtained by applying the pathwise derivative to the functional in Eq. 5.6 yields such an estimator, and this semiparametric estimator can be viewed as augmentation of primal IPW. In this dissertation, we do not focus on the derivation of this nonparametric influence function and the resulting augmented primal IPW (APIPW) (for this we refer the reader to [13] for details) and instead study its efficiency using results we have derived on the tangent space of ADMG models. We present the form of the nonparametric IF in this subsection and then discuss results on its efficiency.

Let $p(V)$ be a distribution that factorizes with respect to an ADMG $\mathcal{G}(V)$ where the treatment T is primal fixable. For notational simplicity, we will assume that the outcome Y has no descendants in $\mathcal{G}$, though our results extend trivially to settings where this is not true. Recall from the previous section, that we use a fixed topological

order τ where T is preceded by all its non-descendants and Y is succeeded by all its non-descendant non-ancestors. The set of variables V can then be partitioned into three disjoint sets $\mathbb{C}, \mathbb{L}$, and $\mathbb{M}$ where,

$$\begin{aligned}
\mathbb{C} &= \{C_i \in V \mid C_i \prec T\}, \\
\mathbb{L} &= \{L_i \in V \mid L_i \in D_T, L_i \succeq T\}, \\
\mathbb{M} &= \{M_i \in V \mid M_i \notin \mathbb{C} \cup \mathbb{L}\}.
\end{aligned} \tag{5.9}$$

Based on the above definitions, the identifying functional for $\psi(t)$ from Eq. 5.6 can be rewritten as,

$$\psi(t) = \sum_{V \setminus T} Y \times \prod_{M_i \in \mathbb{M}} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))\Big|_{T=t} \times \sum_{T} \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)) \times p(\mathbb{C}). \tag{5.10}$$

It was shown in [13] that the nonparametric influence function obtained by applying the pathwise derivative (Eq. 5.3) to Eq. 5.10 is given by the following theorem.

Theorem 11. *Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, the nonparametric influence function U_{ψ_t} for the target parameter $\psi(t)$ is as follows.*

$$\begin{aligned}
U_{\psi_t} = &\sum_{M_i \in \mathbb{M}} \mathbb{E}[\beta_{primal} \mid \{\preceq M_i\}] - \mathbb{E}[\beta_{primal} \mid \{\prec M_i\}] \\
&+ \sum_{L_i \in \mathbb{L}} \mathbb{E}[\beta_{dual} \mid \{\preceq L_i\}] - \mathbb{E}[\beta_{dual} \mid \{\prec L_i\}] \\
&+ \mathbb{E}[\beta_{primal/dual} \mid \mathbb{C}] - \psi(t),
\end{aligned}$$

where β_{primal} and β_{dual} are obtained via Lemmas 5 and 6, respectively, and $\beta_{primal/dual}$ means that we may use either β_{primal} or β_{dual}.

[13] also shows that the above influence function U_{ψ_t} uses information in the models for $M_i \in \mathbb{M}$ and $L_i \in \mathbb{L}$ in order to yield an estimator that is doubly robust in these sets. That is, the estimator obtained by solving the estimating equation $\mathbb{E}[U_{\psi_t}] = 0$, where U_{ψ_t} is given in Theorem 11, is consistent and asymptotically normal if all models in either $\{p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)),\ \forall M_i \in \mathbb{M}\}$ or $\{p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)),\ \forall L_i \in \mathbb{L}\}$ are correctly specified. To make concepts concrete before the discussion on efficiency, we provide a brief example of the application of Theorem 11 to derive the nonparametric IF for the target $\psi(t)$.

Example: Augmented Primal IPW

We derive the nonparametric IF for the ADMG in Fig. 5-3(b). The sets in display (5.9) are $\mathbb{C} = \{C\}, \mathbb{L} = \{T, L\}$, and $\mathbb{M} = \{M, Y\}$. Using Theorem 11, the nonparametric IF for the target is given by,

$$\begin{aligned}\text{(Fig. 5-3b)} \quad U_{\psi_t} = {} & \mathbb{E}[\beta_{\text{primal}} \mid Y, T, M, L, C] - \mathbb{E}[\beta_{\text{primal}} \mid T, M, L, C] \\ & + \mathbb{E}[\beta_{\text{primal}} \mid M, T, C] - \mathbb{E}[\beta_{\text{primal}} \mid T, C] \\ & + \mathbb{E}[\beta_{\text{dual}} \mid L, T, M, C] - \mathbb{E}[\beta_{\text{dual}} \mid T, M, C] \\ & + \mathbb{E}[\beta_{\text{dual}} \mid T, C] \ - \mathbb{E}[\beta_{\text{dual}} \mid C] \\ & + \mathbb{E}[\beta_{\text{dual}} \mid C] \ - \psi(t), \end{aligned} \tag{5.11}$$

where β_{primal} and β_{dual} are the same primal and dual IPW functionals derived for Fig 5-3(b) in the previous subsection. That is,

$$\begin{aligned}\text{(Fig. 5-3b)} \qquad \beta_{\text{primal}} &= \mathbb{I}(T = t) \times \frac{\sum_T \ p(T \mid C) \times p(L \mid T, M, C)}{p(T \mid C) \times p(L \mid T, M, C)} \times Y. \\ \beta_{\text{dual}} &= \frac{p(M \mid T = t, C)}{p(M \mid T, C)} \times \frac{p(Y \mid T = t, M, L, C)}{p(Y \mid T, M, L, C)} \times Y.\end{aligned}$$

Plugging these into Eq. 5.11 allows us to explicitly write out the nonparametric IF in terms of conditional densities appearing in the topological factorization of the ADMG as follows.

(Fig. 5-3b)
$$
\begin{aligned}
U_{\psi_t} = \frac{\mathbb{I}(T=t)}{p(T \mid C) \times p(L \mid T, M, C)} \times \Big(\sum_T \; & p(T \mid C) \times p(L \mid T, M, C) \times Y \\
& - \sum_T \; p(T \mid C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T=t, M, L, C] \Big) \\
+ \frac{\mathbb{I}(T=t)}{p(T \mid C)} \times \Big(\sum_{T,L} \; & p(T \mid C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T=t, M, L, C] \\
& - \sum_{T,L} \; p(T \mid C) \times p(M \mid T=t, C) \times p(L \mid T, M, C) \times \mathbb{E}[Y \mid T=t, M, L, C] \Big) \\
+ \frac{p(M \mid T=t, C)}{p(M \mid T, C)} \times \Big(\mathbb{E}[Y \mid T=t, M, L, C] \; & - \sum_L \; p(L \mid T, M, C) \times \mathbb{E}[Y \mid T=t, M, L, C] \Big) \\
+ \sum_{M,L} \; p(M \mid T=t, C) \times p(L \mid T, M, C) & \times \mathbb{E}[Y \mid T=t, M, L, C] - \psi(t). \qquad (5.12)
\end{aligned}
$$

An estimator for the target $\psi(t)$ is obtained by solving the estimating equation $\mathbb{E}[U_{\psi_t}] = 0$. In the resulting estimator, conditional densities for $p(T \mid C), p(M \mid T, C), p(L \mid T, M, C)$ and the outcome regression $\mathbb{E}[Y \mid T, M, L, C]$ can be fit either parametrically or using flexible models. The outer expectation is then evaluated empirically using the fitted models in order to yield the target parameter. Per the double robustness property of the estimator, the resulting estimate for $\psi(t)$ is consistent as long as one of the sets $\{p(T \mid C), p(L \mid T, M, C)\}$ or $\{p(M \mid T, C), \mathbb{E}[Y \mid T, M, L, C]\}$ is correctly specified while allowing for arbitrary misspecification of the other.

Another estimation strategy that is computationally simpler stems from the usage of Theorem 11 to the ADMG in Fig. 5-3(b). With the simplification that $\mathbb{E}[\beta_{\text{primal}} \mid Y, T, M, L, C] = \beta_{\text{primal}}$, the resulting estimator for the target is,

$$
\begin{aligned}
\text{(Fig. 5-3b)} \quad \psi_{\text{reform}} = \; \mathbb{E}\big[\, & \beta_{\text{primal}} - \mathbb{E}[\beta_{\text{primal}} \mid T, M, L, C] \\
& + \mathbb{E}[\beta_{\text{primal}} \mid M, T, C] - \mathbb{E}[\beta_{\text{primal}} \mid T, C] \\
& + \mathbb{E}[\beta_{\text{dual}} \mid L, T, M, C] - \mathbb{E}[\beta_{\text{dual}} \mid T, M, C] \\
& + \mathbb{E}[\beta_{\text{dual}} \mid T, C] \, \big]. \qquad (5.13)
\end{aligned}
$$

The above can be estimated from finite samples by first obtaining estimates for β_{primal} and β_{dual} for each row in our data and then fitting flexible regressions for each $\mathbb{E}[\cdot \mid \cdot]$ shown in Eq. 5.13 using these estimates as pseudo outcomes. The outer expectation

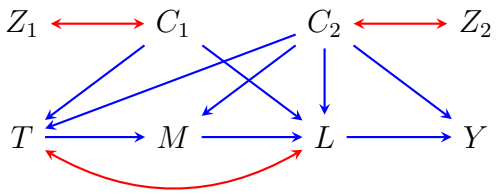


Figure 5-4. An mb-shielded ADMG that is not NPS and where T is primal fixable.

is then evaluated empirically using these fitted models, yielding an estimate for the target parameter $\psi(t)$.

The two estimation strategies described above come with trade-offs. The former approach requires modeling conditional densities and computing sums but preserves the double robustness property and does not face issues of model incompatibility. The latter approach trades model compatibility and double robustness for computational tractability. The latter approach may also face issues in high dimensional settings.

5.3.3 Efficient IF in mb-shielded ADMGs Where T is Primal Fixable

Algorithm 9 served as a means of checking whether the model implied by an ADMG $\mathcal{G}(V)$ is NPS. In an NPS model, there exists a single unique influence function. Hence, the estimator that we obtain by solving $\mathbb{E}[U_\psi] = 0$, where U_ψ is given by Theorem 11 when T is p-fixable, is not only doubly robust but also the most efficient estimator.

On the other hand, constraints in a semiparametric model shrink the tangent space Λ, and thus expand its orthogonal complement $\Lambda^\perp$. As $\Lambda^\perp$ expands, we will have more than one influence function (note that the class of all influence functions is $\{U_\psi + \Lambda^\perp\}$.) We now discuss efficiency results for the class of mb-shielded ADMGs proposed in Theorem 9.

Consider the ADMG shown in Fig. 5-4. Such an ADMG may reflect additional background knowledge or conditional independences known to the analyst. For

example, in Fig. 5-4, $C_1 \perp\!\!\!\perp C_2$ and $M \perp\!\!\!\perp C_1, Z_1, Z_2 \mid T, C_2$. As this model is no longer NPS, the IF obtained via Theorem 11 is not the most efficient. However, it is easy to see that this ADMG is mb-shielded and therefore the efficient IF is given by projection of U_{ψ_t} in Theorem 11 onto the tangent space in Theorem 4. In the following theorem, we provide the general form of the efficient IF in an arbitrary mb-shielded ADMG where T is p-fixable.

Theorem 12. *Given a distribution $p(V)$ that district factorizes with respect to an mb-shielded ADMG $\mathcal{G}(V)$ where T is primal fixable, the efficient influence function for the target parameter $\psi(t)$ is given as follows,*

$$\begin{aligned} U_{\psi_t}^{eff} = &\sum_{M_i \in \mathbb{M}} \mathbb{E}[\beta_{primal} \mid M_i, \text{mp}_{\mathcal{G}}(M_i)] - \mathbb{E}[\beta_{primal} \mid \text{mp}_{\mathcal{G}}(M_i)] \\ &+ \sum_{L_i \in \mathbb{L}} \mathbb{E}[\beta_{dual} \mid L_i, \text{mp}_{\mathcal{G}}(L_i)] - \mathbb{E}[\beta_{dual} \mid \text{mp}_{\mathcal{G}}(L_i)] \\ &+ \sum_{C_i \in \mathbb{C}} \mathbb{E}[\beta_{primal/dual} \mid C_i, \text{mp}_{\mathcal{G}}(C_i)] - \mathbb{E}[\beta_{primal/dual} \mid \text{mp}_{\mathcal{G}}(C_i)] \end{aligned} \tag{5.14}$$

where $\mathbb{C}, \mathbb{L}, \mathbb{M}$ are defined in display (5.9), and β_{primal} and β_{dual} are obtained as in Lemmas 5 and 6 respectively. $\beta_{primal/dual}$ means that we can either use β_{primal} or β_{dual} for the terms in $\mathbb{C}$.

Hence, the primal and dual IPWs comprise the fundamental elements of the efficient influence function in the setting where T is primal fixable. Simplified symbolic representations of the efficient IF in terms of the conditional densities that appear in the topological factorization can be obtained by plugging in the expression from Theorem 12 into computer algebra systems such as [143] and [144].

Example: Efficient APIPW

Applying Theorem 12 to Fig. 5-4 gives us the following efficient estimator. Fix a valid topological order $(C_1, C_2, Z_1, Z_2, T, M, L, Y)$. Then

(Fig. 5-4)
$$\beta_{\text{primal}} = \mathbb{I}(T = t) \times \frac{\sum_T \; p(T \mid C_1, C_2) \times p(L \mid T, M, C_1, C_2)}{p(T \mid C_1, C_2) \times p(L \mid T, M, C_1, C_2)} \times Y,$$
$$\beta_{\text{dual}} = \frac{p(M \mid T = t, C_2)}{p(M \mid T, C_2)} \times Y. \tag{5.15}$$

Define the sets $\mathbb{M} = \{M, Y\}$, $\mathbb{L} = \{T, L\}$, and $\mathbb{C} = \{C_1, C_2\}$. Note that we have dropped terms involving the vertices Z_1 and Z_2 as it is easy to check that $\mathbb{E}[\beta_{\text{dual}} \mid Z_i, \text{mp}_{\mathcal{G}}(Z_i)] = \mathbb{E}[\beta_{\text{dual}} \mid \text{mp}_{\mathcal{G}}(Z_i)]$, resulting in a cancellation of these terms. Then

(Fig. 5-4)
$$\begin{aligned}\psi_{\text{eff}} = \mathbb{E}\big[&\mathbb{E}[\beta_{\text{primal}} \mid Y, L, C_2] - \mathbb{E}[\beta_{\text{primal}} \mid L, C_2] \\ &+ \mathbb{E}[\beta_{\text{primal}} \mid M, T, C_2] - \mathbb{E}[\beta_{\text{primal}} \mid T, C_2] \\ &+ \mathbb{E}[\beta_{\text{dual}} \mid L, M, T, C_1, C_2] - \mathbb{E}[\beta_{\text{dual}} \mid M, T, C_1, C_2] \\ &+ \mathbb{E}[\beta_{\text{dual}} \mid T, C_1, C_2] - \mathbb{E}[\beta_{\text{dual}} \mid C_1, C_2] \\ &+ \mathbb{E}[\beta_{\text{dual}} \mid C_2] + \mathbb{E}[\beta_{\text{dual}} \mid C_1] - \mathbb{E}[\beta_{\text{dual}}]\big]\end{aligned} \tag{5.16}$$

The estimation strategy for the above functional is very similar to the one used for Eq. 5.13. Estimation of the representation of the efficient IF in terms of the original conditional densities as provided by computer algebra systems simply requires fitting models for each conditional density that appears in the functional.

5.4 Estimation of the ACE in Arbitrary ADMGs

So far we have discussed inference of the target $\psi(t)$ in a broad class of ADMGs defined by the primal fixability criterion. However, in arbitrary hidden variable causal models, $\psi(t)$ may be identified even if there exists no valid adjustment set, and T is not p-fixable. The resulting identifying functional is given by truncated factorization of the nested Markov model discussed in Chapter 1 (Eq. 1.9). This strategy for identification of the target is known to be sound and complete [22]. That is, identification of the target parameter $\psi(t)$ in a hidden variable causal model associated with a DAG $\mathcal{G}(V \cup H)$ may be rephrased, without loss of generality, using its corresponding latent projection ADMG $\mathcal{G}(V)$.

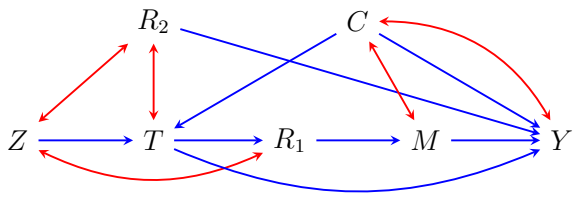


Figure 5-5. An ADMG where the treatment is not p-fixable but $\psi(t)$ is still identified via the truncated nested Markov factorization.

In special cases, when all observed variables are either discrete or multivariate normal, a parametric likelihood can be specified for the nested Markov model [39, 145], which leads naturally to estimation of $\psi(t)$ in Eq. 1.9 by the plug-in principle. However, in applications, assuming a full parametric likelihood is unrealistic. In this section we describe IPW estimators that use only subsets of the likelihood thus reducing the chance of model misspecification. We call this estimator nested IPW and show that its consistency relies only on correct specification of a subset of the nested Markov likelihood that form the district of T.

5.4.1 Nested IPW Estimators

We now describe a general algorithm that yields IPW estimators for *any* $\psi(t)$ that is identifiable from the observed margin $p(V)$ corresponding to an ADMG $\mathcal{G}(V)$. Consider the ADMG shown in Fig. 5-5. Though T is not p-fixable, $\psi(t)$ is still identifiable via the truncated nested Markov factorization as follows. $Y^* = \{Y, M, C, R_1, R_2\}$ and $\mathcal{D}(\mathcal{G}_{Y^*}) = \{\{Y, M, C\}, \{R_1\}, \{R_2\}\}$. Fix a valid topological order $\tau = (Z, C, T, R_1, R_2, M, Y)$. Then from the truncated nested Markov factorization in Eq. 1.9 we have,

$$\begin{aligned}\psi(t) &= \sum_{Y^*} Y \times \phi_{V\setminus\{Y,M,C\}}(p(V);\mathcal{G}) \times \phi_{V\setminus R_1}(p(V);\mathcal{G}) \times \phi_{V\setminus R_2}(p(V);\mathcal{G})\Big|_{T=t} \\ &= \sum_{C,R_1,R_2,M} p(C) \times p(M \mid C, R_1) \times \mathbb{E}[Y \mid M, C, R_1, R_2, t] \times \sum_{Z} p(Z) \times p(R_1 \mid t, Z) \times p(R_2).\end{aligned} \tag{5.17}$$

Algorithm 10 NESTED IPW

Inputs: $\mathcal{G}, p(V)$

Let $Y^* = \text{an}_{\mathcal{G}_{V \setminus T}}(Y)$ and $D_T = \text{dis}_{\mathcal{G}}(T)$ and $\mathcal{D}^* \leftarrow \{D \in \mathcal{D}(\mathcal{G}_{Y^*}) \mid D \cap D_T \neq \emptyset\}$

if $\exists D \in \mathcal{D}^*$ such that D is not intrinsic in $\mathcal{G}$ **then**

    return Fail

Define non-descendants of a vertex V_i as $\text{nd}_{\mathcal{G}}(V_i) \equiv V \setminus \text{de}_{\mathcal{G}}(V_i)$

Fix a topological order τ such that $V_i \succ_\tau \text{nd}_{\mathcal{G}}(V_i) \setminus Y^*, \ \forall V_i \in Y^*$

Define $q_D(D \mid \text{pa}_{\mathcal{G}}(D)) \equiv \phi_{V \setminus D}(p(V); \mathcal{G}(V))$

$\beta_{\text{nested}} \equiv \frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_{\mathcal{G}}(T))} \times \prod_{D \in \mathcal{D}^*} \left(q_D(D \mid \text{pa}_{\mathcal{G}}(D)) \times \prod_{D_i \in D} \frac{1}{p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \right) \times Y$

return $\psi(t)_{\text{nested}} \equiv \mathbb{E}[\beta_{\text{nested}}]$

In the following theorem, we provide the corresponding IPW estimator for all targets $\psi(t)$ that are identifiable from the observed margin of a hidden variable causal DAG $\mathcal{G}(V \cup H)$. As these estimators are derived from the nested Markov factorization of the latent projection ADMG $\mathcal{G}(V)$, we coin the term nested IPW in referring to them. We show that Algorithm 10 which we use to derive such estimators is *sound* and *complete*. That is, when Algorithm 10 returns a nested IPW functional, $\psi(t)_{\text{nested}} = \psi(t)$ and when the algorithm fails to return a functional, $\psi(t)$ is not identifiable within the given model.

