Scores for Learning Discrete Causal Graphs with Unobserved Confounders

Alexis Bellot[1*], **Junzhe Zhang**[2], **Elias Bareinboim**[2]

[1]Google DeepMind
[2]Columbia University, New York
abellot@google.com, junzhez@cs.columbia.edu, 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. In the presence of unobserved confounders, it is understood that non-conditional independence constraints are imposed over the observational distribution, including certain equalities and inequalities between functionals of the joint distribution. In this paper, we develop structural learning methods that leverage additional constraints beyond conditional independencies. Specifically, we first introduce a score 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 us, in principle, to break the problem of structural learning into smaller and more manageable pieces. Third, we implement this score using an MCMC sampling algorithm and test its properties in several simulation scenarios.

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 model represented by a causal graph, where nodes denote random variables (measured or latent) and directed edges denote causal effects from tails to arrowheads (Pearl 2009; Spirtes et al. 2000; Peters, Janzing, and Schölkopf 2017). 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 distribution $P(\boldsymbol{V})$. For example, a d-separation between nodes in a causal graph induces a corresponding conditional independence between variables in $\boldsymbol{V}$. The reverse implication, i.e. that each conditional independence in data implies a corresponding d-separation in the underlying causal graph (known as faithfulness), serves as a statistically testable constraint to narrow the class of compatible graphs (Pearl 1988; Meek 1995; Uhler et al. 2013; Zhang 2006; Marx, Gretton, and Mooij 2021). This is the cornerstone assumption for a plethora of structure learning algorithms (Verma and Pearl 1990; Glymour, Scheines, and Spirtes 2014; Spirtes et al. 2000). When all variables are observable, d-separation statements capture *all* testable constraints implied by the underlying causal model (Verma and Pearl 1990).

This is not the case in the presence of latent variables that are typically used to represent systems involving unobserved confounding. Such causal models are known to induce distributions over observed variables that are defined by more complex statistical constraints, not necessarily of the conditional independence type. The earliest example was given by Verma and Pearl (Verma and Pearl 1990), in which two graphs, shown in Figs. 1c and 1d, imply the same set of conditional independence constraints and yet can be distinguished because they imply an equality between different functionals of $P(\boldsymbol{V})$. In particular, only the Verma graph in Fig. 1c entails the equality

$$\sum_x P(z \mid x, y)P(x) = \sum_x P(z \mid x, y, w)P(x \mid w). \quad (1)$$

Another example is given by the Instrumental Variable (IV) graph in Fig. 1a. While the IV graph does not impose any conditional independencies between variables, compatible data distributions (with discretely-valued observables) $P(x, y, z)$ must satisfy the inequality, first shown by Pearl (Pearl 1995),

$$\sum_y \max_z P(x, y \mid z) < \sum_{u_2, y} \max_z P(x \mid z, u_2)P(y \mid x, u_2)P(u_2) \leqslant 1. \quad (2)$$

The same inequality does not hold in the (otherwise statistically equivalent) unconstrained graph in Fig. 1b. In systems with discrete observables, distributions induced by

[*]Work done primarily while affiliated with Columbia University.

causal graphs are indeed *always* restricted whenever two observed variables are not directly connected, that is are neither adjacent nor subject to unobserved confounding (Evans 2016). For example, it is the structural separation between Z and Y in Fig. 1a that induces an inequality constraint, not present in Fig. 1b due to the bi-directed edge $Z \dashleftarrow\!\dashrightarrow Y$. By adopting the reverse implication, any statistical (in)equality constraint could be used to distinguish between competing causal explanations from observational data.

Early structure learning approaches, starting with the IC/PC algorithms in the context of full observability, and the IC*/FCI algorithms in the presence of unobserved confounding, developed themselves, as well as the causal abstractions involved, around conditional independence testing and faithfulness assumptions (Verma and Pearl 1990; Spirtes et al. 2000). In particular, to reason about unobserved confounding, the latter class of methods considers a special class of graphs, known as Maximal Ancestral Graphs (MAGs), that explicitly associates every separation in the graph with a corresponding conditional independence in $P(\boldsymbol{V})$ (Richardson and Spirtes 2002). The MAG representation of equivalence classes of causal graphs thus loses the finer granularity in induced distributions encoded by (in)equality constraints. For example, both pairs of causal graphs in Fig. 1 are given by the *same* MAGs, as both encode the same set of conditional independencies (and ancestral relations). MAGs are also a popular construct for an alternative class of algorithms known as score-based, that instead search for the MAG $\mathcal{G}$ maximizing the model posterior $P(\mathcal{G} \mid \boldsymbol{V})$ or an approximation thereof (Gelman et al. 1995; Heckerman, Meek, and Cooper 1999; Chickering 2002a,b). The most notable example is the Bayesian Information Criterion (BIC) that can be derived as an asymptotic approximation to $P(\mathcal{G} \mid \boldsymbol{V})$ for distributions defined by MAGs with a Gaussian latent structure (and more general curved exponential models (Haughton 1988; Schwarz 1978)). Several more general causal abstractions, such as discrete chain graph models (Drton 2009a), fully bi-directed graph models (Drton and Richardson 2008), and discrete nested Markov models (Richardson et al. 2017) have also been shown to be curved exponential models and can be scored consistently with the BIC.

Despite the progress achieved so far, there exists no causal discovery algorithm that accounts for inequality constraints in the space of general causal graphs. This paper proposes a new score that distinguishes between causal graphs leveraging both equality and inequality constraints in data and applies to systems with discretely valued observables and arbitrarily defined exogenous variables. Building on Watanabe's asymptotic expansion of the marginal likelihood (Watanabe 2009) and bounds over the cardinality of exogenous variables (Rosset, Gisin, and Wolfe 2018; Zhang, Jin, and Bareinboim 2022), our score generalizes the BIC to the more general class of discrete models with arbitrary latent variables. We further prove the expressiveness power of our score, in the sense that it captures all observable constraints in $P(\boldsymbol{V})$. This implies that, in principle, any two graphs that are distinguishable based on $P(\boldsymbol{V})$ can be distinguished with the proposed score. We show also 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 procedure) and *equivalence* (graphs defining the same family of observational distributions have the same score), and propose 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 simulations using various synthetic datasets.

