TECHNICAL REPORT
R-83
May 2022

Scores for Learning Discrete Causal Graphs with Unobserved Confounders

Alexis Bellot **Junzhe Zhang** **Elias Bareinboim**
CausalAI Lab
Columbia University
{ab5305, junzhez, eb} @cs.columbia.edu

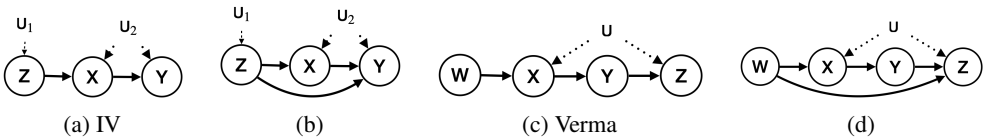
Abstract

Structural learning is arguably one of the most challenging, and pervasive tasks found throughout the data sciences. There exists a growing literature that studies structural learning in non-parametric settings where conditional independence constraints are taken to define the equivalence class. Whenever the underlying causal system is non-Markovian, it is understood that non-conditional independence constraints are imposed over the observational distribution, including certain equalities and inequalities. In this paper, we develop structural learning methods for non-Markovian settings by leveraging additional constraints beyond conditional independences. Specifically, we first introduce a novel Bayesian scoring criterion for arbitrary graphs combining Watanabe's asymptotic expansion of the marginal likelihood and new bounds over the cardinality of the exogenous variables. Second, we show that the new score has desirable properties in terms of expressiveness and computability. In terms of expressiveness, we prove that the score captures distinct constraints imprinted in the data, including Verma's and inequalities'. In terms of computability, we show properties of score equivalence and decomposability, which allows, in principle, to break the problem of structural learning in smaller, more manageable pieces. Third, we implement this score using an MCMC sampling algorithm and test its properties in several simulation scenarios.

1 Introduction

Learning the causal structure underlying a particular phenomenon from data is a fundamental problem across the data sciences. One of the common approaches in the field of causal discovery models the underlying system as a causal graph represented by a Directed Acyclic Graph (DAG), where nodes denote random variables (measured or latent) and directed edges denote causal effects from tails to arrowheads [18, 28, 20]. The task is then to piece together the constraints found in the data (and implied by the underlying, unobserved causal system) to infer the corresponding causal graph.

There are a variety of different types of statistical constraints imposed by the underlying causal system into the observed data with distirbution $P(\mathbf{V})$. One of the most widely used are conditional independences, which can be read off from the causal graph through a criterion known as d-separation [16]. The reverse implication, that each conditional independence in data implies a corresponding separation in the underlying causal graph is known as *faithfulness* [15, 32, 39, 14]. This is the cornerstone assumption for a plethora of algorithms, starting in Markovian models from the IC/PC [33, 9, 28] and non-Markovian models, from the IC*/FCI [33, 28]. In the latter, the set of graphs that entail the same set of conditional independence relations in data can be represented as Partial Ancestral Graphs (PAGs) which defines an equivalence class of Maximal Ancestral Graphs (MAGs) [21]. In practice, an alternative type of algorithm known as score-based, with roots in the wider field of Bayesian model selection [8], have also been popular and instead search for the graph that maximizes the model posterior $P(\mathcal{G} \mid \mathbf{V})$ or an approximation thereof known as the score, without explicitly



(a) IV (b) (c) Verma (d)

Figure 1: Examples of graphs: (a) the instrumental variable graph, (b) an unconstrainted graph, (c) the Verma graph, (d) the Verma graph without the equality constraint.

testing for conditional independencies in data [11, 3, 4]. For instance, the Bayesian Information Criterion (BIC) [25] is a tractable asymptotic approximation to $P(\mathcal{G} \mid \mathbf{V})$ that has been shown to be consistent for Structural Causal Models (SCMs) parameterized by curved exponential models [10], including, e.g., Markovian models and select classes of non-Markovian models such as SCMs compatible with MAGs and with Gaussian exogenous distributions [21].

The constraints in data implied by the SCM and encoded by the presence or absence of edges in the causal graph may be more general equality or inequality relations between marginals of $P(\mathbf{V})$. For instance, for the Instrumental Variable (IV) graph in Fig. 1a, it holds that $\sum_y \max_z P(x, y|z) \leqslant 1$ [19], and for the Verma graph in Fig. 1c an equality constraint not of the conditional independence type holds [34]. In both cases, the MAG representation of these graphs looses this finer granularity, for example graphs in Fig. 1a and Fig. 1b are represented by the same PAG because both encode the same set of conditional independencies. In this line of research, [30] gave a systematic (and complete [6]) algorithm for finding equality constraints implied by SCMs which, in turn, define a class of distributions that agree on both conditional independencies and equality constraints known as Nested Markov models [22]. For discrete observed variables, an explicit smooth parameterization of these models exists and may be consistently scored using the BIC [27].

Despite all the progress achieved so far, no causal discovery algorithm has been developed to account for inequality constraints. This paper proposes a new score that distinguishes between causal graphs leveraging both equality and inequality constraints in data and is applicable to systems with discretely-valued observables and arbitrarily defined exogenous variables. Building on Watanabe's asymptotic expansion of the marginal likelihood [36] and bounds over the cardinality of exogenous variables [23, 40], our score generalizes the BIC to the more general class of singular models with arbitrarily defined latent variables. We then prove the expressiveness power of our score, in the sense that it captures all observable constraints in $P(\mathbf{V})$, and several properties that make the search over the space of causal graphs feasible, such as *decomposability* (only a smaller subgraph needs to be updated in each iteration of the search) and *equivalence* (graphs defining the same family of observational distributions get the same score). We show also that this score has a tractable approximation using an MCMC sampling algorithm and can be plugged into a search procedure for computations in practice. Finally, we evaluate our method through extensive simulations using synthetic data sets. Given the space constraints, all proofs are provided in **??**.

