TECHNICAL REPORT
R-83
First version: May 2022
Last version: Jun 2023

Scores for Learning Discrete Causal Graphs with Unobserved Confounders

Alex Bellot
Deepmind, UK
ab5305@cs.columbia.edu

Junzhe Zhang
Columbia University, USA
junzhez@cs.columbia.edu

Elias Bareinboim
Columbia University, USA
eb@cs.columbia.edu

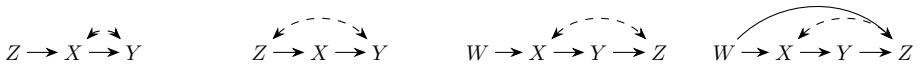
Abstract

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 independences. 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, in principle, to break the problem of structural learning in smaller and 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 model represented by a causal graph, where nodes denote random variables (measured or latent) and directed edges denote causal effects from tails to arrowheads [26, 36, 27]. 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 [24, 23, 39, 46, 22]. This is the cornerstone assumption for a plethora of structure learning algorithms [40, 16, 36]. In fact, when all variables are observable, d-separation statements capture *all* testable constraints implied by the underlying causal model [40].



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

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

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 [40], 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

$$\sum_{x} P(z \mid x, y)P(x) = \sum_{x} P(z \mid x, y, w)P(x \mid w). \tag{1}$$

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 [25],

$$\sum_{y} \max_{z} P(x, y \mid z) = \sum_{u_2, y} \max_{z} P(x \mid z, u_2)P(y \mid x, u_2)P(u_2) \leqslant 1. \tag{2}$$

The same inequality does not hold in the (otherwise statistically equivalent) unconstrained graph in Fig. 1b. In systems with discrete observables, distributions induced by causal graphs are indeed *always* restricted whenever two observed variables are not directly connected, that is are neither adjacent, nor subject to unobserved confounding[13]. 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 in fact 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 [40, 36]. 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 $P(\boldsymbol{V})$ [29]. 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 [15, 19, 6, 7]. 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 [18, 34]). Several more general causal abstractions, such as discrete chain graph models [8], fully bi-directed graph models [11], and discrete nested Markov models [30] 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 is applicable to systems with discretely-valued observables and arbitrarily defined exogenous variables. Building on Watanabe's asymptotic expansion of the marginal likelihood [43] and bounds over the cardinality of exogenous variables [31, 47], 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.

1.1 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, $\mathbb{1}\{\cdot\}$ is the indicator function that equals 1 if the statement in $\{\cdot\}$ evaluates to be true, and equals 0 otherwise.

The basic framework of our analysis rests on *structural causal models* (SCMs) [26, 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, $V \leftarrow f_V(\mathbf{Pa}_V, \boldsymbol{U}_V), \mathbf{Pa}_V \subseteq \boldsymbol{V}, \boldsymbol{U}_V \subseteq \boldsymbol{U}$. Drawing values of exogenous variables $\boldsymbol{U}$ following $P(\boldsymbol{U})$ induces the *observational distribution* over endogenous variables $\boldsymbol{V}$,

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

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 which does not include a particular node more than once. For convenience, we will consider projections of $\mathcal{G}$ onto $\mathbf{V}$, 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 $V_i \dashleftarrow \cdot \dashrightarrow V_j$.

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 [37]. 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 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 readers to [26, 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 $P(\mathcal{G} \mid \bar{\boldsymbol{v}})$ 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,*

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

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 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.

2.1 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 [47])**.** *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*

$$P(\boldsymbol{v} \mid \mathcal{G}, \boldsymbol{\omega}) = \sum_{U \in \boldsymbol{U}} \sum_{u=1,\dots,d_U} \prod_{V \in \boldsymbol{V}} \mathbb{1}\{\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} = v\} \prod_{U \in \boldsymbol{U}} \theta_u, \tag{5}$$

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 Appendix B. 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 a (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

$$\sum_{u_1, u_2} \mathbb{1}\{\xi_Z^{(u_1)} = z\} \mathbb{1}\{\xi_X^{(z,u_2)} = x\} \mathbb{1}\{\xi_Y^{(x,u_2)} = y\} \theta_{u_1} \theta_{u_2}, \tag{6}$$

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 $\{1, \dots, 8\}$ 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,

$$\sum_{u_2, y} \max_z \mathbb{1}\{\xi_X^{(z,u_2)} = x\} \mathbb{1}\{\xi_Y^{(x,u_2)} = y\} \theta_{u_2} \leqslant 1. \tag{7}$$

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

2.2 Singular asymptotics of the marginal likelihood

For marginal likelihood computations in practice, large-sample theory has played an overwhelming role to define 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 [18, 34]. 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 loosing 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.

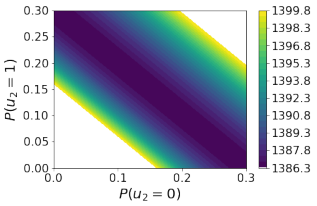


Figure 2: $-\log P(\bar{\boldsymbol{v}} \mid \boldsymbol{\omega}, \mathcal{G})$.

Watanabe reformulated the foundations of asymptotic theory of singular models using the Hironaka resolution on singularities [20, 41, 42, 43]. 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 a general causal graphs with joint distributions parameterized by discrete SCMs.

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

$$-\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}) = -\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}_0) + \lambda_{\mathcal{G}} \log n + \mathcal{O}_p(\log \log n), \tag{8}$$

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

This is a corollary to [41, Thm. 1]. In curved exponential models, $\lambda_{\mathcal{G}}$ is directly proportional the number of free model parameters but it might not be in general (in fact $\lambda_{\mathcal{G}}$ is strictly smaller that 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.

2.3 Approximations to the Bayesian score and consistency for structure learning

A tractable score remains elusive due to 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 [14, 44], the idea is to estimate some expectation $\mathbb{E}_{\boldsymbol{\omega} \sim P}[P(\bar{\boldsymbol{v}} \mid \boldsymbol{\omega}, \mathcal{G})]$ 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

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) := -\mathbb{E}_{\beta} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}) = \frac{\int_{\Omega_{\boldsymbol{\omega}}} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}) P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})^{\beta} dP(\boldsymbol{\omega} \mid \mathcal{G})}{\int_{\Omega_{\boldsymbol{\omega}}} P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})^{\beta} dP(\boldsymbol{\omega} \mid \mathcal{G})}. \tag{9}$$

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 likelihood $P(\bar{\boldsymbol{v}} \mid \mathcal{G})$ 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 [44, Thm. 4] that,

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) = -\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}) + \mathcal{O}_p(\sqrt{\log n}). \tag{10}$$

This result shows that model selection using $\mathcal{S}_{\text{WBIC}}$ approximates a Bayesian procedure seeking the model with highest posterior probability, i.e. Thm. 1. However, $\mathcal{S}_{\text{WBIC}}$ may deviate from the marginal

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

[2] In the Bayesian model selection literature, this expression is known as the Widely applicable Bayesian Information Criterion (WBIC) [44].

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.*

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 complexity of the set of distributions induced by a 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. [10]. 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 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 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 [44])**.** *Let $P(\boldsymbol{V})$ and $\mathcal{G}$ be the joint distribution and causal graph induced by a SCM parameterized by curved exponential models. Then, with probability tending to 1 as $n \to \infty$, $\mathcal{S}_{WBIC}(\mathcal{G}, \bar{\boldsymbol{v}}) = \mathcal{S}_{BIC}(\mathcal{G}, \bar{\boldsymbol{v}}) + \mathcal{O}_p(1)$.*

3 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 causal 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,*

$$\mathcal{S}(\mathcal{G}, \bar{\boldsymbol{v}}) = \sum_{\boldsymbol{C} \in \boldsymbol{C}(\mathcal{G})} \mathcal{S}(\mathcal{G}_{Pa(\boldsymbol{C})}, \bar{\boldsymbol{v}}_{Pa(\boldsymbol{C})}). \tag{11}$$

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 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 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.

3.1 Computing the score

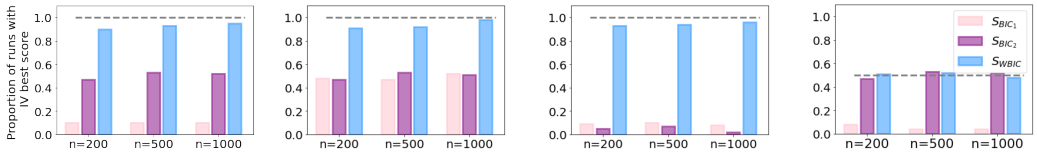
We present in this section a MCMC sampler to approximate the expectation defining $\mathcal{S}_{\text{WBIC}}$ in Eq. (9). Let $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\theta})$, where $\boldsymbol{\xi} = \{\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} : V \in \boldsymbol{V}, Pa_V \subset \boldsymbol{V}, \boldsymbol{U}_V \subset \boldsymbol{U}\}$ 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, $i = 1, \dots, d$. 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 $\mathcal{G}$. $\boldsymbol{\theta}_U$ 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

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

- 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 $P(\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} =$



(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.