Theorem 13. *Let $p(V)$ and $\mathcal{G}(V)$ be the observed marginal distribution and ADMG induced by a hidden variable causal model associated with DAG $\mathcal{G}(V \cup H)$. Then if $\psi(t)$ is identifiable in the model, $\psi(t) = \psi(t)_{nested}$. If $\psi(t)$ is not identifiable in the model, Algorithm 10 returns 'fail'.*

It is easy to see that the nested IPW estimator only requires the specification for parametric models for variables in D_T. That is, the analyst is only required to model distributions related to the treatment assignment and confounding factors related to the treatment assignment.

Example: Nested IPW

We now return to the ADMG shown in Fig. 5-5 and discuss the application of

Theorem 13, in order to obtain an estimator for $\psi(t)$. Recall, $Y^* \equiv \{Y, M, C, R_1, R_2\}$ and $\mathcal{D}(\mathcal{G}_{Y^*}) = \{\{Y, M, C\}, \{R_1\}, \{R_2\}\}$. Note that $\mathcal{D}^*$ simply focuses on the districts related to $\mathcal{G}_{Y^*}$ that do not overlap with D_T. Therefore, $\mathcal{D}^*$ in line 2 of the algorithm is $\{\{R_1\}, \{R_2\}\}$. Since both of these districts are intrinsic in $\mathcal{G}$, Algorithm 10 does not fail. Fix the topological order $(R_2, Z, C, T, R_1, M, Y)$ according to line 6. Then,

$$\begin{aligned}\psi_{\text{nested}} &= \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{p(T \mid Z, C)} \times \frac{\sum_Z p(Z) \times p(R_1 \mid T, Z)}{p(R_1 \mid T, Z, C, R_2)} \times \frac{p(R_2)}{p(R_2)} \times Y\Big] \\ &= \mathbb{E}\Big[\frac{\mathbb{I}(T=t)}{p(T \mid Z, C)} \times \frac{\sum_Z p(Z) \times p(R_1 \mid T, Z)}{p(R_1 \mid T, Z, C, R_2)} \times Y\Big]. \end{aligned} \tag{5.18}$$

The above functional only requires fitting parametric models for conditional densities that appear in the district of T. Thus, the amount of modeling required for the above nested IPW functional is significantly less than parameterizing the full observed data likelihood of the ADMG shown in Fig 5-5.

5.5 Related and Future Work

To the best of our knowledge, prior to the work discussed in this chapter, the *front-door* model [138] was the only graphical model with unmeasured confounders such that no valid covariate adjustment set exists but the ACE is nonparametrically identifiable for which an influence function based estimator had been derived [146]. There also exists a large body of work on semiparametric theory with instrumental variables [147, 148, 149]. However, many of these estimators rely on more assumptions than what is implied by the causal graphical model itself. Weight-based estimators for a subclass of models considered in this paper, were studied in [150]. Other related work includes numerical procedures for approximating the influence function proposed by [151, 152]. However, such methods are either restricted to settings where simple covariate adjustment is valid, or involve numerical approximations of the function itself which may be computationally prohibitive.

The work in this chapter raises several interesting questions for future work. This includes (relatively) simple extensions such as extending the estimators presented here to settings with multiple treatments and outcomes. Other open problems include deriving the nonparametric influence function that serves to augment and improve the efficiency of the nested IPW functional, deriving the tangent space for arbitrary ADMG models which encode generalized equality (Verma) restrictions, and deriving a general algorithm capable of projecting a given nonparametric IF onto the tangent space to obtain the efficient IF.

5.6 Acknowledgements

Chapter 5 is, in part, a reprint of material from [13].

Chapter 6

Conclusion

This dissertation introduced methods for causal inference in the presence of various complications arising from unmeasured confounding, data dependence, missing data, and model misspecification. For the most part, these phenomena were assumed to occur independently of each other. However, there is no substantive reason to believe that these phenomena cannot co-occur. For example, individuals may pressure or influence others in their social network to respond/not respond to surveys. Further, it is well-known that the presence or absence of a friend or partner plays a major role in the outcomes of an individual – whether it be quicker recovery from surgery, or deriving more enjoyment through social bonding over a film. Modeling such phenomena appropriately not only requires new graphical representations, but also definitions of new types of counterfactual random variables, e.g., an individual's outcome from surgery had their partner been present/not present with them for their hospital visits.

Within each sub-problem we discussed there also remain several important unanswered questions. The question of nested Markov equivalence – a graphical characterization of which ADMGs imply the same equality constraints on the observed data distribution – and a description of general nested Markov likelihoods are ones that I hope to continue pursuing past my dissertation work. Efficient semiparametric estimators for causal effects in the presence of arbitrary patterns of confounding and missingness are also problems that interest me a great deal.

Technical issues aside, there remain major challenges in bridging the gap between theory and practice in causal inference. New theoretical results in discovery, identification, and estimation arrive at a much quicker pace than what practictioners are willing to adopt. Some of this, is in part, due to a disconnect between the needs of the practitioner and the technical interests of the theorist. However, a significant portion of this gap may also be attributed to a lack of robust and well-documented statistical software for causal inference. I hope to address both these problems in my future research program. The first by continuing to emphasize the development of practical methods and staying in touch with the problems in computational oncogenomics that motivated them. And second by continuing the development and maintenance of high quality software for causal inference in *Ananke* [9].

Appendix A

Marginalization, Conditioning, and Fixing in Kernels

A kernel $q_V(V \mid W)$ is a mapping from values of W to normalized densities over V. That is, $\sum_V q_V(V \mid W = w) = 1, \forall w \in W$. For any set of variables $X \subseteq V$, marginalization and conditioning in a kernel are defined as follows.

$$q_{V \setminus X}(V \setminus X \mid W) \equiv \sum_X q_V(V \mid W), \text{ and}$$
$$q_V(V \setminus X \mid X, W) \equiv \frac{q_V(V \mid W)}{q_V(X \mid W)}.$$

The notation $q_V(\cdot \mid X)$ makes clear which variables appearing past the "conditioning" bar in a kernel are fixed as opposed to simply conditioned on. That is, if a variable $X_i \notin V$, then it is fixed, else it is conditioned on. Occasionally, fixing operations may also simplify to marginalization or conditioning events. We illustrate these concepts with a simple example.

Consider the ADMG shown in Figure A-1(a) and fix the kernel of interest to be $q_Y(Y \mid T, Z_1, Z_2)$, i.e., a kernel where all other variables except Y are fixed. A valid fixing sequence in order to obtain such a kernel from the joint $p(V)$ is (Z_2, Z_1, T). Fixing Z_2 entails dividing by the simple conditional $p(Z_2 \mid Z_1)$ and yields the CADMG $\phi_{Z_2}(\mathcal{G})$ and corresponding kernel $q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)$ shown in Figure A-1(b). In order to fix Z_1, we must divide by the kernel $q_{Z_1,T,Y}(Z_1 \mid Z_2, T, Y)$. By rules of

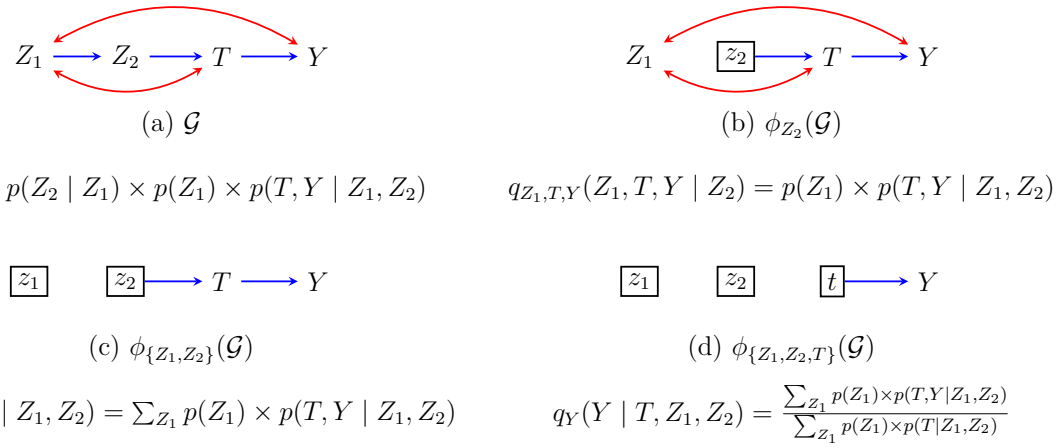


Figure A-1. An example to illustrate fixing and kernel operations.

conditioning and marginalization in kernels,

$$q_{Z_1,T,Y}(Z_1 \mid Z_2, T, Y) \equiv \frac{q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)}{q_{Z_1,T,Y}(T, Y \mid Z_2)} \equiv \frac{q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)}{\sum_{Z_1} q_{Z_1,T,Y}(Z_1, T, Y \mid Z_2)}$$

Fixing Z_1 and evaluating the above expression gives us the CADMG and corresponding kernel shown in Figure A-1(c). That is, fixing Z_1 in the kernel $q_{Z_1,T,Y}(Z_1 \mid Z_2, T, Y)$, simplifies to marginalization of Z_1. Finally, applying rules of conditioning and marginalization to the kernel $q_{T,Y}(T, Y \mid Z_1, Z_2)$ we can obtain the kernel $q_{T,Y}(T \mid Z_1, Z_2, Y)$. Dividing by this corresponds to fixing T, giving us the CADMG and desired kernel shown in Figure A-1(d).

Appendix B

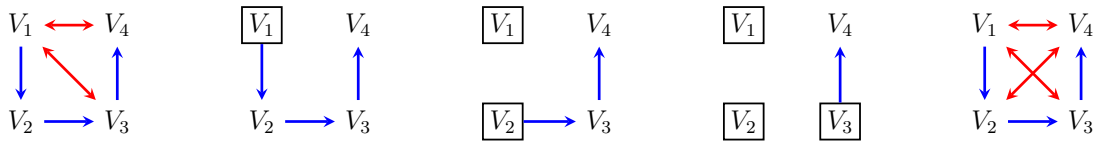
Supplement to Chapter 2

In this Appendix, we first discuss details of the GREENERY algorithm for penalizing c-trees and introduce the formalizations necessary to prove its correctness. We then provide additional comments on the protein expression network learned by applying our method to the data from [1]. We then discuss additional implementation details and choice of hyperparameters for our experiments. Finally we present formal proofs of results in our paper.

B.1 Details of the Greenery Algorithm

[13] introduced a graphical and probabilistic operator called *primal fixing* that can be applied recursively to an ADMG and its statistical model to identify causal parameters of interest. In this section we provide the necessary background on the graphical operator and discuss how it relates to the detection of c-trees. We then show how primal fixing is codified in the steps of Algorithm 1 through an example.

A conditional ADMG (CADMG) $\mathcal{G} = (V, W, E)$ is an ADMG whose vertices can be partitioned into random vertices V and fixed vertices W, with the restriction that no arrowheads point into W [22]. A vertex V_i in a CADMG $\mathcal{G} = (V, W, E)$ is said to be primal fixable if there is no bidirected path from V_i to any of its direct children. The graphical operation of primal fixing V_i in $\mathcal{G}$, denoted by $\phi_{V_i}(\mathcal{G})$, yields a new



(i) $\mathcal{G}^a$ (ii) $\phi_{V_1}(\mathcal{G}^a)$ (iii) $\phi_{\{V_1,V_2\}}(\mathcal{G}^a)$ (iv) $\phi_{\{V_1,V_2,V_3\}}(\mathcal{G}^a)$ (v) $\mathcal{G}^b$

Figure B-1. (i) An arid ADMG; (ii) The CADMG obtained after primal fixing V_1; (iii) The CADMG obtained after primal fixing V_1 and V_2; (iv) The CADMG obtained after primal fixing V_1, V_2, and V_3; (v) A non-arid bow-free ADMG that is a super model of (i).

CADMG $\mathcal{G} = (V \setminus V_i, W \cup V_i, E \setminus \{e \in E \mid e = \circ\!\!\rightarrow V_i \text{ or } \circ\!\!\leftrightarrow V_i\})$ where V_i is now "fixed" (denoted by a square box in figures shown in this Supplement) and incoming edges into V_i are deleted. This can be extended to a set of vertices as follows. A set of k vertices S is said to be primal fixable if there exists an ordering $(S_1, \ldots, S_k)$ such that S_1 is primal fixable in $\mathcal{G}$, S_2 is primal fixable in $\phi_{S_1}(\mathcal{G})$, S_3 is primal fixable in $\phi_{S_2}(\phi_{S_1}(\mathcal{G}))$, and so on. It is easy to see that any such valid ordering on S yields the same final CADMG. Hence, we can denote primal fixing a set of vertices S as simply $\phi_S(\mathcal{G})$. A vertex V_i in an ADMG $\mathcal{G}$ is said to be *reachable* if $V \setminus V_i$ is primal fixable in $\mathcal{G}$. [39] showed that if V_i is reachable in $\mathcal{G}$, then the causal effect of the parents of V_i on V_i itself is identified, and there is no V_i rooted c-tree in $\mathcal{G}$.[1] If no valid primal fixing order exists, V_i along with the unique minimal set of vertices that could not be primal fixed form a V_i-rooted c-tree [39]. That is, an ADMG $\mathcal{G}$ is arid if and only if every vertex $V_i \in V$ is reachable. This forms the basis of Algorithm 1.

We now demonstrate usage of the primal fixing operator to establish that the ADMG $\mathcal{G}^a$ shown in Figure B-1(i) is arid and the ADMG $\mathcal{G}^b$ shown in Figure B-1(v) is not. These are the same graphs shown in Section 2.1 of the paper but we redraw and relabel them here for convenience. The reachability of vertices V_1, V_2, and V_3 in

[1] Actually this was shown with respect to the ordinary fixing operator proposed in [22] which performs the same graphical operation as primal fixing but considers V_i to be fixable when there are no bidirected paths to any descendant (a vertex V_j such that there exists a directed path from V_i to V_j) of V_i. It is easy to see how primal fixing is a strict generalization of fixing by noting that the children of V_i is a subset of its descendants.

$\mathcal{G}^a$ is easily established. In every case, we can primal fix the remaining vertices in a reverse topological order starting with V_4 which has no children. The reachability of V_4 is established by noticing that V_1 is primal fixable in $\mathcal{G}^a$. In the resulting CADMG, shown in Figure B-1(ii), both V_2 and V_3 are primal fixable. Primal fixing V_2 yields the CADMG in Figure B-1(iii) and finally primal fixing V_3 yields the CADMG in Figure B-1(iv). Hence, all vertices in $\mathcal{G}^a$ are reachable. It then follows that $\mathcal{G}^a$ is arid. If we try to apply the same reasoning to the $\mathcal{G}^b$ in Figure B-1(v), we see that V_1, V_2, and V_3 are still reachable as before. However, we cannot establish a sequence of primal fixing operations to reach V_4 as none of the other vertices are primal fixable in the original graph. Hence, there is a V_4-rooted c-tree in $\mathcal{G}^b$ comprised of the arborescence $V_1 \to V_2 \to V_3 \to V_4$ which also forms a bidirected component in $\mathcal{G}^b$.

B.1.1 Example Application of the Greenery Algorithm

We now demonstrate how the above primal fixing steps relate to Algorithm 1. Let the ordering of vertices of entries in the matrix be V_1, V_2, V_3, V_4. The adjacency matrices D and B for $\mathcal{G}^a$ in Figure B-1(i) are as follows.

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

The i^{th} iteration of the outer loop of the algorithm attempts to establish the reachability of V_i, and hence, the presence or absence of a V_i-rooted c-tree. Note that since the primal fixing operation can be applied at most $d-1$ times (where d is the number of vertices in $\mathcal{G}$) to determine the reachability of V_i, the inner loop of Algorithm 1 also executes $d-1$ times. We now focus on the final iteration of the algorithm where it tries to establish the reachability of V_4.

In the first iteration of the inner loop we have $D^f = D$ and $B^f = B$. Therefore we

have,

$$e^{B^f} \circ D = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.59 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad f = \begin{bmatrix} 0 & 0 & 0.53 & 0.76 \end{bmatrix} \qquad F = \begin{bmatrix} 0 & 0 & 0.53 & 0.76 \\ 0 & 0 & 0.53 & 0.76 \\ 0 & 0 & 0.53 & 0.76 \\ 0 & 0 & 0.53 & 0.76 \end{bmatrix}.$$

Each entry i, j of the matrix $e^{B_f} \circ D$ is zero if and only if a bidirected path from V_i to V_j and a directed edge $V_i \to V_j$ do not co-exist in $\mathcal{G}$. The sum of the i^{th} row of this matrix then exactly characterizes the primal fixability criterion. That is, V_i is primal fixable if and only if the sum of the i^{th} row in $e^{B_f} \circ D$ is 0. The above calculations indicate that the vertices V_1, V_2, and V_4 are all primal fixable in $\mathcal{G}^a$, which can be easily confirmed by looking at the graph itself. The vector f then summarizes the primal fixability of each vertex except we add the i^{th} row of an identity matrix to ensure that we do not accidentally primal fix V_i itself when determining its reachability. The matrix F formed by tiling the f vector d times can then be used as a "mask" that implements the primal fixing operation applied to V_1 and V_2 simultaneously, yielding the following updates to D^f and B^f.

$$D^f = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0.53 & 0 \\ 0 & 0 & 0 & 0.76 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad B^f = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

It is easy to confirm that the induced ADMG $\mathcal{G}(D^f, B^f)$ corresponds to the CADMG shown in Figure B-1(iii). Note that a constant positive scaling factor can also be applied to the hyperbolic tangent function to improve the sharpness of the approximation of the primal fixing operator. In the second iteration of the loop, we apply the same

process again and obtain,

$$e^{B^f} \circ D = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad f = \begin{bmatrix} 0 & 0 & 0 & 0.76 \end{bmatrix} \qquad F = \begin{bmatrix} 0 & 0 & 0 & 0.76 \\ 0 & 0 & 0 & 0.76 \\ 0 & 0 & 0 & 0.76 \\ 0 & 0 & 0 & 0.76 \end{bmatrix}.$$

That is, in the second iteration of the algorithm, V_3 becomes primal fixable. Applying the primal fixing operator yields the adjacency matrices,

$$D^f = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.58 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad B^f = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

which induce the CADMG shown in Figure B-1(iv) corresponding to primal fixing V_3. Thus, in this case, reachability of V_4 is established in 2 steps. However, the algorithm will still perform a third step that does not result in any additional primal fixing and does not change the conclusion of reachability of V_4. As there are no vertices that have both a bidirected path and directed path to V_4 in the final CADMG and corresponding adjacency matrices, $C = e^{B^f} \circ e^{D^f}$ is simply the identity matrix. Taking the i^{th} column sum then evaluates to 1 which is subtracted off later in the final "return" step of the algorithm. A similar argument holds for vertices V_1, V_2, and V_3. Thus, applying Algorithm 1 to $\mathcal{G}^a$ in Figure B-1(i) returns a value of 0 confirming that $\mathcal{G}^a$ is arid.