Preliminaries

We use capital letters to denote variables (X), small letters for their values (x), bold letters for sets of variables ($\boldsymbol{X}$) and their values ($\boldsymbol{x}$), and Ω for their domains of definition ($x \in \Omega_X$). The probability distribution over variables $\boldsymbol{X}$ is denoted by $P(\boldsymbol{X})$. We consistently use $P(\boldsymbol{x})$ as abbreviations for probabilities $P(\boldsymbol{X} = \boldsymbol{x})$. Finally, $\mathbf{1}\{\cdot\}$ is the indicator function that equals 1 if the statement in $\{\cdot\}$ evaluates to be true.

The basic framework of our analysis rests on *structural causal models* (SCMs) (Pearl 2009, Def. 7.1.1). An SCM M is a tuple $\langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P \rangle$ where $\boldsymbol{V}$ is a set of endogenous variables and $\boldsymbol{U}$ is a set of exogenous variables. $\mathcal{F}$ is a set of functions where each $f_V \in \mathcal{F}$ decides values of an endogenous variable $V \in \boldsymbol{V}$ taking as argument a combination of other variables in the system. That is, $V \leftarrow f_V(\mathbf{PA}_V, \boldsymbol{U}_V)$ where observed parents $\mathbf{PA}_V \subseteq \boldsymbol{V}$ and unobserved parents $\boldsymbol{U}_V \subseteq \boldsymbol{U}$. Drawing values of exogenous variables $\boldsymbol{U}$ following $P(\boldsymbol{U})$ induces the *observational distribution* over endogenous variables $\boldsymbol{V}$,

$$P(\boldsymbol{v}) = \int_{\Omega_{\boldsymbol{U}}} \prod_{V \in \boldsymbol{V}} \mathbf{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} dP(\boldsymbol{u}). \quad (3)$$

Each SCM M is associated with a *causal graph* $\mathcal{G}$ (e.g., Fig. 1), that is a Directed Acyclic Graph (DAG) where nodes represent endogenous variables $\boldsymbol{V}$ and exogenous variables $\boldsymbol{U}$, and arrows represent the arguments $\mathbf{PA}_V, \boldsymbol{U}_V$ of each function f_V. A path from a node X to a node Y in $\mathcal{G}$ is a sequence of edges that does not include a particular node more than once. For convenience, we will consider projections of $\mathcal{G}$ onto $\boldsymbol{V}$, in which exogenous variables are made implicit. In particular, we represent a path of the form $V_i \leftarrow U_k \rightarrow V_j$ between endogenous $V_i, V_j \in \boldsymbol{V}$ via an exogenous $U_k \in \boldsymbol{U}$ as a *bi-directed edge* between V_i and V_j, denoted by $V_i \dashleftarrow\!\dashrightarrow V_j$. We will leverage a special type of clustering of nodes in the graph $\mathcal{G}$ called the *confounded-component* (or c-component for short) from Tian and Pearl (Tian and Pearl 2002). For a causal graph $\mathcal{G}$, a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ is a c-component if any pair $V_i, V_j \in \boldsymbol{C}$ is connected by a bi-directed path in $\mathcal{G}$. For example, the (implicit) exogenous variables $\boldsymbol{U}_Z$, $\boldsymbol{U}_{XY}$ in the IV graph in Fig. 1a corresponds to c-components $\boldsymbol{C}(\boldsymbol{U}_Z) = \{Z\}$ and $\boldsymbol{C}(\boldsymbol{U}_{XY}) = \{X, Y\}$, respectively. Lastly, we will use standard graph-theoretic family abbreviations to represent graphical relationships. In particular, the set of parent nodes of $\boldsymbol{X}$ in $\mathcal{G}$ is denoted by $pa(\boldsymbol{X})_{\mathcal{G}} = \cup_{X \in \boldsymbol{X}} pa(X)_{\mathcal{G}}$; and its capitalized version Pa

(a) IV (b) Front-door (c) Verma (d)

Figure 1: Example of graphs. Bi-directed edges denote the presence of an unobserved confounder.

includes the argument as well, e.g. $Pa(\boldsymbol{X})_{\mathcal{G}} = pa(\boldsymbol{X})_{\mathcal{G}} \cup \boldsymbol{X}$. For a more detailed survey on SCMs, we refer to (Pearl 2009; Bareinboim et al. 2022).

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 $P(\mathcal{G} \mid \bar{\boldsymbol{v}})$ be the probability that $\mathcal{G}$ defines the causal structure in the underlying SCM given an $i.i.d$ sample $\bar{\boldsymbol{v}} = \left\{\boldsymbol{v}^{(s)} : s = 1, \ldots, n\right\}$.

Definition 1 (Bayesian scoring criterion)**.** *The Bayesian scoring criterion is defined as the posterior,*

$$P(\mathcal{G} \mid \bar{\boldsymbol{v}}) \propto P(\mathcal{G})P(\bar{\boldsymbol{v}} \mid \mathcal{G}) = P(\mathcal{G}) \int_{\Omega_{\boldsymbol{\omega}}} P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}) dP(\boldsymbol{\omega} \mid \mathcal{G}). \quad (4)$$

where $\boldsymbol{\omega}$ refers a particular parameterization, i.e. $\mathcal{F}, P(\boldsymbol{U})$, of the set of SCMs compatible with the functional dependencies specified by $\mathcal{G}$.

In systems described by arbitrary acyclic causal graphs, an explicit approximation of the marginal likelihood $P(\bar{\boldsymbol{v}} \mid \mathcal{G})$ is typically intractable from both a conceptual and computational perspective. From a conceptual perspective, the graph $\mathcal{G}$ does not define a specific latent variable structure, i.e. domain of $\boldsymbol{U}$ and distribution $P(\boldsymbol{U})$, which, in principle, may be arbitrarily complex. The space of distributions $P(\boldsymbol{V})$ encoded by such a system does not necessarily have a systematic, generic parameterization $\boldsymbol{\omega}$ without making strong assumptions on the form of $\mathcal{F}$ and $P(\boldsymbol{U})$. From a computational perspective, for large classes of SCMs, likelihoods are typically multi-modal and complex and are challenging to integrate over potentially high-dimensional parameter spaces. In the following sections, we present several results to consistently *parameterize* and *estimate* marginal likelihoods for arbitrary causal graphs.