$v \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}) = 1$ 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 neighbourhood 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 $\boldsymbol{\theta}_U \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}} \sim Dir(\beta_1, \ldots, \beta_{d_U})$ 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,

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

Finally, $\mathcal{S}_{\text{WBIC}}$'s approximation: $\hat{\mathcal{S}}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) := -\frac{1}{T} \sum_{t=1}^{T} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\xi}_{(t)}, \boldsymbol{\theta}_{(t)})$.

4 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 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 data generated from 100 different SCMs $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P \rangle$ compatible with 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 $V \leftarrow g(f(\beta \boldsymbol{Pa}_V + \alpha \boldsymbol{U}_V)), V \in \boldsymbol{V}$, 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.

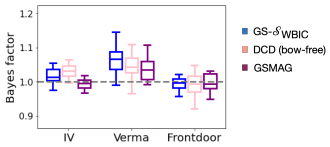
By design, baseline scores do not correctly appreciate the complexity of the class of distributions implied by the graphs which can be illustrated in specific comparisons. For instance, Fig. 3a compares the ground truth IV graph with an unconstrained DAG that voids the inequality constraint while having the same number of edges but fewer parameters. In particular, $\mathcal{S}_{\text{BIC}_1}$ incorrectly favours 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) although IV and unconstrained graphs can be distinguished empirically due to the differing inequality constraint. In contrast, Fig. 3c is also consider an unconstrained graph although this time with fewer edges and fewer parameters: and thus better

[3] Evaluations of decomposability and equivalence, as well as details on data generating mechanisms, causal graphs, algorithm implementations, and run time experiments can be found in Appendices C and D.

[4] Due to space constraints, competing graphs are briefly described for intuition here and plots are given in Fig. 5 (Appendix C). A similar analysis comparing scores on variations of the Verma graph (Fig. 1c), designed to consider the presence and absence of equality constraints, is given in Appendix D.

	n	GS-$\mathcal{S}_{\text{WBIC}}$	DCD (Bow.)	DCD (Anc.)	GSMAG	GES
Sachs	200	1077 (13)	1132 (15)	1157 (15)	1198 (13)	1236 (11)
Sachs	500	2653 (28)	2635 (29)	2643 (31)	2791 (25)	3311 (21)
Sachs	1000	5393 (51)	5401 (40)	5405 (40)	5412 (84)	6610 (32)
Lung	200	350 (18)	347 (16)	360 (20)	387 (10)	329 (5)
Lung	500	825 (31)	827 (31)	852 (32)	856 (27)	821 (13)
Lung	1000	1656 (60)	1653 (56)	1663 (69)	1668 (55)	1656 (12)

(a) Mean score and standard deviation. Lower values indicate better fit.



(b) Bayes factor vs optimal DAG.

Figure 4: Structure learning evaluations.

$\mathcal{S}_{\text{BIC}_1}$ 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 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 empirically, 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.

5 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 adapt a greedy search algorithm to use the decomposable nature of $\mathcal{S}_{\text{WBIC}}$, denoted GS-$\mathcal{S}_{\text{WBIC}}$; pseudocode is given in Appendix C.1. An extensive set of methods exist for searching over spaces of graphs, including greedy search [38], exact dynamic programming [28], integer programming [5], and gradient-based optimization [2] 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* [12]. Empirical comparisons are made with Gaussian-based continuous-optimization algorithm (DCD) [2] for the recovery of ancestral and bow-free graphs, the GES algorithm [6] for the recovery of directed graphs, and the GSMAG algorithm [38] for the recovery of maximally ancestral graphs.

We start by considering graphs returned by each method fit on random datasets from the IV, Verma, and frontdoor models defined in Sec. 4. The objective is 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. Next, we consider comparisons on Sachs [33] and Lung cancer [21] benchmark datasets (with some variables omitted to induce unobserved confounding). Fig. 4 gives mean and standard deviation of $\mathcal{S}_{\text{WBIC}}$ scores of the graph returned by each method on 5 random draws of the simulators. There is variability for all methods on different datasets due to the returned graph and due to the score evaluation. No method significantly outperforms, which is expected as these graphs, to our knowledge, do not entail (in)equality constraints beyond conditional independencies. There is some evidence that greedy search in the space of arbitrary causal graphs can be viable for causal discovery.

6 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

[1] 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.

[2] Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2021.

[3] Blai Bonet. Instrumentality tests revisited. *arXiv preprint arXiv:1301.2258*, 2013.

[4] Constantin Carathéodory. Über den variabilitätsbereich der fourier'schen konstanten von positiven harmonischen funktionen. *Rendiconti Del Circolo Matematico di Palermo (1884-1940)*, 32(1):193–217, 1911.

[5] Rui Chen, Sanjeeb Dash, and Tian Gao. Integer programming for causal structure learning in the presence of latent variables. In *International Conference on Machine Learning*, pages 1550–1560. PMLR, 2021.

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

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

[8] Mathias Drton. Discrete chain graph models. 2009.

[9] Mathias Drton. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2):979–1012, 2009.

[10] Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):323–380, 2017.

[11] Mathias Drton and Thomas S Richardson. Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):287–309, 2008.

[12] Mathias Drton and Thomas S Richardson. Iterative conditional fitting for gaussian ancestral graph models. *arXiv preprint arXiv:1207.4118*, 2012.

[13] Robin J Evans. Graphs for margins of bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

[14] 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.

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

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

[17] Alexander J Hartemink, David K Gifford, Tommi S Jaakkola, and Richard A Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Biocomputing 2001*, pages 422–433. World Scientific, 2000.

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

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

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

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

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

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

[28] Kari Rantanen, Antti Hyttinen, and Matti Järvisalo. Maximal ancestral graph structure learning via exact search. In *Uncertainty in Artificial Intelligence*, pages 1237–1247. PMLR, 2021.

[29] Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[30] 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.

[31] 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.

[32] Herman Rubin and Oscar Wesler. A note on convexity in euclidean n-space. In *Proc. Amer. Math. Soc*, volume 9, pages 522–523, 1958.

[33] 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.

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

[35] Nan Shen and Bárbara González. Bayesian information criterion for linear mixed-effects models. *arXiv preprint arXiv:2104.14725*, 2021.

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

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

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

[39] 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.

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

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

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

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

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

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

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

[47] 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.

Appendix for "Scores for Learning Discrete Causal Graphs with Unobserved Confounders"

This Appendix includes

- Derivation of (in)equality constraints and background on asymptotic theory in Appendix A.
- Proofs in Appendix B.
- Experimental and implementation details in Appendix C.
- Additional experiments in Appendix D.

A Background

A.1 (In)equality derivations

The instrumental variables (IV) model Fig. 1a is perhaps to most extensively studied system in the causal inference literature. It arises naturally in randomized trials with imperfect compliance, in which Z represents a randomized treatment assignment, X the treatment actually taken by the subject, and Y an outcome; U represents unmeasured confounding factors which may affect both the probability of the subject taking the treatment and the outcome of interest.

Making no assumptions on the space of definition of U, and if X is continuous, $P(\boldsymbol{V})$ is unconstrained [3]. However, if the observed variables have finite and discrete state spaces, then the observed distribution obeys the instrumental inequality,

$$\sum_{y} \max_{z} P(x, y \mid z) \leqslant 1.$$

Following [25], it can be shown with the following argument. Note that,

$$P(x, y \mid z) = \sum_{u_2} P(x \mid z, u_2) P(y \mid x, u_2) P(u_2).$$

For a particular value of (x, y), let $z^* = \max_z P(x, y \mid z)$. Thus,

$$\sum_{y} P(x, y \mid z^*) = \sum_{y, u_2} P(x \mid z^*, u_2) P(y \mid x, u_2) P(u_2).$$

For each y, $P(x \mid z^*, u_2) \leqslant 1$,

$$\sum_{y} P(x, y \mid z^*) \leqslant \sum_{y, u_2} P(y \mid x, u_2) P(u_2) \leqslant 1.$$

Substituting z^* and noting that this relationship holds for any x we get,

$$\sum_{y} \max_{z} P(x, y \mid z) \leqslant 1.$$

Verma or "dormant" constraints can be derived by considering statistical independence statements in interventional distributions $P(\boldsymbol{V} \mid do(\boldsymbol{x}))$ that can nevertheless be written as functionals of observational distributions $P(\boldsymbol{V})$. In other words, Verma constraints can be reasoned with by considering d-separation statements in graphs in which incoming edges into selected nodes are removed. For the Verma graph (Fig. 1c) in particular it holds that under intervention on Y, W and Z are d-separated, which implies by Rule 1 of the do-calculus that

$$P(z \mid w, do(y)) = P(z \mid do(y)).$$

Both of this quantities are identifiable,

$$P(z \mid do(y)) = \sum_{x} P(z \mid x, y) P(x)$$

$$P(z \mid w, do(y)) = \sum_{x} P(z \mid x, y, w) P(x \mid w)$$

which delivers the equality constraint.

A.2 Regular models

This section states the asymptotic expansion of the marginal likelihood for regular models. A statistical model is called regular if the parameter which minimizes the Kullback-Leibler (KL) divergence of a true distribution and the statistical model is unique and the Hessian matrix of the KL divergence at the minimum point is regular. The technique that is commonly used is Laplace's method, which is to expand the log likelihood of the data around the maximum likelihood value, and then approximate the peak using a multivariate-normal distribution.

