Biometrika (-), **-**, -, *pp*. 1–20
Printed in Great Britain
Advance Access publication on ...

Bayesian learning of network structures from interventional experimental data

BY F. CASTELLETTI
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan
federico.castelletti@unicatt.it

AND S. PELUSO
Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, Milan
stefano.peluso@unimib.it

SUMMARY

Directed Acyclic Graphs (DAGs) provide an effective framework for learning causal relationships among variables given multivariate observations. Under pure observational data, DAGs encoding the same conditional independencies cannot be distinguished and are collected into Markov equivalence classes. In many contexts however, observational measurements are supplemented by interventional data that improve DAG identifiability and enhance causal effect estimation. We propose a Bayesian framework for multivariate data partially generated after stochastic interventions. To this end, we introduce an effective prior elicitation procedure leading to a closed-form expression for the DAG marginal likelihood and guaranteeing score equivalence among DAGs that are Markov equivalent post intervention. Under the Gaussian setting we show, in terms of posterior ratio consistency, that the true network will be asymptotically recovered, regardless of the specific distribution of the intervened variables and of the relative asymptotic dominance between observational and interventional measurements. We validate our theoretical results in simulation and we implement on both synthetic and biological protein expression data a Markov chain Monte Carlo sampler for posterior inference on the space of DAGs.

Some key words: Bayesian model selection; Causal inference; Directed acyclic graph; Intervention; Markov equivalence; Posterior ratio consistency; Structure learning.

1. INTRODUCTION

Identifying cause-and-effect relations between variables is a fundamental issue in several scientific domains, including medicine, biology and economics (Pingault et al., 2018; Hünermund & Bareinboim, 2019). The objective is to infer these relationships from the data, a task which is not possible in general when only pure observational measurements are given. In some contexts however, one could set up intervention experiments, and jointly model observations that were collected before and after the interventions, or derived under distinct experimental conditions. One example is the analysis of biological protein signalling data, where measurements are typically collected after a series of stimulatory cues and inhibitory interventions obtained from the administration of reagents, responsible of the perturbation of nodes in the pathway (Sachs et al., 2005; Dorel et al., 2018). Another instance is genomics, where transcriptomic gene expression

data can be supplemented by interventional data obtained by performing partial, single, or multiple gene knock-out experiments (Rau et al., 2013; Pinna et al., 2013). When the problem involves several variables, the allied *multivariate* causal structure can be represented through a directed network, namely a Directed Acylic Graph (DAG), with directed edges representing causal relationships between nodes/variables and their parents in the graph. The target is then to infer the network structure, a process known as *structure learning*, based on conditional independence assertions that can be deduced from the joint distribution (Kalisch & Bühlmann, 2007; Friedman & Koller, 2003). Since DAG identification is not guaranteed from observational measurements, the output of the inferential process is a potentially large *Markov equivalence class* of DAGs sharing the same conditional independencies (Andersson et al., 1997). Importantly however, by combining observational and interventional data, one can reduce the Markov equivalence class, in principle up to a single DAG structure.

The literature on structure learning from experimental data has grown surprisingly in the last years, both in the statistical and machine learning community. Several types of interventions have been considered, leading to different characterizations of Markov equivalence under intervention, named *I-Markov* equivalence, and to the development of several dedicated methodologies (Korb et al., 2004; Yang et al., 2018; Jaber et al., 2020); see also Correa & Bareinboim (2020) for a comprehensive treatment. For our purposes, a distinction can be made between *deterministic* and *stochastic* intervention: the former sets each manipulated variable to a given value $\widetilde{x}$, so that the local distribution of each intervened node reduces to a point mass at $\widetilde{x}$; the second type instead replaces the local density with that of a new random variable, say $\widetilde{f}$. In addition, *hard* (also named *perfect*) interventions destroy the dependence of each intervened node from its parents in the DAG; by converse, *soft* interventions preserve the original parent-child relations, but allow for a modification of their "strength" (Yang et al., 2018). Hauser & Bühlmann (2012) provide a characterization of Markov equivalence under stochastic hard interventions and introduce the Greedy Interventional Equivalence Search (GIES) method as a score-based algorithm for structure learning of interventional equivalence classes. Hauser & Bühlmann (2015) present several statistical properties connected to the joint modeling of observational and interventional data and prove consistency of the adopted Bayesian information criterion.

In the Bayesian framework, structure learning is set up as a model selection problem which adopts the DAG marginal likelihood or equivalently the Bayes factor (Kass & Raftery, 1995; Carvalho & Scott, 2009), to derive a posterior distribution over the space of graphs. Posterior approximations are performed through Markov Chain Monte Carlo (MCMC) strategies; see for instance Chickering (2002), Consonni et al. (2017) and Castelletti (2020). To compute analytically the DAG marginal likelihood one needs to assign a suitable parameter prior distribution which is *constrained* to satisfy the conditional independencies encoded by the graph. For undirected decomposable graphs, Dawid & Lauritzen (1993) introduced Hyper Markov laws as a class of conjugate priors for graph-dependent model parameters and specialized them under both a categorical and Gaussian setting; moreover, Roverato (2002) introduced the general G-Wishart distribution for arbitrary, possibly non decomposable, undirected Gaussian graphical models. For DAG models with i.i.d. observational samples, Geiger & Heckerman (2002) proposed an effective method for prior construction, which requires as input the elicitation of a single prior for the parameter of a complete unconstrained DAG model and then derive compatibly the prior for any arbitrary DAG. Importantly, their approach assigns equal marginal likelihoods to Markov equivalent DAGs, besides leading to a closed-form expression in the Gaussian setting. More recently, Ben-David et al. (2015) introduced the DAG-Wishart distribution as a conjugate prior for the parameter of a Gaussian DAG model. Theoretical properties of the DAG-Wishart prior

under the restrictive assumption of known parent ordering are studied by Cao et al. (2019) who established graph selection consistency in high dimensional settings. Peluso & Consonni (2020) then extended the DAG-Wishart to arbitrary DAGs, without a pre-determined node-ordering, by showing that score equivalence is guaranteed only for specific hyperparameter choices.

In this paper we develop a Bayesian framework for the analysis of multivariate experimental data collected under stochastic hard interventions. Our contribution can be summarized as follows: (i) we introduce a new Bayesian model for partially intervened multivariate data, with a theoretically-guaranteed prior elicitation procedure on parameters indexing observational and interventional distributions; (ii) we demonstrate that our method guarantees score equivalence, i.e. same marginal likelihood, for I-Markov equivalent DAGs, thus generalizing the method of Geiger & Heckerman (2002), originally introduced for observational samples, to an interventional setting; (iii) we specialize our model to Gaussian DAGs and prove, up to I-Markov equivalence, the *posterior ratio consistency*: the true network structure can be asymptotically recovered, regardless of the distributional form of the intervened variables, and regardless of the relative asymptotic prevalence of observational or interventional measurements; therefore, we extend the model and results of Cao et al. (2019) designed for pure observational measurements.

2. Background: DAGs, interventions and Markov equivalence

2·1. *Markov properties*

Let $\mathcal{D} = (V, E)$ be a Directed Acyclic Graph (DAG), where $V = \{1, \dots, q\}$ is a finite set of nodes, or vertices, and $E \subset V \times V$ a set of edges. Elements of E are ordered pairs such as (u, v) and corresponding to directed edges of the form $u \to v$. We further assume that $\mathcal{D}$ does not have bi-directed edges, implying that if $(u, v) \in E$, then $(v, u) \notin E$, and cycles, that is paths of the form $u_1 \to u_2 \to \cdots \to u_k$ where $u_1 = u_k$. For a given DAG $\mathcal{D} = (V, E)$, we say that u is a *parent* of v in $\mathcal{D}$ if $(u, v) \in E$; conversely, v is a *child* of u. The set of all parents of u in $\mathcal{D}$ is then $\mathrm{pa}_{\mathcal{D}}(u)$, while $\mathrm{fa}_{\mathcal{D}}(u) = u \cup \mathrm{pa}_{\mathcal{D}}(u)$ is called the *family* of node u. A DAG $\mathcal{D}$ encodes a set of conditional independencies between nodes which can be read-off from the DAG using graphical criteria, such as *d*-separation (Pearl, 2000).

Consider a collection of random variables $X_1, \dots, X_q$ each associated with a node in $\mathcal{D}$, and with joint p.d.f. $f(\cdot)$. The latter factorizes according to $\mathcal{D}$ as

$$f(x_1, \dots, x_q \,|\, \mathcal{D}) = \prod_{j=1}^{q} f(x_j \,|\, \boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}}(j)}), \tag{1}$$

in which case we say that $f(\cdot)$ obeys the *Markov property* of DAG $\mathcal{D}$ (Lauritzen, 1996). Equation (1) is also known as the DAG factorization property and is equivalent to the local and global Markov properties if $f(\cdot)$ is strictly positive (Lauritzen, 1996).

Consider now two distinct DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$. These are called *Markov equivalent* if they encode the *same* set of conditional independencies. For a given DAG $\mathcal{D}$, the set of all DAGs that are Markov equivalent to $\mathcal{D}$ defines its Markov equivalence class, that we denote by $[\mathcal{D}]$. One further result in Andersson et al. (1997) shows that each equivalence class can be uniquely represented by a special partially directed graph named Essential Graph (EG). Most importantly for our purposes, Markov equivalence implies a *partition* of the DAG space into equivalence classes. Markov equivalent DAGs may differ by the orientation of some edges, while they share the same skeleton, that is the underlying undirected graph obtained by disregarding edge orientation; this feature follows from Verma & Pearl (1990) who show that any two Markov equivalent DAGs have the same skeleton and set of *v-structures*.