Parameterizations Capturing All Observational Constraints

We seek to develop general results without (untestable) assumptions over unobserved features of the underlying SCMs, i.e. $P(\boldsymbol{U})$ and $\mathcal{F}$. In systems of discrete observables, $P(\boldsymbol{V})$ has the particularity of being consistently defined by a *finite* set of probabilities, irrespective of the underlying structure $P(\boldsymbol{U})$ and $\mathcal{F}$ from which it is derived. We focus our attention on SCMs with *discrete* endogenous (observed) variables, that is, each $V \in \boldsymbol{V}$ taking values in a finite space of outcomes, while each $U \in \boldsymbol{U}$ is *arbitrarily defined*, e.g. taking values in $\mathbb{R}$, and each $f \in \mathcal{F}$ is similarly arbitrary. For a given arbitrary graph there exists a general parameterization that is expressive enough to model any data distribution $P(\boldsymbol{V})$. Our analysis rests on this special parameterization.

Proposition 1 (Prop. 2.6 (Zhang, Jin, and Bareinboim 2022))**.** *For any causal graph $\mathcal{G}$, let M be an arbitrary SCM compatible with $\mathcal{G}$. The observational distribution $P(\boldsymbol{V})$ induced by M could be parameterized as*

$$P(\boldsymbol{v} \mid \mathcal{G}, \boldsymbol{\omega}) = \sum_{U \in \boldsymbol{U}} \sum_{u=1,\ldots,d_U} \prod_{V \in \boldsymbol{V}} \mathbf{1}\{\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} = v\} \prod_{U \in \boldsymbol{U}} \theta_u,$$

where $\theta_u := P(U = u)$ defines exogenous probabilities of discrete variables $U \in \boldsymbol{U}$ with cardinality $d_U = \left|\Omega_{Pa(\boldsymbol{C}(U))}\right|$; and each $\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)}$ is a deterministic mapping between finite domains $\Omega_{\mathbf{PA}_V} \times \Omega_{\boldsymbol{U}_V} \mapsto \Omega_V$.

For the sake of space, all proofs are provided in the Appendix. In other words, for any SCM M there exists a SCM N defined by $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\theta})$, given by Prop. 1, such that $P_M(\boldsymbol{V}) = P_N(\boldsymbol{V})$. A similar reasoning does not apply for continuously-valued endogenous variables that would require continuously-valued exogenous variables and therefore an (untestable) choice of parametric family for all variables.

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. (3). Prop. 1 implies that any $P(x, y, z)$ can be equivalently expressed as

$$\sum_{u_1, u_2} \mathbf{1}\{\xi_Z^{(u_1)} = z\}\mathbf{1}\{\xi_X^{(z,u_2)} = x\}\mathbf{1}\{\xi_Y^{(x,u_2)} = y\}\theta_{u_1}\theta_{u_2},$$

for some value of $(\xi_Z, \xi_X, \xi_Y, \theta_{u_1}, \theta_{u_2})$. In particular, θ_{u_1} defines a distribution over a binary domain $\{1, 2\}$ since $|\Omega_{U_1}| = |\Omega_X| = 2$; θ_{u_2} defines a discrete distribution over a finite domain $\{1, \ldots, 8\}$ since $|\Omega_{U_2}| = |\Omega_X| \cdot |\Omega_Y| \cdot |\Omega_Z| = 8$; $\xi_Z : \Omega_{U_2} \mapsto \Omega_Z$ is a deterministic mapping between discrete domains, etc. Statistical constraints between functionals of $P(\boldsymbol{V})$, e.g. conditional independencies, *automatically* correspond to explicit constraints on the parameters that define the joint distribution. For example, any parameterization $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\theta})$ of $P(\boldsymbol{V})$ compatible with the IV graph must satisfy,

$$\sum_{u_2, y} \max_z \mathbf{1}\{\xi_X^{(z,u_2)} = x\}\mathbf{1}\{\xi_Y^{(x,u_2)} = y\}\theta_{u_2} \leqslant 1. \quad (5)$$

In turn, in causal graphs such as Fig. 1b the corresponding parameters are *unconstrained*.

Singular Asymptotics of the Marginal Likelihood

For marginal likelihood computations in practice, large-sample theory has played an overwhelming role in defining

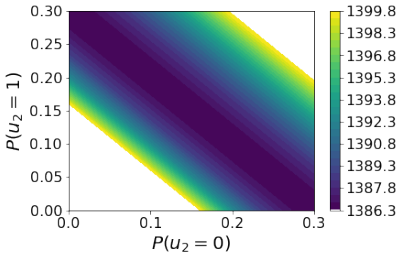


Figure 2: $-\log P(\bar{\boldsymbol{v}} \mid \boldsymbol{\omega}, \mathcal{G})$.

tractable approximations, i.e. scores. Schwarz's Bayesian Information Criterion (BIC), for example, is derived from an asymptotic approximation around maximum likelihood estimates in curved exponential graphical models (Haughton 1988; Schwarz 1978). This asymptotic approximation, however, does not necessarily hold in arbitrary graphs with unobserved confounders; especially those defined by inequality constraints.

In particular, inequalities such as Eq. (2) introduce a boundary in the space of distributions entailed by the underlying graph that induce non-regular likelihood surfaces. For example, in a system described by the IV graph, 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. By Prop. 1, $|\Omega_{U_{XY}}| = 8$, and it can be shown that changing $P(u_{XY})$ while preserving the sums $\sum_{u_2=0,1,2,3} P(u_{XY})$ and $\sum_{u_2=4,5,6,7} P(u_{XY})$ (up to relabelling) does not change the likelihood $P(\bar{\boldsymbol{v}} \mid \boldsymbol{\omega}, \mathcal{G})$. The corresponding log-likelihood, using simulated data from a boundary distribution, is given in Fig. 2 as a function of parameters $P(U_{XY} = 0)$ and $P(U_{XY} = 1)$. The colored pattern represents the likelihood surface that concentrates in a ridge shape along a diagonal line and defines a singular point in the model. In effect, we are losing degrees of freedom in our model and the asymptotic consequences of this fact can be quite severe as approximations can no longer rely on the likelihood around the maximum being a quadratic surface. In general, the BIC will not reflect the asymptotic scaling of $P(\bar{\boldsymbol{v}} \mid \mathcal{G})$ defined by (in)equality constraints.