1.1 Preliminaries

We introduce in this section some basic notations and definitions that will be used throughout the paper. We use capital letters to denote variables (X), small letters for their values (x), bold letters for sets of variables ($\mathbf{X}$) and their values ($\mathbf{x}$), and Ω for their domains of definition ($x \in \Omega_X$). The probability distribution over variables $\mathbf{X}$ is denoted by $P(\mathbf{X})$. Similarly, $P(\mathbf{Y} \mid \mathbf{X})$ represents a set of conditional distributions $P(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ for all realizations $\mathbf{x}$. We consistently use $P(\mathbf{x})$ as abbreviations for probabilities $P(\mathbf{X} = \mathbf{x})$; so does $P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}) = P(\mathbf{y} \mid \mathbf{x})$. Finally, $\mathbb{1}\{\cdot\}$ is the indicator function that equals 1 if the statement in $\{\cdot\}$ evaluates to be true, and equals 0 otherwise.

The basic semantical framework of our analysis rests on *structural causal models* (SCMs) [18, 1]. An SCM M is a tuple $\langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P \rangle$ where $\mathbf{V}$ is a set of endogenous variables and $\mathbf{U}$ is a set of exogenous variables. $\mathbf{F}$ is a set of functions where each $f_V \in \mathbf{F}$ decides values of an endogenous variable $V \in \mathbf{V}$ taking as argument a combination of other variables in the system. That is, $V \leftarrow f_V(\mathbf{Pa}_V, \mathbf{U}_V), \mathbf{Pa}_V \subseteq \mathbf{V}, \mathbf{U}_V \subseteq \mathbf{U}$. Exogenous variables $U \in \mathbf{U}$ are mutually independent, values of which are drawn from the exogenous distribution $P(\mathbf{U})$. Drawing values of exogenous variables

$\mathbf{U}$ following $P(\mathbf{U})$ induces the *observational distribution* $P(\mathbf{V})$ over endogenous variables $\mathbf{V}$.

$$P(\mathbf{v}) = \int_{\Omega_{\mathbf{U}}} \prod_{V \in \mathbf{V}} \mathbb{1}\{f_V(pa_V, \mathbf{u}_V) = v\} dP(\mathbf{u}). \tag{1}$$

Each SCM M is associated with a causal graph $\mathcal{G}$ (e.g., Fig. 1), which is a DAG where nodes represent endogenous variables $\mathbf{V}$ and exogenous variables $\mathbf{U}$, and arrows represent the arguments $\mathbf{Pa}_V, \mathbf{U}_V$ of each function f_V. We will use standard graph-theoretic family abbreviations to represent graphical relationships such as parents, children, descendants, and ancestors. For example, the set of parent nodes of $\mathbf{X}$ in $\mathcal{G}$ is denoted by $pa(\mathbf{X})_{\mathcal{G}} = \cup_{X \in \mathbf{X}} pa(X)_{\mathcal{G}}$; ch, de and an are similarly defined. Capitalized versions Pa, Ch, De, An include the argument as well, e.g. $Pa(\mathbf{X})_{\mathcal{G}} = pa(\mathbf{X})_{\mathcal{G}} \cup \mathbf{X}$. A path from a node X to a node Y in $\mathcal{G}$ is a sequence of edges which does not include a particular node more than once. Two sets of nodes $\mathbf{X}, \mathbf{Y}$ are said to be d-separated by a third set $\mathbf{Z}$ in a DAG $\mathcal{G}$, denoted by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z})_{\mathcal{G}}$, if every edge path from nodes in one set to nodes in another are "blocked". The criterion of blockage follows [18, Def. 1.2.3].

For convenience, we will consistently call a path of the form $V_i \leftarrow U_k \rightarrow V_j$ between endogenous $V_i, V_j \in \mathbf{V}$ via an exogenous $U_k \in \mathbf{U}$ a *bi-directed arrow* between V_i, V_j, denoted by $V_i \leftrightarrow V_j$. A *bi-directed path* is a consecutive sequence of bi-directed arrows. We will leverage a special type of clustering of nodes in the graph $\mathcal{G}$, called the *confounded-component* [29].

Definition 1 (c-component)**.** *For a causal graph $\mathcal{G}$, a subset $\mathbf{C} \subseteq \mathbf{V}$ is a c-component if any pair $V_i, V_j \in \mathbf{C}$ is connected by a bi-directed path in $\mathcal{G}$.*

For example, exogenous variables U_1, U_2 in the IV graph in Fig. 1a corresponds to c-components $\mathbf{C}(U_1) = \{Z\}$ and $\mathbf{C}(U_2) = \{X, Y\}$ respectively. For a more detailed survey on SCMs, we refer readers to [18, 2].

We focus our attention on SCMs with *discrete* endogenous (observed) variables, that is, each $V \in \mathbf{V}$ taking values in a finite space of outcomes, while each $U \in \mathbf{U}$ is *arbitrarily defined*, e.g. taking values in $\mathbb{R}$. Our analysis rests on a special parameterization of these SCMs. It is possible to show that the observational distribution in any causal graph associated with SCMs with discrete observables could be generated using finite exogenous states [23, 40].

Proposition 1. *For any causal graph $\mathcal{G}$, let M be an arbitrary SCM compatible with $\mathcal{G}$, i.e., $\mathcal{G}_M = \mathcal{G}$. The observational distribution $P(\mathbf{V})$ induced by M could be factorized over $\mathcal{G}$ as follows:*

$$P(\mathbf{v}) = \sum_{U \in \mathbf{U}} \sum_{u=1,\dots,d_U} \prod_{V \in \mathbf{V}} \mathbb{1}\{f_V(pa_V, \mathbf{u}_V) = v\} \prod_{U \in \mathbf{U}} P(u), \tag{2}$$

where for every exogenous variable $U \in \mathbf{U}$, its cardinality $d_U = \left|\Omega_{Pa(\mathbf{C}(U))}\right|$; for every endogenous variable $V \in \mathbf{V}$, function f_V is a mapping between finite domains $\Omega_{\mathbf{Pa}_V} \times \Omega_{\mathbf{U}_V} \mapsto \Omega_V$.