2·2. *Interventional Markov equivalence*

We now introduce interventions. Specifically, with a *perfect stochastic* intervention on the set of nodes $I \subset V$, the *intervention target*, we fix each X_j, $j \in I$, to the level of a random variable U_j with density $\widetilde{f}(u_j)$ and such that $\{U_j\}_{j\in I}$ are mutually independent. The *do-operator* (Pearl, 2000) is used to denote such an intervention and the *post-intervention* distribution of $X_1, \ldots, X_q$ can be written as

$$f(x_1, \ldots, x_q \,|\, \text{do}\{X_j = U_j\}_{j\in I}, \mathcal{D}) = \prod_{j\notin I} f(x_j \,|\, \boldsymbol{x}_{\text{pa}_{\mathcal{D}}(j)}) \prod_{j\in I} \widetilde{f}(x_j), \tag{2}$$

where it appears that for each $j \in I$ the original dependence of node j from its parents $\text{pa}_{\mathcal{D}}(j)$ is dropped and $f(x_j \,|\, \boldsymbol{x}_{\text{pa}_{\mathcal{D}}(j)})$ replaced by $\widetilde{f}(x_j)$. Under the specific case $I = \emptyset$, Equation (2) reduces to (1) which is also named the *observational* or pre-interventional distribution.

It is also common to deal with *multiple* and independent intervention experiments, corresponding to a *family* of intervention targets $\mathcal{I} = \{I_1, \ldots, I_K\}$, where $I_k \subset \{1, \ldots, q\}$ and each implying a post-intervention distribution of the form (2). A family of intervention targets is *conservative* (Hauser & Bühlmann, 2012, Definition 6) if for each $j \in \{1, \ldots, q\}$, there exists some $I \in \mathcal{I}$ such that $j \notin I$. This implies that for each node j there exists at least one intervention which does not involve j, a requirement which is always guaranteed whenever observational data are available.

Equation (2) shows that interventions modify the original DAG factorization and the corresponding Markov property. Hauser & Bühlmann (2012) extended the definition of Markov equivalence under interventions. Importantly, under a conservative family of intervention targets $\mathcal{I}$, $\mathcal{I}$-Markov equivalent DAGs are also observationally equivalent. Accordingly, interventions lead to a finer partition of DAGs into equivalence classes. Specifically, given the family of intervention targets $\mathcal{I}$, $\mathcal{D}_1$ and $\mathcal{D}_2$ are *interventionally* Markov, or $\mathcal{I}$-Markov, equivalent if, for each $I \in \mathcal{I}$, $\mathcal{D}_1^I$ and $\mathcal{D}_2^I$ encode the same conditional independencies, where $\mathcal{D}^I = (V, E^I)$, $E^I = \{(u, v) \in E \,|\, v \notin I\}$, is the so-called *intervention DAG* of $\mathcal{D} = (V, E)$ given the target I. Finally, if we let $[\mathcal{D}]_{\mathcal{I}}$ be the $\mathcal{I}$-Markov equivalence class of $\mathcal{D}$, i.e. the set of all DAGs that are $\mathcal{I}$-Markov equivalent to $\mathcal{D}$, each class can be again represented by a chain graph called $\mathcal{I}$-Essential Graph; see also Definition 11 in Hauser & Bühlmann (2012), to which we also refer for further theoretical results and characterizations of $\mathcal{I}$-Markov equivalence.

3. BAYESIAN DAG MODEL COMPARISON

3·1. *Model formulation*

Let $\mathcal{I} = \{I_1, \ldots, I_K\}$ be a family of targets, where $I_k \subset V$ for each $k = 1, \ldots, K$; notice that if $I_k = \emptyset$ the k-th dataset is purely observational. We assume that $f(\cdot)$ in (2) belongs to some parametric family indexed by a global parameter $(\boldsymbol{\theta}, \boldsymbol{\theta}^I)$ and write the post-intervention distribution as

$$f(x_1, \ldots, x_q \,|\, \text{do}\{X_j = U_j\}_{j\in I}, \boldsymbol{\theta}, \boldsymbol{\theta}^I, \mathcal{D}) = \prod_{j\notin I} f(x_j \,|\, \boldsymbol{x}_{\text{pa}_{\mathcal{D}}(j)}, \theta_j) \cdot \prod_{j\in I} \widetilde{f}(x_j \,|\, \boldsymbol{\theta}^I), \tag{3}$$

where in particular $\boldsymbol{\theta} = \{\theta_j\}_{j\notin I}$ is a collection of node-parameters relative to node-conditional observational distributions, while $\boldsymbol{\theta}^I = \{\theta_j^I\}_{j\in I}$ are "interventional" node-parameters. We further assume that for each intervention experiment $k = 1, \ldots, K$, $n^{(k)}$, with target I_k, i.i.d. observations from (3) $\left\{\boldsymbol{x}_i^{(k)}, i = 1, \ldots, n^{(k)}\right\}$ are available, and let $\boldsymbol{X}^{(k)}$ be the corresponding $n^{(k)} \times q$ data matrix. To link observations to intervention targets we introduce the multiset

$\mathcal{T} = (T^{(1)}, \ldots, T^{(n)})$, where $T^{(i)} \in \mathcal{I}$ is the intervention target under which observation i was collected. By assuming independence across interventions, the likelihood function can be written as

$$f(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}, \mathcal{D}) = \prod_{k=1}^{K} \left\{ \prod_{i=1}^{n^{(k)}} \left\{ \prod_{j \notin I_k} f\big(x_{i,j}^{(k)} \mid \boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(j)}^{(k)}, \theta_j\big) \cdot \prod_{j \in I_k} \widetilde{f}_k\big(x_{i,j}^{(k)} \mid \theta_j^{(k)}\big) \right\} \right\}$$
$$= \prod_{k=1}^{K} \left\{ \prod_{j \notin I_k} f\big(\boldsymbol{X}_j^{(k)} \mid \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}^{(k)}, \theta_j\big) \prod_{j \in I_k} \widetilde{f}_k\big(\boldsymbol{X}_j^{(k)} \mid \theta_j^{(k)}\big) \right\}, \qquad (4)$$

where $\boldsymbol{X}_S^{(k)}$ denotes the $n^{(k)} \times |S|$ sub-matrix of $\boldsymbol{X}^{(k)}$ with columns belonging to the set $S \subseteq \{1, \ldots, q\}$ and $\boldsymbol{X} = \big(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}\big)^\top$ the $n \times q$ data matrix, $n = \sum_{k=1}^{K} n^{(k)}$. Notice that in (4) we avoid the conditioning on the do-operator which for clarity was included in the post-intervention distribution (3) since we consider interventions with targets that are known *a priori*; see also Castelletti & Peluso (2022) for a comparison with interventions on uncertain targets. Also notice that $\boldsymbol{\theta}$ is common to all the K terms, while $\boldsymbol{\theta}^{(k)}$ is specific to intervention k. If we now let $\mathcal{A}(j) = \{i \in \{1, \ldots, n\} : j \notin T^{(i)}\}$, then Equation (4) can be written as

$$f\big(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}, \mathcal{D}\big) = \prod_{j=1}^{q} f\big(\boldsymbol{X}_j^{\mathcal{A}(j)} \mid \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)}, \theta_j\big) \cdot \prod_{k=1}^{K} \left\{ \prod_{j \in I_k} \widetilde{f}_k\big(\boldsymbol{X}_j^{(k)} \mid \theta_j^{(k)}\big) \right\} \qquad (5)$$

where now $\boldsymbol{X}^{\mathcal{A}(j)}$ denotes the sub-matrix of $\boldsymbol{X}$ with rows corresponding to $\mathcal{A}(j)$. We emphasize that, as long as we assume a conservative family of targets, all terms $f(\cdot)$ in the first product exist.

3·2. *Prior parameter elicitation*

Prior elicitation for DAG model parameters requires specific care. In particular, an important requirement is that any two DAGs sharing the same I-Markov property, i.e. I-Markov equivalent DAGs, are score equivalent, namely they are assigned the same *marginal likelihood*. The latter corresponds to the likelihood function (4) which is integrated w.r.t. the prior on model parameters $(\boldsymbol{\theta}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)})$,

$$m(\boldsymbol{X} \mid \mathcal{D}) = \int f\big(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}, \mathcal{D}\big)\, p\big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)} \mid \mathcal{D}\big)\, d\big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}\big).$$

Relevant to our method, Heckerman et al. (1995) and Geiger & Heckerman (2002) introduce a set of assumptions that guarantee score equivalence in the case of observational i.i.d. samples. They start by assuming some conditions on the likelihood, namely *complete model equivalence, regularity, likelihood modularity*, which are satisfied by any Gaussian and categorical graphical model. As mentioned however, the distinctive feature of the approach concerns the prior construction, which is based on the following two assumptions. The first assumption of *prior modularity* states that, given two distinct DAG models with the *same* set of parents for vertex j, the prior for the node-parameter θ_j must be the same under both models.

The second assumption of *global parameter independence* assumes that for every DAG model $\mathcal{D}$, the parameters $\{\theta_j;\ j = 1, \ldots, q\}$ should be *a priori* independent. As a result, the parameter prior for any DAG model can be derived from a unique prior on the parameter of an arbitrary unconstrained complete DAG. Moving back to our interventional setting, consider now the likelihood function in (5), which consists of two terms. The first one reflects the DAG factorization which is imposed to the likelihood $f(\cdot)$; also, the node-parameters $\{\theta_j\}_{j=1}^{q}$, each indexing the

conditional distribution of node j in DAG $\mathcal{D}$ are specific to each DAG under consideration and therefore DAG-dependent. We assume for both the likelihood and priors in this term the same assumptions of Geiger & Heckerman (2002), that is Assumptions 1-5 in the original paper. In particular, we first need the prior modularity assumption, exactly as in Geiger & Heckerman (2002), limited to the observational parameters, being the interventional parameters detached from the graphical structures.

On the other hand, the global independence assumption of Geiger & Heckerman (2002) needs now to be extended to the interventional parameters, and to the relationship between observational and interventional parameters. Because of prior independence across parameters θ_j's, we have $p(\boldsymbol{\theta} \mid \mathcal{D}) = \prod_{j=1}^{q} p(\theta_j \mid \mathcal{D})$. Now notice that the second term in (5) corresponds to an unconditional likelihood which does not depend anymore on the DAG and with the node-parameters $\{\theta_j^{(k)}, k = 1, \ldots, K, j \in I_k\}$ indexing an *unconditional* marginal distribution now "free" from the original DAG structure.