Watanabe reformulated the foundations of the asymptotic theory of singular models using the Hironaka resolution on singularities (Hironaka 1964; Watanabe 1999, 2001, 2009). A distinct notion of model dimension emerges in singular models driven by the so-called learning coefficient $\lambda_{\mathcal{G}} > 0$ that describes how fast the posterior distribution shrinks with increasing sample size. In the following corollary, we establish the correct approximation to the log marginal likelihood defined by general causal graphs with joint distributions parameterized by discrete SCMs.

Theorem 1. *In discrete SCMs parameterized by Prop. 1,*

$$-\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}) = \\ -\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}_0) + \lambda_{\mathcal{G}} \log n + \mathcal{O}_p(\log \log n), \tag{6}$$

where $\boldsymbol{\omega}_0$ is a set of parameters that produces the true distribution, and $\lambda_{\mathcal{G}}$, called the learning coefficient, is a scalar.

This is a corollary to (Watanabe 1999, Thm. 1). In curved exponential models, $\lambda_{\mathcal{G}}$ is directly proportional to the number of free model parameters but it might not be in general (in fact $\lambda_{\mathcal{G}}$ is strictly smaller than the penalty given by the BIC in distributions with this parameterization involving (in)equality constraints)[1]). In general, $\lambda_{\mathcal{G}}$ depends on the true (unknown) data generating system $\mathcal{G}$ that makes this particular expression difficult to evaluate in practice.

Approximations to the Bayesian Score and Consistency for Structure Learning

A tractable score remains elusive due to the computational and conceptual challenges of evaluating multi-modal integrals and asymptotic approximations, respectively. This section proposes a compromise that involves sampling based on a tempered, i.e. less modal, version of the likelihood and prior that, however, can be shown to relate directly to Thm. 1 and enjoy consistency guarantees. Following (Friel and Pettitt 2008; Watanabe 2013), the idea is to estimate some expectation $\mathbb{E}_{\boldsymbol{\omega} \sim P(\boldsymbol{\omega} \mid \mathcal{G})}[\log P(\bar{\boldsymbol{v}} \mid \boldsymbol{\omega}, \mathcal{G})]$ by evaluating a less modal distribution P^{β} with $\beta < 1$. We define a score $\mathcal{S}_{\text{WBIC}}$[2] for a causal graph $\mathcal{G}$ and data $\bar{\boldsymbol{v}}$ as

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) := -\mathbb{E}_{\beta} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}) \\ = \frac{\int_{\Omega_{\boldsymbol{\omega}}} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}) P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})^{\beta} dP(\boldsymbol{\omega} \mid \mathcal{G})}{\int_{\Omega_{\boldsymbol{\omega}}} P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})^{\beta} dP(\boldsymbol{\omega} \mid \mathcal{G})}. \tag{7}$$

The significance of this definition lies in the fact that for a consistent parameterization of $P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})$, the marginal log-likelihood $\log P(\bar{\boldsymbol{v}} \mid \mathcal{G})$ is provably equal to $\mathbb{E}_{\beta} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})$ for some value $\beta \in [0, 1]$, with the property that, for the choice $\beta = \frac{1}{\log n}$ it holds, asymptotically by (Watanabe 2013, Thm. 4) that,

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) = -\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}) + \mathcal{O}_p(\sqrt{\log n}). \tag{8}$$

This result shows that model selection using $\mathcal{S}_{\text{WBIC}}$ approximates a Bayesian procedure seeking the model with the highest posterior probability, i.e. Thm. 1. However, $\mathcal{S}_{\text{WBIC}}$ 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 $\log P(\bar{\boldsymbol{v}} \mid \mathcal{G})$ between two different models, which is made precise in the following assumptions.

Assumption 1. *If $\mathcal{G}_1$ is compatible with the data generating distribution P and $\mathcal{G}_2$ is not, then there exists a scalar $c_{12} > 0$ such that $\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}_1) - \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}_2) > c_{12} n$, with probability tending to 1 as $n \to \infty$.*

Assumption 2. *Let causal graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ be defined such that the set of distributions $\mathcal{P}_1$ compatible with $\mathcal{G}_1$ is included in the set of distributions $\mathcal{P}_2$ compatible with $\mathcal{G}_2$. Then, $\lambda_{\mathcal{G}_1} < \lambda_{\mathcal{G}_2}$ with probability tending to 1 as $n \to \infty$, where $\lambda_{\mathcal{G}_1}, \lambda_{\mathcal{G}_2}$ are the learning coefficients in Thm. 1 corresponding to $\mathcal{G}_1, \mathcal{G}_2$ respectively.*

[1] A more detailed exposition of asymptotics in singular models, including of details on thermodynamic integration used in the following section are given in the Appendix.

[2] This expression is also known as the Widely applicable Bayesian Information Criterion (WBIC) (Watanabe 2013).

As the log-likelihood is the sum of logarithmic probabilities for $i.i.d$ observations, if causal graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ encode a similar number of unobserved confounders with a similar underlying parameterization, we can expect the difference in log-likelihoods for $\mathcal{G}_1$ and $\mathcal{G}_2$ to scale linearly with sample size so that Assumption 1 generally holds (if close enough models are compared). The learning coefficient $\lambda_{\mathcal{G}}$ in Thm. 1 acts as a measure of the complexity of the set of distributions induced by an SCM. Assumption 2 states that SCMs inducing more probabilistic constraints also induce families of distributions that are less general and thus an underlying graphical model that is less complex in the sense of $\lambda_{\mathcal{G}}$. Both assumptions can be found in other treatments of model selection, see e.g. (Drton and Plummer 2017). With these assumptions, $\mathcal{S}_{\text{WBIC}}$ coupled with the discrete parameterization of the likelihood assigns the lowest (best) score to the model imposing the fewest constraints that can represent the generative distribution.