This result allows us to consistently parameterize observational distribution associated with an SCM with discrete observables using discrete exogenous variables. In other words, for any SCM M there exists a SCM N given by Prop. 1 such that $P_M(\mathbf{V}) = P_N(\mathbf{V})$. For example, in the IV graph in Fig. 1a, let an observational distribution $P(X, Y, Z)$ over binary variables X, Y, Z be induced by an arbitrary distribution $P(U_1, U_2)$ over a continuous domain of the exogenous variables U_1, U_2, i.e. given by Eq. (1). Prop. 1 implies that any $P(x, y, z)$ can be equivalently expressed as:

$$P(x, y, z) = \sum_{u_1} \mathbb{1}\{f_Z(u_1) = z\} P(u_1) \sum_{u_2} \mathbb{1}\{f_X(z, u_2) = x\} \mathbb{1}\{f_Y(x, u_2) = y\} P(u_2), \tag{3}$$

where $P(u_1)$ is a distribution over a binary domain $\{1, 2\}$ since $|\Omega_{U_1}| = |\Omega_X| = 2$; and $P(u_2)$ is a discrete distribution over a finite domain $\{1, \dots, 8\}$ since $|\Omega_{U_2}| = |\Omega_X| \cdot |\Omega_Y| \cdot |\Omega_Z| = 8$.

In the context of structure learning from data, we will require a one-to-one correspondence between the structure of the causal graph and the distribution of data induced by the SCM. The following assumption is the natural generalization of *faithfulness* of conditional independences [4] to our consideration of general statistical constraints, that includes equality and inequality constraints on margins of $P(\mathbf{V})$.

Assumption 1 (Interventional faithfulness). *Each example in the observed data is an $i.i.d$ sample from a distribution $P(\mathbf{V})$ that is defined by statistical constraints (i.e. equalities or inequalities on*

margins of $P(\mathbf{V})$) that have a corresponding structural explanation in the SCM (e.g. a missing directed or bi-directed arrow in the causal graph induced by the SCM)[1].

2 Expressiveness of Scores in the Presence of Unobserved Confounders

We will focus on Bayesian methods and their asymptotic behaviour for scoring causal graphs $\mathcal{G}$. Let $P(\mathcal{G}|\bar{\mathbf{v}})$ be the probability that $\mathcal{G}$ defines the causal structure in the underlying SCM given an $i.i.d$ sample $\bar{\mathbf{v}} = \left\{\mathbf{v}^{(s)} : s = 1, \ldots, n\right\}$. Let ω denote parameters of the observational distribution $P(\mathbf{V})$ induced by the underlying SCM; its domain $\omega \in \Omega_\omega \subset \mathbb{R}$ is compact.

Definition 2 (Bayesian scoring criterion)**.** *The Bayesian scoring criterion is defined as the posterior $P(\mathcal{G}|\bar{\mathbf{v}})$ which may be computed using the marginal likelihood $P(\bar{\mathbf{v}}|\mathcal{G})$,*

$$P(\mathcal{G}|\bar{\mathbf{v}}) \propto P(\mathcal{G})P(\bar{\mathbf{v}}|\mathcal{G}) = P(\mathcal{G}) \int_{\Omega_\omega} P(\bar{\mathbf{v}}|\omega, \mathcal{G}) dP(\omega|\mathcal{G}). \tag{4}$$

2.1 Constraints on $P(\mathbf{V})$ can lead to singular asymptotics of marginal likelihood

Large-sample theory plays an important role in score-based structure learning because Bayesian scoring criteria often involve high-dimensional integrals that are intractable in practice. For instance, most notably, for the class of curved exponential graphical models [10], a quadratic approximation to the log-likelihood function $\log P(\bar{\mathbf{v}}|\omega, \mathcal{G})$ can be used to relate the marginal likelihood to a Gaussian integral in Eq. (4) which results in Schwarz's Bayesian Information Criterion (BIC) [25]. Curved exponential graphical models include graphical models based on DAGs, Maximal Ancestral graphical models with Gaussian variables [21] and Nested Markov models [22]. However, this asymptotic approximation does not necessarily hold in arbitrary graphs with unobserved confounders.

For instance, the instrumental inequality aluded to in the introduction and described more thoroughly in Eq. (7) introduces a boundary in the space of distributions induced by the IV model in Fig. 1a, e.g. a distribution such that $P(Y = 0, X = 0 \mid Z = z) = P(Y = 1, X = 0 \mid Z = z) = 0.5$ for $z \in \{0, 1\}$, lies on this boundary. Following Prop. 1 $|\Omega_{U_2}| = 8$. In this case, everything else unchanged, it can be shown that changing $P(u_2)$ while preserving the sums $\sum_{u_2=0,1,2,3} P(u_2)$ and $\sum_{u_2=4,5,6,7} P(u_2)$ (up to relabelling) does not change the likelihood $P(\mathbf{v} \mid \omega, \mathcal{G})$. In effect, we are loosing degrees of freedom in our model and the asymptotic consequences of this fact can be quite severe. To witness, we plot on Fig. 2 the negative log-likelihood as a function of parameters $P(U_2 = 0)$ and $P(U_2 = 1)$, everything else being equal, using simulated data from the above boundary distribution of the IV model. The colored pattern represents the likelihood surface that concentrates in a ridge shape along a diagonal line that leaves the sum $\sum_{u_2=0,1} P(u_2)$ unchanged and defines a singular point in the model. In words, asymptotic approximations can no longer rely on the likelihood around the maximum being a quadratic surface. As a consequence, in general, the BIC will not reflect the asymptotic scaling of $P(\bar{\mathbf{v}} \mid \omega, \mathcal{G})$ nor $P(\mathcal{G} \mid \bar{\mathbf{v}})$ defined by inequality constraints.