Each of the terms $\widetilde{f}_k$ is therefore the same across *all* DAG models. For what follows, we only assume that for each $k = 1, \ldots, K$ the "interventional" node-parameters are *a priori* independent, that is $p(\boldsymbol{\theta}^{(k)}) = \prod_{j \in I_k} p\big(\theta_j^{(k)}\big)$, and also global independence among all parameters $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}$. We finally assume that the *joint* prior on $\boldsymbol{\theta}$ and $\big(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}\big)$ factorizes as $p\big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}\big) = p(\boldsymbol{\theta} \mid \mathcal{D})\, p\big(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}\big)$, so that parameters indexing observational and interventional densities are also *a priori* independent. The above line of reasoning is then summarized in the following assumption, that we name of joint, for observational and interventional, global parameter independence.

Assumption 1. (Joint global parameter independence) For a DAG $\mathcal{D}$ and intervention targets $I_1, \ldots, I_K$, we have that $p\big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)} \mid \mathcal{D}\big) = p(\boldsymbol{\theta} \mid \mathcal{D}) \prod_{k=1}^{K} \prod_{j \in I_k} p\big(\theta_j^{(k)}\big)$.

3·3. *Marginal likelihood and Bayes factor computation*

We now focus on the computation of $m(\boldsymbol{X} \mid \mathcal{D})$, the marginal likelihood of DAG $\mathcal{D}$. Because of the assumptions in Section 3·2 we obtain

$$m(\boldsymbol{X} \mid \mathcal{D}) = \prod_{j=1}^{q} \left\{ \frac{m\big(\boldsymbol{X}_{\mathrm{fa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} \mid \mathcal{C}\big)}{m\big(\boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} \mid \mathcal{C}\big)} \prod_{k : j \in I_k} m\big(\boldsymbol{X}_j^{(k)}\big) \right\}, \tag{6}$$

where $m(\cdot \mid \mathcal{C})$ denotes the marginal likelihood computed under any *complete* DAG model $\mathcal{C}$; see also Section 1 of the Supplementary material for full details on the derivation of (6).

Consider now two DAGs $\mathcal{D}_0$ and $\mathcal{D}_1$, differing by one edge, say $u \to v$, which is contained in $\mathcal{D}_1$ and reversed in $\mathcal{D}_0$ (Figure 1). Let also pa be the set of common parents of nodes u and v, possibly an empty set. The Bayes Factor (BF) of $\mathcal{D}_0$ against $\mathcal{D}_1$,

$$\mathrm{BF}(\mathcal{D}_0, \mathcal{D}_1) := \frac{m(\boldsymbol{X} \mid \mathcal{D}_0)}{m(\boldsymbol{X} \mid \mathcal{D}_1)}, \tag{7}$$

then simplifies to

$$\mathrm{BF}(\mathcal{D}_0, \mathcal{D}_1) = \prod_{j \in \{u, v\}} \left\{ \frac{m\big(\boldsymbol{X}_{\mathrm{fa}_{\mathcal{D}_0}(j)}^{\mathcal{A}(j)}\big)}{m\big(\boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}_0}(j)}^{\mathcal{A}(j)}\big)} \cdot \frac{m\big(\boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}_1}(j)}^{\mathcal{A}(j)}\big)}{m\big(\boldsymbol{X}_{\mathrm{fa}_{\mathcal{D}_1}(j)}^{\mathcal{A}(j)}\big)} \right\}, \tag{8}$$

where we omit the conditioning on $\mathcal{C}$ to simplify the notation.

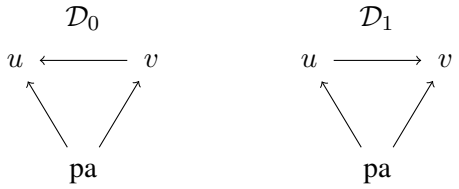

Fig. 1: Two DAGs, $\mathcal{D}_0, \mathcal{D}_1$, differing by an edge reversal between u and v. pa is the set of common parents of nodes u and v.

3·4. *Gaussian DAG models*

In the following we assume that the joint density $f(\cdot)$ in Equation (1) is that of a zero-mean multivariate Normal distribution, namely

$$X_1, \ldots, X_q \,|\, \boldsymbol{\Omega} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Omega}^{-1}), \tag{9}$$

where $\boldsymbol{\Omega} \in \mathcal{P}_{\mathcal{D}}$ corresponds to the precision matrix, inverse of the covariance matrix $\boldsymbol{\Sigma}$, and $\mathcal{P}_{\mathcal{D}}$ is the space of all symmetric positive definite (s.p.d.) matrices Markov w.r.t. DAG $\mathcal{D}$. An equivalent formulation is in terms of the corresponding Structural Equation Model (SEM),

$$\boldsymbol{L}^\top \boldsymbol{x} = \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{D}) \tag{10}$$

where $\boldsymbol{L}$ is a $q \times q$ matrix of regression coefficients such that $\boldsymbol{L}_{u,u} = 1$ for each $u = 1, \ldots, q$ and $\boldsymbol{L}_{u,v} \neq 0$, for each $u \neq v$ if and only if $u \to v$ is in $\mathcal{D}$. Moreover, $\boldsymbol{D}$ is a $q \times q$ diagonal matrix collecting node-conditional variances, $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{D}_{11}, \ldots, \boldsymbol{D}_{qq})$. Equation (10) implies

$$f(x_1, \ldots, x_q \,|\, \boldsymbol{D}, \boldsymbol{L}, \mathcal{D}) = \prod_{j=1}^{q} \phi(x_j \mid -\boldsymbol{L}_{\prec j]}^\top \boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}}(j)}, \boldsymbol{D}_{jj}), \tag{11}$$

where $\prec j] = \mathrm{pa}_{\mathcal{D}}(j) \times j$, $\boldsymbol{L}_{A \times B}$ is the sub-matrix of $\boldsymbol{L}$ with columns and rows indexed by A and B respectively, and $\phi(\cdot \,|\, \mu, \sigma^2)$ denotes the p.d.f. of a $\mathcal{N}(\mu, \sigma^2)$. Equation (11) resembles the DAG factorization in (1) and adopts the re-parameterization $\boldsymbol{\Sigma} \mapsto (\boldsymbol{L}, \boldsymbol{D})$ such that $\boldsymbol{\Sigma} = \boldsymbol{L}^{-\top} \boldsymbol{D} \boldsymbol{L}^{-1}$, where $\boldsymbol{L}^{-\top} := \left(\boldsymbol{L}^\top\right)^{-1}$. Therefore, it corresponds to the observational distribution in the Gaussian setting. The post-intervention distribution for an intervention on $I \subset \{1, \ldots, q\}$ in (3) becomes under the Gaussian assumption

$$\begin{aligned} f(x_1, \ldots, x_q \,|\, \mathrm{do}\{X_j = U_j\}_{j \in I}, \boldsymbol{L}, \boldsymbol{D}, \{\delta_j\}_{j \in I}, \mathcal{D}) = & \prod_{j \notin I} \phi(x_j \mid -\boldsymbol{L}_{\prec j]}^\top \boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}}(j)}, \boldsymbol{D}_{jj}) \\ & \cdot \prod_{j \in I} \phi(x_j \,|\, 0, \delta_j), \end{aligned} \tag{12}$$

where $\{\delta_j\}_{j \in I}$ are the interventional parameters, here corresponding to conditional variances of variables X_j, $j \in I$. The post-intervention covariance matrix, $\widetilde{\boldsymbol{\Sigma}}$, is therefore $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{L}}^{-\top} \widetilde{\boldsymbol{D}} \widetilde{\boldsymbol{L}}^{-1}$ where $\widetilde{\boldsymbol{L}}_{u,v} = 0$ if $v \in I$ and $u \neq v$, while $\widetilde{\boldsymbol{L}}_{u,v} = \boldsymbol{L}_{u,v}$ otherwise, and $\widetilde{\boldsymbol{D}}$ is obtained from $\boldsymbol{D}$ by replacing elements $\boldsymbol{D}_{u,u}$, for each $u \in I$, with δ_u. As in Section 3·1, by assuming K independent

interventions, each with target I_k, the likelihood function in (5) can be written as

$$f\big(\boldsymbol{X} \mid \boldsymbol{L}, \boldsymbol{D}, \boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(K)}, \mathcal{D}\big) = \prod_{j=1}^{q} \phi_{|\mathcal{A}(j)|}\left(\boldsymbol{X}_j^{\mathcal{A}(j)} \mid -\boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} \boldsymbol{L}_{\prec j]}, \boldsymbol{D}_{jj} \boldsymbol{I}_{|\mathcal{A}(j)|}\right) \cdot \prod_{k=1}^{K} \left\{ \prod_{j \in I_k} \phi_{n^{(k)}}\left(\boldsymbol{X}_j^{(k)} \mid \boldsymbol{0}, \delta_j^{(k)} \boldsymbol{I}_{n^{(k)}}\right) \right\}, \tag{13}$$

where $\boldsymbol{\delta}^{(k)} = \{\delta_j^{(k)}\}_{j \in I_k}$, for $k = 1, \ldots, K$, $\phi_d(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p.d.f. of a $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution and $\boldsymbol{I}_d$ is the $d \times d$ identity matrix. To compute the marginal likelihood in (6) and the BF in (8) we only need to specify a prior for the parameter of the observational Gaussian distribution of a complete DAG model (9), with $\boldsymbol{\Omega}$ s.p.d. but otherwise unconstrained. Geiger & Heckerman (2002) show that a Wishart prior assigned to $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ satisfies the assumptions of prior modularity and global parameter independence required to obtain the marginal likelihood in (6). Accordingly we set

$$\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{U}), \tag{14}$$

a Wishart distribution with parameters $a > q - 1$ and $\boldsymbol{U}$, a (q, q) s.p.d. matrix, having expectation $\mathbb{E}(\boldsymbol{\Omega}) = a\boldsymbol{U}^{-1}$. We then write

$$p(\boldsymbol{\Omega}) = c(a, \boldsymbol{U}) |\boldsymbol{\Omega}|^{\frac{a-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{U})\right\}, \qquad c(a, \boldsymbol{U}) = \frac{|\boldsymbol{U}|^{\frac{a}{2}}}{2^{\frac{aq}{2}} \Gamma_q\left(\frac{a}{2}\right)},$$

under which we obtain (Supplementary material, Section 2) the marginal likelihood restricted to $\boldsymbol{X}_B$, a generic submatrix of the $n \times q$ matrix $\boldsymbol{X}$, with columns indexed by $B \subseteq \{1, \ldots, q\}$,

$$m(\boldsymbol{X}_B) = \pi^{-\frac{n|B|}{2}} \frac{|\boldsymbol{U}_{BB}|^{\frac{a-|\bar{B}|}{2}}}{|\boldsymbol{U}_{BB} + \boldsymbol{S}_{BB}|^{\frac{a-|\bar{B}|+n}{2}}} \frac{\Gamma_{|B|}\left(\frac{a-|\bar{B}|+n}{2}\right)}{\Gamma_{|B|}\left(\frac{a-|\bar{B}|}{2}\right)}. \tag{15}$$

This general formula applied to each term $m(\cdot)$ in Equation (8) specializes the BF to the Gaussian setting.