Theorem 3 (Laplace's Approximation)**.** *Suppose that* $\log P(\boldsymbol{v} \mid \mathcal{G}, \boldsymbol{\omega})$ *as a function of* $\boldsymbol{\omega}$ *is twice differentiable and convex, i.e., the Hessian of* $\log P(\boldsymbol{v} \mid \mathcal{G}, \boldsymbol{\omega})$ *is positive definite, the minimum of* $\log P(\boldsymbol{v} \mid \mathcal{G}, \boldsymbol{\omega})$ *on* $\Omega_{\boldsymbol{\omega}}$ *is achieved on a single internal point* $\boldsymbol{\omega}_0$*, and* $P(\boldsymbol{\omega} \mid \mathcal{G})$ *is continuous and* $P(\boldsymbol{\omega}_0 \mid \mathcal{G}) \neq 0$*. The marginal likelihood can be written*

$$-P(\bar{\boldsymbol{v}} \mid \mathcal{G}) = \int_{\Omega_{\boldsymbol{\omega}}} \exp\{-\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})\} dP(\boldsymbol{\omega} \mid \mathcal{G}),$$

If the integral absolutely converges, then, as $n \to \infty$*,*

$$-P(\bar{\boldsymbol{v}} \mid \mathcal{G}) \propto \exp\{-\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}_0)\} n^{d/2}. \tag{12}$$

where $\Omega_{\boldsymbol{\omega}} \subset \mathbb{R}^d$*.*

See for example [35] for a proof of this statement. The Bayesian Information Criterion (BIC) is defined by taking logarithms from this expression [34, 18].

In systems parameterized by Gaussian distributions, Laplace's approximation holds [29] and the BIC can be shown to take the convenient form,

$$-\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}_0) + (|\mathcal{E}| + 2|\boldsymbol{V}|)/2 \log n, \tag{13}$$

where $|\mathcal{E}|$ denotes the number of edges and $|\boldsymbol{V}|$ the number of endogenous variables as the number of parameters correspond to the mean and variance for each node, and one coefficient per directed or bi-directed edge. BIC is an asymptotically consistent scoring criterion for MAGs [29] and returns the same score for all Markov equivalent MAGs, i.e. MAGs that encode the same d-separation statements, as Markov equivalent MAGs share adjacencies. This further justifies the fact that existing scores, most often based on this asymptotic approximation of the marginal likelihood, will not capture differences in more general (in)equality constraints.

A.3 Singular models

A statistical model is singular if either the parameter which minimizes the Kullback-Leibler (KL) divergence of a true distribution and the statistical model is not unique or the Hessian matrix of the KL divergence at the minimum point is singular. One of the difficulties in the analysis of singular models is that the optimal parameter set is not a single point anymore but an analytic set or variety. Such a set usually involves multiple singularities (i.e. points in that set that form a cusp in the manifold) that render the fisher information matrix singular. The log-likelihood can non longer be approximated by a quadratic form of the parameter in the neighbourhood of these singularities. A model is singular if there are parts of the parameter space in which the fisher information is singular. A lot of statistical models are singular, for example, neural networks, reduced rank regressions, normal mixtures, binomial mixtures, hidden Markov models, stochastic context-free grammars, Bayesian networks, and so on. In general, if a statistical model contains hierarchical structure, sub-module, or hidden variables, then it is singular [43]

A.3.1 Example of singularity in graphical model with unobserved confounders

For example, consider a simple graphical model defined by the graph $\{Y \leftarrow U \rightarrow X\}$ where U is an implicit latent variable that causally influences binary observables X and Y. As given by the canonical parameterization in Prop. 1, without loss of generality we may assume the domain of U to be finite and of cardinality 4. The true observational distribution is given by the following expression,

$$P(x, y) = \omega_X^x (1 - \omega_X)^{1-x} \omega_Y^y (1 - \omega_Y)^{1-y}$$

whereas, the joint distribution parameterization according to our latent variable model is given by:

$$
\begin{aligned}
&P(x, y) \\
&= P(U=0) \cdot P(x \mid U=0)^x (1 - P(x \mid U=0))^{1-x} P(y \mid U=0)^y (1 - P(y \mid U=0))^{1-y} \\
&+ P(U=1) \cdot P(x \mid U=1)^x (1 - P(x \mid U=1))^{1-x} \times P(y \mid U=1)^y (1 - P(y \mid U=1))^{1-y} \\
&+ P(U=2) \cdot P(x \mid U=2)^x (1 - P(x \mid U=2))^{1-x} P(y \mid U=2)^y (1 - P(y \mid U=2))^{1-y} \\
&+ P(U=3) \cdot P(x \mid U=3)^x (1 - P(x \mid U=3))^{1-x} P(y \mid U=3)^y (1 - P(y \mid U=3))^{1-y} \\
&= \theta_0 \cdot (\xi_X^{(0)})^x (1 - \xi_X^{(0)})^{1-x} (\xi_Y^{(0)})^y (1 - \xi_Y^{(0)})^{1-y} + \theta_1 \cdot (\xi_X^{(1)})^x (1 - \xi_X^{(1)})^{1-x} (\xi_Y^{(1)})^y (1 - \xi_Y^{(1)})^{1-y} \\
&+ \theta_2 \cdot (\xi_X^{(2)})^x (1 - \xi_X^{(2)})^{1-x} (\xi_Y^{(2)})^y (1 - \xi_Y^{(2)})^{1-y} + \theta_3 \cdot (\xi_X^{(3)})^x (1 - \xi_X^{(3)})^{1-x} (\xi_Y^{(3)})^y (1 - \xi_Y^{(3)})^{1-y}
\end{aligned}
$$

The variety of optimal parameters are given by the union of the following sets:

$$
\begin{aligned}
&\{\theta_0 = 1, \xi_X^{(0)} = \omega_X, \xi_Y^{(0)} = \omega_Y\} \cup \{\theta_1 = 1, \xi_X^{(1)} = \omega_X, \xi_Y^{(1)} = \omega_Y\} \cup \{\theta_2 = 1, \xi_X^{(2)} = \omega_X, \xi_Y^{(2)} = \omega_Y\} \\
&\cup \{\theta_3 = 1, \xi_X^{(3)} = \omega_X, \xi_Y^{(3)} = \omega_Y\} \cup \{\xi_X^{(0)} = \xi_X^{(1)} = \xi_X^{(2)} = \xi_X^{(3)} = \omega_X, \xi_Y^{(0)} = \xi_Y^{(1)} = \xi_Y^{(2)} = \xi_Y^{(3)} = \omega_Y\},
\end{aligned}
$$

which has singularities, for example, at the point:

$$
(\theta_0, \xi_X^{(0)}, \xi_Y^{(0)}, \xi_X^{(1)}, \xi_Y^{(1)}, \xi_X^{(2)}, \xi_Y^{(2)}, \xi_X^{(3)}, \xi_Y^{(3)}) = (1, \omega_X, \omega_Y, \omega_X, \omega_Y, \omega_X, \omega_Y, \omega_X, \omega_Y). \tag{14}
$$

Effectively whenever the parameters of $P(x, y \mid U = u)$ for all u agree we lose a degree of freedom in our model: changing $P(U = u)$ no longer affects the joint distribution of our data. The asymptotic consequences of this behaviour are important.

A.3.2 Asymptotic approximations in singular models

Watanabe reformulated the foundations of asymptotic theory of singular models relying on the [20]'s resolution on singularities. Two distinct concepts of dimension of a model emerge from singular learning theory: the singular fluctuation that shows how strongly the posterior distribution fluctuates, and the learning coefficient and multiplicity that show how fast the posterior distribution shrinks with increasing sample size. The singularities in the parameter space can be analyzed using algebraic geometry with dependencies on the zeta function of the Kullback-Leibler (KL) distance from the true distribution to the model distribution and of the prior parameter distribution [20, 41, 42, 43].

Watanabe's results apply to a large class of models, including reduced-rank regression, factor analysis, Binomial mixtures, and latent class analysis.

For regular models, λ corresponds to an explicit parameter count (recovering Schwarz's Bayesian information Criterion). This is no longer necessarily the case in singular models where λ in general depends on the underlying data generating mechanism which is unknown and in general will be less than Schwarz's factor "half the number of free parameters". Specifically, for priors with smooth and positive densities it holds that $\lambda \leqslant |\Omega_{\boldsymbol{\omega}}|/2$ for any data generating distribution. This implies that,

$$
n^{\lambda} \leqslant n^{|\Omega_{\boldsymbol{\omega}}|/2}. \tag{15}
$$

Consequently, the asymptotic marginal likelihood is of the form

$$
\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}_0) - \text{ penalty}(\mathcal{G}), \tag{16}
$$

where,

$$
\text{penalty}(\mathcal{G}) \leqslant |\Omega_{\boldsymbol{\omega}}|/2, \tag{17}
$$

and is therefore milder than that in the usual BIC.