Theorem 2. *Let $P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})$ be parameterized as in Prop. 1. Under Assumptions 1 and 2, with probability tending to 1 as $n \to \infty$,*

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

The first part of the theorem 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 simpler one will be preferred over the more complex one. This property is also called *consistency* of a score and is key to ensure convergence to the underlying graph that summarizes the SCM that generated the data. As a consequence of the consistency of the score in the space of arbitrary causal graphs, the score captures all statistical constraints over observational probabilities encoded by the structure of the causal graph.

Proposition 2. *$\mathcal{S}_{WBIC}(\mathcal{G}, \bar{\boldsymbol{v}})$ distinguishes between candidate causal graphs differing on an (in)equality constraint between functionals of $P(\boldsymbol{V})$ with probability tending to 1 as $n \to \infty$.*

Intuitively, if Prop. 2 were not to hold, $\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}})$ would not be sound or parsimonious as two candidate graphs that disagree on (in)equality constraints also define two different sets of compatible distributions. If the in(equality) is satisfied in $P(\boldsymbol{V})$, a parsimonious score chooses the graph entailing the (in)equality, else if the inequality is not satisfied a sound score chooses the graph not entailing the (in)equality. It is worth noting also that $\mathcal{S}_{\text{WBIC}}$ may be interpreted as a generalization of the BIC score, denoted $\mathcal{S}_{\text{BIC}}$.

Proposition 3 (Eq. (32) in (Watanabe 2013))**.** *Let $P(\boldsymbol{V})$ and $\mathcal{G}$ be the joint distribution and causal graph induced by an SCM parameterized by curved exponential models. Then, with probability tending to 1 as $n \to \infty$, $\mathcal{S}_{WBIC}(\mathcal{G}, \bar{\boldsymbol{v}}) = \mathcal{S}_{BIC}(\mathcal{G}, \bar{\boldsymbol{v}}) + \mathcal{O}_p(1)$.*

Properties of Score for Causal Discovery and Computation

This section describes 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 graph.

Definition 2 (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 $\boldsymbol{C}$ and its parents,*

$$\mathcal{S}(\mathcal{G}, \bar{\boldsymbol{v}}) = \sum_{\boldsymbol{C} \in \boldsymbol{C}(\mathcal{G})} \mathcal{S}(\mathcal{G}_{Pa(\boldsymbol{C})}, \bar{\boldsymbol{v}}_{Pa(\boldsymbol{C})}). \tag{9}$$

Here $\mathcal{G}_{Pa(\boldsymbol{C})}$ and $\boldsymbol{V}_{Pa(\boldsymbol{C})}$ denote the subgraph and data, respectively, restricted to $Pa(\boldsymbol{C}) \subseteq \boldsymbol{V}$.

Proposition 4. *$\mathcal{S}_{WBIC}$ is decomposable.*

Decomposability will avoid the need to recompute 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 Z only while the second one being a function of $\{X, Y, Z\}$. If we were to add an edge $Z \to Y$ we would only need to recompute the updated c-component $\{X, Y\}$ as the one for $\{Z\}$ can be re-used. An important observation 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 that can be defined by the $\mathcal{S}_{\text{WBIC}}$.

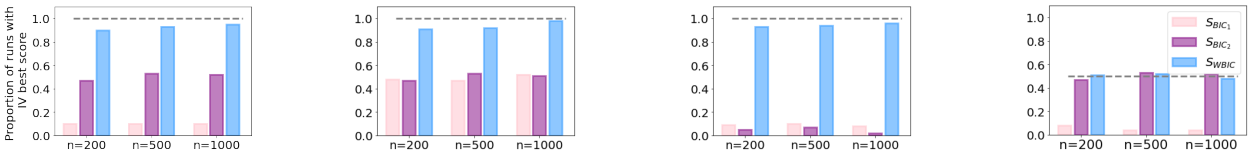
Definition 3 (Score equivalence)**.** *A scoring criterion $\mathcal{S}$ is score equivalent if, for any pair of causal graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ that are compatible with the same family of distributions, $\mathcal{S}(\mathcal{G}_1, \bar{\boldsymbol{v}}) = \mathcal{S}(\mathcal{G}_2, \bar{\boldsymbol{v}})$ with probability tending to 1 as $n \to \infty$.*

Proposition 5. *$\mathcal{S}_{WBIC}$ is score equivalent.*

This proposition formalizes the intuition that if the family of distributions entailed by two graphs are equal then also their scores will be equal. For example, adding a bi-directed edge $Z \leftrightarrow X$ to the graph in Fig. 1a does not remove/add any constraints on the set of induced distributions $P(\boldsymbol{V})$ and has, therefore, the same score.

Computing the Score

We present in this section an MCMC sampler to approximate the expectation defining $\mathcal{S}_{\text{WBIC}}$ in Eq. (7). Let $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\theta})$, where $\boldsymbol{\xi} = \{\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} : V \in \boldsymbol{V}, Pa_V \subset \boldsymbol{V}, \boldsymbol{U}_V \subset \boldsymbol{U}\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_U : U \in \boldsymbol{U}\}$ denote all possible functional assignments and exogenous probabilities, respectively. More specifically, $\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)}$ are parameters that take values in Ω_V and represent the assignment of V given its parents and exogenous variables, $i = 1, \ldots, d$. There is one such parameter of dimensionality $|\Omega_V|$ for each combination of realization of parent variables $\boldsymbol{pa}_V$ and exogenous variables $\boldsymbol{u}_V$ that are defined by the candidate causal graph



(a) IV vs DAG. (b) IV vs Front-door. (c) IV vs Unconstrained. (d) IV vs Equivalent.

Figure 3: Quality of scores. The horizontal gray line indicates the theoretical optimum.

$\mathcal{G}$. $\boldsymbol{\theta}_U$ stands for the vector of probabilities that defines the discrete distribution $P(U = u)$ over its finite domain $u \in \{1, \dots, d_U\}$.