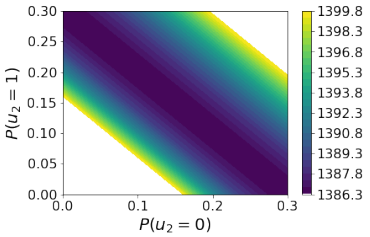


Figure 2: $-\log P(\bar{\mathbf{v}} \mid \omega, \mathcal{G})$.

2.2 Asymptotic approximations capturing all observational constraints

Watanabe reformulated the foundations of asymptotic theory of singular models using the Hironaka resolution on singularities [12, 35, 36]. A distinct notion of model dimension emerges in singular models driven by the learning coefficient $\lambda > 0$ and multiplicity θ that describe how fast the posterior distribution shrinks with increasing sample size. In general, the asymptotic expansion of the log marginal likelihood is given by,

$$-\log P(\bar{\mathbf{v}} \mid \mathcal{G}) = -\log P(\bar{\mathbf{v}} \mid \mathcal{G}, \omega_0) + \lambda \log n + (\theta - 1) \log \log n + \mathcal{O}_p(1), \tag{5}$$

[1] A formal statement of this assumption would require a precise description of equality constraints, which can be done by introducing the *do*-calculus (the converse of each rule (when identified) can be used to define an equality constraint in observational distributions), and a precise description of the influence of inequality constraints on the structure of the graph which are likely to be much complicated to be done in general, see e.g. [5]. We postpone this characterization to future work.

where ω_0 is the parameter that minimizes the Kullback-Leibler distance from a true distribution to the distribution induced by any SCM compatible with $\mathcal{G}$[2]. In general, $\lambda \neq |\Omega_\omega|/2$ and $\theta \neq 1$, which means that the BIC no longer provides the correct scaling of the marginal likelihood. We show next that under appropriate conditions, Watanabe's asymptotic expansion coupled with the discrete parameterization of the likelihood discussed earlier assigns the lowest (best) score to the model imposing the fewest constraints that can represent the generative distribution.

Theorem 1. *Let $P(\bar{\mathbf{v}} \mid \mathcal{G}, \omega)$ be parameterized by a discrete SCM (Prop. 1), and define a score*

$$\mathcal{S}(\mathcal{G}, \bar{\mathbf{v}}) := -\log P(\bar{\mathbf{v}} \mid \mathcal{G}, \hat{\omega}) + \lambda \log n + (\theta - 1) \log \log n, \tag{6}$$

where $\hat{\omega}$ is the maximum likelihood set of parameters of the model in Eq. (2). Under regularity conditions, in the limit as $n \to \infty$,

1. *(Soundness) If the family of distribution compatible with $\mathcal{G}_1$ includes $P(\mathbf{V})$ but the family of distributions compatible with $\mathcal{G}_2$ does not, then $\mathcal{S}(\mathcal{G}_1, \bar{\mathbf{v}}) < \mathcal{S}(\mathcal{G}_2, \bar{\mathbf{v}})$ with probability 1.*
2. *(Parsimony) If the family of distributions compatible with $\mathcal{G}_1$ is included in that compatible with $\mathcal{G}_2$ and both contain $P(\mathbf{V})$, then $\mathcal{S}(\mathcal{G}_1, \bar{\mathbf{v}}) < \mathcal{S}(\mathcal{G}_2, \bar{\mathbf{v}})$ with probability 1.*

The first part of the proposition encodes the soundness of the parametrization, i.e., a graph that encodes the constraints of the original model will have a higher score than a graph that disagrees with these constraints. The second part encodes the idea of simplicity, which means that among two structures that have the same generative capabilities, the one that is simpler will be preferred over the more complex one. This is a key property since, as acknowledged in the field, a complete graph does not impose any statistical constraints but it is not a good representation of the original model.

It follows, as a corollary, that the parametrization in Eq. (2) captures all statistical constraints over observational probabilities encoded by the structure of the causal graph.

Corollary 1. *Under the same conditions as in Theorem 2, score-based learning using $\mathcal{S}(\mathcal{G}, \bar{\mathbf{v}})$ distinguishes between two candidate causal graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ on the basis of both equality and inequality constraints between margins of $P(\mathbf{V})$ implied by any SCM compatible with $\mathcal{G}_1$ and $\mathcal{G}_2$.*

For example, consider the IV graph in Fig. 1a. The factorization in Eq. (3) implies that for any x,

$$\sum_y \max_z P(x, y \mid z) = \sum_{u_2} \max_z \mathbb{1}\{f_X(z, u_2) = x\} P(u_2) \sum_y \mathbb{1}\{f_Y(x, u_2) = y\} \leqslant 1. \tag{7}$$

The last step follows from $\sum_y \mathbb{1}\{f_Y(x, u_2) = y\} = 1$ and the fact that $\sum_{u_2} \max_z \mathbb{1}\{f_X(z, u_2) = x\} P(u_2) \leqslant \sum_{u_2} P(u_2) = 1$. The same inequality, however, does not necessarily hold in the unconstrained model defined by the graph in Fig. 1b because of the directed edge $Z \to Y$, for which,

$$\sum_y \max_z P(x, y \mid z) = \sum_{u_2} \sum_y \max_z \mathbb{1}\{f_Y(x, z, u_2) = y\} \mathbb{1}\{f_X(z, u_2) = x\} P(u_2), \tag{8}$$

which cannot be bounded in the same manner since $\sum_y \max_z \mathbb{1}\{f_Y(x, z, u_2) = y\} \geqslant 1$ [17]. A similar contrast between causal graphs does not hold in the framework of MAGs or Nested Markov models because both graphs correspond to the same parameterization. For example, in the case of MAGs, the parameterization of the likelihood in both graphs is given by,

$$P(x, y, z) = \sum_{u_x, u_y, u_z} \mathbb{1}\{f_Z(u_z) = z\} \mathbb{1}\{f_X(z, u_x) = x\} \mathbb{1}\{f_Y(x, z, u_y) = y\} P(u_x, u_y, u_z). \tag{9}$$

For a different example, using the parameterization of the likelihood in Prop. 1 an equality constraint can be derived for the Verma graph in Fig. 1c due to the absence of an edge $W \to Z$ [30],

$$\sum_{x,u} \mathbb{1}\{f_Z(y, u) = z\} \mathbb{1}\{f_X(w, u\}) = x\} P(u) = \sum_{x,w,u} \mathbb{1}\{f_Z(y, u) = z\} \mathbb{1}\{f_X(w, u) = x\} P(u) P(w),$$

for any value of w in the LHS since the RHS does not depend on w (it is marginalized over). The same equality constraint does not hold for the model in Fig. 1d. One way of seeing this is that in Fig. 1c the LHS and RHS of this equation are equal to the causal effect $P(z|do(x), w)$ (and as mentioned does not depend on w) but the same causal effect in Fig. 1d does depend on w due to the presence of the directed arrow $W \to Z$.