A.4 Path sampling and thermodynamic integration

Other techniques exist for approximating marginal likelihoods $\int_{\Omega_{\boldsymbol{\omega}}} P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}) dP(\boldsymbol{\omega} \mid \mathcal{G})$.

One that is particularly relevant to our discussion and underlies the $\mathcal{S}_{\text{WBIC}}$ is a method inspired by ideas from path sampling and thermodynamic integration that introduces a distribution proportional

to the likelihood raised to a power $\beta \in [0, 1]$ times the prior, called the power posterior. The expected marginal likelihood can then be expressed as an integral with respect to β from 0 to 1, where the expectation is taken with respect to the power distribution at power t. This is useful because of the properties of the value of the integrand at its end points $\beta = 0$ to $\beta = 1$. We describe the argument briefly below and refer readers to [14] for more details.

Consider the integral of a power distribution defined as,

$$f(\beta) = \int_{\Omega_{\boldsymbol{\omega}}} P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})^{\beta} dP(\boldsymbol{\omega} \mid \mathcal{G}). \tag{18}$$

β is also called a temperature parameter. For $\beta = 1$ this expression reduces to the marginal likelihood and for $\beta = 0$ we are simply integrating over the prior which is equal to 1. The key observation is that by explicitly differentiating with respect to β it holds that,

$$\begin{aligned} \frac{d \log f(\beta)}{d\beta} &= \frac{\int_{\Omega_{\boldsymbol{\omega}}} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}) P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})^{\beta} dP(\boldsymbol{\omega} \mid \mathcal{G})}{\int_{\Omega_{\boldsymbol{\omega}}} P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})^{\beta} dP(\boldsymbol{\omega} \mid \mathcal{G})} \\ &= \mathbb{E}_{\beta} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}), \end{aligned} \tag{19}$$

which can also be written as an expectation of the data log-likelihood with respect to the power posterior distribution. By the mean value theorem for differentiable functions, there must exists some temperature $\beta^* \in [0, 1]$ such that,

$$\begin{aligned} \frac{d \log f(\beta^*)}{d\beta} &= \frac{\log(f(1)) - \log(f(0))}{1 - 0} \\ &= \log \int_{\Omega_{\boldsymbol{\omega}}} P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}) dP(\boldsymbol{\omega} \mid \mathcal{G}), \end{aligned} \tag{20}$$

which is the logarithm of the marginal likelihood of interest. With knowledge of this optimal temperature $\beta^* \in [0, 1]$ we could approximate the log marginal likelihood by sampling from the power posterior and approximating the expectation with Monte Carlo samples.

Watanabe's main result in [44] is to show that asymptotically $\beta^* \to \frac{1}{\log n}$ which defines the $\mathcal{S}_{\text{WBIC}}$ as the following approximation to the log marginal likelihood,

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) := -\mathbb{E}_{\beta} \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}), \tag{21}$$

$$\mathbb{E}_{\beta} g(\boldsymbol{\omega}) := \frac{\int_{\Omega_{\boldsymbol{\omega}}} g(\boldsymbol{\omega}) P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})^{\beta} dP(\boldsymbol{\omega} \mid \mathcal{G})}{\int_{\Omega_{\boldsymbol{\omega}}} P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega})^{\beta} dP(\boldsymbol{\omega} \mid \mathcal{G})}, \tag{22}$$

where $\beta = \frac{1}{\log n}$. The accuracy of the approximation can be quantified explicitly and found in [44, Thm. 4],

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) = -\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}) + \mathcal{O}_p(\sqrt{\log n}). \tag{23}$$

An important observation here is that the prior is explicitly required for $\mathcal{S}_{\text{WBIC}}$ whereas it is only used implicitly in the BIC. The performance of WBIC can thus be sensitive to the prior (which is not immediately problematic as it is a basic characteristic of Bayesian model choice which we adhere to in this paper).

B Proofs

We restate statements for convenience.

Prop. 1 restated. *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*

$$P(\boldsymbol{v} \mid \boldsymbol{\omega}, \mathcal{G}) = \sum_{U \in \boldsymbol{U}} \sum_{u=1,\dots,d_U} \prod_{V \in \boldsymbol{V}} \mathbb{1}\{\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} = v\} \prod_{U \in \boldsymbol{U}} \theta_u, \tag{24}$$

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 $\Omega_{\mathbf{Pa}_V} \times \Omega_{\boldsymbol{U}_V} \mapsto \Omega_V$.

Proof. This proposition appears in related formulations in [31] and [47]. For completeness we adapt the proofs to our setting in this section.

We first introduce some necessary notations and concepts. The probability distribution for every exogenous variables $U \subset \boldsymbol{U}$ is characterized with a probability space. It is frequently designated $\langle \Omega_U, \mathcal{F}_U, P_U \rangle$ where Ω_U is a sample space containing all possible outcomes; $\mathcal{F}_U$ is a σ-algebra containing subsets of Ω_U; P_U is a probability measure on $\mathcal{F}_U$ normalized such that $P_U(\Omega_U) = 1$. Elements of $\mathcal{F}_U$ are called events, which are closed under operations of set complement and unions of countably many sets. By means of P_U a real number $P_U(\mathcal{A}) \in [0, 1]$ is assigned to every event $\mathcal{A} \in \mathcal{F}_U$; it is called the probability of event $\mathcal{A}$. For an arbitrary set of exogenous variables $\boldsymbol{U}$, its realization $\boldsymbol{U} = \boldsymbol{u}$ is an element in the Cartesian product $\times_{U \in \boldsymbol{U}} \Omega_U$. We may be interested in inferring whether a sequence of events $\mathcal{A}$ for every $U \in \boldsymbol{U}$ occurs. Such an event is represented by a subset $\times_{U \in \boldsymbol{U}} \mathcal{A}_U \subseteq \times_{U \in \boldsymbol{U}} \Omega_U$ which in turn generate a product of σ-algebras $\bigotimes_{U \in \boldsymbol{U}} \mathcal{F}_U$. Define the product measure $\bigotimes_{U \in \boldsymbol{U}} P_U$ to satisfy the following mutual independence condition given by the definition of the SCM,

$$P\left(\bigtimes_{U \in \boldsymbol{U}} \mathcal{A}_U\right) = \prod_{U \in \boldsymbol{U}} P_U(\mathcal{A}_U). \tag{25}$$

Such P is a probability measure. Moreover,

$$\left\langle \bigtimes_{U \in \boldsymbol{U})} \Omega_U, \bigotimes_{U \in \boldsymbol{U}} \mathcal{F}_U, \bigotimes_{U \in \boldsymbol{U}} P_U \right\rangle, \tag{26}$$

defines a product of probability spaces $\langle \Omega_U, \mathcal{F}_U, P_U \rangle$ that describes measurable events over all exogeneous variables $\boldsymbol{U}$ partitioned into c-components.

Let $\mathbb{C}$ be the collection of all c-components in $\mathcal{G}$. c-components in $\mathbb{C}$ form a partition $\{\bigcup_{V \in \boldsymbol{C}} U_V \mid \boldsymbol{C} \in \mathbb{C}\}$ over exogenous variables U. Therefore, for every $U \in \boldsymbol{U}$, there must exist a unique c-component denoted by $\boldsymbol{C}_U$ containing U. For any c-component $\boldsymbol{C} \in \mathbb{C}$, let $U_{\boldsymbol{C}} = \bigcup_{V \in \boldsymbol{C}} U_V$ the set of exogenous variables affecting (at least one of) endogenous variables in $\boldsymbol{C}$. By the definition of c-components, the exogeneous variables do not overlap between c-components and it holds that,

$$P\left(\bigcap_{U \in \boldsymbol{U}} \mathcal{A}_U\right) = \prod_{C \in \mathcal{C}(\mathcal{G})} P_U\left(\bigcap_{U \in C} \mathcal{A}_U\right).$$

For any SCM M compatible with the causal graph $\mathcal{G}$ the joint distribution may be factorized into c-components [37],

$$P(\boldsymbol{v}) = \prod_{\boldsymbol{C} \in \mathbb{C}} Q[\boldsymbol{C}](\boldsymbol{c}, \boldsymbol{pa}_{\boldsymbol{C}}),$$

where $Q[\boldsymbol{C}]$ is a C-factor and is a function of $(\boldsymbol{c}, \boldsymbol{pa}_{\boldsymbol{C}})$. We often omit the input for readability.

To parameterize this joint distribution it is thus sufficient to look at each C-factor separately. Let $\boldsymbol{C}$ be a generic c-component in $\mathcal{G}$. Denote by $m = |U_{\boldsymbol{C}}|$ the number of exogeneous variables related to

$\boldsymbol{C}$. For convenience, we consistently write $\langle \Omega_i, \mathcal{F}_i, P_i \rangle$ as the probability space of i-th exogeneous variable in C. The product of these probability spaces is thus written,

$$\left\langle \bigtimes_{i=1}^{m} \Omega_i, \bigotimes_{i=1}^{m} \mathcal{F}_i, \bigotimes_{i=1}^{m} P_i \right\rangle.$$

Each C-factor may thus be written,

$$Q[\boldsymbol{C}] = \int_{\times_{i=1}^{m} \Omega_i} \prod_{V \in \boldsymbol{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=1}^{m} P_i.$$

Our goal is to show that all probabilities $Q[\boldsymbol{C}]$, induced by exogenous variables described by arbitrary probability spaces could be produced by a "simpler" generative process with discrete exogenous domains. $Q[\boldsymbol{C}]$ defines a mapping between the space of possible realizations of the variables $Pa(\boldsymbol{C})$ to the $[0, 1]$ interval. Since $Pa(\boldsymbol{C})$ are discrete variables with finite domains, the cardinality of the class of probability assignments that must be defined is also finite. It is given at most by the number of possible combinations of realizations of $Pa(\boldsymbol{C})$ which is given by $\prod_{V \in Pa(\boldsymbol{C})} |\Omega_V|$.

