Biometrika (2024), **111**, 1, *pp.* 195–214 https://doi.org/10.1093/biomet/asad032
Advance Access publication 11 May 2023

Bayesian learning of network structures from interventional experimental data

By F. CASTELLETTI

Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo A. Gemelli 1, 20123 Milan, Italy
federico.castelletti@unicatt.it

and S. PELUSO

Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milan, Italy
stefano.peluso@unimib.it

Summary

Directed acyclic graphs provide an effective framework for learning causal relationships among variables given multivariate observations. Under pure observational data, directed acyclic graphs 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 directed acyclic graph 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 directed acyclic graph marginal likelihood and guaranteeing score equivalence among directed acyclic graphs 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 via simulation and we implement a Markov chain Monte Carlo sampler for posterior inference on the space of directed acyclic graphs on both synthetic and biological protein expression data.

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 2023). The objective is to infer these relationships from the data, a task that 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 for 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 (Pinna et al., 2013; Rau 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 (Friedman & Koller, 2003; Kalisch & Bühlmann, 2007). 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 few years, both in the statistical and machine learning communities. 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 interventions: the former sets each manipulated variable to a given value $\tilde{x}$, so that the local distribution of each intervened node reduces to a point mass at $\tilde{x}$; stochastic intervention instead replaces the local density with that of a new random variable, say $\tilde{f}$. In addition, hard, also called perfect, interventions destroy the dependence of each intervened node from its parents in the DAG; conversely, soft interventions preserve the original parent-child relations, but allow for a modification of their 'strength' (Yang et al., 2018). Hauser & Bühlmann (2012) provided a characterization of Markov equivalence under stochastic hard interventions and introduced the greedy interventional equivalence search method as a score-based algorithm for structure learning of interventional equivalence classes. Hauser & Bühlmann (2015) presented several statistical properties connected to the joint modelling of observational and interventional data and proved consistency of the adopted Bayesian information criterion.

In the Bayesian framework, structure learning is set up as a model selection problem that 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 strategies; see, for instance, Chickering (2002), Consonni et al. (2017) and Castelletti (2020). To analytically compute the DAG marginal likelihood, one needs to assign a suitable parameter prior distribution that 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 nondecomposable, undirected

Gaussian graphical models. For DAG models with independent and identically distributed 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 compatibly derived 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 were 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 predetermined 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 DAG, where $V = \{1, \ldots, q\}$ 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 bidirected 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 that can be read-off from the DAG using graphical criteria, such as d-separation (Pearl, 2000).

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

$$f(x_1, \ldots, x_q \mid \mathcal{D}) = \prod_{j=1}^{q} f(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)}), \tag{1}$$

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, which 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 the essential graph. 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 showed 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 $\tilde{f}(u_j)$ 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

$$f(x_1, \ldots, x_q \mid \mathrm{do}\{X_j = U_j\}_{j\in I}, \mathcal{D}) = \prod_{j\notin I} f(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)}) \prod_{j\in I} \tilde{f}(x_j), \tag{2}$$

where it appears that, for each $j \in I$, the original dependence of node j from its parents $\mathrm{pa}_{\mathcal{D}}(j)$ is dropped and $f(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)})$ is replaced by $\tilde{f}(x_j)$. Under the specific case $I = \emptyset$, (2) reduces to (1), which is also named the observational or pre-intervention 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 $I_k \subset \{1, \ldots, q\}$ with 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 that does not involve j, a requirement that 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 $\mathcal{D}^I = (V, E^I)$, $E^I = \{(u, v) \in E \mid v \notin I\}$, is the so-called *intervention DAG* of $\mathcal{D} = (V, E)$ given 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 again be represented by a chain graph called the $\mathcal{I}$-essential graph; see also Definition 11 of Hauser & Bühlmann (2012), to which we also refer the reader 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 $k = 1, \ldots, K$; note that if $I_k = \emptyset$, the kth dataset is purely observational. We assume that $f(\cdot)$ in (2) belongs to some parametric family indexed by a global parameter (θ, θ^I) and write the post-intervention distribution as