4. THEORETICAL PROPERTIES

In this section we present our main results, for which our model is able to consistently detect *a posteriori* the true DAG, or, more precisely, the interventional equivalence class $[\mathcal{D}_0]_{\mathcal{I}}$ to which the true DAG $\mathcal{D}_0$ belongs. Furthermore, it guarantees score equivalence among graphs within the same interventional equivalence class, under a conservative family of intervention targets $\mathcal{I}$. We need to distinguish three settings, according to the relative asymptotic dominance between observational and interventional measurements, since they correspond to three distinct behaviours of the posterior ratio of the graphs under comparison. Specifically, we say that a node $u \in I$ is a *balanced target* if $n^{\mathcal{A}(u)}/n \to \alpha \in (0, 1)$, where $n^{\mathcal{A}(u)}$ is the number of observational measurements of X_u, with the meaning that asymptotically a proportion $(1 - \alpha)$ of measurements will be interventional for this variable; furthermore, we say that $u \in I$ is an *observationally dominant target*, or an *interventionally dominant target*, if, respectively, $n^{\mathcal{A}(u)}/n \to 1$ or $n^{\mathcal{A}(u)}/n \to 0$.

We will first show that the prior elicitation procedure introduced in Section 3·2 guarantees *score equivalence*, namely that any two DAGs $\mathcal{D}_0$ and $\mathcal{D}_1$ that are $\mathcal{I}$-Markov equivalent, are assigned the same marginal likelihood. For this purpose, recall Theorem 2 in Chickering (1995)

which shows that two DAGs $\mathcal{D}_0$, $\mathcal{D}_1$ are Markov equivalent if and only if there exists a sequence of edge reversals transforming $\mathcal{D}_0$ into $\mathcal{D}_1$ having the following properties: i) after each reversal the resulting graph is a DAG belonging to the Markov equivalence class of $\mathcal{D}_0$ and $\mathcal{D}_1$; ii) each reversed arc is *covered*. In particular, an arc $u \to v$ is covered in $\mathcal{D}$ if $\mathrm{pa}_{\mathcal{D}}(v) = \mathrm{pa}_{\mathcal{D}}(u) \cup u$. Moreover, the length of the sequence is $|\Delta(\mathcal{D}_0, \mathcal{D}_1)|$, where $\Delta(\mathcal{D}_0, \mathcal{D}_1)$ is the set of edges in $\mathcal{D}_0$ that have opposite orientation in $\mathcal{D}_1$. In the Supplementary material, we show that such a sequence of graphs also exists within an $\mathcal{I}$-Markov equivalent class, and we use this result to further show the following proposition on score equivalence of $\mathcal{I}$-Markov equivalent graphs.

PROPOSITION 1. *(Score equivalence) Let $\mathcal{I}$ be a conservative family of targets, $\mathcal{D}_0$ and $\mathcal{D}_1$ two $\mathcal{I}$-Markov equivalent DAGs, Then, $\mathcal{D}_0$ and $\mathcal{D}_1$ have the same marginal likelihood, namely $m(\boldsymbol{X} \mid \mathcal{D}_0) = m(\boldsymbol{X} \mid \mathcal{D}_1)$, with $m(\boldsymbol{X} \mid \mathcal{D})$ as in* (6)*.*

Proof. See the Supplementary material. □

To understand now how we correctly identify $[\mathcal{D}_0]_{\mathcal{I}}$, the $\mathcal{I}$-Markov equivalence class of the true DAG $\mathcal{D}_0$, it is crucial to focus on the comparison among graphs which are equivalent before the interventions, but whose equivalence is broken after the interventions. In other words, these graphs are observationally, but not interventionally, equivalent: this occurs when $\mathcal{D}_0$ is compared with $\mathcal{D} \in [\mathcal{D}_0]$ but $\mathcal{D} \notin [\mathcal{D}_0]_{\mathcal{I}}$. For such two graphs $\mathcal{D}_0$ and $\mathcal{D}$, we find the exact rate of convergence of the posterior ratio when the edges involved in the intervention are not *strong protected* in $\mathcal{D}_0$, i.e their reversal does not break observational Markov equivalence; see also Definition 3.3 in Andersson et al. (1997). This rate is at least the rate of the posterior ratio when graphs neither observationally nor interventionally equivalent are compared, in line with the intuition that these latter graphs tend to be more easily discriminated. When these observationally equivalent graphs are compared, we know from Chickering (1995) that there exists a graph sequence in which *adjacent* graphs are equal, with the exception of $u \to v$ reversed as $v \to u$, with u and v having the same parents pa, of cardinality $|\mathrm{pa}| = p \geq 0$, in both graphs.

It turns out that in our setting it is relevant to analyze the matrix

$$\boldsymbol{A}(u \mid \mathcal{D}) = \left(\boldsymbol{\Sigma}_{0,\mathrm{fa}_{\mathcal{D}}(u)}\right)^{-1} \widetilde{\boldsymbol{\Sigma}}_{0,\mathrm{fa}_{\mathcal{D}}(u)}$$

for the intervened node $u \in V$, where $\boldsymbol{\Sigma}_0$ and $\widetilde{\boldsymbol{\Sigma}}_0$ denote respectively pre- and post-intervention true covariance matrices; see also Section 3·4. For brevity, we may omit the dependence on the graphs when clear from the context. If in $\mathcal{D}_0$ there are some edges involving the intervened node u that are not strong protected in $\mathcal{D}_0$, it exists a pair of adjacent graphs $\mathcal{D}_j$ and $\mathcal{D}_{j+1}$ in the sequence of Chickering (1995) for which u is part of a covered edge. We therefore study how the intervention on node u breaks the equivalence between the two adjacent graphs, and how this is reflected in terms of posterior ratio. If on the other hand all edges of u are strong protected, v has to be intended as an empty set, and $\boldsymbol{A}(u \mid \mathcal{D}_j) = \boldsymbol{A}(u \mid \mathcal{D}_{j+1})$. We define the ordered, from smallest to largest, eigenvalues of $\boldsymbol{A}(u \mid \mathcal{D})$ as $\lambda_j(u \mid \mathcal{D})$ for $j = 1, \ldots, |\mathrm{fa}_{\mathcal{D}}(u)|$. In particular the minimum and maximum eigenvalues are denoted respectively as $\underline{\lambda}(u \mid \mathcal{D})$ and $\overline{\lambda}(u \mid \mathcal{D})$.

We are then ready in the following propositions to show posterior ratio consistency to the true interventional equivalence class when all targets are interventionally dominant, observationally dominant, or balanced. Their proofs are contained within the Supplementary material, in a more general proposition that includes the three cases. In the first of these results, we show that the true interventional equivalence class is consistently identified, when the oservational measurements asymptotically dominate the interventions, as long the number of interventions $n - n^{\mathcal{A}(u)}$ increases in n.

PROPOSITION 2. *(Observationally dominant setting) Let $\mathcal{D}_0$ be the true DAG. Under the assumptions of Section 3·2, consider a prior $\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{U})$ and the likelihood function in Equation* (13)*. For conservative interventions on u, and with $n^{\mathcal{A}(u)}/n \to 1$ as $n \to \infty$, we have, for all $\mathcal{D} \notin [\mathcal{D}_0]_{\mathcal{I}}$,*

$$\frac{p(\mathcal{D} \mid \boldsymbol{X}, \mathcal{I})}{p(\mathcal{D}_0 \mid \boldsymbol{X}, \mathcal{I})} = O_{\bar{P}}\left(C_\alpha \frac{p(\mathcal{D})}{p(\mathcal{D}_0)} \prod_{j=0}^{J-1} \left(\frac{\boldsymbol{\Sigma}_{0,u \mid \mathrm{pa}_{\mathcal{D}_{j+1}}(u)}}{\boldsymbol{\Sigma}_{0,u \mid \mathrm{pa}_{\mathcal{D}_j}(u)}} \frac{|e^{\boldsymbol{A}(u \mid \mathcal{D}_{j+1})}|}{|e^{\boldsymbol{A}(u \mid \mathcal{D}_j)}|} \right)^{\frac{1}{2}(n - n^{\mathcal{A}(u)})} \right),$$

where $\{\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_J = \mathcal{D}\}$ is a sequence of observationally equivalent adjacent graphs, C_α some constant. Furthermore, for all $\mathcal{D} \in [\mathcal{D}_0]_{\mathcal{I}}$, we have $p(\mathcal{D} \mid \boldsymbol{X}, \mathcal{I})/p(\mathcal{D}_0 \mid \boldsymbol{X}, \mathcal{I}) = p(\mathcal{D})/p(\mathcal{D}_0)$ $\bar{P}-$almost surely.