Let $\bar{P}$ be a vector representing probabilities $Q[\boldsymbol{C}](\boldsymbol{c}, \boldsymbol{pa_C})$. Counting all possible combinations of outcomes for all possible conditioning sets, $\bar{P}$ is therefore a vector of at most size $d = \prod_{V \in Pa(\boldsymbol{C})} |\Omega_V|$. And since $Q[\boldsymbol{C}](\boldsymbol{c}, \boldsymbol{pa_C})$ is a probability mass function, it only takes a vector with $d - 1$ dimensions to uniquely determine it. $\bar{P}$ may thus be interpreted as a point in the $(d-1)$-dimensional real space. Similarly, $(P, 1)$ is vector in d-dimensional space where the d-th element is equal to 1.

Now consider sampling a value $U_1 = u_1$ from the underlying SCM and let Q_{u_1} be the probability model with $U_1 = u_1$.

$$Q_{u_1}[\boldsymbol{C}] = \left[\int_{\times_{i=2}^{m} \Omega_i} \prod_{V \in \boldsymbol{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=2}^{m} P_i \right]_{U_1 = u_1},$$

and $\bar{P}_{u_1}$ is a $(d-1)$-dimensional probability vector representing the probabilities of each one of the combinations $Pa(\boldsymbol{C})$ given that $U_1 = u_1$. We will show that P_1 may equally well be represented by a discrete distribution. For this, let $\mathcal{U} = \{\bar{P}_{u_1} : u_1 \in \Omega_1\} \subset \mathbb{R}^d$ be the set of probability points that can be constructed as u_1 varies in Ω_1. The average $\int_{\Omega_1} \bar{P}_{u_1} dP_1$ is a convex mixture of points in $\mathcal{U}$ by [32] that equals $\bar{Q}$ since,

$$\bar{P} = \int_{\Omega_1} \left[\int_{\times_{i=2}^{m} \Omega_i} \prod_{V \in \boldsymbol{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=2}^{m} P_i \right]_{U_1 = u_1} dP_1.$$

By construction, $\bar{P}$ itself is a convex mixture of at most $d + 1$ points in $\mathcal{U}$. That is, by using Carathéodory's theorem [4],

$$\bar{P} = \sum_{k=1}^{d+1} w_k \bar{P}_{u_{1,k}}.$$

Replacing the definition of $\bar{P}_{u_{1,k}}$ we obtain $\bar{P}$ equal to,

$$\sum_{k=1}^{d+1} w_k \left[\int_{\times_{i=2}^{m} \Omega_i} \prod_{V \in \boldsymbol{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=2}^{m} P_i \right]_{U_1 = u_{1,k}}.$$

This means that we can replace the continuous measure P_1 with a discrete probability set with outcomes $\{u_{1,1}, \ldots, u_{1,d}\}$ and corresponding probabilities $\{w_1, \ldots, w_d\}$ with cardinality d and obtain a probability model that is equivalent to the original $\bar{P}$. This procedure can be repeated for all m exogeneous variables in the c-component $\boldsymbol{C}$. We are thus left with a model,

$$Q[\boldsymbol{C}] = \int_{\times_{i=1}^{m} \Omega_i} \prod_{V \in \boldsymbol{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=1}^{m} P_i,$$

equivalent to its discrete counterpart,

$$Q[\boldsymbol{C}] = \sum_{u \in \boldsymbol{u_c}} \sum_{u=1,\dots,d} \prod_{V \in \boldsymbol{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} \prod_{u \in \boldsymbol{u_c}} P(u),$$

where $d = \prod_{V \in Pa(\boldsymbol{C})} |\Omega_V|$.

This process may now be applied to each C-factor separately to obtain a parameterization for the joint distribution $P(\boldsymbol{v})$ given by,

$$P(\boldsymbol{v}) = \sum_{U \in \boldsymbol{U}} \sum_{u=1,\dots,d_U} \prod_{V \in \boldsymbol{V}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} \prod_{U \in \boldsymbol{U}} P(u),$$

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

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

to stress the underlying parameterization. □

Thm. 1 restated. *In discrete SCMs parameterized by Prop. 1,*

$$-\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}) = -\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}, \boldsymbol{\omega}_0) + \lambda_{\mathcal{G}} \log n + \mathcal{O}_p(\log\log n), \tag{27}$$

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

Proof. This result is a consequence on Watanabe's asymptotic expansion of the marginal likelihood [41, Thm. 1]. We prove that its conditions, also stated as [43, Definitions 6.1 and 6.3], apply to singular models parameterized by Prop. 1. We require that,

1. The distributions in all candidate graphs share a common support and have densities with respect to a dominating measure.
2. The parameter space Ω_ω is compact and defined by real analytic constraints.
3. The log-likelihood ratios of $P(\bar{\boldsymbol{v}} \mid \boldsymbol{\omega}_0, \mathcal{G})$, the true parameter, with respect to the distributions $P(\bar{\boldsymbol{v}} \mid \boldsymbol{\omega}, \mathcal{G})$ can be bounded by a function that is square integrable under $P(\boldsymbol{\omega} \mid \mathcal{G})$ and satisfy a requirement of analyticity that allows for power series expansions in $\boldsymbol{\omega}$.
4. The prior distribution $P(\boldsymbol{\omega} \mid \mathcal{G})$ has a density that is the product of a smooth positive function and a non-negative analytic function.

For (1), in systems of latent variables paramterized by Prop. 1 both the latent and observed variables are discrete and share a common support which is the set of all possible values that the variables can take. In this case, local conditional distributions are discrete probability distributions and have a density with respect to the counting measure.

For (2), the importance of compactness of the parameter space comes from the need to define neighbourhoods around each value in the parameter space allowing for local analysis. The parameter space $(\Omega_{\boldsymbol{\omega}}, \Omega_{\boldsymbol{\xi}})$ is a subset of a finite-dimensional Euclidean space which is itself a compact set, and therefore the parameter space is also compact and, specifically, it is closed and bounded.

For (3), note that in structural causal models defined by discretely-valued parameters, the corresponding probability distribution over all variables can be represented as a discrete exponential family. As some variables are latent, the joint probability distribution over observed variables corresponds to a marginalization which might result in a singular submodel of an exponential family with a non-invertible natural parameter function. For this class of systems, [9] showed that the sequence of likelihood ratios $P(\bar{\boldsymbol{v}} \mid \mathcal{G}_1, \boldsymbol{\omega}_0)/P(\bar{\boldsymbol{v}} \mid \mathcal{G}_2, \boldsymbol{\omega}_0) = \mathcal{O}_p(1)$ for two causal graphs $\mathcal{G}_1, \mathcal{G}_2$.

For (4) the prior density $P(\boldsymbol{\omega} \mid \mathcal{G})$ can be chosen by the investigator. For example, for the prior Dirichlet distribution used to parameterize distributions of exogenous probabilities, it holds that the probability density function is given by a product of gamma functions and a power function. The product of gamma functions is a smooth positive function, and the power function is a non-negative analytic function (depending on the value of concentration parameters). □

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 corresponding to $\mathcal{G}_1, \mathcal{G}_2$ respectively.*

Thm. 2 restated. *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 $P(\boldsymbol{V})$, $\mathcal{S}_{WBIC}(\mathcal{G}_1, \bar{\boldsymbol{v}}) < \mathcal{S}_{WBIC}(\mathcal{G}_2, \bar{\boldsymbol{v}})$.*

Proof. For the soundness part, let $\mathcal{G}_1$ be compatible with a SCM that is able to generate the data distribution P and assume $\mathcal{G}_2$ is not compatible with any SCM that is able to generate P. By Prop. 1, the discrete parameterization of SCMs compatible with $\mathcal{G}_1$ and $\mathcal{G}_2$ is rich enough to define a set of distributions that contains any distribution compatible with $\mathcal{G}_1$ and $\mathcal{G}_2$. By [44, Thm. 4], for $\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}})$ defined by the proposed discrete parameterization it holds then that there exists a constant $c > 0$ such that,

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) = -\log P(\mathcal{G} \mid \bar{\boldsymbol{v}}) + c\sqrt{\log n}. \tag{28}$$

with probability 1 as $n \to \infty$.