We now consider application of the algorithm to the ADMG $\mathcal{G}^b$ shown in Figure B-1(v). We will apply a scaling constant of 10 to the hyperbolic tangent function, i.e., we use $tanh(10x)$, so that the values are large enough to illustrate the main concept. We again focus on the reachability of V_4. The adjacency matrices for $\mathcal{G}^b$ are:

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

In the first iteration of the inner loop we have,

$$e^{B^f} \circ D = \begin{bmatrix} 0 & 0.64 & 0 & 0 \\ 0 & 0 & 0.19 & 0 \\ 0 & 0 & 0 & 0.64 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad f = \begin{bmatrix} 1 & 0.96 & 1 & 1 \end{bmatrix} \qquad F = \begin{bmatrix} 1 & 0.96 & 1 & 1 \\ 1 & 0.96 & 1 & 1 \\ 1 & 0.96 & 1 & 1 \\ 1 & 0.96 & 1 & 1 \end{bmatrix}.$$

That is, we see that none of the vertices in $\mathcal{G}^b$ are primal fixable. Therefore applying the primal fixable operator through the matrix F results in adjacency matrices,

$$D^f = \begin{bmatrix} 0 & 0.96 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad B^f = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0.96 \\ 1 & 0 & 0 & 0 \\ 1 & 0.96 & 0 & 0 \end{bmatrix},$$

which induce a "CADMG" that has the same edges as the original graph $\mathcal{G}^b$. Repeated applications of this in the second and third iterations do not change the structure of the induced graph. Therefore, upon termination of the inner loop, there remains a directed path from every vertex in $V \setminus V_4$ to V_4 and the vertices still form a bidirected connected component. That is, there is a V_4-rooted c-tree in $\mathcal{G}^b$. This is confirmed when we evaluate the sum of the i^{th} column of $C = e^{D^f} \circ e^{B^f}$ to 2.34. The other vertices V_1, V_2, and V_3 are still reachable and their respective column sums upon termination of the inner loop yield a value of 1 each. Subtracting d at the end of the algorithm still leaves a positive remainder of 1.34. Hence, Algorithm 1 returns a positive quantity when applied to $\mathcal{G}^b$, confirming that it is not arid.

B.2 Comments on Protein Expression Analysis

In this section we discuss the Verma restriction that allows us to establish that Erk is *not* a cause of PKA. The importance of this relation stems from manipulation of Erk by the authors of [1] and establishing that no downstream change was observed in PKA.

We first point out that there is no ordinary conditional independence constraint between Akt and PKC in the learned structure shown in the right panel of Figure 2-3, despite the absence of an edge between the two. This can be confirmed by noting the presence of an *inducing path* between Akt and PKC. An inducing path between V_i and V_j is a path from V_i to V_j where every non-endpoint is both a collider ($\rightarrow \circ \leftarrow$, $\leftrightarrow \circ \leftarrow$, or $\leftrightarrow \circ \leftrightarrow$) and has a directed path to either V_i or V_j. It is well-known that the presence of such a path precludes the possibility of an ordinary conditional independence of the form $V_i \perp\!\!\!\perp V_j \mid Z$ for any $Z \subseteq V \setminus \{V_i, V_j\}$ [23]. In our analysis it can be confirmed that Akt $\rightarrow$ Erk $\leftrightarrow$ PKA $\leftrightarrow$ PKC is an inducing path between Akt and PKC. Thus, there is no ordinary conditional independence between these two proteins under our learned model. However, under the faithfulness assumption, the absence of the edge between Akt and PKC implies an equality restriction. We now provide the non-parametric form of the corresponding Verma constraint.

Consider the ADMG and corresponding distribution obtained by recursively marginalizing out all vertices (except PKC) with no outgoing directed edges in Figure 2-3. In performing this graphical operation, none of the variables removed act as a latent confounder for the remaining variables in the problem. Therefore, by rules of latent projection described in [23], we simply obtain a subgraph of the original network as shown in Figure B-2(i). Note that the inducing path between Akt and PKC is still preserved. Let $p(V^s)$ be the corresponding marginal distribution on the remaining subset of variables. The Verma constraint is then given by,

$$\text{Akt} \perp\!\!\!\perp \text{PKC in } \frac{p(V^s)}{p(\text{Jnk} \mid \text{Erk}, \text{PKA})}.$$

Intuitively, one can view the independence between Akt and PKC as manifesting in a post-intervention distribution obtained after intervening on Jnk, resulting in the CADMG (or truncated ADMG) shown in Figure B-2(ii) where incoming edges to Jnk are removed. The resulting independence is then easily read off from the CADMG via the m-separation criterion [29]. See [30] and [22] for more details on

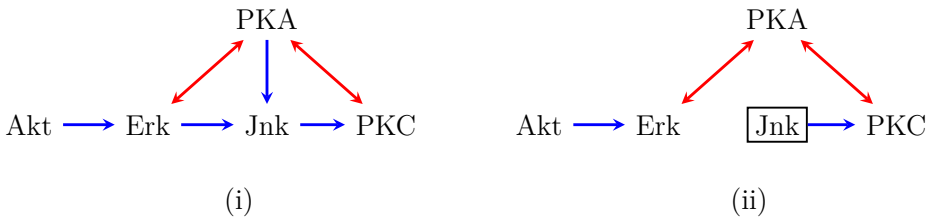


Figure B-2. (i) A subgraph of the protein network in Figure 2-3 that we use to highlight the Verma constraint between Akt and PKC; (ii) A CADMG corresponding to the post-intervention distribution that would be obtained by intervening on Jnk.

how to derive such constraints in general. Orienting the Erk $\leftrightarrow$ PKA edge as either Erk $\leftarrow$ PKA or Erk $\rightarrow$ PKA breaks the inducing path between Akt and PKC, meaning that either orientation produces a different independence model implying an ordinary independence constraint instead of the Verma restriction. We evaluated the BIC scores with either orientation and confirm that they both yield an increase in the score. This indicates that our learned model which posits that Erk is correlated with PKA through unmeasured confounding is the preferred causal explanation. This explanation is consistent with experiments performed in [1], and we are able to arrive at the same conclusion from purely observational data. Moreover, this explanation was differentiated from others via the Verma restriction between Akt and PKC, highlighting the value of considering general equality restrictions beyond ordinary conditional independence.

B.3 Implementation Details

In this section we discuss implementation details of our procedure that were not included in the main chapter.

B.3.1 Implementation of Constraints

As mentioned in Chapter 2, we use the representation of constraints in Table 2-I obtained by replacing each matrix exponential e^A with $(I + cA)^d$. We have two

primary reasons for doing so. First, as pointed out by [45], the latter representation is numerically more stable. Second, by evaluating the binomial expansion $(I + cA)^d = I + \sum_{k=1}^{d} \binom{d}{k} c^k A^k$ explicitly, we are able to obtain analytic gradients for our constraints automatically via the HIPS Autograd package [153, 154]. Analytic gradients for the matrix exponential on the other hand are not easily obtained and the function itself is not implemented in many popular computing libraries. In our implementation we use a value of $c = 1$ when computing portions of the constraint related to directed edges and a value of $c = 2$ when computing portions of the constraint related to bidirected edges. As the constraints in Theorem 1 are valid for any $c > 0$, these values were chosen only to make values of $h(\theta)$ under violations of ancestrality, aridity, and bow-freeness to be larger than the tolerance level (10^{-8}) of the augmented Lagrangian procedure. A scaling factor applied to the hyperbolic tangent function controls the sharpness of approximation of the primal fixing operator. In our experiments we use a scaling factor of $\ln(5000)$, but any sufficiently large value suffices as long as the penalty $h(\theta)$ computed for c-trees is above the tolerance level of the augmented Lagrangian procedure. Finally symmetry of the matrix β is enforced by requiring each off-diagonal entry β_{ij} and β_{ji} are tied to a single free parameter. Positive-definiteness of β is guaranteed by construction in the RICF procedure [66].

B.3.2 Choice of Hyperparameters

We summarize our choice of hyperparameters and justification for these choices in Table B-I. Choice of some hyperparameters, such as tolerance levels for RICF and increments in RICF iterations, require little justification as lower tolerance and more iterations can only improve approximation. We set specific values only to cap the run time of our procedure. Choices for most other hyperparameters are based on prior literature.

Hyperparameter	Setting	Justification
Tolerance for $h(\theta)$	10^{-8}	Numerically close enough to 0 – the lower the better.
Max dual ascent iterations	100	Same value as in [44]; convergence is typically achieved within 10 iterations.
RICF increment s	1	RICF often converges in 10 steps [66, 68]. Higher values should be used for larger graphs.
Regularization strength λ	0.05	Obtained through manual testing on held-out data derived from Figure 2-1(b,c).
Progress rate r	0.25	Same value as in [44]; [45].
Tolerance for RICF	10^{-4}	Numerically close enough to 0 – the lower the better.

Table B-I. Hyperparameter settings used for our experiments.

B.3.3 Converting Estimates of θ to an ADMG $\mathcal{G}(\theta)$

The final step of Algorithm 3 returns an ADMG $\mathcal{G}(\theta)$ as follows. We first derive the matrices δ and β from θ. The structure of the induced ADMG is then given by: $V_i \to V_j$ exists in $\mathcal{G}$ if $|\delta_{ij}| > \omega$ and $V_i \leftrightarrow V_j$ exists in $\mathcal{G}$ if $|\beta_{ij}| > \omega$ for all $i \neq j$. Such thresholding is standard in similar continuous optimization structure learning methods, such as [44] and [45], and the threshold can be made arbitrarily small as long as tolerance to $h(\theta)$ is also small. In our experiments we use $\omega = 0.05$.

B.4 Proofs

Theorem 1 *The constraints shown in Table 2-I are satisfied if and only if the adjacency matrices satisfy the relevant property of ancestrality, aridity, and bow-freeness respectively.*

Proof. We use the following facts for all of our proofs. The matrix exponential of a

square matrix A is defined as the infinite Taylor series,

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k. \tag{B.1}$$

For a binary square matrix A, corresponding to a directed/bidirected adjacency matrix, the entry A^k_{ij} counts the number of directed/bidirected walks of length k from vertex i to vertex j; see for example [53].

Ancestral ADMGs

Consider the constraint shown in Table 2-I. That is,

$$\text{trace}(e^D) - d + \text{sum}(e^D \circ B) = 0.$$

It is easy to see from results in [44] that the constraint $\text{trace}(e^D) - d = 0$ is satisfied if and only if the induced graph $\mathcal{G}(D, B)$ is acyclic. We now show that $\text{sum}(e^D \circ B) = 0$ if and only if $\mathcal{G}$ is ancestral.

By definition of the matrix exponential,

$$\begin{aligned}\text{sum}(e^D \circ B) &= \text{sum}\left(I \circ B + \sum_{k=1}^{\infty} \frac{1}{k!} D^k \circ B\right) \\ &= \text{sum}\left(I \circ B\right) + \sum_{k=1}^{\infty} \frac{1}{k!} \text{sum}\left(D^k \circ B\right),\end{aligned}$$

where the second equality follows from basic matrix properties.

The first term in the series, $\text{sum}(I \circ B)$, counts the number of self bidirected edges $V_i \leftrightarrow V_i$ which is a special-case violation of ancestrality. This term is zero if no such edges exist. An entry i, j in the matrix $D^k \circ B$ counts the number of occurences of directed paths from V_i to V_j of length k such that V_i and V_j are also connected via a bidirected edge. Therefore, all remaining terms of the form $\frac{1}{k!}\text{sum}(D^k \circ B)$ count the number of directed paths of length k that violate the ancestrality property rescaled by a positive factor of $\frac{1}{k!}$. That is, these terms are all ≥ 0 and equal to zero only when no such paths exist, i.e., $\mathcal{G}$ is ancestral.

Arid ADMGs

Consider the constraint shown in Table 2-I. That is,

$$\text{trace}(e^D) - d + \textsc{Greenery}(D, B) = 0.$$

The terms $\text{trace}(e^D) - d$ capture the acyclicity constraint as before. We now show that the output of Algorithm 1 is zero if and only if $\mathcal{G}$ satisfies the arid property. That is, $\textsc{Greenery}(D, B) = 0$ is satisfied if and only if $\mathcal{G}$ is arid. The background required for this proof was laid out at the beginning of this Appendix.

The outer loop of Algorithm 1 iterates over each vertex V_i in order to evaluate its reachability, or equivalently, the presence/absence of a V_i-rooted c-tree [39]. The inner loop achieves this as follows.

Reachability of V_i can be determined in at most $d - 1$ primal fixing operations. Therefore, the inner loop executes $d - 1$ times. On each iteration, the algorithm considers the primal fixability of vertices by effectively treating the matrices D^f and B^f as adjacency matrices of a CADMG. In the first iteration, D^f and B^f are initialized with values from the directed and bidirected adjacency matrices respectively. The sum of the j^{th} row in the matrix $e^{B^f} \circ D^f$ evaluates to zero if and only if there are no bidirected paths from V_j to any of its direct children V_k, which exactly corresponds to the graphical criterion for determining primal fixability of V_j. The addition of the i^{th} row of an identity matrix to t ensures that V_i itself is not treated as primal fixable when evaluating its reachability. Therefore, in the first iteration, the vector f encodes a smoothened version (due to the application of the hyperbolic tangent function) of the usual primal fixability criterion for all vertices $V \setminus V_i$ in the original graph $\mathcal{G}$. Tiling the vector f to form the $d \times d$ matrix F allows us to apply the softened version of primal fixing to the adjacency matrices, which is performed in lines 7-9 of the algorithm. On the next iteration, the matrices D^f and B^f can then be treated as adjacency matrices of a CADMG obtained by primal fixing a set of vertices, say S_1,

that satisfied the primal fixability criterion in $\mathcal{G}$. The same logic can be applied to subsequent iterations of the algorithm where we determine the primal fixability of a set of vertices $V \setminus (S_1 \cup V_i)$ in $\phi_{S_1}(\mathcal{G})$, denote the primal fixable vertices as S_2, and then proceed to do the same for $V \setminus (S_1 \cup S_2 \cup V_i)$ in $\phi_{S_1 \cup S_2}(\mathcal{G})$, and so on.

On termination of the inner loop, we have that $S_1 \cup S_2, \ldots, \cup S_{d-1} \subseteq V \setminus V_i$. We first consider the case when equality holds. In this case, V_i is reachable, from which it follows that there is no V_i-rooted c-tree in $\mathcal{G}$ [39]. The final matrices D^f and B^f then correspond to a CADMG where all vertices except V_i have been primal fixed. In such a CADMG the only edges that may be present are directed edges into V_i due to the removal of incoming edges to all other vertices in the graph. Thus, e^{B^f} evaluates to an identity matrix as there are no bidirected edges. Assuming $\mathcal{G}$ is a graph with no directed cycles (which is already enforced by the first two terms in the arid constraint), the Hadamard product $C = e^{B^f} \circ e^{D^f}$ is then also an identity matrix. Taking the sum of the i^{th} column of C then simply evaluates to 1. If every vertex $V_i \in V$ is reachable in this manner, it implies that the graph is arid, and the greenery quantity will then evaluate to d. The subtraction of d in the "return" statement of Algorithm 1 then returns a value of 0 for arid graphs. Now we consider the case when equality does not hold, i.e., there exists a set of vertices $X = V \setminus V_i \setminus (S_1 \cup S_2 \cdots \cup S_{d-1})$ that could not be primal fixed. This implies that V_i is not reachable and there exists a V_i-rooted c-tree. By definition, the structure of this c-tree comprises of directed and bidirected paths from vertices in X to V_i. The sum of the i^{th} column in $C = e^{B^f} \circ e^{D^f}$ then provides a weighted count of these paths. Subtracting off d in the final "return" statement then yields a positive quantity that provides a weight for each V_i-rooted c-tree detected in a non-arid graph $\mathcal{G}$.

Bow-free ADMGs

Consider the constraint shown in Table 2-I. That is,

$$\text{trace}(e^D) - d + \text{sum}(D \circ B) = 0.$$

The terms $\text{trace}(e^D) - d$ capture the acyclicity constraint as before. It is easy to see that the term $\text{sum}(D \circ B)$ counts the number of bows in the induced graph $\mathcal{G}$. Hence, $\text{sum}(D \circ B)$ is zero if and only if $\mathcal{G}$ is bow-free.

□

Theorem 2 *Let $p(V; \theta^*)$ be a distribution in the curved exponential family that is Markov and faithful with respect to an arid ADMG $\mathcal{G}^*$. Finding the global optimum of the continuous program in display (2.1) with $f \equiv BIC$ yields an ADMG $\mathcal{G}(\theta)$ that implies the same equality restrictions as $\mathcal{G}^*$.*

Proof. This follows immediately from the validity of the constraints in Theorem 1 and the consistency of the BIC score for model selection in curved exponential families [55].

□

Corollary 1.2 *The results in Theorem 1 and Corollary 1.1 hold if every occurrence of a matrix exponential e^A is replaced with the matrix power $(I + cA)^d$ for any $c > 0$, where I is the identity matrix.*

Proof. The proof is straightforward by noting that the binomial expansion of $(I + cA)^d = I + \sum_{k=1}^{d} \binom{d}{k} c^k A^k$ which is similar to the infinite series expansion of the matrix exponential truncated to d terms. As paths greater than length d are irrelevant in a system with d vertices, these terms are sufficient.

□

Appendix C

Supplement to Chapter 3

In this Appendix, we first describe how CG models may be viewed as a set of conditional MRFs. We then provide an informal analysis of the computational cost incurred by computing the PBIC for CG models. Finally we provide longer technical proofs that were excluded from the main chapter.

C.1 Conditional MRFs

A CG model can be viewed as a set of conditional MRFs. A conditional MRF corresponds to a graph whose vertices can be partitioned into two disjoint sets: W, corresponding to non-random variables whose values are fixed; and V, corresponding to random variables. The only edges allowed in a conditional MRF are directed edges $W_i \to V_i$ and undirected edges $V_i - V_j$ for $W_i \in W$ and $V_i, V_j \in V$. A statistical model associated with a conditional MRF $\mathcal{G}$ is a set of densities that factorize as:

$$p(V \mid W) = \frac{\prod_{C \in \left\{ \mathcal{C}\left(\mathcal{G}^a_{\mathrm{bd}_{\mathcal{G}}(B)}\right) \,:\, C \not\subseteq W \right\}} \kappa_C(C)}{Z\big(W\big)}.$$

It is easy to see that the above factorization is analogous to the second level of CG factorization found in Eq. 1.10 where V is a block, and W are its parents.

C.2 Computational Complexity of Computing Scores of a CG Model

In blocks of a CG, the number of local terms that need to be computed corresponds to the number of vertices present in cliques containing the edge of interest in the augmented subgraph of the block and its parents. A term for V_j requires an $O(|\operatorname{bd}_{\mathcal{G}}(V_j)|)$ computation to update, which in the worst case may be exponential in the number of vertices if the graph is not sparse. In search problems, restrictions can be made on the maximum size of the boundary set, sacrificing accuracy for tractability. For a block in a CG corresponding to a conditional MRF in the exponential family, and an edge that is present in a set of cliques spanning all vertices, we will have a local set of size $O(d)$ in the worst case, with each local term requiring an $O(\text{clique size})$ computation. Thus, limiting the maximum clique size may speed up the computation of each local term, but in many cases we may be unable to avoid an $O(d)$ number of such terms. In other words, our scoring method for CG models where blocks correspond to conditional MRFs in the exponential family may not scale to very large graphs, even if such graphs are sparse. Achieving such a scaling will entail making additional assumptions, such as Gaussianity, or non-existence of higher order interaction terms in log-linear models. We contrast this with DAG models, where the local set is of constant size regardless of parametric assumptions made.

C.3 Proofs

Lemma 1 *With dimension fixed and sample size increasing to infinity, the PBIC is a consistent score for curved exponential families whose natural parameter space Θ forms a compact set.*

Proof. Let $\mathcal{M}_0$ denote the true model and $\mathcal{M}_1$, $\mathcal{M}_2$ two candidate models. A scoring

criterion $S(X;\mathcal{M})$ is said to be *consistent* if:

$$\lim_{n\to\infty} P_n(S(X;\mathcal{M}_1) < S(X;\mathcal{M}_2)) \to 1 \text{ when}$$

$$\mathcal{M}_1 \not\supseteq \mathcal{M}_0 \text{ and } \mathcal{M}_2 \supseteq \mathcal{M}_0 \text{ or} \tag{*}$$

$$\mathcal{M}_1, \mathcal{M}_2 \supseteq \mathcal{M}_0 \text{ and } k_1 > k_2. \tag{**}$$

To prove consistency of the PBIC we need to show that,

$$\lim_{n\to\infty} P_n(PBIC(X;\mathcal{M}_1) < PBIC(X;\mathcal{M}_2)) \to 1 \tag{C.1}$$

when (*) or (**).