Proof. See the Supplementary material. □

The above proposition tells that the posterior ratio consistency rate to the true $[\mathcal{D}_0]_{\mathcal{I}}$ depends on $n - n^{\mathcal{A}(u)}$. In particular, $\log \mathrm{BF}(\mathcal{D}, \mathcal{D}_0) \leq -\frac{C}{2}(n - n^{\mathcal{A}(u)})$, $\bar{P}$ almost surely, for some C not dependent on n and for a sufficiently large n. When $n^{\mathcal{A}(u)} \sim n - \log n$, or $n^{\mathcal{A}(u)} \sim n - n^\beta$, for some $\beta \in (0, 1)$, we have Bayes factor consistency, and posterior ratio consistency, respectively at rate at least n and $\exp\{n^\beta\}$, even if asymptotically the observational part of the data will dominate. On the other hand, if $n^{\mathcal{A}(u)} \sim n - k$, for some constant number of interventions $k > 0$ not dependent on n, there is not necessarily consistency: still, $\log \mathrm{BF}(\mathcal{D}, \mathcal{D}_0) \leq -Ck/2$ with $\bar{P}$ probability 1 for n sufficiently large, suggesting evidence in favour of the true graph $\mathcal{D}_0$.

Proposition 2 also suggests that, after an intervention on u, the convergence rate to $[\mathcal{D}_0]_{\mathcal{I}}$ is not affected by $v \to u$ or $u \to v$ being present in the true edge set E_0. Nevertheless, the constant C_α is affected: for those targets with $v \to u \in E_0$ and $\mathcal{I}-$strong protected, as in Definition 14 of Hauser & Bühlmann (2012), it is more beneficial to intervene on nodes with $\boldsymbol{\Sigma}_{u \mid \mathrm{pa},v} |\boldsymbol{A}(u \mid \mathcal{D}_j)| >> \boldsymbol{\Sigma}_{u \mid \mathrm{pa}} |\boldsymbol{A}(u \mid \mathcal{D}_{j+1})|$; on the other hand, for those targets with $u \to v \in E_0$ and $\mathcal{I}-$strong protected, from the proof of Proposition 1 we have $|\boldsymbol{A}(u \mid \mathcal{D}_j)| = |\boldsymbol{A}(u \mid \mathcal{D}_{j+1})|$ and then targets with $\boldsymbol{\Sigma}_{u \mid \mathrm{pa},v} << \boldsymbol{\Sigma}_{u \mid \mathrm{pa}}$ should be privileged, looking only at the pre-intervention covariance matrix.

In the next result, we demonstrate the asymptotic correct identification of the true class when interventions asymptotically dominate observational measurements. Relative to the previous scenario and in line with intuition, it is now easier to find the correct graph class, and the requirement of an $n^{\mathcal{A}(u)}$ increasing in n is not needed.

PROPOSITION 3. *(Interventionally dominant setting) In the setting of Proposition* 2*, for conservative interventions on u, and with $n^{\mathcal{A}(u)}/n \to 0$ as $n \to \infty$, we have, for all $\mathcal{D} \notin [\mathcal{D}_0]_{\mathcal{I}}$,*

$$\frac{p(\mathcal{D} \mid \boldsymbol{X}, \mathcal{I})}{p(\mathcal{D}_0 \mid \boldsymbol{X}, \mathcal{I})} = O_{\bar{P}}\left(C_\alpha \frac{p(\mathcal{D})}{p(\mathcal{D}_0)} \prod_{j=0}^{J-1} \left(\frac{n^{\mathcal{A}(u)}}{n} \right)^{k_j(u)} \left(\frac{\boldsymbol{\Sigma}_{0,u \mid \mathrm{pa}_{\mathcal{D}_{j+1}}(u)}}{\boldsymbol{\Sigma}_{0,u \mid \mathrm{pa}_{\mathcal{D}_j}(u)}} \right)^{\frac{1}{2}(n - n^{\mathcal{A}(u)})} \left(\frac{|\boldsymbol{A}(u \mid \mathcal{D}_{j+1})|}{|\boldsymbol{A}(u \mid \mathcal{D}_j)|} \right)^{\frac{n}{2}} \right),$$

where $\{\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_J = \mathcal{D}\}$ is a sequence of observationally equivalent adjacent graphs, C_α some constant, $k_j(u) = \frac{1}{2}(|\mathrm{pa}_{\mathcal{D}_j}(u)| - |\mathrm{pa}_{\mathcal{D}_{j+1}}(u)|)$. Furthermore, for all $\mathcal{D} \in [\mathcal{D}_0]_{\mathcal{I}}$, we have $p(\mathcal{D} \mid \boldsymbol{X}, \mathcal{I})/p(\mathcal{D}_0 \mid \boldsymbol{X}, \mathcal{I}) = p(\mathcal{D})/p(\mathcal{D}_0)$ $\bar{P}-$almost surely.

Proof. See the Supplementary material. □

The implications of Proposition 3 in the interventionally dominant setting are considerably different than in the previous scenario: we now have

$$\log \mathrm{BF}(\mathcal{D}, \mathcal{D}_0) \leq -\frac{1}{2} \sum_j C_j \left[n \mathbb{1}(u \to v_j \in E_0) + (n^{A(u)} + \log n) \mathbb{1}(v_j \to u \in E_0) \right]$$

$\bar{P}$ almost surely, for a sufficiently large n, constant C_j, and the sum is intended over $j = 1, \ldots, J-1$ for which $u \to v_j$ or $v_j \to u$ is $\mathcal{I}-$strong protected. This shows that the convergence rate can be better for those targets u for which $u \to v$ is in the true edge set, always at rate at least $\exp\{n\}$, regardless of the specific behaviour of $n^{\mathcal{A}(u)}$. On the other hand, for a target u with $v_j \to u \in E_0$ for all $j = 0, \ldots, J-1$, $\mathcal{I}-$strong protected for some j, the identification of the true graph depends on $n^{\mathcal{A}(u)}$: if $n^{\mathcal{A}(u)} \sim \log n$, or $n^{\mathcal{A}(u)} \sim n^{\beta}$, for some $\beta \in (0,1)$, we have a rate at least $\sqrt{n}$ and $\exp\{n^{\beta}/2\}$, respectively. If $n^{\mathcal{A}(u)} \sim k > 0$ independent of n, we have Bayes factor consistency at rate at least $\sqrt{n}$. Therefore, Proposition 3 suggests to choose as targets in an interventionally dominant setting, those u with $u \to v$ in E_0 and, among these, those nodes showing $\boldsymbol{\Sigma}_{0,u \,|\, \mathrm{pa},v} << \boldsymbol{\Sigma}_{0,u \,|\, \mathrm{pa}}$.

In the next result we balance purely observational and interventional data, and show that we always identify the correct graph, at a rate that increases in the proportion of interventional measurements. Convergence is better than in the obervationally dominant setting, in line with intuition, but not always worse than in the interventionally dominant case.

PROPOSITION 4. *(Balanced setting) In the setting of Proposition* 2*, for conservative interventions on* u*, and with* $n^{\mathcal{A}(u)}/n \to \alpha \in (0,1)$ *as* $n \to \infty$*, we have, for all* $\mathcal{D} \notin [\mathcal{D}_0]_{\mathcal{I}}$*,*

$$\frac{p(\mathcal{D} \,|\, \boldsymbol{X}, \mathcal{I})}{p(\mathcal{D}_0 \,|\, \boldsymbol{X}, \mathcal{I})} = O_{\bar{P}} \left(C_\alpha \frac{p(\mathcal{D})}{p(\mathcal{D}_0)} \prod_{j=0}^{J-1} \left(\frac{\boldsymbol{\Sigma}_{0,u \,|\, \mathrm{pa}_{\mathcal{D}_{j+1}}(u)}}{\boldsymbol{\Sigma}_{0,u \,|\, \mathrm{pa}_{\mathcal{D}_j}(u)}} \right)^{\frac{n}{2}(1-\alpha)} M_{j,\alpha}(u)^{\frac{n}{2}} \right),$$

where $\{\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_J = \mathcal{D}\}$ *is a sequence of observationally equivalent adjacent graphs,* C_α *some constant, and*

$$M_{j,\alpha}(u) = \frac{\alpha + (1-\alpha)\underline{\lambda}(u \,|\, \mathcal{D}_{j+1})}{\alpha + (1-\alpha)\underline{\lambda}(u \,|\, \mathcal{D}_j)} \frac{\alpha + (1-\alpha)\overline{\lambda}(u \,|\, \mathcal{D}_{j+1})}{\alpha + (1-\alpha)\overline{\lambda}(u \,|\, \mathcal{D}_j)}.$$

Furthermore, for all $\mathcal{D} \in [\mathcal{D}_0]_{\mathcal{I}}$*, we have* $p(\mathcal{D} \,|\, \boldsymbol{X}, \mathcal{I})/p(\mathcal{D}_0 \,|\, \boldsymbol{X}, \mathcal{I}) = p(\mathcal{D})/p(\mathcal{D}_0)$ $\bar{P}-$*almost surely.*

Proof. See the Supplementary material. □

In the balanced setting, we have

$$\log \mathrm{BF}(\mathcal{D}, \mathcal{D}_0) \leq -\frac{1}{2} \sum_j C_j \left[n(1-\alpha) \mathbb{1}(u \to v_j \in E_0) + n \mathbb{1}(v_j \to u \in E_0) \right]$$

$\bar{P}$ almost surely, for a sufficiently high n, constant C_j, and the sum again intended over $j = 1, \ldots, J-1$ for which $u \to v_j$ or $v_j \to u$ is $\mathcal{I}-$strong protected. Therefore the Bayes factor consistency rate is $\exp\{n(1-\alpha)\}$ or $\exp\{n\}$, depending on the direction in $\mathcal{D}_0$ of the edges involving the target u. Intuitively, convergence improves for lower values of α, that is for a higher proportion of interventional measurements, but only when $u \to v_j$, whilst the convergence is faster when $v_j \to u$. Interestingly, relative to the interventionally dominant setting, balanced interventions induce a worsening in the convergence rate that depends on the direction of the $\mathcal{I}-$strong protected edges of u in $\mathcal{D}_0$: for $v_j \to u \in E_0$ and $\mathcal{I}-$strong protected, the rate remains

$\exp\{n\}$ in both cases; for $u \to v_j \in E_0$ and $\mathcal{I}-$strong protected, we assist on the contrary to an improvement in rate for balanced interventions, to $\exp\{n(1-\alpha)\}$. Finally, in line with the observationally equivalent setting, for those targets with $v \to u \in E_0$ and $\mathcal{I}-$strong protected, the result suggests to intervene on nodes with $\boldsymbol{\Sigma}_{u\,|\,\mathrm{pa},v}|\boldsymbol{A}(u\,|\,\mathcal{D}_j)| >> \boldsymbol{\Sigma}_{u\,|\,\mathrm{pa}}|\boldsymbol{A}(u\,|\,\mathcal{D}_{j+1})|$; whilst, if $u \to v \in E_0$ and $\mathcal{I}-$strong protected, targets with $\boldsymbol{\Sigma}_{u\,|\,\mathrm{pa},v} << \boldsymbol{\Sigma}_{u\,|\,\mathrm{pa}}$ are preferred.