Then, for constants $c_1, c_2 > 0$,

$$\begin{aligned}
\mathcal{S}_{\text{WBIC}}(\mathcal{G}_1, \bar{\boldsymbol{v}}) - \mathcal{S}_{\text{WBIC}}(\mathcal{G}_2, \bar{\boldsymbol{v}}) &= -\log P(\mathcal{G} \mid \bar{\boldsymbol{v}}) + \log P(\mathcal{G} \mid \bar{\boldsymbol{v}}) + (c_1 - c_2)\sqrt{\log n} \\
&= -\log \frac{P(\bar{\boldsymbol{v}} \mid \mathcal{G}_1) P(\mathcal{G}_1)}{P(\bar{\boldsymbol{v}})} + \log \frac{P(\bar{\boldsymbol{v}} \mid \mathcal{G}_2) P(\mathcal{G}_2)}{P(\bar{\boldsymbol{v}})} + (c_1 - c_2)\sqrt{\log n}, \\
&= -\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}_1) + \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}_2) - \log P(\mathcal{G}_1) + \log P(\mathcal{G}_2) + (c_1 - c_2)\sqrt{\log n} \\
&= -c_3 n + (c_1 - c_2)\sqrt{\log n} + c_4,
\end{aligned}$$

with probability tending to 1 as $n \to \infty$. $c_3 > 0$ is a constant corresponding to Assumption 1. Therefore, with sufficiently large values of n, $\mathcal{S}_{\text{WBIC}}(\mathcal{G}_1, \bar{\boldsymbol{v}}) < \mathcal{S}_{\text{WBIC}}(\mathcal{G}_2, \bar{\boldsymbol{v}})$.

For the parsimony part, note that in structural causal models defined by discretely-valued parameters, the likelihood function can be expressed as a singular submodel of an exponential family. Therefore, for any two graphical models $\mathcal{G}_1$ and $\mathcal{G}_2$ that induce the data generating distribution, the sequence of likelihood ratios $P(\bar{\boldsymbol{v}} \mid \mathcal{G}_1, \boldsymbol{\omega}_0)/P(\bar{\boldsymbol{v}} \mid \mathcal{G}_2, \boldsymbol{\omega}_0) = \mathcal{O}_p(1)$ [9].

Fix a data generating distribution. 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 $\mathcal{G}_2$. Given Assumption 2, SCMs inducing more probabilistic constraints are also less complex in this sense, which yields $\lambda_{\mathcal{G}_1} < \lambda_{\mathcal{G}_2}$.

Following a similar decomposition of the $\mathcal{S}_{\text{WBIC}}$ together with Thm. 1,

$$\begin{aligned}
&\mathcal{S}_{\text{WBIC}}(\mathcal{G}_1, \bar{\boldsymbol{v}}) - \mathcal{S}_{\text{WBIC}}(\mathcal{G}_2, \bar{\boldsymbol{v}}) \\
&= -\log P(\mathcal{G} \mid \bar{\boldsymbol{v}}) + \log P(\mathcal{G} \mid \bar{\boldsymbol{v}}) + (c_1 - c_2)\sqrt{\log n} \\
&= -\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}_1) + \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}_2) + (c_1 - c_2)\sqrt{\log n} + c_3 \\
&= -\log P(\bar{\boldsymbol{v}} \mid \mathcal{G}_1, \boldsymbol{\omega}_0) + \log P(\bar{\boldsymbol{v}} \mid \mathcal{G}_1, \boldsymbol{\omega}_0) + \lambda_{\mathcal{G}_1} \log n - \lambda_{\mathcal{G}_2} \log n + (c_1 - c_2)\sqrt{\log n} + c_3 \\
&= (\lambda_{\mathcal{G}_1} - \lambda_{\mathcal{G}_2}) \log n + (c_1 - c_2)\sqrt{\log n} + c_3 + c_4
\end{aligned}$$

asymptotically with probability tending to 1, where $c_1, c_2, c_3, c_4 > 0$ are constants. For sufficiently large n, therefore, $\mathcal{S}_{\text{WBIC}}(\mathcal{G}_1, \bar{\boldsymbol{v}}) < \mathcal{S}_{\text{WBIC}}(\mathcal{G}_2, \bar{\boldsymbol{v}})$. □

Prop. 2 restated. *Under the assumption of extended faithfulness, as $n \to \infty$, $\mathcal{S}_{WBIC}(\mathcal{G}, \bar{\boldsymbol{v}})$ distinguishes between candidate causal graphs differing on a (in)equality constraint between margins of $P(\boldsymbol{V})$ with probability 1.*

Proof. For a contradiction, assume that $\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}})$ does not distinguish between two graphs, $\mathcal{G}_1$ and $\mathcal{G}_2$ that differ on a (in)equality constraint. By Prop. 1, the family of observational distributions defined by discrete parameterizations of SCMs will be different for $\mathcal{G}_1$ and $\mathcal{G}_2$. Thus also the scores $\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}})$ must be different if $\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}})$ is sound and parsimonious. This is a contradiction of Thm. 2. □

Prop. 3 restated. *Let $P(\boldsymbol{V})$ and $\mathcal{G}$ be the joint distribution and causal graph implied by an SCM parameterized by curved exponential models. Then, asymptotically,*

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) = \mathcal{S}_{\text{BIC}}(\mathcal{G}, \bar{\boldsymbol{v}}) + \mathcal{O}_p(1).$$

Proof. The proof can be found in [44, Eq. (32)]. □

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

Proof. We use the concept of C-factors. Following [37], for any $\boldsymbol{C} \subseteq \boldsymbol{V}$, we define function $Q[\boldsymbol{C}](\boldsymbol{v}) = P(\boldsymbol{c} \mid do(\boldsymbol{v} \backslash \boldsymbol{c}))$. For convenience, we omit input $\boldsymbol{v}$ and write $Q[\boldsymbol{C}]$. In particular, $Q[\boldsymbol{v}] = P(P(\boldsymbol{v} \mid \mathcal{G}, \boldsymbol{\omega}))$ if parameterized by $\boldsymbol{\omega}$, and any $Q[\boldsymbol{C}]$ is a function of $\boldsymbol{C}$ and its parents $Pa(\boldsymbol{C})$. Recall that $\mathcal{S}_{\text{WBIC}}$ is defined as,

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

which can be re-written as

$$\mathcal{S}_{\text{WBIC}}(\mathcal{G}, \boldsymbol{v}) := -\frac{\int_{\Omega_\omega} \log(Q[\boldsymbol{v}]) Q[\boldsymbol{v}]^\beta dP(\boldsymbol{\omega} \mid \mathcal{G})}{\int_{\Omega_\omega} Q[\boldsymbol{v}]^\beta dP(\boldsymbol{\omega} \mid \mathcal{G})},$$

in terms of C-factors, where the dependence of Q on $\boldsymbol{\omega}$ is implicit. Let $\boldsymbol{C}_i$ denote the set of endogenous variables contained in the i-th c-component, $i = 1, \ldots, m$. Following [37], the likelihood can then be decomposed in terms of C-factors associated to c-components in the graph,

$$P(\boldsymbol{v} \mid \boldsymbol{\omega}, \mathcal{G}) = \prod_{i=1}^{m} Q[\boldsymbol{C}_i]. \tag{29}$$

Notice that the parameter space is similarly partitioned across C-factors as each exogenous variable is associated to a single c-component. Let $\omega_i \in \Omega_i$ be the parameters that define the probability of exogenous variables in c-component C_i and functional assignments of endogeneous variables in c-component C_i. $\mathcal{S}_{\text{WBIC}}$ may then be written as a sum of integrals, where the numerator takes the form,

$$-\sum_i \int_{\times_{j=1}^m \Omega_j} \log Q[\boldsymbol{C}_i] \times \prod_{j=1}^{m} Q[\boldsymbol{C}_j]^\beta d \bigotimes_{j=1}^{m} P(\boldsymbol{\omega}_j \mid \mathcal{G}).$$

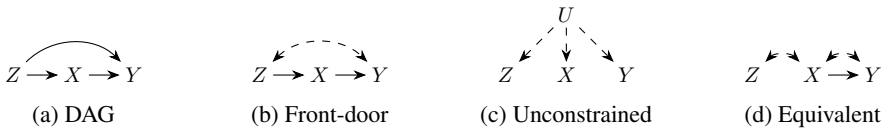
With the assumption that the prior on $\boldsymbol{\omega}$ factors similarly across c-components, all terms that correspond to c-components other than $\boldsymbol{C}_i$ can be taken out of the integral with respect to Ω_i and cancel with equal integrals of the denominator of the definition of $\mathcal{S}_{\text{WBIC}}$. Thus,

$$\mathcal{S}_{\text{WBIC}} = \sum_{i=1}^{m} \frac{-\int_{\Omega_i} \log(Q[\boldsymbol{C}_i]) Q[\boldsymbol{C}_i]^\beta dP(\boldsymbol{\omega}_i \mid \mathcal{G})}{\int_{\Omega_i} Q[\boldsymbol{C}_i]^\beta dP(\boldsymbol{\omega}_i \mid \mathcal{G})},$$

which we recognise as $\mathcal{S}_{\text{WBIC}}(\mathcal{G}_{Pa(\boldsymbol{C}_i)}, \boldsymbol{v}_{Pa(\boldsymbol{C}_i)})$ where $\mathcal{G}_{Pa(\boldsymbol{C}_i)}$ and $\boldsymbol{v}_{Pa(\boldsymbol{C}_i)}$ denote the subgraph and data, respectively, with restriction to the variables in $Pa(\boldsymbol{C}_i) \subseteq \boldsymbol{V}$. □

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

Proof. $\mathcal{S}_{\text{WBIC}}$ is defined based on model likelihoods and parameter priors. Under assumptions 1 and 2. If the prior probability of two graphical models are equal and they encode exactly the same probabilistic constraints that the likelihoods are equal up to constant terms and thus also the values of $\mathcal{S}_{\text{WBIC}}$ are equal up to constant terms asymptotically. □



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

Figure 5: Graphs used in Sec. 3.