$$f(x_1, \ldots, x_q \mid \mathrm{do}\{X_j = U_j\}_{j\in I}, \theta, \theta^I, \mathcal{D}) = \prod_{j\notin I} f(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)}, \theta_j) \cdot \prod_{j\in I} \tilde{f}(x_j \mid \theta_j^I), \tag{3}$$

where, in particular, $\theta = \{\theta_j\}_{j\notin I}$ is a collection of node parameters relative to node-conditional observational distributions, while $\theta^I = \{\theta_j^I\}_{j\in I}$ are interventional node parameters. We further assume that, for each intervention experiment $k = 1, \ldots, K$, $n^{(k)}$, with target I_k, independent and identically distributed observations from (3) $\{x_i^{(k)}, i = 1, \ldots, n^{(k)}\}$ are available, and let $X^{(k)}$ 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

$$\begin{aligned} &f(X \mid \theta, \theta^{(1)}, \ldots, \theta^{(K)}, \mathcal{D}) \\ &\quad= \prod_{k=1}^{K} \left\{ \prod_{i=1}^{n^{(k)}} \left(\prod_{j\notin I_k} f(x_{i,j}^{(k)} \mid x_{i,\mathrm{pa}_{\mathcal{D}}(j)}^{(k)}, \theta_j) \cdot \prod_{j\in I_k} \tilde{f}_k(x_{i,j}^{(k)} \mid \theta_j^{(k)}) \right) \right\} \\ &\quad= \prod_{k=1}^{K} \left\{ \prod_{j\notin I_k} f(X_j^{(k)} \mid X_{\mathrm{pa}_{\mathcal{D}}(j)}^{(k)}, \theta_j) \prod_{j\in I_k} \tilde{f}_k(X_j^{(k)} \mid \theta_j^{(k)}) \right\}, \end{aligned} \tag{4}$$

where $X_S^{(k)}$ denotes the $n^{(k)} \times |S|$ submatrix of $X^{(k)}$ with columns belonging to the set $S \subseteq \{1, \ldots, q\}$ and $X = (X^{(1)}, \ldots, X^{(K)})^{\mathrm{T}}$ the $n \times q$ data matrix, $n = \sum_{k=1}^{K} n^{(k)}$. 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, θ is common to all the K terms, while $\theta^{(k)}$ is specific to intervention k. If we now let $\mathcal{A}(j) = \{i \in \{1, \ldots, n\} : j \notin T^{(i)}\}$ then (4) can be written as

$$f(X \mid \theta, \theta^{(1)}, \ldots, \theta^{(K)}, \mathcal{D}) = \prod_{j=1}^{q} f(X_j^{\mathcal{A}(j)} \mid X_{\mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)}, \theta_j) \cdot \prod_{k=1}^{K} \left\{ \prod_{j\in I_k} \tilde{f}_k(X_j^{(k)} \mid \theta_j^{(k)}) \right\}, \tag{5}$$

where now $X^{\mathcal{A}(j)}$ denotes the submatrix of 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) that is integrated with respect to the prior on model parameters $(\theta, \theta^{(1)}, \ldots, \theta^{(K)})$,

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

Relevant to our method, Heckerman et al. (1995) and Geiger & Heckerman (2002) introduced a set of assumptions that guarantee score equivalence in the case of observational independent and identically distributed samples. They started 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, now consider the likelihood function in (5), which consists of two terms. The first one reflects the DAG factorization that is imposed on 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 as in 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) now needs to be extended to the interventional parameters, and to the relationship between observational and interventional parameters. Because of prior independence across parameters θ_j, we have $p(\theta \mid \mathcal{D}) = \prod_{j=1}^{q} p(\theta_j \mid \mathcal{D})$. The second term in (5) corresponds to an unconditional likelihood that no longer depends on the DAG and with 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 $\tilde{f}_k$ terms 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, $p(\theta^{(k)}) = \prod_{j \in I_k} p(\theta_j^{(k)})$, and also global independence among all parameters $\theta^{(1)}, \ldots, \theta^{(K)}$. We finally assume that the joint prior on θ and $(\theta^{(1)}, \ldots, \theta^{(K)})$ factorizes as $p(\theta, \theta^{(1)}, \ldots, \theta^{(K)}) = p(\theta \mid \mathcal{D}) p(\theta^{(1)}, \ldots, \theta^{(K)})$, 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 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