[2]The presence of the $\mathcal{O}_p(1)$ error means that, as n increases, the approximation can differ from the true log marginal likelihood by a constant term. Note also that the prior term $P(\mathcal{G})$ does not depend on the data, it does not grow with n and therefore can also be absorbed into an error term $\mathcal{O}_p(1)$. In essence, asymptotically $\log P(\bar{\mathbf{v}} \mid \mathcal{G}) = \log P(\mathcal{G} \mid \bar{\mathbf{v}})$ up to constant terms.

3 Computation of Scores and their Properties for Causal Discovery

An explicit approximation of the marginal likelihood $P(\bar{\mathbf{v}} \mid \mathcal{G})$ using Eq. (5) is intractable because, in general, λ and θ depend on the true data generating distribution and have only been explicitly computed for a handful of models using Hironaka's resolution of singularities theory [12]. One challenge for evaluating Eq. (5), and obtaining all properties discussed before, is therefore that the true model has to be known beforehand. Following Watanabe [37], we use a different asymptotic approximation to the marginal likelihood using techniques from path sampling and thermodynamic integration [7]. Our proposed score $\mathcal{S}_{\text{WBIC}}$ for a candidate causal graph $\mathcal{G}$ and data $\bar{\mathbf{v}}$ is defined as,

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\mathbf{v}}) := -\mathbb{E}_\beta \log P(\bar{\mathbf{v}} \mid \mathcal{G}, \omega), \quad \mathbb{E}_\beta g(\omega) := \frac{\int_{\Omega_\omega} g(\omega) P(\bar{\mathbf{v}} \mid \omega, \mathcal{G})^\beta dP(\omega \mid \mathcal{G})}{\int_{\Omega_\omega} P(\bar{\mathbf{v}} \mid \omega, \mathcal{G})^\beta dP(\omega \mid \mathcal{G})}. \tag{10}$$

The significance of this definition lies in the fact that the marginal likelihood $P(\bar{\mathbf{v}} \mid \mathcal{G})$ is equal to $-\mathbb{E}_\beta \log P(\bar{\mathbf{v}} \mid \mathcal{G}, \omega)$ for some value $\beta^* \in [0, 1]$. More specifically, $\mathcal{S}_{\text{WBIC}}$ is defined as the derivative of $\int_{\Omega_\omega} P(\bar{\mathbf{v}} \mid \omega, \mathcal{G})^\beta dP(\omega \mid \mathcal{G})$ with respect to β, which by the mean value theorem must be equal to the marginal likelihood $P(\bar{\mathbf{v}} \mid \mathcal{G}) = \int_{\Omega_\omega} P(\bar{\mathbf{v}} \mid \omega, \mathcal{G}) dP(\omega \mid \mathcal{G})$ for some value $\beta^* \in [0, 1]$. For the choice $\beta = \frac{1}{\log n}$ it holds, asymptotically by Theorem 4 in [37], that,

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\mathbf{v}}) = -\log P(\mathcal{G} \mid \bar{\mathbf{v}}) + \mathcal{O}_p(\sqrt{\log n}). \tag{11}$$

This result shows that model selection using $\mathcal{S}_{\text{WBIC}}$ approximates a Bayesian procedure seeking the model with highest posterior probability. However, the $\mathcal{S}_{\text{WBIC}}$ may deviate from the marginal likelihood by a constant term times $\sqrt{\log n}$. For consistency of model selection this difference must be of lower order than the difference in $\mathcal{S}_{\text{WBIC}}$ between two different models. This this additional assumption, a consistency result analogous to Thm. 1 can be derived.

Corollary 2. *Let $\mathcal{G}_1$ be compatible with $P(\mathbf{V})$ but not $\mathcal{G}_2$. With the assumption that $\log P(\mathcal{G}_1 \mid \bar{\mathbf{v}}) - \log P(\mathcal{G}_2 \mid \bar{\mathbf{v}})$ is asymptotically larger than $\mathcal{O}_p(\log n)$, $\mathcal{S}_{WBIC}$ is sound and parsimonious with probability 1, as defined by Thm. 1 under the same regularity conditions.*

3.1 Properties of $\mathcal{S}_{\text{WBIC}}$ for causal discovery

Next we describe some properties of the proposed score $\mathcal{S}_{\text{WBIC}}$ which will be desirable for causal discovery. Our next result shows that $\mathcal{S}_{\text{WBIC}}$ decomposes over c-components in the causal graph.

Definition 3 (Decomposability)**.** *The score $\mathcal{S}$ is decomposable if it can be written as a sum of measures, each of which is a function only of the variables in the c-component and its parents,*

$$\mathcal{S}(\mathcal{G}, \bar{\mathbf{v}}) = \sum_{\mathbf{C} \in \mathbf{C}(\mathcal{G})} \mathcal{S}(\mathcal{G}_{Pa(\mathbf{C})}, \bar{\mathbf{v}}_{Pa(\mathbf{C})}). \tag{12}$$

Here $\mathcal{G}_{Pa(\mathbf{C})}$ and $\mathbf{v}_{Pa(\mathbf{C})}$ denote the subgraph and data, respectively, with restriction to the variables in $Pa(\mathbf{C}) \subseteq \mathbf{V}$.