Note in all following steps, we assume dependence on the dataset X to be implicit in the calculation of the likelihoods and pseudolikelihoods.

To prove (C.1) holds under the scenario (*), it is sufficient to show that the following is true for some $\epsilon > 0$

$$\frac{1}{n}(\ln \mathcal{PL}_n(\hat{\theta}_2) - \ln \mathcal{PL}_n(\hat{\theta}_1)) > \epsilon \tag{C.2}$$

It was shown in [55] that for any $\mathcal{M}_1$ outside of a neighborhood N of θ_0, and $\mathcal{M}_2$ containing this neighborhood, we can pick a $\delta > 0$ such that:

$$\frac{1}{n}(\ln \mathcal{L}_n(\hat{\theta}_2) - \ln \mathcal{L}_n(\hat{\theta}_1)) > \delta \tag{C.3}$$

In order to extend this result to (C.2), we invoke a result from [155] stating that

$$\mathcal{PL}_n(\theta) \geq d\mathcal{L}_n(\theta) + \sum_{i=1}^{d} H_i(\widetilde{P}_n) \tag{C.4}$$

where d is the dimensionality of the data, and $H_i(\widetilde{P}_n)$ is the Shannon entropy of the empirical distribution. It then follows that (C.2) holds when (C.3) is true.

Showing that (C.1) holds under the scenario (**) is equivalent to showing that the following quantity is $O_p(1/n)$:

$$\frac{1}{n}|\ln \mathcal{PL}_n(\hat{\theta}_1) - \ln \mathcal{PL}_n(\hat{\theta}_2)| \tag{C.5}$$

Consider the difference between the full log-likelihoods:

$$\frac{1}{n}|\ln \mathcal{L}_n(\hat{\theta}_1) - \ln \mathcal{L}_n(\hat{\theta}_2)|. \tag{C.6}$$

We first closely follow the proof in [55] to show that the quantity in (C.6) is $O_p(1/n)$. Consider data drawn from a curved exponential family density $p(X;\theta) = h(X)\exp(\theta T(X) - Z(\theta))$, where $\theta \in \mathbb{R}^k$ is a set of canonical parameters in the natural parameter space Θ, $T(X)$ is a set of sufficient statistics, and $Z(\theta)$ is a normalizing function. For a particular choice of a model $\mathcal{M}$ in this setting, the BIC can be written as $\ln\mathcal{L}_n(D;\hat{\theta}) - \frac{k}{2}\ln(n)$ or equivalently,

$$\sup_{\theta\in\mathcal{M}\cap\Theta} \sum_{i=1}^{n} \theta T(X_i) - Z(\theta) - \frac{k}{2}\ln(n), \tag{C.7}$$

Note that for simplicity of notation and without loss of generality, we set $h(X) = 1$. Now consider $T_n = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$, the sample average of the sufficient statistics. We can then express (C.7) as

$$n \sup_{\theta\in\mathcal{M}\cap\Theta} \theta T_n - Z(\theta) - \frac{k}{2}\ln(n). \tag{C.8}$$

Define the quantities $S_{n,i}$ and U_n as,

$$\begin{aligned} S_{n,i} &\equiv \sup_{\theta_i\in\mathcal{M}_i\cap\Theta} \theta_i T_n - Z(\theta_i) = \hat{\theta}_{n,i} T_n - Z(\hat{\theta}_{n,i}), \\ U_n &\equiv \theta_0 T_n - Z(\theta_0), \end{aligned}$$

where $\hat{\theta}_{n,i}$ is the MLE. We now show that $S_{n,i} - U_n$ and by extension each term in (C.6) is $O_p(1/n)$. Since θ_0 lies in both model spaces under scenario (**),

$$S_{n,i} - U_n = (\hat{\theta}_{n,i} - \theta_0)T_n - Z(\hat{\theta}_{n,i}) + Z(\theta_0) \geq 0. \tag{C.9}$$

Considering the Taylor expansion of Z about θ_0, we have that $Z(\hat{\theta}_{n,i}) - Z(\theta_0) = (\hat{\theta}_{n,i} - \theta_0)\nabla Z(\theta_0) + O_p(1/n)$, where the $O_p(1/n)$ term comes from the efficiency of MLE [156]. Plugging this into (C.9) we get,

$$S_{n,i} - U_n = (T_n - \nabla Z(\theta_0))(\hat{\theta}_{n,i} - \theta_0) + O_p(1/n). \tag{C.10}$$

By the Central Limit Theorem, $T_n - \nabla Z(\theta_0)$ is $O_p(1/\sqrt{n})$ and by the efficiency of MLE, $\hat{\theta}_{n,i} - \theta_0$ is also $O_p(1/\sqrt{n})$. Thus, $S_{n,i} - U_n$ is $O_p(1/n)$, and we have our result.

In order to extend this result to (C.5), we once again invoke the result from [155] that

$$\mathcal{PL}_n(\theta) \geq d\mathcal{L}_n(\theta) + \sum_{i=1}^{d} H_i(\widetilde{P}_n) \tag{C.11}$$

where $H_i(\widetilde{P}_n)$ is the Shannon entropy of the empirical distribution. We see that as long $d \ll n$ (which in our setting we assume to be true), (C.6) being $O_p(1/n)$ implies that (C.5) is as well.

□

Lemma 2 *Let $\mathcal{G}$ and $\mathcal{G}'$ be graphs which differ by a single edge between V_i and V_j. For conditional MRFs in the exponential family, the local score difference between $\mathcal{G}$ and $\mathcal{G}'$ is given by: $\sum_{V_k \in \text{loc}(V_i,V_j;\mathcal{G}) \cap B_{\text{loc}}} \{s_{V_k}\Big(X;\mathcal{G}\Big) - s_{V_k}\Big(X;\mathcal{G}'\Big)\}$, where $s_{V_k}(.)$ denotes the component of the score for V_k.*

Proof. A conditional MRF corresponding to $p(B \mid \text{pa}_{\mathcal{G}}(B))$ for a block B in a CG $\mathcal{G}$ in the (conditional) exponential family has a probability distribution of the general form:

$$p(B \mid \text{pa}_{\mathcal{G}}(B); \psi) = \quad \exp\left(\sum_{\{C \in \mathcal{C}((\mathcal{G}_{\text{bd}_{\mathcal{G}}(B)})^a) : C \not\subseteq \text{pa}_{\mathcal{G}}(B)\}} \psi_C T(C) - Z(\psi, \text{pa}_{\mathcal{G}}(B)) \right) \tag{C.12}$$

where

$$\Big\{\psi_C : C \in \mathcal{C}((\mathcal{G}_{\text{bd}_{\mathcal{G}}(B)})^a), C \not\subseteq \text{pa}_{\mathcal{G}}(B)\Big\}$$

is a set of canonical parameters associated with potential functions $\kappa_(C)$ in the CG factorization,

$$\Big\{T(C) : C \in \mathcal{C}((\mathcal{G}_{\text{bd}_{\mathcal{G}}(B)})^a), C \not\subseteq \text{pa}_{\mathcal{G}}(B)\Big\}$$

is a set of sufficient statistics for ψ_C, and $Z(\theta, \text{pa}_{\mathcal{G}}(B))$ is a normalizing function.

Assume V_k is in a clique C that contains the edge $V_i - V_j$ in $\mathcal{G}$, and let $\mathcal{G}^-$ be the edge subgraph of $\mathcal{G}$ with that edge removed. Then $p(V_k \mid \text{bd}_{\mathcal{G}}(V_k))$ will only

be a function of clique parameters ψ_S, where $S \subseteq \mathcal{C}((\mathcal{G}_{\text{bd}_\mathcal{G}(B)})^a) : C \not\subseteq \text{pa}_\mathcal{G}(B)$ and $V_k \in S$. All others terms in the factorization cancel by definition of conditioning. As a consequence, $p(V_k \mid \text{bd}_\mathcal{G}(V_k))$ will be a function of ψ_C.

However, after $V_i - V_j$ is removed, C will no longer be a clique in $\mathcal{G}^-$, by definition, but will instead decompose into two cliques, say C_1 and C_2. By following the above reasoning, $p(V_k \mid \text{bd}_{\mathcal{G}^-}(V_k))$ will be function of all clique parameters $\{\psi_S : S \subseteq \mathcal{C}((\mathcal{G}_{\text{bd}_\mathcal{G}(B)})^a), C \not\subseteq \text{pa}_\mathcal{G}(B), V_k \in S\}$, which will include ψ_{C_1} and ψ_{C_2}. Since the parameterization for $p(V_k \mid \text{bd}_{\mathcal{G}^-}(V_k))$ is thus different in models for $\mathcal{G}$ and $\mathcal{G}^-$, the contribution to the score associated with this term will also be different.

Now assume V_k is not in a clique that contains the edge $V_i - V_j$ in $\mathcal{G}$, and let $\mathcal{G}^-$ be the edge subgraph of $\mathcal{G}$ with that edge removed, as before. Then $p(V_k \mid \text{bd}_\mathcal{G}(V_k))$ will only be a function of clique parameters ψ_S, where S contains V_k, all others will cancel by definition of conditioning. Note that since no such S contains the edge $V_i - V_j$ in $\mathcal{G}$, the set of cliques S in $\mathcal{G}$ is the same as the set of cliques S in $\mathcal{G}^-$. Moreover, since $\mathcal{G}^-$ is an edge subgraph of $\mathcal{G}$, no new cliques are introduced. As a result, $p(V_k \mid \text{bd}_{\mathcal{G}^-}(V_k))$ will be parameterized by the same set of ψ_S in the model for $\mathcal{G}^-$ as it was in the model for $\mathcal{G}$.

Our conclusion then follows because by properties of the exponential family, the sufficient statistics for a clique parameter ψ_S are functions of only S. Since draws from $p(S)$ are fixed, the estimates for ψ_S will coincide if the data is evaluated under the model for $\mathcal{G}$, and the model for $\mathcal{G}^-$. Furthermore, the number of parameters in $p(V_k \mid \text{bd}_\mathcal{G}(V_k))$ and $p(V_k \mid \text{bd}_{\mathcal{G}^-}(V_k))$ is the same. This implies the score contribution for $p(V_k \mid \text{bd}_\mathcal{G}(V_k))$ in $\mathcal{G}$ will equal the score contribution of $p(V_k \mid \text{bd}_{\mathcal{G}^-}(V_k))$ in $\mathcal{G}^-$. The only terms remaining in the score difference between $\mathcal{G}$ and $\mathcal{G}'$ are then local scores for $V_k \in \text{loc}(V_i, V_j; \mathcal{G})$. This implies the conclusion.

□

Lemma 3 *If the generating distribution is Markov to a CG satisfying tier symmetry and the causal ordering assumption, then the search space of* Greedy Network Search *consists of graphs belonging to their own equivalence classes of size 1.*

Proof. Under the restrictions listed above, the only moves in our search procedure are edge deletions or additions of the form $L_i - L_j$, $A_i - A_j$, $Y_i - Y_j$, $L_i \to A_j$, $L_i \to Y_j$, $A_i \to Y_j$. Since chain graphs are maximal in the sense that every missing edge corresponds to conditional independence relations via c-separation [25], it immediately follows that two CGs $\mathcal{G}$ and $\mathcal{G}'$ that differ by an edge will imply different restrictions on the observed data distribution. Thus, in general, an edge deletion or addition in our search space gives rise to graphs that are not Markov equivalent and reside in their own equivalence classes of size 1.

□

Theorem 3 *If the generating distribution is in the exponential family (with compact natural parameter space Θ) and is Markov and faithful to a CG satisfying tier symmetry and causal ordering then* Greedy Network Search *is consistent.*

Proof. The algorithm begins with a complete conditional MRF that contains the true underlying distribution. We are guaranteed that the truth is contained in every state through the entirety of the algorithm by the following argument. Consider the first edge deletion performed by GNS to a conditional MRF that does not contain the true model. It follows from consistency of the PBIC that any such deletion would decrease the score. Choosing such an edge deletion would contradict the greediness of the algorithm.

Now assume the algorithm stops at a sub optimal conditional MR $\mathcal{G}$ that contains the truth but has more parameters than the true model $\mathcal{G}^*$. We know there exists a series of single edge deletions in $\mathcal{E}_{\mathcal{N}}$ that takes us from $\mathcal{G}$ to $\mathcal{G}^*$. By Lemma 3, each of these edge deletions yield graphs in separate equivalence classes. It follows then

from the consistency of the PBIC that each of these edge deletions strictly increases the score (each edge deletion yields a smaller model containing the truth) and thus, a local optimum found by greedily maximizing the PBIC corresponds to finding the global optimum $\mathcal{G}^*$. □

Corollary 3.1 *The* HETEROGENOUS *procedure is consistent.*

Proof. By consistency of GNS, each conditional MRF returned for L, A, and Y corresponds to the true model. The union of these will then produce the true CG on V. □

Corollary 3.2 *When the true network ties are homogenous,* HOMOGENOUS *network search is consistent.*

Proof. Each of the homogenous procedures described above can be decomposed into a series of single edge deletions that we have shown to be consistent. □

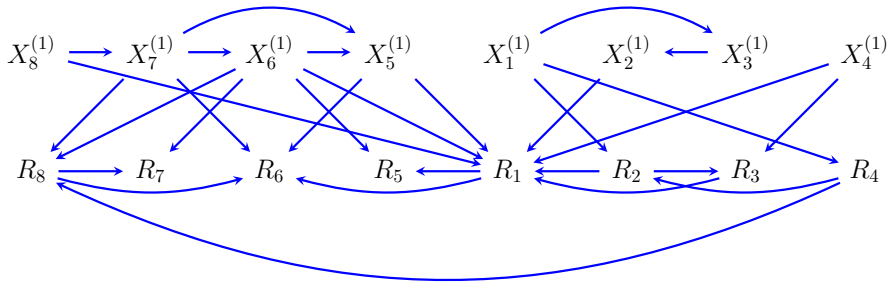
Appendix D

Supplement to Chapter 4

In this Appendix, we first provide an example that showcases the missing data identification algorithm we have developed in its full generality. We then provide longer technical proofs for results in the main chapter.

D.1 An Example to Illustrate the Algorithm

We walk the reader through identification of the target law for the missing data DAG shown in Figure D-1(a) in order to demonstrate the full generality of our missing ID algorithm. As a reminder, the target law is identified by Eq. 4.2 if we are able to identify $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))|_{R=1}$ for each $R_i \in R$. The identification of these conditional



(a) $\mathcal{G}$

Figure D-1. A complex missing data DAG used to illustrate the general techniques used in our algorithm

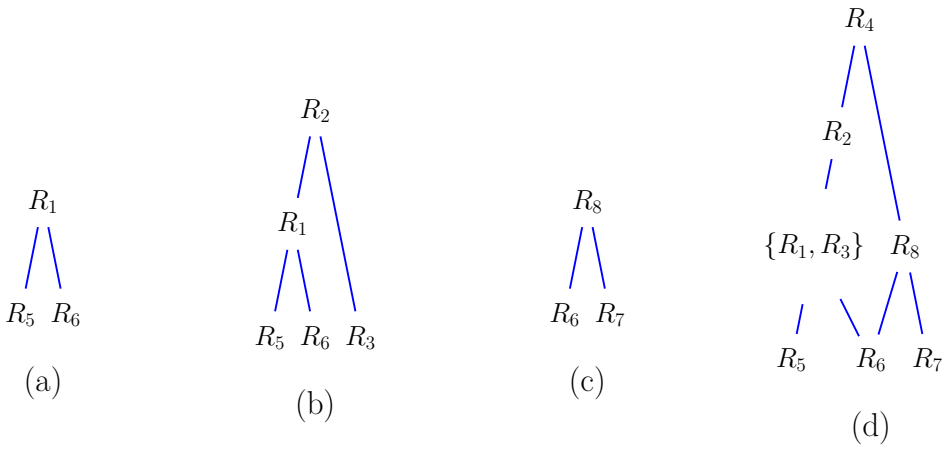


Figure D-2. (a-d) The corresponding fixing schedules of Rs.

densities are shown in equations (i) through (viii).

We start with $\{R_3, R_5, R_6, R_7\}$. The fixing schedules for these are empty and we obtain the following immediately from the original distribution.

(i) $p(R_3 \mid \mathrm{pa}(R_3)) = p(R_3 \mid R_2, X_4^{(1)}) = p(R_3 \mid R_2, X_4, 1_{R_4})$,

(ii) $p(R_5 \mid \mathrm{pa}(R_5)) = p(R_5 \mid R_1, X_6^{(1)}) = p(R_5 \mid R_1, X_6, 1_{R_6})$,

(iii) $p(R_6 \mid \mathrm{pa}(R_6)) = p(R_6 \mid R_1, R_8, X_5^{(1)}, X_7^{(1)}) = p(R_6 \mid R_1, R_8, X_5, X_7, 1_{R_5,R_7})$,

(iv) $p(R_7 \mid \mathrm{pa}(R_7)) = p(R_7 \mid R_8, X_6^{(1)}) = p(R_7 \mid R_8, X_6, 1_{R_6})$.

For R_1, we choose $Z = \{R_1, R_5, R_6\}$, and no equivalence relations. Thus, ${}^{Z}\!/_{\sim} = \{\{R_1\}, \{R_5\}, \{R_6\}\}$. The fixing schedule $\lhd$ is a partial order shown in Figure D-1(b) where R_5 and R_6 are incompatible, and $R_5 \prec R_1$, $R_6 \prec R_1$. Starting with the original $\mathcal{G}$ in Figure D-1(a), fixing R_5 and R_6 in parallel yields the following kernel.

$$q_{r_1}(X \setminus \{X_5, X_6\}, X_5^{(1)}, X_6^{(1)}, R \setminus \{R_5, R_6\} \mid 1_{R_5,R_6}) = \frac{p(X, R=1)}{p(R_5 \mid R_1, X_6^{(1)})\; p(R_6 \mid R_1, R_8, X_5^{(1)}, X_7^{(1)})|_{R=1}}, \tag{D.1}$$

where the propensity scores in the denominator are identified using (ii) and (iii). The CADMG corresponding to this fixing operation is shown in Figure D-3(a).

$$\begin{aligned}
\text{(v)}\quad p(R_1 \mid \text{pa}(R_1))|_{R=1} &= p(R_1 \mid R_2, R_3, X_2^{(1)}, X_4^{(1)}, X_5^{(1)}, X_6^{(1)})|_{R=1} \\
&= q_{r_1}(R_1 \mid R_2, R_3, X_2^{(1)}, X_4^{(1)}, X_5, X_6, 1_{R_5,R_6})|_{R=1} \\
&= q_{r_1}(R_1 \mid R_3, X_2, X_4^{(1)}, X_5, X_6, 1_{R_2,R_5,R_6})|_{R=1} \\
&= q_{r_1}(R_1 \mid R_3, X_2, X_4, X_5, X_6, 1_{R_2,R_4,R_5,R_6})|_{R=1} \quad \text{(by d-sep)}
\end{aligned}$$

where the last term can be obtained using kernel operations (conditioning+marginalization) on $q_{r_1}(\cdot \mid \cdot)$ defined in (D.1).