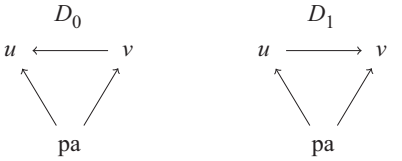

Fig. 1. Two DAGs, $\mathcal{D}_0, \mathcal{D}_1$, differing by an edge reversal between u and v, where pa is the set of common parents of nodes u and v.

$$p(\theta, \theta^{(1)}, \ldots, \theta^{(K)} \mid \mathcal{D}) = p(\theta \mid \mathcal{D}) \prod_{k=1}^{K} \prod_{j \in I_k} p(\theta_j^{(k)}).$$

3.3. *Marginal likelihood and Bayes factor computation*

We now focus on the computation of $m(X \mid \mathcal{D})$, the marginal likelihood of DAG $\mathcal{D}$. Because of the assumptions in § 3.2, we obtain

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

where $m(\cdot \mid \mathcal{C})$ denotes the marginal likelihood computed under any complete DAG model $\mathcal{C}$; see also 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 $\mathcal{D}_0$; see Fig. 1. Also, let pa be the set of common parents of nodes u and v, possibly an empty set. The Bayes factor of $\mathcal{D}_0$ against $\mathcal{D}_1$,

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

then simplifies to

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

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

3.4. *Gaussian DAG models*

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

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

where $\Omega \in \mathcal{P}_{\mathcal{D}}$ corresponds to the precision matrix, inverse of the covariance matrix Σ, and $\mathcal{P}_{\mathcal{D}}$ is the space of all symmetric positive definite matrices Markov with respect to DAG $\mathcal{D}$. An equivalent formulation is in terms of the corresponding structural equation model,

$$L^{\mathrm{T}} x = \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_q(0, D), \tag{10}$$

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

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

where $\prec j] = \mathrm{pa}_{\mathcal{D}}(j) \times j$, $L_{A \times B}$ is the submatrix of L with columns and rows indexed by A and B, respectively, and $\phi(\cdot \mid \mu, \sigma^2)$ denotes the probability density function of an $\mathcal{N}(\mu, \sigma^2)$ distribution. Equation (11) resembles the DAG factorization in (1) and adopts the reparameterization $\Sigma \mapsto (L, D)$ such that $\Sigma = L^{-\mathrm{T}} D L^{-1}$, where $L^{-\mathrm{T}} := (L^{\mathrm{T}})^{-1}$. 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,

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

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, $\tilde{\Sigma}$, is therefore $\tilde{\Sigma} = \tilde{L}^{-\mathrm{T}} \tilde{D} \tilde{L}^{-1}$, where $\tilde{L}_{u,v} = 0$ if $v \in I$ and $u \neq v$, while $\tilde{L}_{u,v} = L_{u,v}$ otherwise, and $\tilde{D}$ is obtained from D by replacing elements $D_{u,u}$, for each $u \in I$, with δ_u. As in § 3.1, by assuming K independent interventions, each with target I_k, the likelihood function in (5) can be written as

$$\begin{aligned} f(X \mid L, D, \delta^{(1)}, \ldots, \delta^{(K)}, \mathcal{D}) = \prod_{j=1}^{q} \phi_{|\mathcal{A}(j)|}(X_j^{\mathcal{A}(j)} \mid -X_{\mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} L_{\prec j]}, D_{jj} I_{|\mathcal{A}(j)|}) \\ \cdot \prod_{k=1}^{K} \left\{ \prod_{j \in I_k} \phi_{n^{(k)}}(X_j^{(k)} \mid 0, \delta_j^{(k)} I_{n^{(k)}}) \right\}, \end{aligned} \tag{13}$$

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

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

a Wishart distribution with parameters $a > q - 1$ and U, a (q, q) symmetric positive definite matrix, having expectation $\mathbb{E}(\Omega) = aU^{-1}$. We then write

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

under which we obtain the marginal likelihood restricted to X_B, a generic submatrix of the $n \times q$ matrix X, with columns indexed by $B \subseteq \{1, \ldots, q\}$,

$$m(X_B) = \pi^{-n|B|/2} \frac{|U_{BB}|^{(a-|\bar{B}|)/2}}{|U_{BB} + S_{BB}|^{(a-|\bar{B}|+n)/2}} \frac{\Gamma_{|B|}\{(a - |\bar{B}| + n)/2\}}{\Gamma_{|B|}\{(a - |\bar{B}|)/2\}}.$$

This general formula applied to each term $m(\cdot)$ in (8) specializes the Bayes factor 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 $n^{\mathcal{A}(u)}/n \to 1$ or $n^{\mathcal{A}(u)}/n \to 0$, respectively.