Proposition 2. *$\mathcal{S}_{WBIC}$ is decomposable.*

The decomposability will avoid the need of recomputing the entire score when examining a new graphical structure, which makes the search feasible in principle. For example, to score the IV graph in Fig. 1a, we may separately score c-components $\{Z\}$ and $\{X, Y\}$, the first one being a function of X only while the second one being a function of $\{X, Y, Z\}$. If we were to add an edge $Z \to Y$ to arrive at Fig. 1b we only need to recompute the updated c-component $\{X, Y\}$ and compare its score with the previous parameterization to infer which is more likely to include $P(\mathbf{V})$.

An important observations is that statistical constraints in data are usually not sufficient to narrow down a unique causal graph and, in practice, multiple graphs may encode the same constraints as those of the true graph. This set forms an equivalence class.

Definition 4 (Score equivalence)**.** *A scoring criterion $\mathcal{S}$ is score equivalent if, for any pair of causal graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ that are compatible with the exact same family of distributions, $\mathcal{S}(\mathcal{G}_1, \bar{\mathbf{v}}) = \mathcal{S}(\mathcal{G}_2, \bar{\mathbf{v}})$ asymptotically with probability 1.*

Proposition 3. *Under regularity conditions, $\mathcal{S}_{WBIC}$ is score equivalent.*

For example, adding a bi-directed edge $Z \leftrightarrow X$ to the graph in Fig. 1b does not remove any constraints on the set of induced observational distributions $P(\mathbf{V})$ and has therefore the same score even though its parameterization and number of parameters differs, highlighting again that the number of parameters does not capture the complexity of the model in general.

The space of arbitrary causal graphs is different than that of MAGs or DAGs, and in particular the consistency of search methods and characterizations of equivalence classes do not necessarily hold more generally. We propose a (heuristic) greedy search algorithm that uses the decomposable nature of $\mathcal{S}_{\text{WBIC}}$. The greedy search starts from an empty graph and proceeds iteratively. At each stage, $\mathcal{S}_{\text{WBIC}}$ evaluates neighbouring graphs by considering every pair of variables to which one can remove, change, or add a directed, bi-directed, edge, or expanding a bi-directed edge denoting an unobserved confounder to have three or more children without violating the acyclicity constraint. In each step of the search, all the graphs that occur with single changes of the current graph are considered but one only needs to recompute the scores of c-components that are affected by the change. Greedy search implemented with $\mathcal{S}_{\text{WBIC}}$ is denoted GS-$\mathcal{S}_{\text{WBIC}}$.

3.2 Computing $\mathcal{S}_{\text{WBIC}}$

We first recall that given a graph $\mathcal{G}$, the parameterization of $P(\mathbf{V})$ is given as,

$$P(\mathbf{v}) = \sum_{U \in \mathbf{U}} \sum_{u=1,\dots,d_U} \prod_{V \in \mathbf{V}} \mathbb{1}\{\xi_V^{(\mathbf{u}_V, \mathbf{pa}_V)} = v\} \prod_{U \in \mathbf{U}} \theta_u, \tag{13}$$

where $\xi_V^{(\mathbf{u}_i, \mathbf{pa}_i)}$ are parameters that take values in the range of V and represent the assignment of V given its parents and exogenous variables, $i = 1, \dots, d$. There is one such parameter of dimensionality $|\Omega_V|$ for each combination of realization of parent variables $\mathbf{pa}_V$ and exogenous variables $\mathbf{u}_V$ that are defined by the candidate causal graph $\mathcal{G}$. θ_u stands for the vector of probabilities that defines the discrete distribution $P(U = u)$ over its finite domain $u \in \{1, \dots, d_U\}$. For convenience, we group parameters into a single symbol $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\theta})$ where $\boldsymbol{\xi} = \{\xi_V^{(\mathbf{u}_V, \mathbf{pa}_V)} : V \in \mathbf{V}, \mathbf{Pa}_V \subset \mathbf{V}, \mathbf{U}_V \subset \mathbf{V}\}$ and $\boldsymbol{\theta} = \{\theta_u : U \in \mathbf{U}\}$.

$\mathcal{S}_{\text{WBIC}}$ is computed by setting the temperature $\beta := 1/\log n$ and prior over parameters $\boldsymbol{\omega}$ given $\mathcal{G}$ (possibly uninformative), and drawing Monte Carlo samples of the posterior distribution $P(\boldsymbol{\omega}|\bar{\mathbf{v}}, \mathcal{G})^\beta$ at temperature β. All parameters, their dimensionalities, and space of potential values are determined by the structure of the candidate graph and the observed data $\bar{\mathbf{v}}$, but also depend on (unobserved) exogenous variables $\bar{\mathbf{u}} = \{\mathbf{u}^{(s)} : s = 1, \dots, n\}$. We approximate the posterior $P(\boldsymbol{\omega}|\bar{\mathbf{v}}, \mathcal{G})$ by marginalizing over $\bar{\mathbf{u}}$. We begin with some initial value for all unobserved quantities $(\mathbf{u}, \boldsymbol{\xi}, \boldsymbol{\theta})$, and sample each one iteratively conditioned on the current values of the remaining terms in this vector before introducing a Metropolis step.

The posterior distribution of exogenous variables is given by,

$$P(\mathbf{u}^{(i)}|\mathbf{v}^{(i)}, \boldsymbol{\xi}, \boldsymbol{\theta}) \propto P(\mathbf{u}^{(i)}, \mathbf{v}^{(i)}|\boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{V \in \mathbf{V}} \mathbb{1}\{\xi_V^{(\mathbf{u}_V, \mathbf{pa}_V)} = v\} \prod_{U \in \mathbf{U}} \theta_{\mathbf{u}}. \tag{14}$$

The posterior distribution of functional assignment variables is given by,

$$P(\xi_V^{(\mathbf{u}_V, \mathbf{pa}_V)} = v) = 1, \quad \text{if} \quad \exists s : (\mathbf{u}_V^{(s)}, v_V^{(s)}, \mathbf{pa}_V^{(s)}) = (\mathbf{u}_V, v, \mathbf{pa}_V), \tag{15}$$

$$P(\xi_V^{(\mathbf{u}_V, \mathbf{pa}_V)} = v) = q, \quad \text{otherwise}, \tag{16}$$

where q is a proposal distribution that samples $\xi_V^{(\mathbf{u}_V, \mathbf{pa}_V)}$ in Ω_V with probabilities that are uniformly updated in a small neighbourhood of the previous parameter value in each iteration of the sampler.