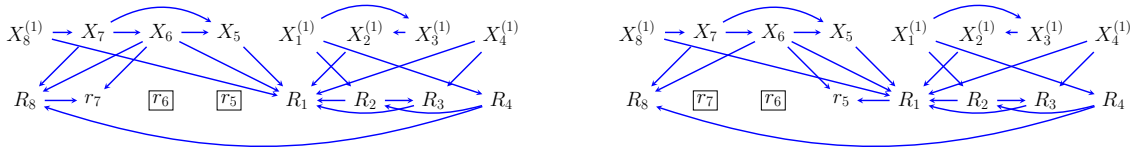
A similar procedure is applicable to R_8, where ${}^{Z}\!/\!_{\sim} = \{\{R_8\}, \{R_7\}, \{R_6\}\}$; Figure D-1(d). Starting with the original $\mathcal{G}$ in Figure D-1(a), fixing R_6 and R_7 in parallel yields the following kernel.

$$q_{r_8}(X \setminus \{X_6, X_7\}, X_6^{(1)}, X_7^{(1)}, R \setminus \{R_6, R_7\} \mid 1_{R_6,R_7}) = \frac{p(X, R=1)}{p(R_6 \mid R_1, R_8, X_5^{(1)}, X_7^{(1)})\; p(R_7 \mid R_8, X_6^{(1)})|_{R=1}}, \tag{D.2}$$

where the propensity scores in the denominator are identified using (iii) and (iv). The CADMG corresponding to this fixing operation is shown in Figure D-3(b).

$$\begin{aligned}
\text{(vi)}\quad p(R_8 \mid \text{pa}(R_8))|_{R=1} &= p(R_8 \mid R_4, X_6^{(1)}, X_7^{(1)})|_{R=1} \\
&= q_{r_8}(R_8 \mid R_4, X_6^{(1)}, X_7^{(1)}, 1_{R_6,R_7})|_{R=1} \\
&= q_{r_8}(R_8 \mid R_4, X_6, X_7, 1_{R_6,R_7})|_{R=1}
\end{aligned}$$

where the last term can be obtained using kernel operations (conditioning+marginalization) on $q_{r_8}(\cdot \mid \cdot)$ defined in (D.2).



(a) $\phi_{\{R_5,R_6\}}(\mathcal{G})$ (b) $\phi_{\{R_6,R_7\}}(\mathcal{G})$

Figure D-3. (a) Graph corresponding to the kernel obtained in (D.1) (b) Graph corresponding to the kernel obtained in (D.2).

For R_2, we choose $Z = \{R_1, R_2, R_3, R_5, R_6\}$, and no equivalence relations. Thus, ${}^{Z}/_{\sim} = \{\{R_1\}, \{R_2\}, \{R_3\}, \{R_5\}, \{R_6\}\}$. The fixing schedule $\lhd$ is a partial order where R_3, R_5, R_6 are incompatible and $R_5, R_6 \prec R_1 \prec R_2$ and $R_3 \prec R_2$ as shown in Figure D-1(c). In addition, the portion of the fixing schedule involving R_1, R_5, and R_6 is executed in a latent projection ADMG where we treat $X_2^{(1)}$ as being hidden as shown in Figure D-4(a), while the portion of the fixing schedule involving R_3 is executed in the original graph, Figure D-1(a).

$$\text{(vii)} \quad p(R_2 \mid R_4, X_1^{(1)}) = q_{r_2}(R_2 \mid R_4, X_1^{(1)}, 1_{R_1,R_3}), \tag{D.3}$$

where q_{r_2} corresponds to the kernel obtained by following the partial order of fixing R_3 and R_1, separately. That is,

$$q_{r_2}(\cdot \mid 1_{R_1,R_3}) = \frac{p(X, R=1)}{q^1_{r_2}(R_1 \mid R_2, R_3, X_2, X_5, X_6, X_3^{(1)}, X_8^{(1)}, 1_{R_5,R_6})\; p(R_3 \mid R_2, X_4^{(1)})}. \tag{D.4}$$

The propensity score for R_3 is obtained from (i) and $q^1_{r_2}$ is the kernel obtained by fixing R_5 and R_6 in parallel in a graph where $X_2^{(1)}$ is treated as hidden, as shown in Figures D-4(a) and (b). That is,

$$q^1_{r_2}(X \setminus \{X_5, X_6\}, X_5^{(1)}, X_6^{(1)}, R \setminus \{R_5, R_6\} \mid 1_{R_5,R_6}) = \frac{p(X, R=1)}{p(R_5 \mid R_1, X_6^{(1)})\; p(R_6 \mid R_1, R_8, X_5^{(1)}, X_7^{(1)})|_{R=1}}.$$

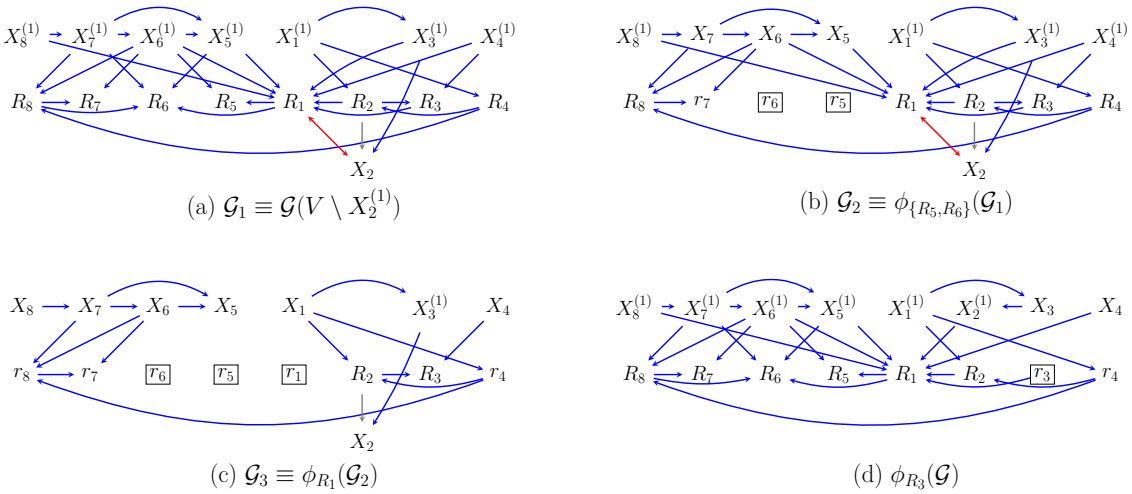


Figure D-4. Execution of the fixing schedule to obtain the propensity score for R_1 (a) Latent projection ADMG obtained by projecting out $X_2^{(1)}$ (b) Fixing R_5 and R_6 in $\mathcal{G}_1$ (c) Fixing R_1 in $\mathcal{G}_2$ (d) Fixing R_3 in the original graph.

The propensity scores in the denominator above are identified using (ii) and (iii). For clarity, the CADMGs corresponding to fixing R_1 and R_3 are illustrated in Figures D-4(c) and (d).

Finally, for R_4, we choose $Z = \{R\}$ and equivalence relation $R_1 \sim R_3$. Thus, ${}^{Z}/_{\sim} = \{\{R_1, R_3\}, \{R_2\}, \{R_4\}, \{R_5\}, \{R_6\}, \{R_7\}, \{R_8\}\}$. The fixing schedule $\lhd$ is a partial order where $R_5, R_6 \prec \{R_1, R_3\} \prec R_2 \prec R_4$ and $R_6, R_7 \prec R_8 \prec R_4$ as shown in Figure D-1(e). In addition, the portion of the fixing schedule involving R_5, R_6, $\{R_1, R_3\}$, and R_2 is executed in a latent projection ADMG where we treat $X_2^{(1)}$ and $X_4^{(1)}$ as hidden variables, shown in Figure D-5(b), while the portion of the fixing schedule involving R_6, R_7, and R_8 is executed in the original graph, Figure D-1(a).

$$\text{(viii)} \quad p(R_4 \mid X_1^{(1)}) = q_{r_4}(R_4 \mid X_1^{(1)}, 1_{R_2,R_8}), \tag{D.5}$$

where q_{r_4} corresponds to the kernel obtained by following the partial order of fixing

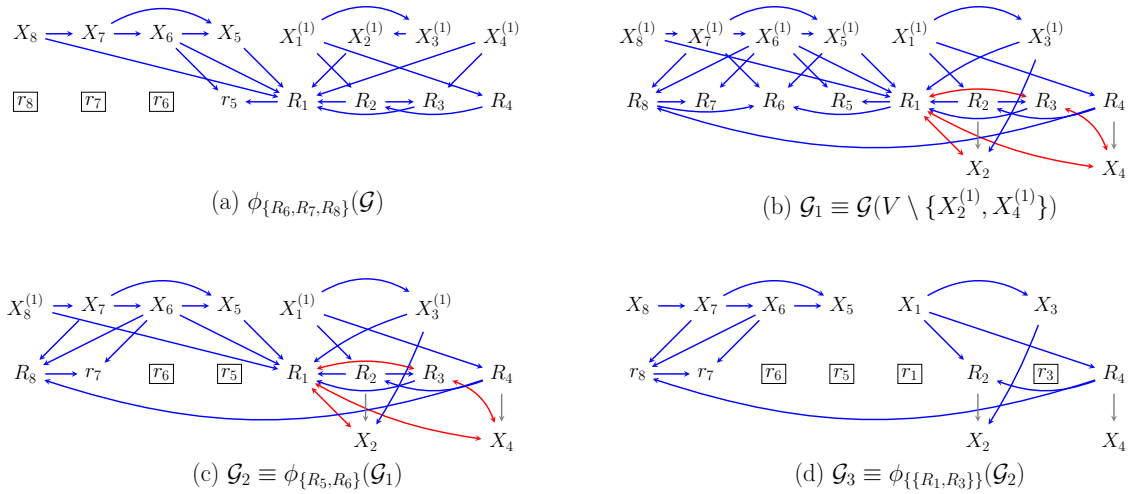


Figure D-5. Execution of the fixing schedule to obtain the propensity score for R_4 (a) CADMG obtained by following the schedule to get the propensity score for R_8 (b) Latent projection ADMG obtained by projecting out $X_2^{(1)}$ and $X_4^{(1)}$ (c) Fixing R_5 and R_6 in $\mathcal{G}_1$ (d) Fixing R_1 in $\mathcal{G}_2$.

R_2 and R_8, separately. That is,

$$q_{r_4}(\cdot \mid 1_{R_2,R_8}) = \frac{p(X, R=1)}{q^1_{r_4}(R_2 \mid R_4, X_2)\; q^2_{r_4}(R_8 \mid R_4, X_6, X_7)}. \tag{D.6}$$

$q^1_{r_4}$ is the kernel obtained by fixing the set $\{R1, R3\}$ in graph $\mathcal{G}_2$ shown in Figure D-5(c). That is,

$$\begin{aligned} q^1_{r_4}(\cdot \mid 1_{R_1,R_3,R_5,R_6}) &= \frac{q^3_{r_4}(\cdot \mid 1_{R_5,R_6})}{q^3_{r_4}(R_1, R_3 \mid R_2, R_4, X_2, X_3^{(1)}, X_4)} \\ &= \frac{q^3_{r_4}(. \mid 1_{R_5,R_6})}{q^3_{r_4}(R_1 \mid R_2, R_4, X_2, X_3, X_4, 1_{R_3})\; q^3_{r_4}(R_3 \mid R_2, R_4, X_2, X_4)} \end{aligned}$$

$q^3_{r_4}$ is the kernel obtained by fixing R_5 and R_6 in parallel in the graph $\mathcal{G}_1$ shown in Figure D-5(b). That is,

$$q^3_{r_4}(\cdot \mid 1_{R_5,R_6}) = \frac{p(X, R=1)}{p(R_5 \mid R_1, X_6^{(1)})\; p(R_6 \mid R_1, R_8, X_5^{(1)}, X_7^{(1)})|_{R=1}}.$$

The propensity scores in the denominator above are identified using (ii) and (iii).

Finally, $q^2_{r_4}$ is the kernel obtained by fixing R_6 and R_7 in parallel in the original graph $\mathcal{G}$, shown in Figure D-1(a). That is,

$$q^2_{r_4}(\cdot \mid 1_{R_6,R_7}) = \frac{p(X, R=1)}{p(R_6 \mid R_1, R_8, X_5^{(1)}, X_7^{(1)})\; p(R_7 \mid R_8, X_6^{(1)})|_{R=1}}.$$

The propensity scores in the denominator above are identified using (iii) and (iv). For clarity, the CADMG corresponding to fixing R_8 is illustrated in Figures D-5(a).

D.2 Proofs

Theorem 4 *Given a DAG $\mathcal{G}(X^{(1)}, R, O, X)$, the distribution $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))|_{\mathrm{pa}_{\mathcal{G}}(R_i)\cap R=1}$ is identifiable from $p(R, O, X)$ if there exists*

(i) $Z \subseteq X^{(1)} \cup R \cup O$,

(ii) an equivalence relation $\sim$ *on* Z *such that* $\{R_i\} \in {}^{Z}\!/\!_{\sim}$,

(iii) a set of elements $X^{(1)}_{\tilde{Z}}$ *such that* $X^{(1)}_{\{\lhd \tilde{Z}\}} \subseteq X^{(1)}_{\tilde{Z}} \subseteq X^{(1)}$ *for each* $\tilde{Z} \in {}^{Z}\!/\!_{\sim}$,

(iv) $X^{(1)} \cap \mathrm{pa}_{\mathcal{G}}(R_i) \subseteq (Z \setminus \{R_i\}) \cup X^{(1)}_{\{R_i\}}$,

(v) and a valid fixing schedule $\lhd$ *for* ${}^{Z}\!/\!_{\sim}$ *in* $\mathcal{G}$ *such that for each* $\tilde{Z} \in {}^{Z}\!/\!_{\sim}$, $\tilde{Z} \lhd \{R_i\}$.

Moreover, $p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i))|_{\mathrm{pa}_{\mathcal{G}}(R_i)\cap R=1}$ is equal to $q_{\{R_i\}}$, defined inductively as the denominator of Eq. 4.4 for $\{R_i\}$, $\phi_{\lhd\{R_i\}}(\mathcal{G})$ and $\phi_{\lhd\{R_i\}}(p; \mathcal{G})$, and evaluated at $\mathrm{pa}_{\mathcal{G}}(R_i) \cap R = 1$.

Proof. We first outline the essential argument made in this proof. We will reformulate the process of fixing according to a partial order in a missing data problem as a problem of ordinary fixing based on a total order in a causal inference problem where, previously missing variables are in fact observed. If we are able to show this, we can

invoke results from [22], that guarantee that we obtain the desired conditional for each R_i.

Consider $\widetilde{Z} \in {}^{Z}/_{\sim}$, and define $X^{(1)}_{\{\trianglelefteq\widetilde{Z}\}} \equiv \bigcup_{Z\in\{\trianglelefteq\widetilde{Z}\}} X^{(1)}_Z$, and $R_{\{\trianglelefteq\widetilde{Z}\}} \equiv \{R_k \mid X^{(1)}_k \in X^{(1)}_{\{\trianglelefteq\widetilde{Z}\}}\}$, and similarly for $X^{(1)}_{\{\triangleleft\widetilde{Z}\}}$ and $R_{\{\triangleleft\widetilde{Z}\}}$.

We first note that any total ordering $\prec$ on $\{\triangleleft\widetilde{Z}\}$ consistent with $\triangleleft$ yields a valid fixing sequence on sets in $\{\triangleleft\widetilde{Z}\}$ in $\mathcal{G}(R, O, X^{(1)}, X))$, where $X^{(1)}_{\{\triangleleft\widetilde{Z}\}}, R, O, X$ are observed. The total ordering $\prec$ can be refined to operate on single variables where each set $\widetilde{Z}$ is fixed as singletons following a topological total order where variables with no children in $\widetilde{Z}$ would be fixed first. Such a total order is also valid and follows from the validity of $\triangleleft$ and the fact that at each step of the fixing operation in the total order, the Markov blanket of each Z contains only observed variables; hence no selection bias is induced on any singleton variables $\{\succ \widetilde{Z}\}$.

We now show, by induction on the structure of the partial order $\triangleleft$, that for a particular $\widetilde{Z} \in {}^{Z}/_{\sim}$, $q_{\widetilde{Z}}$ is equal to

$$\prod_{Z\in\mathcal{Z}} \prod_{Z\in Z} \tilde{q}(Z \mid \mathrm{mb}_{\tilde{\mathcal{G}}}(Z; \mathrm{an}_{\tilde{\mathcal{G}}}(D_Z) \cap \prec_{\tilde{\mathcal{G}}} \{Z\}, R_Z)|_{(R\cap Z)\cup R_Z=1}, \tag{D.7}$$

obtained from a kernel

$$\tilde{q} \equiv \phi_{\{\triangleleft\widetilde{Z}\}}(p(R, O, X^{(1)}_{\{\triangleleft\widetilde{Z}\}}, X); \mathcal{G}),$$

and CADMG

$$\tilde{\mathcal{G}} \equiv \phi_{\{\triangleleft\widetilde{Z}\}}(\mathcal{G}(R, O, X^{(1)}_{\{\triangleleft\widetilde{Z}\}}, X)),$$

where $X^{(1)}_{\{\triangleleft\widetilde{Z}\}}, R, O, X$ are observed.

For any $\triangleleft$-smallest $\widetilde{Z}$, $\widetilde{Z}$ is independent of $R_{\{\trianglelefteq\widetilde{Z}\}}$ given its Markov blanket; therefore treating $X^{(1)}_{\{\trianglelefteq\widetilde{Z}\}}$ as observed results in the same kernel as $q_{\widetilde{Z}}$.

We now show that the above is also true for any $\widetilde{Z} \in {}^{Z}/_{\sim}$. Assume the inductive

hypothesis holds for all $\widetilde{Y} \in \{\lhd \widetilde{Z}\}$. Since $\lhd$ is valid, we obtain $q_{\widetilde{Z}}$ by applying

$$\phi_{\trianglelefteq_{\widetilde{Z}}}(q; \mathcal{G}) \equiv \\ \phi_{\widetilde{Z}}\Big(\frac{p(O, X, R \setminus R_{\{\lhd \widetilde{Z}\}}, R_{\{\lhd \widetilde{Z}\}} = 1)}{\prod_{\widetilde{Y} \in \{\lhd \widetilde{Z}\}} q_{\widetilde{Y}}}; \phi_{\lhd_{\widetilde{Z}}}(\mathcal{G})\Big), \tag{D.8}$$

where $q_{\widetilde{Y}}$ are defined by the inductive hypothesis, and $\phi_{\widetilde{Z}}$ is defined via

$$\frac{q(V \setminus ((X^{(1)} \setminus X^{(1)}_{\{\lhd \widetilde{Z}\}}) \cup R_Z), R_Z = 1 \mid W)}{\prod\limits_{Z \in \mathcal{Z}} \prod\limits_{Z \in Z} q(Z \mid \mathrm{mb}_{\tilde{\mathcal{G}}}(Z; \mathrm{an}_{\tilde{\mathcal{G}}}(D_Z) \cap \prec_{\tilde{\mathcal{G}}} (Z)), R_Z)|_{(R \cap Z) \cup R_Z = 1}}, \tag{D.9}$$

where

$$q(V \setminus (X^{(1)} \setminus X^{(1)}_{\{\lhd \widetilde{Z}\}}) \mid W) \equiv \frac{p(O, X, R \setminus R_{\{\lhd \widetilde{Z}\}}, R_{\{\lhd \widetilde{Z}\}} = 1)}{\prod_{\widetilde{Y} \in \{\lhd \widetilde{Z}\}} q_{\widetilde{Y}}}.$$

Consider the equivalent functional in the model where we observe $X^{(1)}_{\{\lhd \widetilde{Z}\}}$

$$\frac{q^\dagger(V \setminus ((X^{(1)} \setminus X^{(1)}_{\{\lhd \widetilde{Z}\}}) \cup R_Z), R_Z = 1 \mid W)}{\prod\limits_{Z \in \mathcal{Z}} \prod\limits_{Z \in Z} q^\dagger(Z \mid \mathrm{mb}_{\tilde{\mathcal{G}}}(Z; \mathrm{an}_{\tilde{\mathcal{G}}}(D_Z) \cap \prec_{\tilde{\mathcal{G}}} (Z)), R_Z)|_{(R \cap Z) \cup R_Z = 1}}, \tag{D.10}$$

where

$$q^\dagger(V \setminus (X^{(1)} \setminus X^{(1)}_{\{\lhd \widetilde{Z}\}}) \mid W) \equiv \\ \frac{p(O, X, X^{(1)}_{\{\lhd \widetilde{Z}\}}, R \setminus \widetilde{R}_{\{\lhd \widetilde{Z}\}}, \widetilde{R}_{\{\lhd \widetilde{Z}\}} = 1)}{\prod_{\widetilde{Y} \in \{\lhd \widetilde{Z}\}} q_{\widetilde{Y}}},$$

and $\widetilde{R}_{\{\lhd \widetilde{Z}\}}$ is defined as the subset of $R_{\{\lhd \widetilde{Z}\}}$ that is fixed in $\{\lhd \widetilde{Z}\}$.