$\mathcal{S}_{\text{WBIC}}$ is computed by setting the tempering temperature $\beta := 1/\log n$ and prior over parameters given $\mathcal{G}$ (possibly uninformative), and drawing Monte Carlo samples of the posterior distribution $P(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \mathcal{G})^\beta$ 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 $\bar{\boldsymbol{v}}$ but also depend on (unobserved) exogenous variables $\bar{\boldsymbol{u}} = \{\boldsymbol{u}^{(s)} : s = 1, \dots, n\}$. For every $V \in \boldsymbol{V}, \forall \boldsymbol{pa}_V, \boldsymbol{u}_V$, the functional assignment parameters $\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)}$ are drawn uniformly in the discrete domain Ω_V. For every $U \in \boldsymbol{U}$, exogenous probabilities $\boldsymbol{\theta}_U$ with dimension $d_U = \prod_{V \in \boldsymbol{C}_U} |\Omega_{Pa(V)}|$ are drawn from a prior Dirichlet distribution $\boldsymbol{\theta}_U = (\theta_1, \dots, \theta_{d_U}) \sim Dir(\alpha_1, \dots, \alpha_{d_U})$, with hyperparameters $\alpha_1, \dots, \alpha_{d_U}$. Fix some initial value for all unobserved quantities $(\boldsymbol{u}, \boldsymbol{\xi}, \boldsymbol{\theta})$, and sample each one iteratively conditioned on the current values of the remaining terms with a Metropolis step.

- Exogenous variables $\boldsymbol{U}^{(s)}$ are mutually independent given $\boldsymbol{V}^{(s)}, \boldsymbol{\xi}, \boldsymbol{\theta}$ and thus we can sample each separately using the conditional

$$P(\boldsymbol{u}^{(s)} \mid \boldsymbol{v}^{(s)}, \boldsymbol{\xi}, \boldsymbol{\theta}) \propto P(\boldsymbol{u}^{(s)}, \boldsymbol{v}^{(s)} \mid \boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{V \in \boldsymbol{V}} \mathbf{1}\{\xi_V^{(\boldsymbol{u}_V^{(s)}, \boldsymbol{pa}_V^{(s)})} = v^{(s)}\} \prod_{U \in \boldsymbol{U}} \theta_{u^{(s)}}.$$

- Similarly, for fixed $\boldsymbol{pa}_V, \boldsymbol{u}_V$, parameters $\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)}$ are mutually independent given $\bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}, \boldsymbol{\theta}$. As they represent a mapping between variables, its conditional distribution is given by $P(\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} = v \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}) = 1$ if there exists a sample $(v^{(s)}, \boldsymbol{pa}_V^{(s)}, \boldsymbol{u}_V^{(s)})$ that fixes the mapping $\boldsymbol{pa}_V^{(s)}, \boldsymbol{u}_V^{(s)} \mapsto v^{(s)}$. Otherwise, $P(\xi_V^{(\boldsymbol{u}_V, \boldsymbol{pa}_V)} = v) = q_v$, where $\boldsymbol{q} = \{q_v : v \in \Omega_V\}$ is a proposal distribution that samples $\xi_V^{(\boldsymbol{u}_V, \boldsymbol{pa}_V)}$ in Ω_V with probabilities that are uniformly updated in a small neighborhood of the previous parameter value in each iteration of the sampler.
- Fix $U \in \boldsymbol{U}$. Given $\bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}, \boldsymbol{\theta}_U$ is independent of $\boldsymbol{\xi}$ and is given by a Dirichlet distribution $\boldsymbol{\theta}_U \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}} \sim Dir(\beta_1, \dots, \beta_{d_U})$ where $\beta_j := \alpha_j + c_j$ where c_j is updated in each iteration of the sampler using a uniform proposal distribution, e.g. $c_j \sim \text{Uniform}(c_j - \epsilon, c_j + \epsilon)$ 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 $P(\boldsymbol{\xi}_{(t+1)}, \boldsymbol{\theta}_{(t+1)} \mid \bar{\boldsymbol{v}}, \mathcal{G})^\beta / P(\boldsymbol{\xi}_{(t)}, \boldsymbol{\theta}_{(t)} \mid \bar{\boldsymbol{v}}, \mathcal{G})^\beta$ where,

$$P(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \mathcal{G})^\beta \propto \exp\{-\beta \log P(\bar{\boldsymbol{v}} \mid \boldsymbol{\xi}, \boldsymbol{\theta}, \mathcal{G}) + \log P(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \mathcal{G})\}.$$

Finally, $\mathcal{S}_{\text{WBIC}}$'s approximation:

$$\hat{\mathcal{S}}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) := -\frac{1}{T} \sum_{t=1}^{T} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\xi}_{(t)}, \boldsymbol{\theta}_{(t)}). \quad (10)$$

Our next results establish finite-sample deviation bounds for empirical estimates $\hat{\mathcal{S}}_{\text{WBIC}}$ defined in Eq. (10). In particular, we apply standard concentration inequalities (Hoeffding 1994, Thm. 2) in Prop. 6 to determine a sufficient number of *independent* posterior samples T required for obtaining accurate estimates of the "population-level" score $\mathcal{S}_{\text{WBIC}}$ defined in Eq. (7). For this, in addition, we will require a positivity constraint on the likelihood.

Proposition 6. Assume that the absolute value of $\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\xi}_{(t)}, \boldsymbol{\theta}_{(t)}), t = 1, \dots, T$ be bounded from above by $M > 0$. Then, for $\delta \in (0, 1)$ we have that with probability at least $1 - \delta$,

$$\left| \hat{\mathcal{S}}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) - \mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) \right| \leqslant \sqrt{\frac{1}{2T} M^2 \log(2/\delta)}. \quad (11)$$

Experiments: Quality of Scores

This section evaluates the ability of the proposed score to distinguish between graphs that differ in equality and inequality constraints[3].