The posterior distribution of exogenous probabilities is given by $\theta_u \sim \text{Dir}(c_1, \dots, c_{d_u})$, whose concentration parameters are updated in each iteration of the sampler using a uniform proposal distribution, e.g. $c_i \sim \text{Uniform}(c_i - \epsilon, c_i + \epsilon)$ and $\epsilon > 0$ a small scalar.

Let $(\boldsymbol{\xi}_{(t)}, \boldsymbol{\theta}_{(t)})$ be the t-th sample in the Markov chain. A new sample $(\boldsymbol{\xi}_{(t+1)}, \boldsymbol{\theta}_{(t+1)})$ is recorded with an acceptance ratio given by $P(\boldsymbol{\xi}_{(t+1)}, \boldsymbol{\theta}_{(t+1)}|\bar{\mathbf{v}}, \mathcal{G})^\beta / P(\boldsymbol{\xi}_{(t)}, \boldsymbol{\theta}_{(t)}|\bar{\mathbf{v}}, \mathcal{G})^\beta$ where,

$$P(\boldsymbol{\xi}, \boldsymbol{\theta}|\bar{\mathbf{v}}, \mathcal{G})^\beta \propto \exp\left\{-\beta \log P(\bar{\mathbf{v}}|\boldsymbol{\xi}, \boldsymbol{\theta}, \mathcal{G}) + \log P(\boldsymbol{\xi}, \boldsymbol{\theta}|\mathcal{G})\right\}. \tag{17}$$

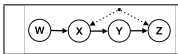
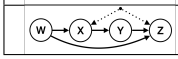
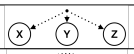
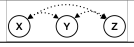
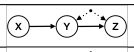
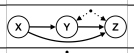
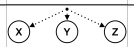
Graph	Parameterization of $P(\mathbf{v})$	$\boldsymbol{\lambda}$	$\mathcal{S}_{\text{WBIC}}$	True Graph?
	$P(w)P(y\|x)\sum_u P(x\|w,u)P(z\|y,u)P(u)$	6.78	2770.8	✓
	$P(w)P(y\|x)\sum_u P(x\|w,u)P(z\|w,y,u)P(u)$	7.87	2778.7	×
	$\sum_u P(x\|u)P(y\|u)P(z\|u)P(u)$	1.83	698.1	✓
	$\sum_{\mathbf{u}} P(x\|u_1,u_2)P(y\|u_1,u_3)P(z\|u_2,u_3)P(u_1,u_2,u_3)$	1.05	1399.3	×
	$P(x)\sum_u P(y\|x,u)P(z\|y,u)P(u)$	2.08	1654.6	✓
	$P(x)\sum_u P(y\|x,u)P(z\|x,y,u)P(u)$	2.47	1659.8	×
	$\sum_u P(x\|u)P(y\|u)P(z\|u)P(u)$	2.51	1660.3	×

Table 1: Experiments to illustrate the behavior of $\mathcal{S}_{\text{WBIC}}$ on graphs imposing different constraints on data.

Finally,

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G},\bar{\mathbf{v}}) := -\mathbb{E}_\beta \log P(\bar{\mathbf{v}}|\mathcal{G},\boldsymbol{\omega}) \approx -\frac{1}{T}\sum_{t=1}^{T}\log P(\bar{\mathbf{v}}|\mathcal{G},\boldsymbol{\xi}_{(t)},\boldsymbol{\theta}_{(t)}). \tag{18}$$

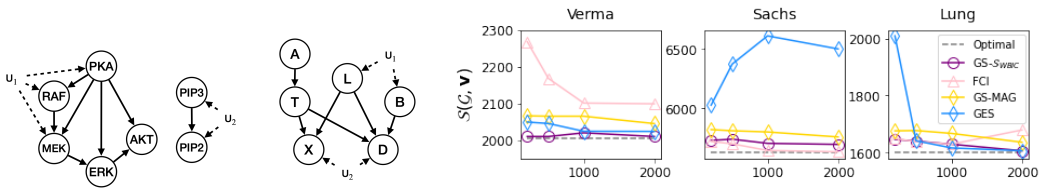
4 Experiments

This section evaluates the ability of the newly proposed score to distinguish between graphs that differ in equality and inequality constraints, as well as test the quality of solutions returned by greedy search in comparison with alternative causal discovery algorithms.

Our first experiment is summarized in Table 1, itself sub-divided into 3 sections. Each section involves data sampled from a different SCM shown in the row labeled "✓" that is to be compared with alternative (erroneous "×") causal graphs given in the rows below that differ only in one equality or inequality constraint. For illustration, we give the likelihood parameterization of each candidate causal graph, the estimated score $\mathcal{S}_{\text{WBIC}}$ computed following Sec. 3.2, and the approximated learning coefficient "$\boldsymbol{\lambda}$" for each pair of model and data generating mechanism. (Note in particular that it is not necessarily equal to the BIC's "half the number of parameters".)