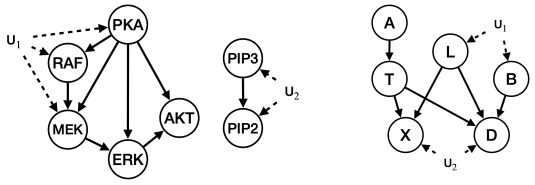


Figure 6: $\mathcal{G}_{Sachs}$ and $\mathcal{G}_{Lung}$.

C Experimental and implementation details

In this section we give details of the data generating mechanisms for the synthetic simulations and real data.

The graphs used in Sec. 3 are given in Fig. 5.

Sachs data is downloaded with the discretization procedure of [17], with 3 levels, from the `bnlearn` data repository. The corresponding graph is given in Fig. 6.

The lung cancer data is downloaded from the `bnlearn` data repository. The corresponding graph is given in Fig. 6.

Algorithm 1 Hill climbing greedy search

Input: A dataset $\bar{v}$, a score function $\mathcal{S}$.
Output: The graph $\mathcal{G}^*$ that maximizes the score.
Initialize: Set $\mathcal{G}^*$ to the empty graph and compute $\mathcal{S}$ on $\mathcal{G}^*$, denoted $\mathcal{S}(\mathcal{G}^*)$.

1. *While* $\mathcal{S}(\mathcal{G}^*)$ decreases, *for* every possible edge addition, deletion or modification that does not prevent acyclicity.
   (a) Let $\mathcal{G}$ be the updated graph.
   (b) *If* $\mathcal{S}(\mathcal{G}) < \mathcal{S}(\mathcal{G}^*)$ *then* set $\mathcal{G}^* = \mathcal{G}$.
2. *Tabu list.* Repeat step 1 but choose the graph $\mathcal{G}$ with highest score that has not been considered in the last steps.
3. *Random restart.* Repeat step 1 a fixed number of times by adding or removing multiple random edges to $\mathcal{G}^*$.

C.1 Implementation details

All experiments were run on a 3.2 GHz M1 Apple processor with 8 cores under 16-GB memory limit.

We use a standard Hill climbing greedy search implementation that is given in Alg. 1. 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 or bi-directed edge, or expand 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. One only needs to recompute the scores of c-components that are affected by the change. The algorithm terminates whenever no change can be found that improves the score. Note that greedy search in the space of arbitrary graphs, even with an oracle scoring method is not known to converge to a globally optimal graph, and may get stuck in local optima.

C.2 Example parameterization and MCMC

In this example, we show how to compute all steps of the MCMC for a specific graph and joint distribution of data. We consider the IV graph presented in Fig. 1a.

The parameterization of the joint distribution is given by

$$P(z,x,y) = \sum_{u_{xy},u_z} \mathbb{1}\{\xi_Y^{(x,u_{xy})} = y\}\mathbb{1}\{\xi_X^{(z,u_{xy})} = x\}\mathbb{1}\{\xi_Z^{(u_z)} = z\}\theta_{u_{xy}}\theta_{u_z}$$

where e.g. $\xi_Y^{(x,u_{xy})}$ represents the causal assignment of Y given its observed and latent parents and $\theta_{u_{xy}}$ represents the exogenous probability $P(U_{XY} = u_{xy})$. To compute $\mathcal{S}_{\text{WBIC}}$ we approximate the power posterior of all relevant parameters, that is $\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{u}$ given $\bar{\boldsymbol{v}}$, with a metropolis step.

1. Sampling from $P(u_{xy}^{(n)}, u_z^{(n)} \mid \bar{\boldsymbol{v}}, \boldsymbol{\xi}, \boldsymbol{\theta})$. The complete conditional can be derived following the functional dependencies in the underlying SCM given by the causal graph,

$$\begin{aligned}
P(u_{xy}^{(n)}, u_z^{(n)} \mid \bar{\boldsymbol{v}}, \boldsymbol{\xi}, \boldsymbol{\theta}) &= P(u_{xy}^{(n)}, u_z^{(n)} \mid \boldsymbol{v}^{(n)}, \boldsymbol{\xi}, \boldsymbol{\theta}) \\
&\propto P(u_{xy}^{(n)}, u_z^{(n)}, \boldsymbol{v}^{(n)} \mid \boldsymbol{\xi}, \boldsymbol{\theta}) \\
&= P(y^{(n)} \mid x^{(n)}, u_{xy}^{(n)})P(x^{(n)} \mid z^{(n)}, u_{xy}^{(n)})P(z^{(n)} \mid u_z^{(n)})P(u_z^{(n)})P(u_{xy}^{(n)}) \\
&= \mathbb{1}\{\xi_Y^{(x^{(n)},u_{xy})} = y^{(n)}\}\mathbb{1}\{\xi_X^{(z^{(n)},u_{xy}^{(n)})} = x^{(n)}\}\mathbb{1}\{\xi_Z^{(u_z^{(n)})} = (z^{(n)}\}\theta_{u_{xy}^{(n)}}\theta_{u_z^{(n)}},
\end{aligned}$$

   where we have replaced the probabilities with the corresponding parameters that are used to define them. Let $\bar{\boldsymbol{u}}$ denote n instantiations of latent variables sampled according to the probabilities above.

2. Sampling from deterministic causal mechanisms. We consider $P(\xi_Y^{(x,u_{xy})} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}, \boldsymbol{\theta})$ for illustration as other parameters are sampled similarly. For fixed x, u_{xy}, parameter $\xi_Y^{(x,u_{xy})}$ is mutually independent of any other parameter in $\boldsymbol{\xi}$ given $\bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}, \boldsymbol{\theta}$ and can be sampled separately. Recall that by definition of the underlying SCM $\xi_Y^{(x,u_{xy})}$ represent a deterministic mapping between inputs x, u_{xy} and output $y \in \Omega_Y$. The value $\xi_Y^{(x,z,u_{xy})} \in \Omega_Y$ is therefore implicitly determined by the current values $\bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}$: if there exists a tuple $(x^{(n)} = x, u_{xy}^{(n)} = u_{xy}, y^{(n)} = y)$ for some $n = 1 \ldots, N$, then by definition $\xi_Y^{(x,u_{xy})} := y$ with probability 1. If no such tuple exist, then $\xi_Y^{(x,z,u_{xy})}$ is sampled from its domain Ω_Y with proposal probability $\boldsymbol{q} = \{q_y : y \in \Omega_Y\}$ that are uniformly updated in a small neighbourhood of the previous parameter value.

3. Sampling from $P(\theta_{u_{xy}} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}, \boldsymbol{\theta})$. The conditional distribution over $\theta_{u_{xy}}$ given $\bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}$ is given by a Dirichlet distribution $\boldsymbol{\theta}_U \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}} \sim Dir(\beta_1, \ldots, \beta_{d_U})$ 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.

This process eventually forms a chain of samples from the correct posterior distribution of each parameter. At this stage, we record the current $t+1$-th sample $(\boldsymbol{\xi}_{(t+1)}, \boldsymbol{\theta}_{(t+1)})$ 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,

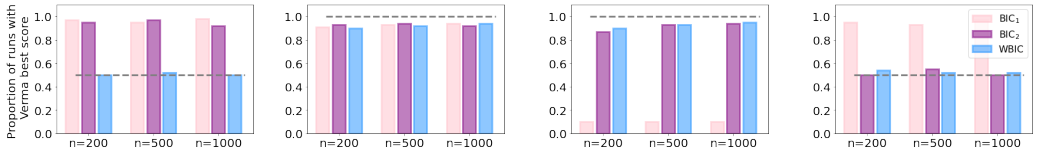
$$P(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \mathcal{G})^\beta \ \exp\left\{-\beta \log P(\bar{\boldsymbol{v}} \mid \boldsymbol{\xi}, \boldsymbol{\theta}, \mathcal{G}) + \log P(\boldsymbol{\xi}, \boldsymbol{\theta} | \mathcal{G})\right\},$$

where $P(\bar{\boldsymbol{v}} \mid \boldsymbol{\xi}, \boldsymbol{\theta}, \mathcal{G})$ evaluates to,

$$\prod_{i=1}^{n} \sum_{u_{xy},u_z} \mathbb{1}\{\xi_Y^{(x^{(i)},u_{xy})} = y^{(i)}\}\mathbb{1}\{\xi_X^{(z^{(i)},u_{xy})} = x^{(i)}\}\mathbb{1}\{\xi_Z^{(u_z} = z^{(i)}\}\theta_{u_{xy}}\theta_{u_z},$$

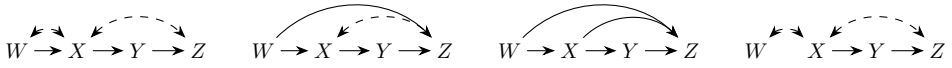
And finally, we approximate $\mathcal{S}_{\text{WBIC}}$ with MCMC samples,

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



(a) Verma vs Equivalent 1. (b) Verma vs Unconst. (c) Verma vs DAG. (d) Verma vs Equivalent 2.