In Section 3 of the Supplementary material we provide a first result and related considerations for the high-dimensional case, where the number of nodes q increases with n. In particular, we show that, under an additional assumption on neighbourhood sparsity, posterior ratio consistency outside the equivalence class is still valid, for the balanced setting, in the particular case of $n/q_n \to \beta > 1$. We conjecture that similar results exist in the observationally dominant and interventionally dominant settings, and when $n/q_n \to \beta < 1$ or $n/q_n \to 0$, but the whole treatment of all relevant cases is beyond the scope of the current work and is left as future research.

5. EMPIRICAL ANALYSES

5·1. *Simulated validations*

In this section we investigate the asymptotic behaviour of the Bayes Factor (BF) through simulation experiments. Specifically, we consider the BF of $\mathcal{D}_0$ against $\mathcal{D}_1$, the two DAGs in Figure 1, where for simplicity we also assume that the set of common parents of u and v consists of a single node z, i.e. $\mathrm{pa} \equiv \{z\}$. Notice that the two DAGs, that differ by the orientation of $u \leftarrow v$ in $\mathcal{D}_0$ which is reversed in $\mathcal{D}_1$, are *observationally* Markov equivalent. We consider a family of intervention targets $\mathcal{I} = \{\emptyset, u\}$ corresponding to a dataset which consists of observational data and interventional data produced from an intervention with target $I_2 = u$. Under the conservative family of targets $\mathcal{I}$, $\mathcal{D}_0$ and $\mathcal{D}_1$ are *not* $\mathcal{I}$-Markov equivalent.

Assuming the true DAG is $\mathcal{D}_0$ we then proceed by randomly generating the parameters $(\boldsymbol{D}, \boldsymbol{L}, \delta_u)$ of the underlying SEM; see also Equations (11) and (12). Specifically, similarly to Hauser & Bühlmann (2012), we draw independently the non-zero elements of $\boldsymbol{L}$ in $[-2,-0.1] \cup [0.1,2]$, while we fix $\boldsymbol{D} = \mathrm{diag}(1,1,1)$ and $\delta_u = 0.1$, the conditional variance of X_u in the post-intervention distribution. Under this setting, a dataset $\boldsymbol{X}$ combines $n^{\emptyset}$ observational data $\boldsymbol{X}^{(1)}$, corresponding to $I_1 = \emptyset$, and n^{int} interventional data $\boldsymbol{X}^{(2)}$ for $I_2 = u$. Letting $n = n^{\emptyset} + n^{int}$, we also have $n^{\mathcal{A}(v)} = n$ and $n^{\mathcal{A}(u)} = n^{\emptyset}$. In the following we build different scenarios with respect to the sample size n which we vary in a grid within $[10, 1000]$ and $\alpha = n^{\mathcal{A}(u)}/n$, the proportion of observational data over the total sample size n. Under each scenario defined by (n, α) we generate independently $N = 100$ datasets. We also fix the hyperparameters of the Wishart prior (14) as $a = q = 4$ and $\boldsymbol{U} = \boldsymbol{I}_q$, the (q,q) identity matrix, a weakly informative prior with weight corresponding to a prior sample of size one. Given each dataset, we then compute the BF of $\mathcal{D}_0$ against $\mathcal{D}_1$ as in Equation (8).

We fist consider *balanced* settings in which the $N = 100$ datasets are generated for values of $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$. The distribution of the logBF across the $N = 100$ simulations, under each scenario defined by α and for increasing sample sizes n, is summarized in the box-plots of Figure 2. The theoretical values, represented as red dots in the figure, are computed in accordance with Proposition 4, and show perfect harmony to the empirically evaluated BFs. Next, we consider an observationally dominant setting, corresponding to the BF limiting value $\alpha = 1$. Accordingly, we take $n^{\mathcal{A}(u)} = n^{\emptyset} = n - n^{\beta}$, for $\beta \in \{0.2, 0.4, 0.6, 0.8\}$ which indeed implies $n^{\mathcal{A}(u)}/n \to \alpha = 1$. Results are reported in Figure 3 where the four panels refer to the four levels of β. As in previous case, there is accordance between the theoretical values of Proposition 2 and the empirical results. The correct graph is always preferred, and we assist to an ameliora-

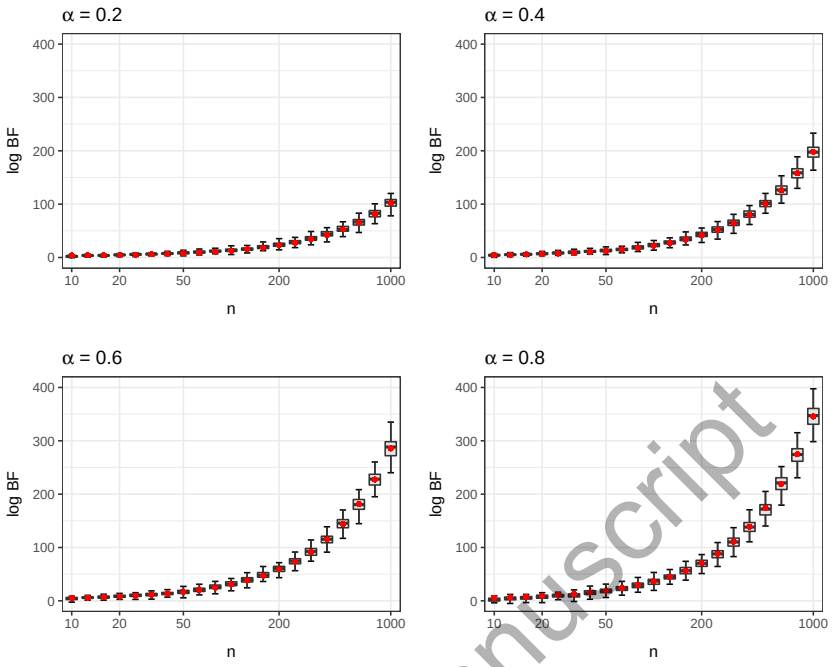

Fig. 2: Simulated data. Balanced setting. Distribution, over $N = 100$ simulated datasets, of the log Bayes Factor (BF) of $\mathcal{D}_0$ (true DAG) against $\mathcal{D}_1$ for increasing sample sizes n and values of $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$, corresponding to increasing balanced proportions of observational measurements.

tion of the BF value, with more and more detachment from score equivalence as we move to higher and higher amounts of interventional data. We finally consider the interventionally dominant case corresponding to $\alpha = 0$. We fix $n^{\mathcal{A}(u)} = n^{\emptyset} = n^{\beta}$, for $\beta \in \{0.2, 0.4, 0.6, 0.8\}$, so that $n^{\mathcal{A}(u)}/n \to \alpha = 0$. Results are reported in Figure 4 where the four panels refer to the four levels of β, again validating the theoretical results of Proposition 3.

5·2. *Posterior sampling scheme*

In this section we implement a Bayesian posterior sampler for DAG structure learning, and we apply it to two public synthetic datasets. A further application to the protein signalling data of Sachs et al. (2005) is provided in the Supplementary material. Our sampler is based on a Metropolis Hastings MCMC scheme which adopts the BF in (8) to compute the acceptance ratio between any two competing DAGs $\mathcal{D}$, $\widetilde{\mathcal{D}}$. More specifically, we consider as a target distribution the marginal posterior $p(\mathcal{D} \mid \boldsymbol{X}) \propto m(\boldsymbol{X} \mid \mathcal{D})\, p(\mathcal{D}), \mathcal{D} \in \mathcal{S}_q$, where $\mathcal{S}_q$ is the set of all DAGs on q vertices and $p(\mathcal{D})$ is a prior on DAG $\mathcal{D}$ that we specify through independent Bernoulli random variables on the collection of $q(q-1)/2$ 0-1 elements indicating the absence/presence of a link between two nodes; see also Castelletti (2020). At each step of the MCMC scheme, corresponding to a current DAG $\mathcal{D}$, a new DAG $\widetilde{\mathcal{D}}$ is proposed from a suitable proposal distribution $q(\widetilde{\mathcal{D}} \mid \mathcal{D})$ based on a local modification of $\mathcal{D}$ through insertion, deletion or reversal of a single edge; see Castelletti (2020, Algorithm 1) for full details. The acceptance probability for $\widetilde{\mathcal{D}}$ under