We first show that the prior elicitation procedure introduced in § 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 of Chickering (1995), which shows that two DAGs $\mathcal{D}_0$ and $\mathcal{D}_1$ 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, and let $\mathcal{D}_0$ and $\mathcal{D}_1$ be two $\mathcal{I}$-Markov equivalent DAGs. Then, $\mathcal{D}_0$ and $\mathcal{D}_1$ have the same marginal likelihood, namely, $m(X \mid \mathcal{D}_0) = m(X \mid \mathcal{D}_1)$, with $m(X \mid \mathcal{D})$ as in (6).*

The proofs of Proposition 1 and all subsequent propositions can be found in 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 that 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 two such 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 strongly protected in $\mathcal{D}_0$, i.e., their reversal does not break observational Markov equivalence; see also Definition 3.3 of 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 $|\text{pa}| = p \geqslant 0$, in both graphs.

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

$$A(u \mid \mathcal{D}) = (\Sigma_{0,\,\text{fa}_{\mathcal{D}}(u)})^{-1} \tilde{\Sigma}_{0,\,\text{fa}_{\mathcal{D}}(u)}$$

for the intervened node $u \in V$, where Σ_0 and $\tilde{\Sigma}_0$ respectively denote pre- and post-intervention true covariance matrices; see also § 3.4. For brevity, we may omit the dependence on the graphs when it is clear from the context. If in $\mathcal{D}_0$ there are some edges involving the intervened node u that are not strongly protected in $\mathcal{D}_0$, there 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 a posterior ratio. If, on the other hand, all edges of u are strongly protected, v has to be intended as an empty set, and $A(u \mid \mathcal{D}_j) = A(u \mid \mathcal{D}_{j+1})$. We define the ordered, from smallest to largest, eigenvalues of $A(u \mid \mathcal{D})$ as $\lambda_j(u \mid \mathcal{D})$ for $j = 1, \ldots, |\text{fa}_{\mathcal{D}}(u)|$. In particular, the minimum and maximum eigenvalues are respectively denoted as $\underline{\lambda}(u \mid \mathcal{D})$ and $\overline{\lambda}(u \mid \mathcal{D})$.

We are now ready in the following three 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 given in 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 observational measurements asymptotically dominate the interventions, as long as the number of interventions $n - n^{A(u)}$ increases in n.

PROPOSITION 2 (OBSERVATIONALLY DOMINANT SETTING). *Let $\mathcal{D}_0$ be the true DAG. Under the assumptions of § 3.2, consider a prior $\Omega \sim \mathcal{W}_q(a, U)$ and the likelihood function in (13). For conservative interventions on u, and with $n^{A(u)}/n \to 1$ as $n \to \infty$, we have, for all $\mathcal{D} \notin [\mathcal{D}_0]_{\mathcal{I}}$,*

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

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

The above proposition tells us that the posterior ratio consistency rate to the true $[\mathcal{D}_0]_{\mathcal{I}}$ depends on $n - n^{A(u)}$. In particular, $\log \text{BF}(\mathcal{D}, \mathcal{D}_0) \leqslant -C(n - n^{A(u)})/2$, $\bar{P}$-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 rates 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, $\log \mathrm{BF}(\mathcal{D}, \mathcal{D}_0) \leqslant -Ck/2$ with $\bar{P}$ probability 1 for sufficiently large n, 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 Hauser & Bühlmann (2012, Definition 14), it is more beneficial to intervene on nodes with $\Sigma_{u\,|\,\mathrm{pa},v}|A(u\,|\,\mathcal{D}_j)| \gg \Sigma_{u\,|\,\mathrm{pa}}|A(u\,|\,\mathcal{D}_{j+1})|$; 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 $|A(u\,|\,\mathcal{D}_j)| = |A(u\,|\,\mathcal{D}_{j+1})|$ and then targets with $\Sigma_{u\,|\,\mathrm{pa},v} \ll \Sigma_{u\,|\,\mathrm{pa}}$ should be privileged, looking only at the pre-intervention covariance matrix.