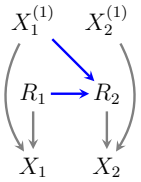
The only difference between Eq. D.9 and Eq. D.10 for the purposes of the denominator is the variables in $R_{\{\lhd \widetilde{Z}\}} \setminus \widetilde{R}_{\{\lhd \widetilde{Z}\}}$. But the denominator is independent of these variables, by assumption. Thus, it follows that fixing on a valid partial order with missing data and fixing on a total order consistent with this partial order, as in causal inference, yield equivalent kernels.

The conclusion follows by Lemma 55 in [22].

□

Theorem 6 *In a DAG $\mathcal{G}(X^{(1)}, R, O, X)$, if there exists $R_i, R_j \in R$ such that $\{R_j, X_j^{(1)}\} \in \text{pa}_{\mathcal{G}}(R_i)$, then $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))|_{R_j=0}$ is not identified. Hence, the full law $p(X^{(1)}, R)$ is not identified.*

Proof. Proven by providing two full laws that agree on the observed data law as in the tables below.



R_1	$p(R_1)$
0	a
1	$1-a$

$X_1^{(1)}$	$p(X_1^{(1)})$
0	b
1	$1-b$

$X_2^{(1)}$	$p(X_2^{(1)})$
0	c
1	$1-c$

R_2	R_1	$X_1^{(1)}$	$p(R_2 \mid R_1, X_1^{(1)})$
0	0	0	d
1	0	0	$1-d$
0	1	0	e
1	1	0	$1-e$
0	0	1	f
1	0	1	$1-f$
0	1	1	g
1	1	1	$1-g$

R_1	R_2	$X_1^{(1)}$	$X_2^{(1)}$	p(Full Law)	X_1	X_2	p(Observed Law)
0	0	0	0	$abcd$	?	?	$a\Big[db + f(1-b))\Big]$
		1	0	$a(1-b)cf$			
		0	1	$ab(1-c)d$			
		1	1	$a(1-b)(1-c)f$			
1	0	0	0	$(1-a)ebc$	0	?	$(1-a)eb$
		1	0	$(1-a)g(1-b)c$			
		0	1	$(1-a)eb(1-c)$	1		$(1-a)g(1-b)$
		1	1	$(1-a)g(1-b)(1-c)$			
0	1	0	0	$abc(1-d)$	?	0	$ac\Big[1-\Big(db + f(1-b)\Big)\Big]$
		1	0	$a(1-b)c(1-f)$			
		0	1	$ab(1-c)(1-d)$		1	$a(1-c)\Big[1-\Big(db + f(1-b)\Big)\Big]$
		1	1	$a(1-b)(1-c)(1-f)$			
1	1	0	0	$(1-a)(1-e)bc$	0	0	$(1-a)(1-e)bc$
		1	0	$(1-a)(1-g)(1-b)c$	1	0	$(1-a)(1-g)(1-b)c$
		0	1	$(1-a)(1-e)b(1-c)$	0	0	$(1-a)(1-e)b(1-c)$
		1	1	$(1-a)(1-g)(1-b)(1-c)$	1	1	$(1-a)(1-g)(1-b)(1-c)$

Table D-I. Table for proof of non-identifiability of the full law in missing data DAG models full with colluder structures.

Any pair of $\{d, f\}$ would lead to different full laws. However, as long as $db + f(1-b)$ stays constant, the observe law would agree across all different full laws (which include infinitely many models). This is a general characterization of non-identifiable models with two binary random variables.

□

Theorem 7 *Consider a DAG $\mathcal{G}(X^{(1)}, R, O, X)$ such that for every $R_i \in R$, $\{R_j \mid X_j^{(1)} \in \text{pa}_{\mathcal{G}}(R_i)\} \cap \text{an}_{\mathcal{G}}(R_i) = \emptyset$. Then for every $R_i \in R$, a fixing schedule $\lhd$ for $\{\{R_j\} \mid R_j \in \mathcal{G}_{R \cap \text{de}_{\mathcal{G}}(R_i)}\}$ given by the partial order induced by the ancestrality relation on $\mathcal{G}_{R \cap \text{de}_{\mathcal{G}}(R_i)}$ is valid in $\mathcal{G}(X^{(1)}, R, O, X)$, by taking each $X^{(1)}_{\widetilde{Z}} = \bigcup_{Z \in \{\trianglelefteq \widetilde{Z}\}} X^{(1)}_Z$, for every $\widetilde{Z} \in \{\trianglelefteq \{R_i\}\}$. Thus the target law is identified.*

Proof. In order to prove that the target law is identified, we demonstrate that conditions (i-v) in Theorem 4 are satisfied for each R_i.

Conditions (i) and (ii) are trivially satisfied as we choose to fix $Z \subseteq R$, and we choose no equivalence relation, thus ${}^{Z}\!/_{\sim}$ consists of singleton sets of Rs. Condition (iii) is also trivial as each $X^{(1)}_{\widetilde{Z}}$ is a union of the corresponding sets $X^{(1)}_{\widetilde{Y}}$, for $\widetilde{Y}$ earlier in the partial order. In the proposed order we never fix elements in $X^{(1)}$, and propose to keep elements in $X^{(1)} \cap \text{pa}_{\mathcal{G}}(R_j)$ for every $R_j \in Z$. In particular, this also includes R_i, satisfying condition (iv).

Finally, we show that the proposed schedule $\lhd$ is valid by showing that each $\widetilde{Z} \in {}^{Z}\!/_{\sim}$ is fixable. There are 3 conditions for an element $\widetilde{Z}$ to be fixable as mentioned in the missing data setting. We go through each of these conditions and demonstrate each $\widetilde{Z}$ in ${}^{Z}\!/_{\sim}$ is a valid fixing in $\phi_{\lhd_{\widetilde{Z}}}(\mathcal{G})$ where $\lhd$ is the proposed fixing schedule above.

In the proposed schedule each $\widetilde{Z}$ is a singleton $R_j \in {}^{Z}\!/_{\sim}$ that we are trying to fix in a graph $\phi_{\lhd_{R_j}}(\mathcal{G})$. Since $X^{(1)}_{R_j} = X^{(1)}$, $\phi_{\lhd_{R_j}}(\mathcal{G})$ is a CDAG. Thus, $\mathcal{D}(\phi_{\lhd_{R_j}}(\mathcal{G}))$ is just sets of singleton vertices. In particular, $D_{R_j} = \{R_j\}$. Further, by definition of the schedule, it must be that $\text{de}_{\phi_{\lhd_{R_j}}(\mathcal{G})}(R_j) = \{R_j\}$. This satisfies condition (i).

For condition (ii), we note that $S \subseteq \mathrm{nd}_{\phi_{\lhd R_j}(\mathcal{G})}(R_j)$ else, S contains some $R_k \in \mathrm{de}_{\mathcal{G}}(R_j)$ which should have been fixed prior to R_j by the proposed partial order. Thus, it follows that $S \cap \{R_j\} = \emptyset$.

Finally, following the partial order, and under the assumption stated in the theorem, $R_{\{R_j\}} \subseteq \{\lhd R_j\}$. We have also proved that $S \subseteq \mathrm{nd}_{\phi_{\lhd R_j}(\mathcal{G})}(R_j)$. Therefore, $R_j \perp\!\!\!\perp (S \cup R_{\{R_j\}}) \setminus \mathrm{mb}_{\phi_{\lhd R_j}(\mathcal{G})}(R_j) \mid \mathrm{mb}_{\phi_{\lhd R_j}(\mathcal{G})}(R_j)$.

Since each $\widetilde{Z}$ is fixable, the proposed partial order $\lhd$ for each R_i is valid. Therefore, all five conditions in Theorem 4 are satisfied concluding the target law is ID.

□

Appendix E

Supplement to Chapter 5

This Appendix contains longer proofs for results in the main chapter.

E.1 Proofs

Theorem 8 *Algorithm 9 is sound and complete for deciding the nonparametric saturation status of the model implied by an ADMG $\mathcal{G}(V)$ by determining the absence of equality constraints.*

Proof. The construction of Algorithm 9 is closely related to the *maximal arid projection* described in [39]. MArGs were proposed as a more general analogue of maximal ancestral graphs typically used in the context of causal discovery and where the absence of edges may only imply ordinary conditional independence constraints [21, 157, 158]. The absence of an edge between two vertices in a MArG rule out the presence of certain paths between them known as *dense inducing paths* resulting in the so called *maximality* property. We now show that Algorithm 9 declares an input ADMG to be NPS if it is equivalent to a MArG with no missing edges, and not NPS if it is equivalent to one with at least one missing edge. We then use the maximality property to derive the form of the implied equality constraint.

Given *any* ADMG $\mathcal{G}(V)$, there exists a nested Markov equivalent MArG $\mathcal{G}^a(V)$ that implies the same set of conditional and generalized independence constraints and

can be obtained via the maximal arid projection as follows [39]. Recall the definition of $\mathrm{pa}^{\mathrm{d}}_{\mathcal{G}}(S)$ as $\bigcup_{S_i \in S} \mathrm{pa}_{\mathcal{G}}(S_i)$.

- For $V_i \in V$, the edge $V_i \to V_j$ exists in $\mathcal{G}^a(V)$ if $V_i \in \mathrm{pa}^{\mathrm{d}}_{\mathcal{G}}(\langle V_j \rangle_{\mathcal{G}})$.
- For $V_i, V_j \in V$, the edge $V_i \leftrightarrow V_j$ exists in $\mathcal{G}^a(V)$ if neither $V_i \in \mathrm{pa}^{\mathrm{d}}_{\mathcal{G}}(\langle V_j \rangle_{\mathcal{G}})$ nor $V_j \in \mathrm{pa}^{\mathrm{d}}_{\mathcal{G}}(\langle V_i \rangle_{\mathcal{G}})$ but $\langle V_i, V_j \rangle_{\mathcal{G}}$ is a bidirected connected set.

Soundness

We prove soundness by showing that Algorithm 9 declares the model to be nonparametric saturated (NPS) only when the input ADMG $\mathcal{G}(V)$ is nested Markov equivalent to a MArG $\mathcal{G}^a(V)$ where all vertices in V are pairwise adjacent. If all vertices are pairwise adjacent, this immediately rules out the possibility of equality constraints.

For each pair of vertices (V_i, V_j) either of the first two conditions in line 4 of Algorithm 9 evaluates to True precisely when the MArG projection operator adds a directed edge between V_i and V_j. Further, the third condition in line 5 evaluates to True when the MArG projection adds a bidirected edge between V_i and V_j. Thus, as long as the MArG projection operator continues to require the presence of an edge between each pair (V_i, V_j) the negation of all the conditions makes it so that line 6 of the algorithm is never executed. Once all pairs have been checked, the model is declared to be nonparametrically saturated in line 7.

Completeness

We prove completeness by showing that Algorithm 9 declares the model to be not NPS only when the input ADMG is nested Markov equivalent to a MArG $\mathcal{G}^a(V)$ that has a pair of vertices (V_i, V_j) that are not connected by a directed or bidirected edge. We then explicate the equality constraint implied by this missing edge.

It is clear from previous arguments in the proof of soundness that the negation of the conditions in line 2 evaluates to True only when the MArG projection operator

fails to add an edge between a pair of vertices (V_i, V_j). As soon as this occurs, it is also clear that the resulting MArG $\mathcal{G}^a(V)$ obtained by executing the full projection will still have a missing edge between V_i and V_j. We now show that this missing edge corresponds to an equality constraint involving V_i and V_j.

A path $(V_i, X_1, \ldots, X_p, V_j)$ is said to be inducing if every non-endpoint node X_i is both a collider on this path as well as an ancestor of at least one of the vertices V_i or V_j. Such paths are important because it has been shown that the absence of an inducing path between two non-adjacent vertices V_i and V_j implies the existence of a set Z such that V_i and V_j are m-separated given Z [23]. That is, when V_i and V_j are not connected by an inducing path in $\mathcal{G}^a(V)$, there exists a set Z such that $V_i \perp\!\!\!\perp V_j \mid Z$ and this is an equality constraint that rules out nonparametric saturation of $\mathcal{G}$.

Consider the case when there does exist an inducing path between V_i and V_j. By definition of the maximality property of MArGs, there exists a valid fixing sequence for some $S \subset V$ such that this path is no longer inducing in $\phi_S(\mathcal{G}^a(V))$. We now discuss all possible cases of inducing paths between V_i and V_j and the corresponding equality constraint obtained after fixing some subset of vertices in $\mathcal{G}^a(V)$. Note it is sufficient for us to focus on the subgraph $\mathcal{G}^{ant} \equiv \mathcal{G}^a_{\text{an}_\mathcal{G}(V_i \cup V_j)}$ [29]. This subgraph also preserves the inducing path as all ancestors of V_i and V_j are included.

Consider the case when the inducing path consists of only bidirected edges i.e., $V_i \leftrightarrow X_1 \leftrightarrow, \ldots, \leftrightarrow X_p \leftrightarrow V_j$. Note that none of the vertices X_i in this path are fixable in $\mathcal{G}^{ant}$ as by definition of an inducing path, X_i is either an ancestor of V_i or of V_j. Thus, $\text{dis}_{\mathcal{G}^{ant}}(X_i) \cap \text{de}_{\mathcal{G}^{ant}}(V_i) \neq \{X_i\}$. However, the construction of the MArG $\mathcal{G}^a$ guarantees that V_i and V_j are not bidirected connected in $\langle V_i, V_j \rangle_\mathcal{G}$ and consequently not bidirected connected in the ancestral subgraph $\langle V_i, V_j \rangle_{\mathcal{G}^{ant}}$. In order for this to be true, at least one vertex X_i must become fixable after a sequence of fixing on some vertices S that are descendants of X and ancestors of V_i and V_j (excluding

X, V_i, and V_j). In the graph $\phi_S(\mathcal{G}^{ant})$, X_i is fixable precisely because it is no longer an ancestor of either V_i or V_j. Therefore, the path $V_i \leftrightarrow X_1 \leftrightarrow, \ldots, \leftrightarrow X_p \leftrightarrow V_j$ is no longer inducing in $\phi_S(\mathcal{G}^{ant})$. Thus, there exists a set Z such that V_i and V_j can be m-separated in $\phi_S(\mathcal{G}^{ant})$, and the corresponding equality constraint is $V_i \perp\!\!\!\perp V_j \mid Z$ in $\phi_S(p(\text{an}_{\mathcal{G}}(V_i \cup V_j)); \mathcal{G}^{ant})$.

Consider the case when the inducing path is of the form $V_i \rightarrow X_1 \leftrightarrow, \ldots, \leftrightarrow X_p \leftrightarrow V_j$. As the graph $\mathcal{G}^{ant}$ is an ancestral subgraph of $\mathcal{G}^a(V)$, we can apply the district factorization to $\mathcal{G}^{ant}$. Define $X \equiv \{X_1, \ldots, X_p\}$, and let V^{ant} denote all vertices in $\mathcal{G}^{ant}$ and D_X denote the district in $\mathcal{G}^{ant}$ that contains $\{X, V_j\}$ or $\{V_i, X, V_j\}$ if V_i is also in the same district. Then, $q_{D_X}(D_X \mid \text{pa}_{\mathcal{G}^{ant}}(D_X))$ is identified and district factorizes with respect to the CADMG $\phi_{V \setminus D_X}(\mathcal{G}^{ant})$ [22]. In such a CADMG, the only possible directed paths from any vertex X_i to V_i or V_j are through vertices in D_X as these are the only random vertices that remain in $\phi_{V \setminus D_X}(\mathcal{G}^{ant})$. First consider the case when V_i is not in D_X. Then V_i is fixed in $\phi_{V \setminus D_X}(\mathcal{G}^{ant})$ and has no ancestors so the path $V_i \rightarrow X_1 \leftrightarrow, \ldots, X_p \leftrightarrow V_j$ remains inducing only if all vertices $X_i \in X$ have a directed path to V_j. If such a path exists for every $X_i \in X$ then no X_i is fixable in $\phi_{V \setminus D_X}(\mathcal{G}^{ant})$. Further, no $D_i \in D_X \setminus V_j$ is fixable either as they are all within the same district and have directed paths to V_j. Thus, the reachable closure of V_j in $\mathcal{G}^{ant}$ and as a consequence in $\mathcal{G}^a$, contains X_1. Since V_i is a parent of X_1, the MArG projection should have yielded an edge $V_i \rightarrow V_j$ which is a contradiction. Similarly, if V_i is in D_X, and all $X_i \in X$ have directed paths to either V_i or V_j, then $\langle V_i, V_j \rangle_{\mathcal{G}^a}$ would remain a bidirected connected set and the MArG projection would have yielded an edge $V_i \leftrightarrow V_j$ which is also a contradiction. Therefore, in either case, there exists at least one $X_i \in X$ such that X_i is neither an ancestor of V_i nor V_j in $\phi_{V \setminus D_X}(\mathcal{G}^{ant})$. Thus, the path $V_i \rightarrow X_1 \leftrightarrow, \ldots, \leftrightarrow X_p \leftrightarrow V_j$ is no longer inducing and we have the equality constraint, given some set $Z \subset V^{ant}$ that $V_i \perp\!\!\!\perp V_j \mid Z$ in $\phi_{V \setminus D_X}(\mathcal{G}^{ant})$.

Thus, it must be true that the input ADMG $\mathcal{G}(V)$ implies at least one equality

constraint, specifically between the variables V_i and V_j, as it is nested Markov equivalent to a MArG with a missing edge between these two vertices, and we have provided a form for the implied equality constraint. Hence, whenever line 6 is executed in Algorithm 9, the model is truly *not* nonparametrically saturated. □

Theorem 9 *Consider a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where an edge between two vertices is absent only if $V_i \notin \mathrm{mb}_{\mathcal{G}}(V_j)$ and $V_j \notin \mathrm{mb}_{\mathcal{G}}(V_i)$. Then, given any valid topological order on V, all equality constraints in $p(V)$ are implied by the set of restrictions: $V_i \perp\!\!\!\perp \{\prec V_i\} \setminus \mathrm{mp}_{\mathcal{G}}(V_i) \mid \mathrm{mp}_{\mathcal{G}}(V_i)$, $\forall V_i \in V$.*

Proof. The proof relies on the fact that the constraint finding algorithm provided in [30] finds a list of equality constraints that is sufficient to define the nested Markov model of an ADMG (this was shown by [22]). Here we show that the only non-trivial equality constraints found by applying this algorithm to an arbitrary mb-shielded ADMG $\mathcal{G}$ are of the form $V_i \perp\!\!\!\perp \{\prec V_i\} \mid \mathrm{mp}_{\mathcal{G}}(V_i)$ thus implying that all equality constraints in the nested Markov model of such an ADMG are implied by ordinary conditional independences of that form.

Given a valid topological order on the vertices, the constraint finding algorithm in [30] iterates over each vertex V_i in the order and attempts to find constraints between V_i and $\{\prec V_i\}$. In substep (A1) of the algorithm (see [30] for details), it identifies constraints of the form $V_i \perp\!\!\!\perp \{\prec V_i\} \mid \mathrm{mp}_{\mathcal{G}}(V_i)$. Substep (A2) and recursive applications of it, attempts to find constraints between V_i and subsets of $\{\prec V_i\} \cap \mathrm{mb}(V_i)$. Since the mb-shielded criterion enforces that if some vertex V_j is in the Markov blanket of V_i, they must be adjacent, there can be no non-trivial equality constraint between V_i and any subset of $\{\prec V_i\} \cap \mathrm{mb}(V_i)$. Hence, this step never returns any new constraints (compared to those found in (A1)) for an mb-shielded ADMG.