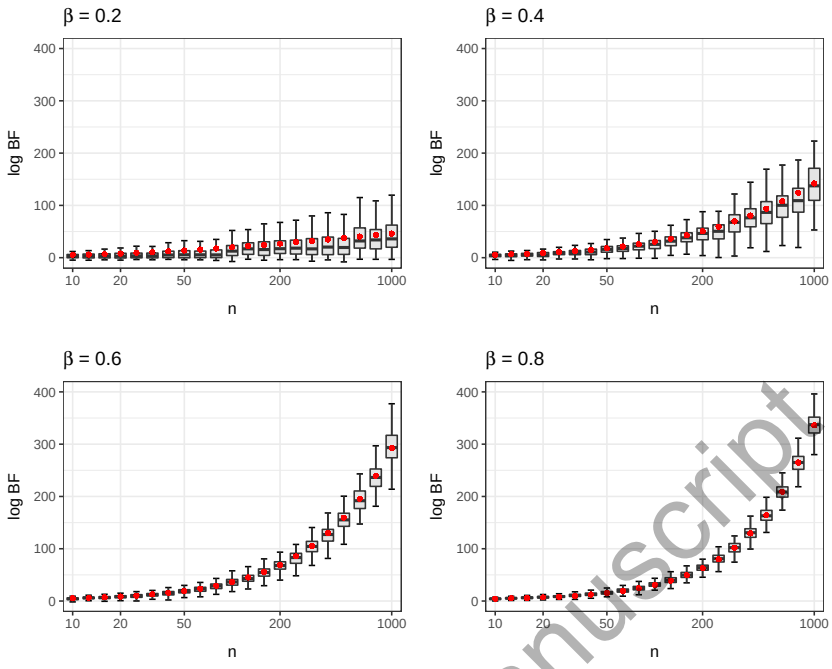

Fig. 3: Simulated data. Observationally dominant setting. Distribution, over $N = 100$ simulated datasets, of the log Bayes Factor (BF) of $\mathcal{D}_0$ (true DAG) against $\mathcal{D}_1$ for increasing sample sizes n and values of $\beta \in \{0.2, 0.4, 0.6, 0.8\}$, corresponding to increasing proportions of interventional measurements.

a Metropolis Hastings algorithm is given by

$$\alpha_{\widetilde{\mathcal{D}}} = \min\left\{1; \mathrm{BF}(\widetilde{\mathcal{D}}, \mathcal{D}) \cdot \frac{p(\widetilde{\mathcal{D}})}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D} \,|\, \widetilde{\mathcal{D}})}{q(\widetilde{\mathcal{D}} \,|\, \mathcal{D})}\right\}, \tag{16}$$

with $\mathrm{BF}(\widetilde{\mathcal{D}}, \mathcal{D})$ as in (7). Output of the MCMC is a collection of DAGs $\{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(S)}\}$ visited by the chain, where S is the number of final iterations. Let now $\boldsymbol{\mathcal{D}}$ be the set of distinct DAGs visited by the MCMC chain. The posterior probability of $\mathcal{D} \in \boldsymbol{\mathcal{D}}$ can be approximated as

$$\widehat{p}(\mathcal{D} \,|\, \boldsymbol{X}) = \frac{m(\boldsymbol{X} \,|\, \mathcal{D}) p(\mathcal{D})}{\sum_{\mathcal{D} \in \boldsymbol{\mathcal{D}}} m(\boldsymbol{X} \,|\, \mathcal{D}) p(\mathcal{D})} = \left\{1 + \sum_{\mathcal{D}^* \neq \mathcal{D}} \frac{p(\mathcal{D}^*)}{p(\mathcal{D})} \mathrm{BF}(\mathcal{D}^*, \mathcal{D})\right\}^{-1} \tag{17}$$

while it is $\widehat{p}(\mathcal{D} \,|\, \boldsymbol{X}) = 0$ if $\mathcal{D} \notin \boldsymbol{\mathcal{D}}$; see also García-Donato & Martínez-Beneito (2013) for a comparison with frequency-based approximations of posterior probabilities in large model spaces. To evaluate the performance of our methodology in recovering the underlying DAG, we then consider the Maximum A Posteriori (MAP) DAG, $\widehat{\mathcal{D}}$, corresponding to the DAG with the highest estimated posterior probability. As a further summary of the MCMC output, we can also compute, for each edge $u \to v$, its marginal posterior probability of inclusion

$$\widehat{p}(u \to v \,|\, \boldsymbol{X}) = \sum_{\mathcal{D} \in \boldsymbol{\mathcal{D}}} \widehat{p}(\mathcal{D} \,|\, \boldsymbol{X}) \mathbb{1}(u \to v \in \mathcal{D}) \tag{18}$$

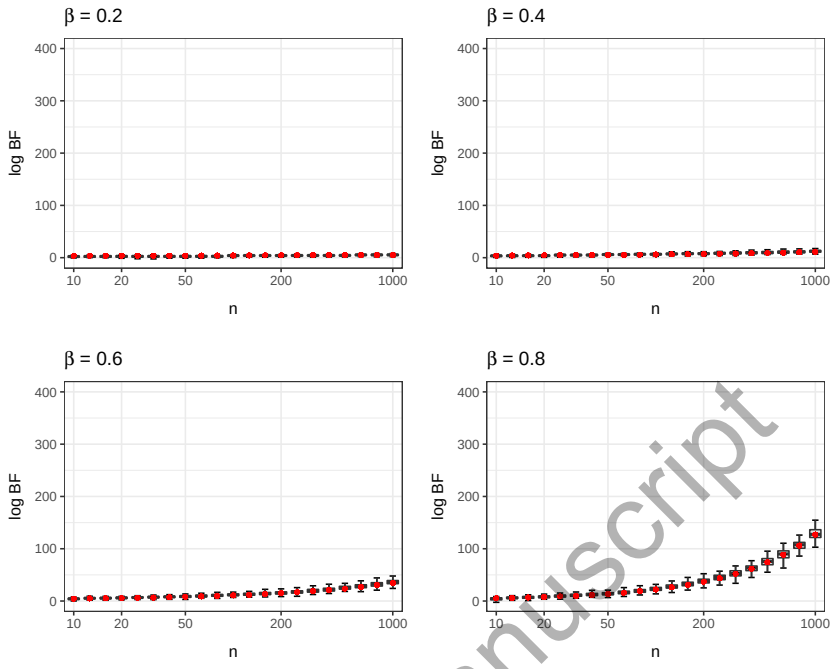

Fig. 4: Simulated data. Interventionally dominant setting. Distribution, over $N = 100$ simulated datasets, of the log Bayes Factor (BF) of $\mathcal{D}_0$ (true DAG) against $\mathcal{D}_1$ for increasing sample sizes n and values of $\beta \in \{0.2, 0.4, 0.6, 0.8\}$, corresponding to increasing proportions of observational measurements.

where $\mathbb{1}(u \to v \in \mathcal{D}) = 1$ if $\mathcal{D}$ contains $u \to v$, 0 otherwise.

Dataset 1: gmInt data. We first apply our MCMC scheme to the `gmInt` data of Kalisch et al. (2012), which consists of an ensemble of observational and interventional measurements simulated from a 8-dimensional Gaussian DAG model. The family of targets is $\mathcal{I} = \{\emptyset, \{3\}, \{5\}\}$ and the corresponding sample sizes are $n^{\emptyset} = 3000, n^{\{3\}} = n^{\{5\}} = 1000$, so that the overall proportions of observational and interventional data are almost balanced. We apply our MCMC scheme for a number of iterations $S = 10000$. Given the output, we first recover the MAP DAG estimate, whose representative $\mathcal{I}$-EG coincides with the true $\mathcal{I}$-EG in Figure 5 (b). The true DAG is reported in Figure 5 (a), together with the representative true $\mathcal{I}$-EG. The latter contains one undirected, i.e. bidirected, edge between nodes Author and Bar; accordingly there are two DAGs in the true $\mathcal{I}$-Markov equivalence, corresponding to the two possible orientations of the undirected edge. In addition, we estimate the posterior probability of inclusion as in Equation (18) for each possible edge (u, v), $u, v \in \{1, \dots, q\}$. Results are summarized in the heat map of Figure 5 (c). It appears that the posterior probability of inclusion is approximately one for all directed edges that are included in the estimated $\mathcal{I}$-EG and zero otherwise, with the exception of the two directed edges Author $\to$ Bar and Author $\leftarrow$ Bar whose posterior probabilities are approximately equal to 0.5. Indeed, the posterior distribution of DAGs computed from Equation (17) is concentrated around the two DAGs belonging to the true $\mathcal{I}$-Markov equivalence class, which divide into two equal parts the overall posterior over the DAG space.

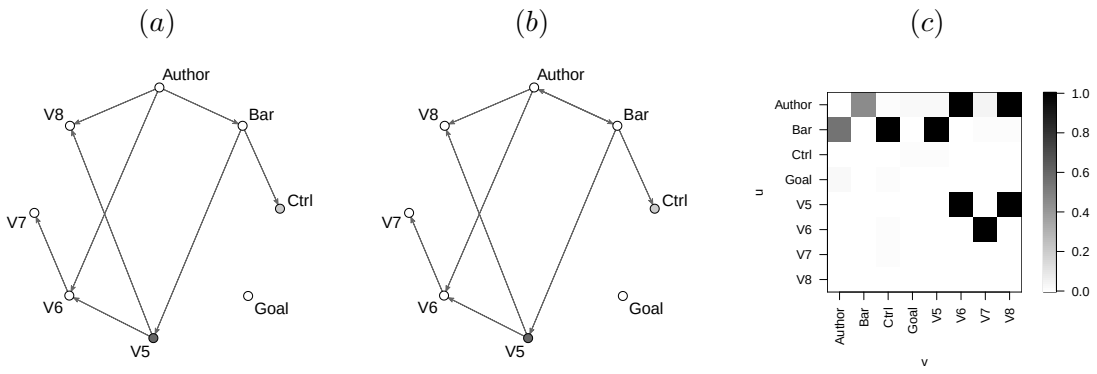

Fig. 5: gmInt data. True DAG (a), true and estimated $\mathcal{I}$-essential graph (b) and heat map collecting the estimated posterior probabilities of edge inclusion $\widehat{p}(u \to v \mid \boldsymbol{X})$ (c).