We consider variations of the IV (Fig. 1a) graph that are designed to consider the presence and absence of inequality constraints[4]. The task is to score these variations and compare them to scores of the ground truth IV graph, based on data generated from 100 different SCMs $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P \rangle$ compatible with the ground truth graph. Each SCM is specified as follows. Exogenous distributions $P(U), U \in \boldsymbol{U}$ are randomly chosen from a set of continuous distributions {Gaussian, Exponential, Gumbel, Uniform}; functional associations are defined by $V \leftarrow g(f(\beta \boldsymbol{Pa}_V + \alpha \boldsymbol{U}_V)), V \in \boldsymbol{V}$,

[3]Further algorithmic details, as well as evaluations of decomposability and equivalence, can be found in the Appendix.

[4]Plots of graphs are given in Fig. 5 in the Appendix. A similar analysis on the Verma graph (Fig. 1c), designed to consider the presence and absence of equality constraints, is given in the Appendix.

with f randomly chosen as a linear, trigonometric $(\cos, \sin)$, or logarithmic function; α, β uniformly chosen in $[0, 1]$ with the required dimensionality; and g a step function used to define a binary outcome. For comparison, we consider two implementations of the BIC used in the literature: $\mathcal{S}_{\text{BIC}_1} := -2 \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \hat{\boldsymbol{\omega}}) + |\Omega_{\boldsymbol{\omega}}| \log n$, and $\mathcal{S}_{\text{BIC}_2} := -2 \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \hat{\boldsymbol{\omega}}) + (2|\boldsymbol{V}| + |\mathcal{E}|) \log n$, where $|\mathcal{E}|$ denotes the number of directed and bi-directed edges. Our results are summarized in Fig. 3. Each bar gives the proportion of experiments (out of 100) in which the correct causal explanation, i.e. the IV graph, is scored better than a competing graph that differs in subtle ways.

By design, baseline scores do not correctly appreciate the complexity of the class of distributions implied by the graphs which can be illustrated in specific comparisons. For instance, Fig. 3a compares the ground truth IV graph with an unconstrained DAG that voids the inequality constraint while having the same number of edges but fewer parameters. In particular, $\mathcal{S}_{\text{BIC}_1}$ incorrectly favors the DAG in most cases as a consequence of its lower complexity term $|\Omega_{\boldsymbol{\omega}}| \log n$, and $\mathcal{S}_{\text{BIC}_2}$ scores both graph equally on average as both graphs have equal fit and number of edges $|\mathcal{E}|$. In turn, Fig. 3b considers an unconstrained graph with both the same number of edges and parameters as the IV model (therefore equal, on average, $\mathcal{S}_{\text{BIC}_1}$ and $\mathcal{S}_{\text{BIC}_2}$ scores). The IV and unconstrained graphs can be distinguished empirically due to the differing inequality constraints. In contrast, Fig. 3c is also considered an unconstrained graph although this time with fewer edges and fewer parameters: and thus better $\mathcal{S}_{\text{BIC}_1}$ and $\mathcal{S}_{\text{BIC}_2}$ scores. Fig. 3d considers a model for $P(\boldsymbol{V})$ that is equivalent to the IV model, i.e. $Z \to X$ is replaced with $Z \dashleftarrow\dashrightarrow X$ (and thus have a different number of parameters) as both induce a single inequality constraint. Theoretically, the two alternatives cannot be distinguished and we would expect scores to be equal on average. We conclude with the observation that across variations of different graphs and sample sizes, $\mathcal{S}_{\text{WBIC}}$ correctly scores graphs based on inequality constraints and appreciates equivalence in the space of distributions $P(\boldsymbol{V})$ induced by graphs even if those have differing number of edges or parameters.

Experiments: Structure Learning

This section explores the use of $\mathcal{S}_{\text{WBIC}}$ within search algorithms to recover the causal graph that best describes the statistical constraints found in data. We adopt a greedy search algorithm to use the decomposable nature of $\mathcal{S}_{\text{WBIC}}$, denoted GS-$\mathcal{S}_{\text{WBIC}}$; pseudocode is given in the Appendix. Several methods exist for searching over spaces of graphs, including greedy search (Triantafillou and Tsamardinos 2016), exact dynamic programming (Rantanen, Hyttinen, and Järvisalo 2021), integer programming (Chen, Dash, and Gao 2021), and gradient-based optimization (Bhattacharya et al. 2021; Bellot and van der Schaar 2021) methods. Existing implementations rely on Drton's Residual Iterative Conditional Fitting algorithm for maximum likelihood estimation of the BIC score which applies to *linear Gaussian models* (Drton and Richardson 2012). Empirical comparisons are made with Gaussian-based continuous-optimization algorithm

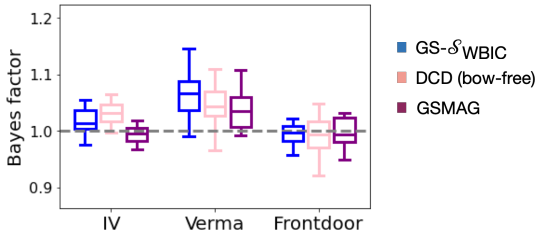


Figure 4: Bayes factor vs optimal DAG.

(DCD) (Bhattacharya et al. 2021) for recovering ancestral and bow-free graphs, the GES algorithm (Chickering 2002a) for recovering directed graphs, and the GSMAG algorithm (Triantafillou and Tsamardinos 2016) for recovering MAGs.

We start by considering graphs returned by each method fit on random datasets from the IV, Verma, and front door models defined in the previous section. The objective is to understand the relative gain of searching over larger spaces of graphs, beyond the spaces of bow-free, ancestral, and directed graphs considered in the literature. The IV graph is in neither of these classes, the Verma graph is bow-free, and the frontdoor graph is bow-free. Fig. 4 plots Bayes factors in comparison to the optimal DAG (inferred with GES). There is some variation over different datasets although we observe that on average searching over larger spaces eventually returns graphs that are more likely for the IV and Verma models (Bayes factor larger than 1). The frontdoor graph is the only model that is empirically indistinguishable from a fully connected DAG, which sets a bound of 1 in theory on the Bayes factor. In the Appendix, we further compare structure learning baselines on Sachs (Sachs et al. 2005) and Lung cancer (Lauritzen and Spiegelhalter 1988) benchmark datasets (with some variables omitted to induce unobserved confounding). There we show that greedy search in the space of arbitrary causal graphs with the proposed score can be competitive for causal discovery on real datasets.