Thus, running the algorithm on an mb-shielded ADMG returns a list of constraints

consisting of only ordinary conditional independence constraints (those that are found by substep (A1)), and specifically ones that are of the form $V_i \perp\!\!\!\perp \{\prec V_i\} \mid \text{mp}_{\mathcal{G}}(V_i)$.

□

Lemma 5 *Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, $\psi(t) = \psi(t)_{primal} \equiv \mathbb{E}[\beta(t)_{primal}]$ where*

$$\begin{aligned}\beta(t)_{primal} &\equiv \frac{\mathbb{I}(T=t)}{q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T))} \times Y \\ &= \mathbb{I}(T=t) \times \frac{\sum_T \prod_{V_i \in D_T \cap \{\succeq T\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}{\prod_{V_i \in D_T \cap \{\succeq T\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))} \times Y.\end{aligned}$$

Proof. Our goal is to demonstrate that the primal IPW formulation is equivalent to the identifying functional of the target parameter $\psi(t)$ shown in Eq. 5.6 and restated below.

$$\psi(t) = \sum_{V \setminus T} \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t} \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y.$$

The primal IPW formulation for the target $\psi(t)$ is,

$$\mathbb{E}[\beta_{\text{primal}}(t)] \equiv \mathbb{E}\left[\frac{\mathbb{I}(T=t)}{q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T))} \times Y\right]$$

where $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T)) = \prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))$, and

$$\begin{aligned}q_{D_T}(T \mid \text{mb}_{\mathcal{G}}(T)) &= q_{D_T}(T \mid D_T \cup \text{pa}_{\mathcal{G}}(D_T) \setminus T) = \frac{q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))}{q_{D_T}(D_T \setminus T \mid \text{pa}_{\mathcal{G}}(D_T))} \\ &= \frac{q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))}{\sum_T q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))} = \frac{\prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}{\sum_T \prod_{V_i \in D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))} \\ &= \frac{\prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}{\sum_T \prod_{V_i \in \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))}.\end{aligned}$$

The last equality holds because the conditional densities of $V_i \in \mathbb{C}$, does not depend on T, and they cancel out from the numerator and denominator. Therefore, product in the ratio is over the variables in $D_T \cap \{\succeq T\}$ which we have denoted by $\mathbb{L}$. Therefore,

$$
\begin{aligned}
\mathbb{E}[\beta_{\text{primal}}(t)] &= \mathbb{E}\Big[\mathbb{I}(T=t) \times \frac{\sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}{\prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \times Y \Big] \\
&= \sum_V \prod_{V_i \in V} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \mathbb{I}(T=t) \times \frac{\sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}{\prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \times Y \\
&= \sum_V \mathbb{I}(T=t) \times \prod_{V_i \in V \setminus \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times \frac{\sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))}{\prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))} \times Y \\
&= \sum_V \mathbb{I}(T=t) \times \prod_{V_i \in V \setminus \mathbb{L}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_T \prod_{D_i \in \mathbb{L}} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y.
\end{aligned}
$$

In the second equality, we evaluated the outer expectation with respect to the joint $p(V)$. In the third equality, we partitioned the joint into factors for the set $\mathbb{L}$ and factors for $V \setminus \mathbb{L}$. In the fourth equality, we canceled out the the factors involved in the denominator of the primal IPW with the corresponding terms in the joint.

We can then move the conditional factors of pre-treatment variables in the district of T past the summation over T as these factors are not functions of T. Finally, we evaluate the indicator function, concluding the proof. That is,

$$
\begin{aligned}
\psi_{\text{primal}} &= \sum_V \mathbb{I}(T=t) \times \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_T \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y \\
&= \sum_{V \setminus T} \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))\Big|_{T=t} \times \sum_T \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y = \psi(t)
\end{aligned}
$$

□

Lemma 6 *Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, $\psi(t) = \psi(t)_{dual} \equiv \mathbb{E}[\beta(t)_{dual}]$ where*

$$
\beta(t)_{dual} = \frac{\prod_{V_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \mid_{T=t}}{\prod_{V_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i))} \times Y.
$$

Proof. The proof strategy is similar to the one used for the primal IPW. The dual IPW formulation for the target $\psi(t)$ is,

$$
\begin{aligned}
\mathbb{E}[\beta_{\text{dual}}(t)] &= \mathbb{E}\Big[\frac{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} \; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t}}{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} \; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y\Big] \\
&= \sum_{V} \prod_{V_i \in V} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \frac{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} \; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t}}{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} \; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y \\
&= \sum_{V} \prod_{V_i \in V \setminus \text{mp}_{\mathcal{G}}^{-1}(T)} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \\
&\qquad \times \prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \times \frac{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} \; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t}}{\prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} \; p(M_i \mid \text{mp}_{\mathcal{G}}(M_i))} \times Y \\
&= \sum_{V} \prod_{V_i \in V \setminus \text{mp}_{\mathcal{G}}^{-1}(T)} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \times Y \\
&= \sum_{V \setminus T} \prod_{V_i \in V \setminus \{\text{mp}_{\mathcal{G}}^{-1}(T) \cup D_T\}} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \prod_{M_i \in \text{mp}_{\mathcal{G}}^{-1}(T)} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \mid_{T=t} \\
&\qquad \times \sum_{T} \prod_{D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y.
\end{aligned}
$$

In the above derivation, we first evaluated the outer expectation with respect to the joint $p(V)$. We then partitioned the joint into factors corresponding to $\text{mp}_{\mathcal{G}}^{-1}(T)$ and $V \setminus \text{mp}_{\mathcal{G}}^{-1}(T)$. The factors involved in the denominator of the dual IPW then canceled out with the corresponding terms in the joint. The last equality holds because by the definition of the inverse Markov pillow, $\text{mp}_{\mathcal{G}}^{-1}(T)$ contains all variables not in the district of T such that T is a member of its Markov pillow. In the above expression, factors corresponding to the inverse Markov pillow of T are evaluated at $T = t$. Consequently, the only factors above that are still functions of T are the ones corresponding to the district of T. This allows us to push the summation over T.

Finally, since the summation over T will prevent factors within the district of T from being evaluated at $T = t$, we can simply apply the evaluation to the entire functional and merge the sets not involved in the district of T above. That is,

$$
\psi_{\text{dual}} = \sum_{V \setminus T} \prod_{V_i \in V \setminus D_T} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)) \times \sum_{T} \prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i)) \times Y \Big|_{T=t} = \psi(t).
$$

□

Theorem 10 *Given a distribution $p(V)$ that district factorizes with respect to an ADMG $\mathcal{G}(V)$ where T is primal fixable, the IPW estimators ψ_{primal} and ψ_{dual} proposed in Lemmas 5 and 6 respectively, use variationally independent components of the observed distribution $p(V)$.*

Proof. Consider the topological factorization of the observed distribution $p(V)$ for the ADMG as shown in Eq. 1.6.

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{mp}_{\mathcal{G}}(V_i)).$$

Note by definition, the inverse Markov pillow of T does not contain elements in the district of T, i.e., $\text{mp}_{\mathcal{G}}^{-1}(T) \cap D_T = \emptyset$. Thus, we can partition V into three disjoint sets as follows:

$$\mathbb{L} = D_T \cap \{\succeq T\}, \qquad \mathbb{M}^* = \text{mp}_{\mathcal{G}}^{-1}(T), \qquad \mathbb{R} = V \setminus (\mathbb{L} \cup \mathbb{M}^*)$$

The set $\mathbb{L}$ is the same as what we defined earlier at the beginning of this proof section. $\mathbb{M}^*$ is a subset of $\mathbb{M}$, and the remainder terms $\mathbb{R} = \mathbb{C} \cup \{\mathbb{M} \setminus \mathbb{M}^*\}$. The topological factorization of the observed joint can then be restated as,

$$p(V) = \prod_{R_i \in \mathbb{R}} p(R_i \mid \text{mp}_{\mathcal{G}}(R_i)) \prod_{M_i \in \mathbb{M}^*} p(M_i \mid \text{mp}_{\mathcal{G}}(M_i)) \prod_{L_i \in \mathbb{L}} p(L_i \mid \text{mp}_{\mathcal{G}}(L_i)).$$

It is then clear from the above factorization that the components of the primal IPW estimator which sit in $\mathbb{L}$, and the components of the dual IPW estimator which sit in $\mathbb{M}$, form congenial and variationally independent pieces of the joint distribution $p(V)$. □

Theorem 12 *Given a distribution $p(V)$ that district factorizes with respect to an mb-shielded ADMG $\mathcal{G}(V)$ where T is primal fixable, the efficient influence function*

for the target parameter $\psi(t)$ is given as follows,

$$
\begin{aligned}
U_{\psi_t}^{\mathit{eff}} = &\sum_{M_i \in \mathbb{M}} \mathbb{E}[\beta_{\mathit{primal}} \mid M_i, \mathrm{mp}_{\mathcal{G}}(M_i)] - \mathbb{E}[\beta_{\mathit{primal}} \mid \mathrm{mp}_{\mathcal{G}}(M_i)] \\
&+ \sum_{L_i \in \mathbb{L}} \mathbb{E}[\beta_{\mathit{dual}} \mid L_i, \mathrm{mp}_{\mathcal{G}}(L_i)] - \mathbb{E}[\beta_{\mathit{dual}} \mid \mathrm{mp}_{\mathcal{G}}(L_i)] \\
&+ \sum_{C_i \in \mathbb{C}} \mathbb{E}[\beta_{\mathit{primal/dual}} \mid C_i, \mathrm{mp}_{\mathcal{G}}(C_i)] - \mathbb{E}[\beta_{\mathit{primal/dual}} \mid \mathrm{mp}_{\mathcal{G}}(C_i)]
\end{aligned}
$$

where $\mathbb{C}, \mathbb{L}, \mathbb{M}$ are defined in display (5.9), and β_{primal} and β_{dual} are obtained as in Lemmas 5 and 6 respectively. $\beta_{\mathit{primal/dual}}$ means that we can either use β_{primal} or β_{dual} for the terms in $\mathbb{C}$.

Proof. Consider the reformulated IF in Theorem 11. In order to get the efficient IF, we project the reformulated IF onto the tangent space Λ^* given by Lemma 4. We first note that we can rewrite the term $\sum_{\mathbb{C}} \mathbb{E}[\beta_{\text{primal/dual}} \mid \mathbb{C}] - \psi(t)$ in the reformulated IF as $\sum_{C_i \in \mathbb{C}} \mathbb{E}[\beta_{\text{primal/dual}} \mid \{\preceq C_i\}] - \mathbb{E}[\beta_{\text{primal/dual}} \mid \{\prec C_i\}]$, where $\beta_{\text{primal/dual}}$ means that we can use either β_{primal} or β_{dual} for the $\mathbb{C}$ term. We have,

$$
\begin{aligned}
\pi[U_{\psi_t}^{\text{reform}} \mid \Lambda^*] = &\sum_{M_i \in \mathbb{M}} \pi\Big[\mathbb{E}[\beta_{\text{primal}} \mid \{\preceq M_i\}] - \mathbb{E}[\beta_{\text{primal}} \mid \{\prec M_i\}] \Big| \Lambda^*\Big] \\
&+ \sum_{L_i \in \mathbb{L}} \pi\Big[\mathbb{E}[\beta_{\text{dual}} \mid \{\preceq L_i\}] - \mathbb{E}[\beta_{\text{dual}} \mid \{\prec L_i\}] \;\Big|\; \Lambda^*\Big] \\
&+ \sum_{C_i \in \mathbb{C}} \pi\Big[\mathbb{E}[\beta_{\text{primal/dual}} \mid \preceq C_i] - \mathbb{E}[\beta_{\text{primal/dual}} \mid \prec C_i] \;\Big|\; \Lambda^*\Big].
\end{aligned}
$$

Let β be either β_{primal} or β_{dual} or $\beta_{\text{primal/dual}}$. Note that $\Big\{\mathbb{E}\Big[\beta \mid \{\preceq V_i\}\Big] - \mathbb{E}\Big[\beta_{\text{primal}} \mid \{\prec V_i\}\Big]\Big\}$ lives in Λ_{V_i}, and $\Lambda_{V_i} \perp\!\!\!\perp \Lambda^* \setminus \Lambda^*_{V_i}$. Therefore, their projection onto $\Lambda^* \setminus \Lambda^*_{V_i}$ is zero. We have,

$$
\begin{aligned}
&\pi\Big[\mathbb{E}[\beta \mid \{\preceq V_i\}] - \mathbb{E}[\beta \mid \{\prec V_i\}] \Big| \Lambda^*_{V_i}\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}[\beta \mid \{\preceq V_i\}] - \mathbb{E}[\beta \mid \{\prec V_i\}] \;\Big|\; V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\Big] - \mathbb{E}\Big[\mathbb{E}[\beta \mid \{\preceq V_i\}] - \mathbb{E}[\beta \mid \{\prec V_i\}] \;\Big|\; \mathrm{mp}_{\mathcal{G}}(V_i)\Big] \\
&\quad = \mathbb{E}[\beta \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)] - \mathbb{E}\Big[\mathbb{E}[\beta \mid \prec V_i] \;\Big|\; V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\Big] - \mathbb{E}[\beta \mid \mathrm{mp}_{\mathcal{G}}(V_i)] + \mathbb{E}[\beta \mid \mathrm{mp}_{\mathcal{G}}(V_i)] \\
&\quad = \mathbb{E}[\beta \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)] - \mathbb{E}\Big[\mathbb{E}[\beta \mid \prec V_i] \;\Big|\; V_i, \mathrm{mp}_{\mathcal{G}}(V_i)\Big] \\
&\quad = \mathbb{E}[\beta \mid V_i, \mathrm{mp}_{\mathcal{G}}(V_i)] - \mathbb{E}[\beta \mid \mathrm{mp}_{\mathcal{G}}(V_i)].
\end{aligned}
$$

Therefore, the efficient IF is as follows.

$$\begin{aligned}\pi[U_{\psi_t}^{\text{reform}} \mid \Lambda^*] = &\sum_{M_i \in \mathbb{M}} \mathbb{E}[\,\beta_{\text{primal}} \mid M_i, \text{mp}_{\mathcal{G}}(M_i)] - \mathbb{E}[\,\beta_{\text{primal}} \mid \text{mp}_{\mathcal{G}}(M_i)] \\ &+ \sum_{L_i \in \mathbb{L}} \mathbb{E}[\,\beta_{\text{dual}} \mid L_i, \text{mp}_{\mathcal{G}}(L_i)] - \mathbb{E}[\,\beta_{\text{dual}} \mid \text{mp}_{\mathcal{G}}(L_i)] \\ &+ \sum_{C_i \in \mathbb{C}} \mathbb{E}[\beta_{\text{primal/dual}} \mid C_i, \text{mp}_{\mathcal{G}}(C_i)] - \mathbb{E}[\beta_{\text{primal/dual}} \mid \text{mp}_{\mathcal{G}}(C_i)].\end{aligned}$$

□

Theorem 13 *Let $p(V)$ and $\mathcal{G}(V)$ be the observed marginal distribution and ADMG induced by a hidden variable causal model associated with DAG $\mathcal{G}(V \cup H)$. Then if $\psi(t)$ is identifiable in the model, $\psi(t) = \psi(t)_{nested}$. If $\psi(t)$ is not identifiable in the model, Algorithm 10 returns 'fail'.*

Proof. Soundness of the algorithm implies that when our algorithm succeeds, the subsequent identifying functional for $\psi(t)$ is correct. Completeness implies, that when the algorithm fails, the target parameter $\psi(t)$ is not identifiable within the model.

Soundness

We first prove soundness of the algorithm. That is, when Algorithm 10 does not fail, $\psi(t)$ is indeed equal to $\psi(t)_{\text{primal}}$ and $\psi(t)_{\text{dual}}$. The algorithm does not fail when all districts $D \in \mathcal{D}^*$ are intrinsic in $\mathcal{G}$. Note that $\mathcal{D}^*$ is a subset of the districts in $\mathcal{G}_{Y^*}$. However, by construction of $\mathcal{D}^*$, the remaining districts in $\mathcal{G}_{Y^*}$ are those that do not have any overlap with D_T. We now show that such districts are always intrinsic in $\mathcal{G}$.

Consider a district $D \in \mathcal{D}(\mathcal{G}_{Y^*})$ such that $D \cap D_T = \emptyset$. The district D forms a subset of a larger district in $\mathcal{G}$, say $D' \in \mathcal{D}(\mathcal{G})$. Due to results in [26], we know that D' is always intrinsic. If $D = D'$ then the result immediately follows. Otherwise, In the CADMG $\phi_{V \setminus D'}(\mathcal{G})$, there exists at least one vertex D_i in D' not in Y^*, that has no children. This is because all directed paths from D_i to vertices in Y^* must go through T and since T is not in D', all incoming edges to T have been deleted.

The only other way D_i may not be childless is if there existed a cycle in $\mathcal{G}$, which is a contradiction. Thus, such a vertex D_i is always fixable and furthermore, fixing it corresponds to the marginalization operation $\sum_{D_i} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D'))$ [22]. Once D_i is fixed, another vertex D_j that is in D' but not in Y^* becomes childless. Applying this argument inductively, we see that all $D_i \in D'$ such that $D_i \notin Y^*$ are fixable through marginalization under a reverse topological order. Hence for districts D in $\mathcal{G}_{Y^*}$ that do not overlap with D_T, the set $D = D' \setminus \{D_i \in D' \mid D_i \notin Y^*\}$ is always intrinsic. Thus, Algorithm 10 succeeds when all districts in $\mathcal{G}_{Y^*}$ are intrinsic.