- **Verma graph.** The Verma graph in the first row specifies an equality constraint that is violated in the second graph, while both specify the same set of inequality constraints (and also conditional independencies) over $P(\mathbf{v})$. $\mathcal{S}_{\text{WBIC}}$ gives a lower score to the Verma graph, correctly inferring the better fit given that the equality constraint in data is not accounted for in the second graph.
- **Unconstrained graph.** The graph in the third row is compatible with any probability distribution $P(x,y,z)$. We generate data in a manner that $P(x=y=z) = P(u) \sim \text{Bernoulli}(0.5)$ chosen because it cannot be generated by the triple bi-directed graph in the fourth row, see e.g. [38], even though both models specify exactly the same constraints otherwise. $\mathcal{S}_{\text{WBIC}}$ correctly infers (gives a lower score) to the true causal graph.
- **IV graph.** The last section of Table 1 considers data from the IV graph that encodes an inequality constraint in $P(x,y,z)$. The last two graphs are both compatible with any distribution $P(x,y,z)$ which we include here both to demonstrate that $\mathcal{S}_{\text{WBIC}}$ correctly infers the data generating mechanism but also that $\mathcal{S}_{\text{WBIC}}$ gives the exact same score to equivalent graphs. We further illustrate equivalence and decomposability features of $\mathcal{S}_{\text{WBIC}}$ in **??**.

Our second experiment evaluates the quality of solutions returned by the proposed greedy search algorithm GS-$\mathcal{S}_{\text{WBIC}}$ in comparison with popular MAG constraint-based algorithms: FCI [28], and MAG and DAG score-based algorithms using the BIC: GS-MAG [31] and GES [3], respectively. We consider two real-world examples: the Sachs dataset for the discovery of a protein signaling network [24], and the Lung cancer dataset [13], both available through the bnlearn network repository [26], and one synthetic example with data sampled from the Verma graph that can be found in Fig. 1c. To account for unobserved confounding and make the setting more realistic, we omitted four and two



(a) $\mathcal{G}_{\text{Sachs}}$ (b) $\mathcal{G}_{\text{Lung}}$ (c) $\mathcal{S}_{\text{WBIC}}$ of estimated graph as a function of sample size.

Figure 3: Graphs and performance comparisons.

variables from the Sachs and Lung Cancer datasets, respectively. Their graphs can be found in Fig. 3. Please refer to **??** for further details on the experiments.

Fig. 3 presents score $\mathcal{S}_{\text{WBIC}}$ comparisons (Structural Hamming Distance comparisons given in **??**) of the returned graph $\mathcal{G}$ (or members of the equivalence class if appropriate) as a function of the number of samples n used by each one of the methods under consideration. On average GS-$\mathcal{S}_{\text{WBIC}}$ outperforms on a majority of runs, especially in the small sample regime although there is notable variation between datasets. For example, in the large sample regime, FCI outperforms in the Sachs dataset and GSMAG and GES are competitive in the Lung cancer dataset.

5 Conclusions

We investigated the problem of learning the causal structure underlying a phenomenon of interest from discretely-valued observational data with arbitrary latent dependencies. In this paper, we defined a score based on the asymptotic expansion of the marginal likelihood using a parameterization that is expressive enough to capture consistently both equality and inequality constraints in the observational data. We then proposed a tractable approximation to this score that involves a posterior sampling algorithm using power posteriors and that enjoys desirable properties for causal discovery such as score decomposition and score equivalence that make searching over the space of causal graphs feasible. These results extend score-based causal discovery based on Maximal Ancestral Graphs and Nested Markov models based on equality constraints.

Acknowledgments and Disclosure of Funding

This research was supported in part by the NSF, ONR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

[1] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345–7352, 2016.

[2] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 507–556. Association for Computing Machinery, NY, USA, 1st edition, 2022.

[3] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

[4] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[5] Robin J Evans. Graphical methods for inequality constraints in marginalized dags. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.

[6] Robin J Evans. Margins of discrete bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, 2018.

[7] Nial Friel and Anthony N Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607, 2008.

[8] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[9] Clark Glymour, Richard Scheines, and Peter Spirtes. *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press, 2014.

[10] Dominique MA Haughton. On the choice of a model to fit data from an exponential family. *The annals of statistics*, pages 342–355, 1988.

[11] David Heckerman, Christopher Meek, and Gregory Cooper. A bayesian approach to causal discovery. *Computation, causation, and discovery*, 19:141–166, 1999.

[12] Heisuke Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero: Ii. *Annals of Mathematics*, pages 205–326, 1964.

[13] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

[14] Alexander Marx, Arthur Gretton, and Joris M. Mooij. A weaker faithfulness assumption based on triple interactions. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, pages 451–460. AUAI Press, 2021.

[15] Christopher Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 411–418, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[16] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[17] J. Pearl. On the testability of causal models with latent and instrumental variables. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 435–443. Morgan Kaufmann, San Francisco, CA, 1995.

[18] Judea Pearl. *Causality*. Cambridge university press, 2009.

[19] Judea Pearl. On the testability of causal models with latent and instrumental variables. *arXiv preprint arXiv:1302.4976*, 2013.

[20] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[21] Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The 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] Denis Rosset, Nicolas Gisin, and Elie Wolfe. Universal bound on the cardinality of local hidden variables in networks. *Quantum Information and Computation*, 18(11&12):0910–0926, 2018.

[24] 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.

[25] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[26] Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

[27] Ilya Shpitser, Thomas S Richardson, James M Robins, and Robin Evans. Parameter and structure learning in nested markov models. *arXiv preprint arXiv:1207.5058*, 2012.

[28] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[29] Jin Tian and Judea Pearl. *A general identification condition for causal effects*. eScholarship, University of California, 2002.

[30] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. *arXiv preprint arXiv:1301.0608*, 2012.

[31] Sofia Triantafillou and Ioannis Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In *CFA@ UAI*, pages 59–67, 2016.

[32] Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

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

[34] Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 221–236. 2022.

[35] Sumio Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13(4):899–933, 2001.

[36] Sumio Watanabe. *Algebraic geometry and statistical learning theory*. Number 25. Cambridge university press, 2009.

[37] Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897, 2013.

[38] Elie Wolfe, Robert W Spekkens, and Tobias Fritz. The inflation technique for causal inference with latent variables. *Journal of Causal Inference*, 7(2), 2019.

[39] Jiji Zhang. *Causal inference and reasoning in causally insufficient systems*. PhD thesis, Department of Philosophy, Carnegie Mellon University, 2006.

[40] Junzhe Zhang, Tian Jin, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning (ICML-22)*, 2022.