Figure 7: Quality of scores. The horizontal gray line indicates the theoretical optimum.



(a) Equivalent 1. (b) Unconstrained. (c) DAG. (d) Equivalent 2

Figure 8: Graphs used in Appendix D.1.

D Additional experiments

This section provides additional experiments to illustrate score consistency in the presence of equality constraints, score decomposability, an an illustration of the expressiveness of discrete SCMs, and an empirical run time analysis of the proposed score.

D.1 Scores on variations of the Verma graph

We consider variations of the Verma graph (Fig. 1c) given in Fig. 8. The task is to score these variations, and compare them to scores of the ground truth IV graph, based data generated from different random SCMs $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P \rangle$ compatible with ground truth graph.

Comparisons between Verma graphs and its variations emphasize the trends observed in the main body of this paper. For instance, we highlight Fig. 7a that considers an equivalent graph that adds a bi-directed edge $W \dashleftarrow\dashrightarrow X$ to the Verma graph in Fig. 1c, and results in a model with more edges and parameters which has worse $\mathcal{S}_{\text{BIC}_1}, \mathcal{S}_{\text{BIC}_2}$ scores even though defining the same model for $P(\boldsymbol{V})$. Figs. 7b and 7c both consider unconstrained variations of the Verma graph with more edges (but Fig. 7c fewer parameters) and Fig. 7d considers an equivalent graph with the same number of edges but more parameters which results in the expected scoring pattern observed in Figs. 7b to 7d. We conclude with the observation that empirically, across variations of different graphs and sample sizes, $\mathcal{S}_{\text{WBIC}}$ correctly scores graphs based on (in)equality 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.

D.2 Illustration of score consistency and equivalence

We consider in this subsection additional experiments to illustrate the consistency of our score in systems that differ on an (in)equality constraint but also more exotic constraints that have been studied in the literature.

Our results are 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 in terms of $\mathcal{S}_{\text{WBIC}}$ and $\mathcal{S}_{\text{BIC}}$ with alternative (erroneous "×") causal graphs. In particular, the Verma graph in the first row specifies an equality constraint over $P(\boldsymbol{v})$ that is violated in the second graph. The graph in the third row is unconstrained, i.e. 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. [45], while both models specify exactly the same constraints otherwise. 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 compatible with any distribution $P(x, y, z)$ which we include here to demonstrate that $\mathcal{S}_{\text{WBIC}}$ gives the same score to equivalent graphs. We observe that in all examples, $\mathcal{S}_{\text{WBIC}}$ gives a lower score to the correct graph, illustrating empirically that the proposed score is able to leverage (in)equality as well as more general constraints to correctly infer the correct graph. This is not the case for $\mathcal{S}_{\text{BIC}_1}$, especially

in comparisons to the IV example where the free parameter count does not reflect the asymptotic complexity $P(\boldsymbol{v} \mid \mathcal{G})$.

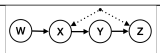
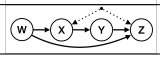
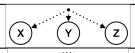
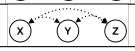
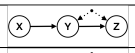
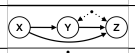
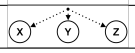
Graph	$\mathcal{S}_{\text{WBIC}}$	$\mathcal{S}_{\text{BIC}_1}$	True Graph?
	2770.8	2707.7	✓
	2778.7	2709.0	×
	697.7	709.9	✓
	1293.2	1453.3	×
	1557.7	1570.8	✓
	1559.8	1578.4	×
	1559.7	1564.2	×

Table 1: $\mathcal{S}_{\text{WBIC}}$ on graphs imposing different constraints on data. Lower values indicate a better fit.

D.3 Illustration of score decomposability and equivalence

We consider in this subsection additional experiments to illustrate the decomposability features of the score

Table 2 exemplifies these facts by showing that scores of separate c-components can be added to generate a total score, that equivalent graphs have equivalent scores, and that incorrectly adding or removing statistical independencies worsens the score due to the worse fit of the resulting graph to the data generating distribution.

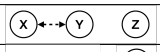
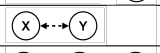
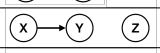
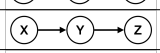
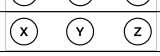
Graph	$\mathcal{S}_{\text{WBIC}}$	Interpretation
	3426.1	Data generating graph $\mathcal{G}$
	1388.1	c-component of $\mathcal{G}$
	2038.3	c-component of $\mathcal{G}$
	3426.4	Equivalent graph to $\mathcal{G}$
	3431.0	$\mathcal{G}$ with incorrect dependence
	4167.9	$\mathcal{G}$ with incorrect independence

Table 2: Examples of decomposition, equivalence, and consistency.

D.4 Collider graph to illustrate upperbound on cardinality of exogenous

We use the graph illustrated in Table 3 to show that the upperbound on the dimensionality of exogenous variables in an observational equivalent discrete SCM correctly encodes the required complexity to generate the class of observational distributions implied by the underlying SCM (with continuously-valued exogeneous variables). Data is generated according to the graph with the following parameterization: $x \leftarrow \mathbb{1}\{u_x > 0\}, y \leftarrow \mathbb{1}\{x > 0.5, -0.5 < u_{yz} < 1\}, z \leftarrow \mathbb{1}\{u_{yz} > 0\}$.

Observe that for parameterizations of the likelihood given by the upper-bound in Prop. 1, i.e. $|\Omega_{U_{yz}}| = |\Omega_X| \cdot |\Omega_Y| \cdot |\Omega_Z| = 8$, $\mathcal{S}_{\text{WBIC}}$ reports a score of 2003 which is the same as that given for any model with a larger dimensionality of exogeneous variables but that the score worsens for models with a dimensionality $|\Omega_{U_{yz}}| = 4$ for example, with score 2033. Prop. 1 only specifies an upper-bound to the dimensionality of exogeneous variables such that we are able to reproduce any observational distribution given by the continuous SCM but these may be over-parameterized as can be seen by computing $\mathcal{S}_{\text{WBIC}}$ for a model with $|\Omega_{U_{yz}}| = 6$ that, it turns out, is expressive enough to encode the observed data distribution. The score $\mathcal{S}_{\text{WBIC}}$ penalizes based on the *effective* dimensionality of the parameter space and is thus insensitive to increasing the dimensionality of

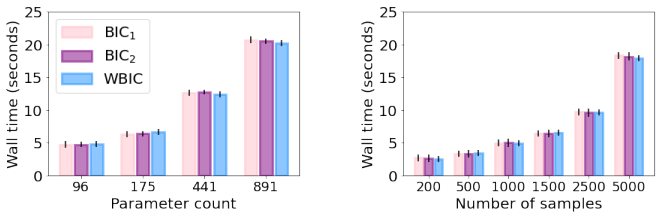


Figure 9: Run time experiments.

exogenous variables if this does not change the family of distributions that such a parameterization induces.

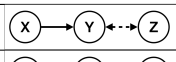
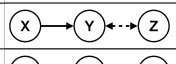
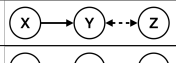
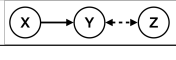
Graph	$\lvert\Omega_{U_{YZ}}\rvert$	$\mathcal{S}_{\text{WBIC}}$
	8	2003.2
	10	2003.4
	6	2003.2
	4	2033.4

Table 3: Varying the dimensionality of u_{YZ}.

D.5 Run time performance

This section describes the run time complexity (run here for illustration on a standard 3.2 GHz M1 Apple processor with 8 cores under 16-GB memory limit) of scoring causal graphs with $\mathcal{S}_{\text{WBIC}}, \mathcal{S}_{\text{BIC}_1}$ and $\mathcal{S}_{\text{BIC}_2}$ as a function of the number of parameters that define the underlying model and as a function of the number of samples. Fig. 9 (LHS) gives the time in seconds on this machine needed to score a graph $\mathcal{G} = \{X \to Y, X \leftrightarrow Y\}$ in which we set the cardinality of X and Y to $4, 5, 7, 9$ which results in a total of $96, 175, 441, 891$ parameters. We use a sample size of 1000 and 5000 iterations of the MCMC sampler. Fig. 9 (RHS) gives the time in seconds on this machine needed to score a graph $\mathcal{G} = \{X \to Y, X \dashleftarrow\!\!\dashrightarrow Y\}$ with X and Y of cardinality 3 with increasing sample size and 5000 iterations of the MCMC sampler. Due to the decomposable nature of the proposed score all relevant c-components may be scored in parallel.