Conclusions

We investigated the problem of learning the causal structure underlying a phenomenon of interest in discrete models with arbitrary latent dependencies. Our contribution is a new 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. To our knowledge, this score is the first to apply to arbitrary models of unobserved confounding. 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.

References

Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. 2022. On Pearl's Hierarchy and the Foundations of Causal Infer-

ence. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 507–556. NY, USA: Association for Computing Machinery, 1st edition.

Bellot, A.; and van der Schaar, M. 2021. Deconfounded score method: Scoring DAGs with dense unobserved confounding. *arXiv preprint arXiv:2103.15106*.

Bhattacharya, R.; Nagarajan, T.; Malinsky, D.; and Shpitser, I. 2021. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, 2314–2322. PMLR.

Bonet, B. 2013. Instrumentality tests revisited. *arXiv preprint arXiv:1301.2258*.

Carathéodory, C. 1911. Über den Variabilitätsbereich der Fourier'schen Konstanten von positiven harmonischen Funktionen. *Rendiconti Del Circolo Matematico di Palermo (1884-1940)*, 32(1): 193–217.

Chen, R.; Dash, S.; and Gao, T. 2021. Integer Programming for Causal Structure Learning in the Presence of Latent Variables. In *International Conference on Machine Learning*, 1550–1560. PMLR.

Chickering, D. M. 2002a. Learning equivalence classes of Bayesian-network structures. *The Journal of Machine Learning Research*, 2: 445–498.

Chickering, D. M. 2002b. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov): 507–554.

Drton, M. 2009a. Discrete chain graph models.

Drton, M. 2009b. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2): 979–1012.

Drton, M.; and Plummer, M. 2017. A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2): 323–380.

Drton, M.; and Richardson, T. S. 2008. Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2): 287–309.

Drton, M.; and Richardson, T. S. 2012. Iterative conditional fitting for Gaussian ancestral graph models. *arXiv preprint arXiv:1207.4118*.

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

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

Gelman, A.; Carlin, J. B.; Stern, H. S.; and Rubin, D. B. 1995. *Bayesian data analysis*. Chapman and Hall/CRC.

Glymour, C.; Scheines, R.; and Spirtes, P. 2014. *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press.

Hartemink, A. J.; Gifford, D. K.; Jaakkola, T. S.; and Young, R. A. 2000. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Biocomputing 2001*, 422–433. World Scientific.

Haughton, D. M. 1988. On the choice of a model to fit data from an exponential family. *The annals of statistics*, 342–355.

Heckerman, D.; Meek, C.; and Cooper, G. 1999. A Bayesian approach to causal discovery. *Computation, causation, and discovery*, 19: 141–166.

Hironaka, H. 1964. Resolution of singularities of an algebraic variety over a field of characteristic zero: II. *Annals of Mathematics*, 205–326.

Hoeffding, W. 1994. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, 409–426.

Lauritzen, S. L.; and Spiegelhalter, D. J. 1988. 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.

Marx, A.; Gretton, A.; and Mooij, J. M. 2021. A weaker faithfulness assumption based on triple interactions. In de Campos, C. P.; Maathuis, M. H.; and Quaeghebeur, E., eds., *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*, 451–460. AUAI Press.

Meek, C. 1995. Strong Completeness and Faithfulness in Bayesian Networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, 411–418. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

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

Pearl, J. 1995. On the Testability of Causal Models with Latent and Instrumental Variables. In Besnard, P.; and Hanks, S., eds., *Uncertainty in Artificial Intelligence 11*, 435–443. San Francisco, CA: Morgan Kaufmann.

Pearl, J. 2009. *Causality*. Cambridge university press.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. MIT press.

Rantanen, K.; Hyttinen, A.; and Järvisalo, M. 2021. Maximal ancestral graph structure learning via exact search. In *Uncertainty in Artificial Intelligence*, 1237–1247. PMLR.

Richardson, T.; and Spirtes, P. 2002. Ancestral graph Markov models. *The Annals of Statistics*, 30(4): 962–1030.

Richardson, T. S.; Evans, R. J.; Robins, J. M.; and Shpitser, I. 2017. Nested Markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*.

Rosset, D.; Gisin, N.; and Wolfe, E. 2018. Universal bound on the cardinality of local hidden variables in networks. *Quantum Information and Computation*, 18(11&12): 0910–0926.

Rubin, H.; and Wesler, O. 1958. A note on convexity in Euclidean n-space. In *Proc. Amer. Math. Soc*, volume 9, 522–523.

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

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*, 461–464.

Shen, N.; and González, B. 2021. Bayesian information criterion for linear mixed-effects models. *arXiv preprint arXiv:2104.14725*.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Tian, J.; and Pearl, J. 2002. *A general identification condition for causal effects*. eScholarship, University of California.

Triantafillou, S.; and Tsamardinos, I. 2016. Score-based vs Constraint-based Causal Learning in the Presence of Confounders. In *CFA@ UAI*, 59–67.

Uhler, C.; Raskutti, G.; Bühlmann, P.; and Yu, B. 2013. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 436–463.

Verma, T.; and Pearl, J. 1990. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, 69–76. Elsevier.

Watanabe, S. 1999. Algebraic analysis for singular statistical estimation. In *Algorithmic Learning Theory: 10th International Conference, ALT'99 Tokyo, Japan, December 6–8, 1999 Proceedings 10*, 39–50. Springer.

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

Watanabe, S. 2009. *Algebraic geometry and statistical learning theory*. Cambridge university press.

Watanabe, S. 2013. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar): 867–897.

Wolfe, E.; Spekkens, R. W.; and Fritz, T. 2019. The inflation technique for causal inference with latent variables. *Journal of Causal Inference*, 7(2).

Zhang, J. 2006. *Causal inference and reasoning in causally insufficient systems*. Ph.D. thesis, Department of Philosophy, Carnegie Mellon University.

Zhang, J.; Jin, T.; and Bareinboim, E. 2022. Partial Counterfactual Identification from Observational and Experimental Data. In *Proceedings of the 39th International Conference on Machine Learning (ICML-22)*.