Dataset 2: DREAM4 data. We now consider synthetic gene expression data from the DREAM4 *in silico* challenge (Marbach et al., 2009, 2010). DREAM4 provides five datasets with an ensemble of interventional and observational data simulated from five biologically plausible, possibly cyclic gene regulatory networks with 10 genes. Each dataset contains both observational measurements, as well as measurements from single-gene knockdowns, single-gene knockouts on each gene and time series data simulated from an unknown change of parameters in the first half and unperturbed data in the second half. We follow the procedure of Hauser & Bühlmann (2012, Section 5.3.1) for the construction of the datasets here analyzed. Most importantly, given the composition of each dataset, a DAG structure is fully identifiable. We then implement the MCMC scheme for a number of iterations $S = 10000$ on each of the five datasets to approximate the posterior (17) from which we recover, as a summary of the entire output, the MAP DAG estimate. The latter is compared with the true regulatory network by means of the structural Hamming distance between the two graphs (Kalisch & Bühlmann, 2007). As a benchmark we also include the Greedy Interventional Equivalence Search (GIES) method of Hauser & Bühlmann (2012), a search-and-score method based on penalized maximum likelihood estimation which jointly model observational and interventional data. The Greedy Interventional Equivalence Search is implemented for three different optimization criteria: the BIC (GIES 0) and the Extended BIC with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5 and GIES 1, respectively); see also Foygel & Drton (2010). The BIC and the Extended BIC correspond to Laplace approximations of the marginal likelihood based on differently regularized likelihood functions. Differences between our closed-form expression for the DAG marginal likelihood and the BIC are therefore expected at small sample sizes, where the Laplace approximation can be less accurate (Konishi & Kitagawa, 2008), and in contexts more sensitive to changes in the likelihood penalty tuning. Results are reported in Table 1, where it is clear that, with few exceptions, we favourably compare with the alternative methodologies, and that overall our performance shows a lower error.

6. DISCUSSION

Our method has been specifically constructed for assumed hard stochastic interventions that destroy the relations between intervened variables and their parents, but it can be useful with no further adaptations to hard deterministic interventions, by fixing the $\tilde{f}$ density to a degenerate delta function $\delta_{\tilde{x}}$ with total mass on the chosen fixed value $\tilde{x}$ of the intervened node. In this spe-

Dataset	Bayes MAP	GIES 0	GIES 0.5	GIES 1
1	10	10	9	10
2	12	12	11	12
3	12	16	15	13
4	6	10	6	7
5	6	8	8	7
Average	9.2	11.2	9.8	9.8

Table 1: DREAM4 data. Structural Hamming Distance (SHD) between true and estimated DAG for each of the five datasets. Methods under comparison are: our Bayesian approach with the Maximum A Posteriori DAG estimate (Bayes MAP); the Greedy Interventional Equivalence Search (GIES) method implementing the Bayesian Information Criterion (GIES 0) and the Extended Bayesian Information Criterion with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5 and GIES 1).

cial case, the whole prior elicitation construction is trivially respected, since for all $k = 1, \ldots, K$, there are no interventional parameters $\boldsymbol{\theta}^{(k)}$. More strongly, any choice of $\tilde{f}$ for the generic node $j \in V$ is compatible with our framework, as long as it implies the same set of post-intervention dependencies towards j for any couple of graphs under comparison: when this happens, the related interventional part of the likelihood ratio will not affect the Bayes factor and the posterior ratio, exactly in finite sample sizes. On the contrary, with soft interventions (Yang et al., 2018) which only weaken the strength of the parent-child relationships, or with general interventions (Correa & Bareinboim, 2020) that allow for local modifications of the DAG structure, the Bayes factor would be explicitly affected by the choice of $\tilde{f}$. We conjecture that our theoretical results on score equivalence and posterior ratio consistency of, respectively, equivalent and nonequivalent graphs, can be extended to soft interventions and to general interventions, by imposing some constraints on the hyperparameters of the interventional prior distributions, along the same line of those constraints suggested by Geiger & Heckerman (2002) and elicited by Peluso & Consonni (2020), in the context of multivariate data with no interventions, on the hyperparameters of the observational prior distributions.

Active learning methods implement experimental design techniques to identify a family of intervention targets that guarantee full DAG identification *via* the smallest number of intervention experiments (Eberhardt, 2008; He & Geng, 2008). To this purpose, He & Geng (2008) propose two kinds of optimal intervention strategies: a *batch* intervention, which identifies upfront, before any intervention, the minimum set of variables to manipulate leading to DAG identification, and a *sequential* approach which iteratively selects an optimal target variable and collects interventional data gathered from a manipulation of the selected node. The procedure is repeated until the estimated interventional Markov equivalence class consists of a single DAG. Based on these premises, Castelletti & Consonni (2022) propose a Bayesian methodology for *sample size determination*, which computes at each intervention the minimal sample size guaranteeing a pre-experimental overall probability of decisive and correct evidence in favour of a correct DAG identification. To discriminate between competing DAG structures, their method adopts a Bayes Factor (BF) computed from observational data, and therefore is limited to targets designed by a batch strategy. On the other hand, our methodology is based on a BF that integrates observational and interventional data, and it can therefore be used to extend the method of Castelletti & Consonni (2022) to sample size determination for sequentially designed interventions.

ACKNOWLEDGEMENT

F.C. was partially supported by UCSC (D1 research grants). S.P. acknowledges partial support of the Swiss National Science Foundation project CRSK-2_190656. We thank the Editor, the Associate Editor and two anonymous Reviewers for their comments and suggestions to improve the paper.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes (i) additional material on the DAG marginal likelihood in the general and Gaussian case, (ii) proofs of propositions with related auxiliary results and discussion, (iii) a first result and discussion in the high-dimensional case with an increasing number of nodes, (iv) an empirical application to protein signalling data. Code implementing our methodology is publicly available at https://github.com/FedeCastelletti/bayes_learning_networks_interventional.

REFERENCES

ANDERSSON, S. A., MADIGAN, D. & PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics* **25**, 505–541.

BEN-DAVID, E., LI, T., MASSAM, H. & RAJARATNAM, B. (2015). High dimensional Bayesian inference for Gaussian directed acyclic graph models. *arXiv pre-print* .

CAO, X., KHARE, K. & GHOSH, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *The Annals of Statistics* **47**, 319–348.

CARVALHO, C. M. & SCOTT, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96**, 497–512.

CASTELLETTI, F. (2020). Bayesian model selection of Gaussian directed acyclic graph structures. *International Statistical Review* **88**, 752–775.

CASTELLETTI, F. & CONSONNI, G. (2022). Bayesian sample size determination for causal discovery. *arXiv pre-print* .

CASTELLETTI, F. & PELUSO, S. (2022). Network structure learning under uncertain interventions. *Journal of the American Statistical Association* **NA**, NA–NA.

CHICKERING, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research* **2**, 445–498.

CONSONNI, G., LA ROCCA, L. & PELUSO, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. *Scandinavian Journal of Statistics* **44**, 741–764.

CORREA, J. & BAREINBOIM, E. (2020). A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, NY: AAAI Press.

DAWID, A. P. & LAURITZEN, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **21**, 1272–1317.

DOREL, M., KLINGER, B., GROSS, T., SIEBER, A., PRAHALLAD, A., BOSDRIESZ, E., WESSELS, L. F. A. & BLÜTHGEN, N. (2018). Modelling signalling networks from perturbation data. *Bioinformatics* **34**, 4079–4086.

EBERHARDT, F. (2008). Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI '08. Arlington, Virginia, USA: AUAI Press.

FOYGEL, R. & DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems 23*.

FRIEDMAN, N. & KOLLER, D. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* **50**, 95–125.

GARCÍA-DONATO, G. & MARTÍNEZ-BENEITO, M. A. (2013). On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association* **108**, 340–352.

GEIGER, D. & HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics* **30**, 1412–1440.

HAUSER, A. & BÜHLMANN, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research* **13**, 2409–2464.

HAUSER, A. & BÜHLMANN, P. (2015). Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**, 291–318.

HE, Y. & GENG, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research* **9**, 2523–2547.

HECKERMAN, D., GEIGER, D. & CHICKERING, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20**, 197–243.

HÜNERMUND, P. & BAREINBOIM, E. (2019). Causal inference and data fusion in econometrics. Tech. rep.

JABER, A., KOCAOGLU, M., SHANMUGAM, K. & BAREINBOIM, E. (2020). Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in Neural Information Processing Systems 33*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin, eds. Vancouver, Canada: Curran Associates, Inc.

KALISCH, M. & BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* **8**, 613–636.

KALISCH, M., MÄCHLER, M., COLOMBO, D., MAATHUIS, M. H. & BÜHLMANN, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* **47**, 1–26.

KASS, R. E. & RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

KONISHI, S. & KITAGAWA, G. (2008). *Information Criteria and Statistical Modeling*. Springer New York.

KORB, K. B., HOPE, L. R., NICHOLSON, A. E. & AXNICK, K. (2004). Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*. Springer.

LAURITZEN, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.

MARBACH, D., PRILL, R. J., SCHAFFTER, T., MATTIUSSI, C., FLOREANO, D. & STOLOVITZKY, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences* **107**, 6286–6291.

MARBACH, D., SCHAFFTER, T., MATTIUSSI, C. & FLOREANO, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology* **16**, 229–239.

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

PELUSO, S. & CONSONNI, G. (2020). Compatible priors for model selection of high-dimensional Gaussian DAGs. *Electronic Journal of Statistics* **14**, 4110–4132.

PINGAULT, J.-B., O'REILLY, P. F., SCHOELER, T., PLOUBIDIS, G. B., RIJSDIJK, F. & DUDBRIDGE, F. (2018). Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics* **19**, 566–580.

PINNA, A., HEISE, S., FLASSIG, R. J., , SORANZO, N., DE LA FUENTE, A. & KLAMT, S. (2013). Reconstruction of large-scale regulatory networks based on perturbation graphs and transitive reduction: improved methods and their evaluation. *BMC Systems Biology* **7**, 1–19.

RAU, A., JAFFRÉZIC, F. & NUEL, G. (2013). Joint estimation of causal effects from observational and intervention gene expression data. *BMC Systems Biology* **7**, 1–12.

ROVERATO, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* **29**, 391–411.

SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. & NOLAN, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529.

VERMA, T. & PEARL, J. (1990). Equivalence and Synthesis of Causal Models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI 90. New York, NY, USA: Elsevier Science Inc.

YANG, K., KATCOFF, A. & UHLER, C. (2018). Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proceedings of the 35th International Conference on Machine Learning*, J. Dy & A. Krause, eds., vol. 80 of *Proceedings of Machine Learning Research*. PMLR.

[*Received* to *be* defined. *Editorial decision on* to *be* defined]