We now show that under this condition, $\psi(t)_{\text{nested}} \equiv \mathbb{E}_{p^\dagger}\left[\frac{\mathbb{I}(T=t)}{p(T|\text{mp}_{\mathcal{G}}(T))} \times Y\right] = \psi(t)$. By definition, we have

$$\psi(t)_{\text{nested}} = \sum_V p(V) \times \prod_{D^* \in \mathcal{D}^*} \frac{q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*))}{\prod_{D_i^* \in D^*} p(D_i^* \mid \text{mp}_{\mathcal{G}}(D_i^*))} \times \frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_{\mathcal{G}}(T))} \times Y.$$

The districts of $\mathcal{G}$ can be partitioned into three sets. $\mathcal{D}_T$ is the district in $\mathcal{G}$ that contains T (with all elements in $\mathcal{D}^*$, if any, subsets of D_T). $\mathcal{D}'$ is the set of districts in $\mathcal{G}$, excluding D_T, that overlap with Y^*. $\mathcal{D}^z$ is the set of districts in $\mathcal{G}$, excluding D_T, that do not overlap with Y^*. The observed distribution $p(V)$ then district factorizes as,

$$p(V) = \prod_{D^z \in \mathcal{D}^z} q_{D^z}(D^z \mid \text{pa}_{\mathcal{G}}(D^z)) \times \prod_{D' \in \mathcal{D}'} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) \times q_{D_T}(D_T | \text{pa}_{\mathcal{G}}(D_T)).$$

By results in [26], $q_{D_T}(D_T \mid \text{pa}_{\mathcal{G}}(D_T))$ is identified as $\prod_{D_i \in D_T} p(D_i \mid \text{mp}_{\mathcal{G}}(D_i))$ (for any topological ordering). Since every element in $\mathcal{D}^*$ is a subset of D_T, and since vertices in $D_T \setminus \bigcup_{D^* \in \mathcal{D}^*}$ precede vertices $D_T \cap \bigcup_{D^* \in \mathcal{D}^*} = D_T \cap Y^*$ in the ordering, we have

$$\begin{aligned}\psi(t)_{\text{nested}} = \sum_V \prod_{D^z \in \mathcal{D}^z} q_{D^z}(D^z \mid \text{pa}_{\mathcal{G}}(D^z)) \times \prod_{D' \in \mathcal{D}'} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) \times \prod_{D^* \in \mathcal{D}^*} q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*)) \\ \times \sum_{D_T \cap Y^*} q_{D_T}(D_T | \text{pa}_{\mathcal{G}}(D_T)) \times \frac{\mathbb{I}(T=t)}{p(T \mid \text{mp}_{\mathcal{G}}(T))} \times Y.\end{aligned}$$

Since T is the last element in the ordering in $D_T \setminus Y^*$, we further have:

$$\psi(t)_{\text{nested}} = \sum_{Y^*} \sum_{V \setminus Y^*} \prod_{D^z \in \mathcal{D}^z} q_{D^z}(D^z \mid \text{pa}_{\mathcal{G}}(D^z)) \times \prod_{D' \in \mathcal{D}'} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) \times \prod_{D^* \in \mathcal{D}^*} q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*))$$
$$\times \sum_{(D_T \cap Y^*) \cup \{T\}} q_{D_T}(D_T | \text{pa}_{\mathcal{G}}(D_T)) \times \mathbb{I}(T = t) \times Y.$$

Consider applying marginalization of elements in $V \setminus Y^*$ to $\psi(t)_{\text{nested}}$ above in the reverse topological ordering on $V \setminus Y^*$. Districts in $\mathcal{G}$ partition V and so, by definition of $\mathcal{D}^*, \mathcal{D}'$ and D_T, elements in $\mathcal{D}^z \cup \{D' \setminus Y^* : D' \in \mathcal{D}'\} \cup \{D_T \setminus (Y^* \cup \{T\})\}$ partition $V \setminus Y^*$. This partition, and the fact that marginalizations are processed in reverse topological order, means that at every stage, the variable to be summed occurs in precisely one place in the expression. This implies that the result of the overall summation of $V \setminus Y^*$ yields:

$$\psi(t)_{\text{nested}} = \sum_{Y^*} \prod_{D' \in \mathcal{D}'} \sum_{D' \setminus Y^*} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) \times \prod_{D^* \in \mathcal{D}^*} q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*)) \times \mathbb{I}(T = t) \times Y$$

By definition, $q_{D^*}(D^* \mid \text{pa}_{\mathcal{G}}(D^*)) \equiv \phi_{V \setminus D^*}(p(V); \mathcal{G}(V))$. Since every D' in $\mathcal{D}'$ is a top level district in $\mathcal{G}$, there exists a valid fixing sequence on $V \setminus D'$. Further, in the CADMG $\phi_{V \setminus D'}(\mathcal{G}(V))$, any element in $D' \setminus Y^*$ cannot be an ancestor of an element in $D' \cap Y^*$ (if a directed path not through T existed from an element V_i in D' to an element in $D' \cap Y^*$, then V_i must itself be in $D' \cap Y^*$, while a directed path from V_i to $D' \cap Y^*$ through T disappears in $\phi_{V \setminus D'}(\mathcal{G}(V))$ since T is outside D'. Consequently fixing elements $D' \setminus Y^*$ in reverse topological order in $\phi_{V \setminus D'}(\mathcal{G}(V))$ and $\phi_{V \setminus D'}(p(V), \mathcal{G}(V))$ is equivalent to marginalizing those variables. As a result, for every $D' \in \mathcal{D}'$, $\sum_{D' \setminus Y^*} q_{D'}(D' \mid \text{pa}_{\mathcal{G}}(D')) = \phi_{V \setminus (D' \cap Y^*)}(p(V); \mathcal{G}(V))$. Our conclusion follows:

$$\psi(t)_{\text{nested}} = \sum_{Y^*} \prod_{D \in \mathcal{D}(\mathcal{G}_{Y^*})} \phi_{V \setminus D}(p(V); \mathcal{G}) \times Y \Big|_{T=t} = \psi(t).$$

Completeness

Follows trivially as we have shown the failure condition of Algorithm 10 to be equivalent to the failure condition of the identification algorithm in [22] which is known to be sound and complete. □

Bibliography

[1] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[2] Nancy Cartwright. What are randomised controlled trials good for? *Philosophical Studies*, 147(1):59, 2010.

[3] James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[4] Peter Spirtes, Clark N. Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas S. Richardson. *Causation, Prediction, and Search*. MIT press, 2000.

[5] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

[6] Miguel A. Hernán and James M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016.

[7] Shona Fielding, Peter M. Fayers, Alison McDonald, Gladys McPherson, Marion K. Campbell, et al. Simple imputation methods were inadequate for missing

not at random (MNAR) quality of life data. *Health and Quality of Life Outcomes*, 6(1):57, 2008.

[8] Elizabeth L. Ogburn, Oleg Sofrygin, Iván Diaz, and Mark J. van der Laan. Causal inference for social network data. *arXiv preprint arXiv:1705.08527*, 2017.

[9] Rohit Bhattacharya, Jaron J. R. Lee, Razieh Nabi, and Ilya Shpitser. *Ananke*: A Python package for causal inference with graphical models.

[10] Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

[11] Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2019.

[12] Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James M. Robins. Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2019.

[13] Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

[14] Sander Greenland, Judea Pearl, and James M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

[15] Thomas C. Williams, Cathrine C. Bach, Niels B. Matthiesen, Tine B. Henriksen, and Luigi Gagliardi. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatric Research*, 84(4):487–493, 2018.

[16] Paul Hünermund and Elias Bareinboim. Causal inference and data-fusion in econometrics. *arXiv preprint arXiv:1912.09104*, 2019.

[17] Daniel Malinsky, Ilya Shpitser, and Thomas S. Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088, 2019.

[18] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kauffman, 1988.

[19] Mathias Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009.

[20] Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In *Machine Intelligence and Pattern Recognition*, volume 9, pages 69–76. Elsevier, 1990.

[21] Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.

[22] Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.

[23] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, 1990.

[24] Robin J. Evans. Margins of discrete Bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, 2018.

[25] Steffen L. Lauritzen. *Graphical Models*. Oxford, U.K.: Clarendon, 1996.

[26] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pages 567–573. American Association for Artificial Intelligence, 2002.

[27] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 13–16, 2006.

[28] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

[29] Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[30] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 519–527, 2002.

[31] Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011.

[32] Eli Sherman and Ilya Shpitser. Identification and estimation of causal effects from dependent data. *Advances in Neural Information Processing Systems*, 2018.

[33] Eric J. Tchetgen Tchetgen, Isabel R. Fulcher, and Ilya Shpitser. Auto-g-computation of causal effects on a network. *Journal of the American Statistical Association*, pages 1–12, 2020.

[34] Elizabeth L. Ogburn, Ilya Shpitser, and Youjin Lee. Causal inference, social networks and chain graphs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1659–1676, 2020.

[35] Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002.

[36] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[37] Milan Studenỳ. On separation criterion and recovery algorithm for chain graphs. In *Proceedings of the 12th International Conference on Uncertainty in Artificial Intelligence*, pages 509–516, 1996.

[38] Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *Annals of Statistics*, 39(2):865–886, 2011.

[39] Ilya Shpitser, Robin J. Evans, and Thomas S. Richardson. Acyclic linear SEMs obey the nested Markov property. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence*, 2018.

[40] Carlos Brito and Judea Pearl. A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9(4):459–474, 2002.

[41] Juan M. Ogarrio, Peter L. Spirtes, and Joseph D. Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the 8th International Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

[42] Daniel Bernstein, Basil Saeed, Chandler Squires, and Caroline Uhler. Ordering-based causal structure learning in the presence of latent variables. In *Inter-*

national Conference on Artificial Intelligence and Statistics, pages 4098–4108. PMLR, 2020.

[43] Robin J. Evans and Thomas S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, pages 1452–1482, 2014.

[44] Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483, 2018.

[45] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7154–7163, 2019.

[46] Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, and Yixin Chen. D-VAE: A variational autoencoder for directed acyclic graphs. In *Advances in Neural Information Processing Systems*, pages 1588–1600, 2019.

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

[48] Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–580, 1921.

[49] Sewall Wright. The method of path coefficients. *Annals of Mathematical Statistics*, 5(3):161–215, 1934.

[50] Brendan D. McKay, Frédérique E. Oggier, Gordon F. Royle, N. J. A. Sloane, Ian M. Wanless, and Herbert S. Wilf. Acyclic digraphs and eigenvalues of (0, 1)-matrices. *Journal of Integer Sequences*, 7(2):3, 2004.

[51] Bohao Yao and Robin J. Evans. Constraints in Gaussian graphical models. *arXiv preprint arXiv:1911.12754*, 2019.

[52] AmirEmad Ghassami, Alan Yang, Negar Kiyavash, and Kun Zhang. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[53] Steven Kay Butler. *Eigenvalues and structures of graphs*. PhD thesis, UC San Diego, 2008.

[54] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[55] Dominique M.A. Haughton. On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16(1):342–355, 1988.

[56] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[57] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[58] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[59] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[60] Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

[61] Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! varsortability in additive noise models. *arXiv preprint arXiv:2102.13647*, 2021.

[62] Jian Huang, Yuling Jiao, Yanyan Liu, and Xiliang Lu. A constructive approach to L_0 penalized regression. *The Journal of Machine Learning Research*, 19(1):403–439, 2018.

[63] Xiaogang Su, Chalani S. Wijayasinghe, Juanjuan Fan, and Ying Zhang. Sparse estimation of Cox proportional hazards models via approximated information criteria. *Biometrics*, 72(3):751–759, 2016.

[64] Razieh Nabi and Xiaogang Su. coxphMIC: An R package for sparse estimation of Cox proportional hazards models via approximated information criteria. *R Journal*, 9(1), 2017.

[65] Dimitri P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[66] Mathias Drton, Michael Eichler, and Thomas S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(10), 2009.

[67] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

[68] Christopher Nowzohour, Marloes H. Maathuis, Robin J. Evans, Peter Bühlmann, et al. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11(2):5342–5374, 2017.

[69] Chi Zhang, Bryant Chen, and Judea Pearl. A simultaneous discover-identify approach to causal inference in linear models. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 10318–10325, 2020.

[70] Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39, 2014.

[71] Robin J Evans. Model selection and local geometry. *arXiv preprint arXiv:1801.08364*, 2018.

[72] Joseph Ramsey and Bryan Andrews. FASK with interventional knowledge recovers edges from the Sachs model. *arXiv preprint arXiv:1805.03108*, 2018.

[73] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, pages 294–321, 2012.

[74] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 340–349, 2014.

[75] Y. Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-Gaussian data. *arXiv preprint arXiv:2007.11131*, 2020.

[76] Ricardo Silva, Richard Scheines, Clark Glymour, and Peter L. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(Feb):191–246, 2006.

[77] Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. In *Advances in Neural Information Processing Systems*, pages 12883–12892. Curran Associates, Inc., 2019.

[78] Limor Gultchin, Matt Kusner, Varun Kanade, and Ricardo Silva. Differentiable causal backdoor discovery. In *International Conference on Artificial Intelligence and Statistics*, pages 3970–3979. PMLR, 2020.

[79] Hao-Chih Lee, Matteo Danieletto, Riccardo Miotto, Sarah T. Cherng, and Joel T. Dudley. Scaling structural learning with NO-BEARS to infer causal transcriptome networks. In *Pacific Symposium on Biocomputing*, volume 25, pages 391–402, 2020.

[80] Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1949.

[81] Kevin Lewis, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1):68–72, 2012.

[82] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, pages 8788–8790, 2014.

[83] Joshua D. Angrist and Kevin Lang. Does school integration generate peer effects? evidence from Boston's METCO program. *American Economic Review*, 94(5):1613–1634, 2004.

[84] Guanglei Hong and Stephen W. Raudenbush. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101(475):901–910, 2006.

[85] Elizabeth L. Ogburn and Tyler J. VanderWeele. Causal diagrams for interference. *Statistical Science*, 29(4):559–578, 2014.

[86] Michael E. Sobel. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

[87] Paul R. Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007.

[88] Michael G. Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[89] Eric J. Tchetgen Tchetgen and Tyler J. VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

[90] Razieh Nabi, Joel Pfeiffer, Murat Ali Bayir, Denis Charles, and Emre Kıcıman. Causal inference in the presence of interference in sponsored search advertising. *arXiv preprint arXiv:2010.07458*, 2020.

[91] Forrest W. Crawford, Peter M. Aronow, Li Zeng, and Jianghong Li. Identification of homophily and preferential recruitment in respondent-driven sampling. *American Journal of Epidemiology*, 187(1):153–160, 2017.

[92] Yann Bramoullé, Andrea Galeotti, and Brian Rogers. *The Oxford Handbook of the Economics of Networks*. Oxford University Press, 2016.

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

[94] Shohei Shimizu. LiNGAM: non-Gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

[95] Marc Maier, Katerina Marazopoulou, David Arbour, and David Jensen. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 371–380. AUAI Press, 2013.

[96] Sanghack Lee and Vasant Honavar. On learning causal models from relational data. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3263–3270, 2016.

[97] Noa Novershtern, Aviv Regev, and Nir Friedman. Physical module networks: an integrative approach for reconstructing transcription regulation. *Bioinformatics*, 27(13):i177–i185, 2011.

[98] Sharon Stancliff, Bruce Agins, Josiah D. Rich, and Scott Burris. Syringe access for the prevention of blood borne infections among injection drug users. *BMC Public Health*, 3(1):37, 2003.

[99] Zongming Ma, Xianchao Xie, and Zhi Geng. Structural learning of chain graphs via decomposition. *Journal of Machine Learning Research*, 9(Dec):2847–2880, 2008.

[100] Jose Peña, Dag Sonntag, and Jens Nielsen. An inclusion optimal algorithm for chain graph structure learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 778–786, 2014.

[101] Mohammad Ali Javidian, Marco Valtorta, and Pooyan Jamshidi. Learning LWF chain graphs: A Markov blanket discovery approach. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, pages 1069–1078. PMLR, 2020.

[102] Mohammad Ali Javidian, Marco Valtorta, and Pooyan Jamshidi. Learning LWF chain graphs: an order independent algorithm. *arXiv preprint arXiv:2005.14037*, 2020.

[103] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological Statistics)*, 36(2):192–236, 1974.

[104] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014.

[105] Victor Chernozhukov, Chris Hansen, and Martin Spindler. High-dimensional metrics in R. *arXiv preprint arXiv:1603.01700*, 2016.

[106] Aliye Atay-Kayis and Hélene Massam. A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92(2):317–335, 2005.

[107] Susan Athey, Dean Eckles, and Guido W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.

[108] Georgia Papadogeorgou, Fabrizia Mealli, and Corwin M. Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787, 2019.

[109] Ilya Shpitser. Segregated graphs and marginals of chain graph models. *Advances in Neural Information Processing Systems*, 28:1720–1728, 2015.

[110] Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag New York, 1st edition edition, 2006.

[111] Eric J. Tchetgen Tchetgen, Linbo Wang, and BaoLuo Sun. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4):2069–2088, 2018.

[112] James M. Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16:21–37, 1997.

[113] Yan Zhou, Roderick J. A. Little, and John D Kalbfleisch. Block-conditional missing at random models for missing data. *Statistical Science*, 25(4):517–532, 2010.

[114] Ilya Shpitser. Consistent estimation of functions of data missing non-monotonically and not at random. In *Advances in Neural Information Processing Systems*, pages 3144–3152, 2016.

[115] Mauricio Sadinle and Jerome P. Reiter. Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104(1):207–220, 2017.

[116] Rhian M. Daniel, Michael G. Kenward, Simon N. Cousens, and Bianca L. De Stavola. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256, 2012.

[117] Felix Thoemmes and Norman Rose. Selection of auxiliary variables in missing data problems: Not all auxiliary variables are created equal. Technical report, R-002, Cornell University, 2013.

[118] Fernando Martel García. Definition and diagnosis of problematic attrition in randomized controlled experiments. *Working paper. Available at SSRN 2302735*, 2013.

[119] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems*, pages 1277–1285. 2013.

[120] Karthika Mohan and Judea Pearl. Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems*, pages 1520–1528. 2014.

[121] Mojdeh Saadati and Jin Tian. Adjustment criteria for recovering causal effects from missing data. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2019.

[122] Ilya Shpitser, Karthika Mohan, and Judea Pearl. Missing data as a causal and probabilistic problem. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 802–811. AUAI Press, 2015.

[123] Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7153–7163. PMLR, 2020.

[124] Daniel Malinsky, Ilya Shpitser, and Eric J. Tchetgen Tchetgen. Semiparametric inference for non-monotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pages 1–22, 2020.

[125] Eric V. Strobl, Shyam Visweswaran, and Peter L. Spirtes. Fast causal inference with non-random missingness by test-wise deletion. *International Journal of Data Science and Analytics*, 6(1):47–62, 2018.

[126] Alex Gain and Ilya Shpitser. Structure learning under missing data. In *Proceedings of the 9th International Conference on Probabilistic Graphical Models*, pages 121–132, 2018.

[127] Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770, 2019.

[128] Aad W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[129] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[130] Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*, 2019.

[131] Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *arXiv preprint arXiv:1912.00306*, 2019.

[132] Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

[133] Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[134] Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

[135] David G. Luenberger. *Optimization By Vector Space Methods*. John Wiley & Sons, 1997.

[136] Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.

[137] Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media, 2003.

[138] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[139] George B. Dantzig, Lester R. Ford Jr., and Delbert R. Fulkerson. A primal-dual algorithm. *Linear Equalities and Related Systems, Annals of Mathematics Study*, 38:171–181, 1956.

[140] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[141] Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30), 2013.

[142] Masashi Sugiyama, Motoaki Kawanabe, and Pui Ling Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, 23(1):44–59, 2010.

[143] Santtu Tikka and Juha Karvanen. Simplifying probabilistic expressions in causal inference. *Journal of Machine Learning Research*, 18(1):1203–1232, 2017.

[144] Maxima. Maxima: a computer algebra system, 2020.

[145] Robin J. Evans and Thomas S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 25(2):848–876, 2019.

[146] Isabel R. Fulcher, Ilya Shpitser, Stella Marealle, and Eric J. Tchetgen Tchetgen. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):199–214, 2020.

[147] Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

[148] Ryo Okui, Dylan S. Small, Zhiqiang Tan, and James M. Robins. Doubly robust instrumental variable regression. *Statistica Sinica*, pages 173–205, 2012.

[149] Linbo Wang and Eric J. Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B, Statistical Methodology*, 80(3):531, 2018.

[150] Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating causal effects using weighting-based estimators. In *Proceedings of the 34th Conference on Artificial Intelligence*. AAAI Press, 2020.

[151] Constantine E. Frangakis, Tianchen Qian, Zhenke Wu, and Iván Díaz. Deductive derivation and Turing-computerization of semiparametric efficient estimation. *Biometrics*, 71(4):867–874, 2015.

[152] Marco Carone, Alexander R. Luedtke, and Mark J. van der Laan. Toward computerized efficient estimation in infinite-dimensional models. *Journal of the American Statistical Association*, 114(527):1174–1190, 2019.

[153] Dougal Maclaurin, David Duvenaud, and Ryan P. Adams. Autograd: Effortless gradients in Numpy. In *ICML 2015 AutoML Workshop*, volume 238, page 5, 2015.

[154] Dougal Maclaurin. *Modeling, inference, and optimization with composable differentiable procedures*. PhD thesis, 2016.

[155] Alexander Mozeika, Onur Dikmen, and Joonas Piili. Consistent inference of a general model using the pseudolikelihood method. *Physical Review E*, 90(1):010101, 2014.

[156] Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233. University of California Press, 1967.

[157] Hui Zhao, Zhongguo Zheng, and Baijun Liu. On the Markov equivalence of maximal ancestral graphs. *Science in China Series A: Mathematics*, 48(4):548–562, 2005.

[158] R. Ayesha Ali, Thomas S. Richardson, and Peter L. Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B):2808–2837, 2009.

Biographical Sketch

Rohit is a PhD candidate in Computer Science at Johns Hopkins University and completed his bachelor's and master's degree in Biomedical Engineering and Computer Science respectively from Johns Hopkins University. His research focuses on the development of statistical and machine learning methods to infer causal relations from unstructured data as well as the application of such methods to complex genome wide association studies to improve patient outcomes. Rohit is passionate about sharing his love for causal inference and computational genomics through pedagogy. This means that during his PhD, when Rohit was not hunched over his laptop writing research papers, he was instead hunched over his laptop developing pedagogical materials and projects to engage undergraduate students in learning scientific concepts and conducting research at the intersection of these rapidly growing fields.