In the next result, we demonstrate the asymptotically 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 $\mathcal{D} \notin [\mathcal{D}_0]_{\mathcal{I}}$,*

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

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

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

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

$\bar{P}$-almost surely for a sufficiently large n and a constant C_j, with the sum 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 a 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 rates 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 a rate at least $\sqrt{n}$. Therefore, Proposition 3 suggests choosing as targets in an interventionally dominant setting those u with $u \to v$ in E_0 and, among these, those nodes showing $\Sigma_{0,u\,|\,\mathrm{pa},v} << \Sigma_{0,u\,|\,\mathrm{pa}}$.

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 observationally 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* $\mathcal{D} \notin [\mathcal{D}_0]_{\mathcal{I}}$,

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

where $\{\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_J = \mathcal{D}\}$ *is a sequence of observationally equivalent adjacent graphs,* C_α *is 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* $p(\mathcal{D}\,|\,X,\mathcal{I})/p(\mathcal{D}_0\,|\,X,\mathcal{I}) = p(\mathcal{D})/p(\mathcal{D}_0)$, $\bar{P}$*-almost surely.*

In the balanced setting, we have

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

$\bar{P}$-almost surely for a sufficiently high n and a constant C_j, with 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 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, our method improves the 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 $\Sigma_{u\,|\,\mathrm{pa},v}|A(u\,|\,\mathcal{D}_j)| \gg \Sigma_{u\,|\,\mathrm{pa}}|A(u\,|\,\mathcal{D}_{j+1})|$, whilst, if $u \to v \in E_0$ and $\mathcal{I}$-strong protected, targets with $\Sigma_{u\,|\,\mathrm{pa},v} \ll \Sigma_{u\,|\,\mathrm{pa}}$ are preferred.

In 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 through simulation experiments. Specifically, we consider the Bayes factor of $\mathcal{D}_0$ against $\mathcal{D}_1$, the two DAGs in Fig. 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., $\text{pa} \equiv \{z\}$. 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 that 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 that the true DAG is $\mathcal{D}_0$, we then proceed by randomly generating the parameters (D, L, δ_u) of the underlying structural equation model; see also (11) and (12). Specifically, similarly to Hauser & Bühlmann (2012), we independently draw the nonzero elements of L in $[-2, -0.1] \cup [0.1, 2]$, while we fix $D = \text{diag}(1, 1, 1)$ and $\delta_u = 0.1$, the conditional variance of X_u in the post-intervention distribution. Under this setting, a dataset X combines $n^{\emptyset}$ observational data $X^{(1)}$, corresponding to $I_1 = \emptyset$, and n^{int} interventional data $X^{(2)}$ for $I_2 = u$. Letting $n = n^{\emptyset} + n^{\text{int}}$, 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 that 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 independently generate $N = 100$ datasets. We also fix the hyperparameters of the Wishart prior (14) as $a = q = 4$ and $U = I_q$, 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 Bayes factor of $\mathcal{D}_0$ against $\mathcal{D}_1$ as in (8).

We first 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 log BF across the $N = 100$ simulations, under each scenario defined by α and for increasing sample sizes n, is summarized in the box plots of Fig. 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 Bayes factors. Next, we consider an observationally dominant setting, corresponding to the Bayes factor 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 that $n^{\mathcal{A}(u)}/n \to \alpha = 1$. The results are reported in Fig. 3 where the four panels refer to the four levels of β. As in the previous case, there is accordance between the theoretical values of Proposition 2 and the empirical results. The correct graph is always preferred, and we see an amelioration of the Bayes factor 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 $n^{\mathcal{A}(u)}/n \to \alpha = 0$. The results are reported in Fig. 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 Markov chain Monte Carlo, MCMC, scheme that

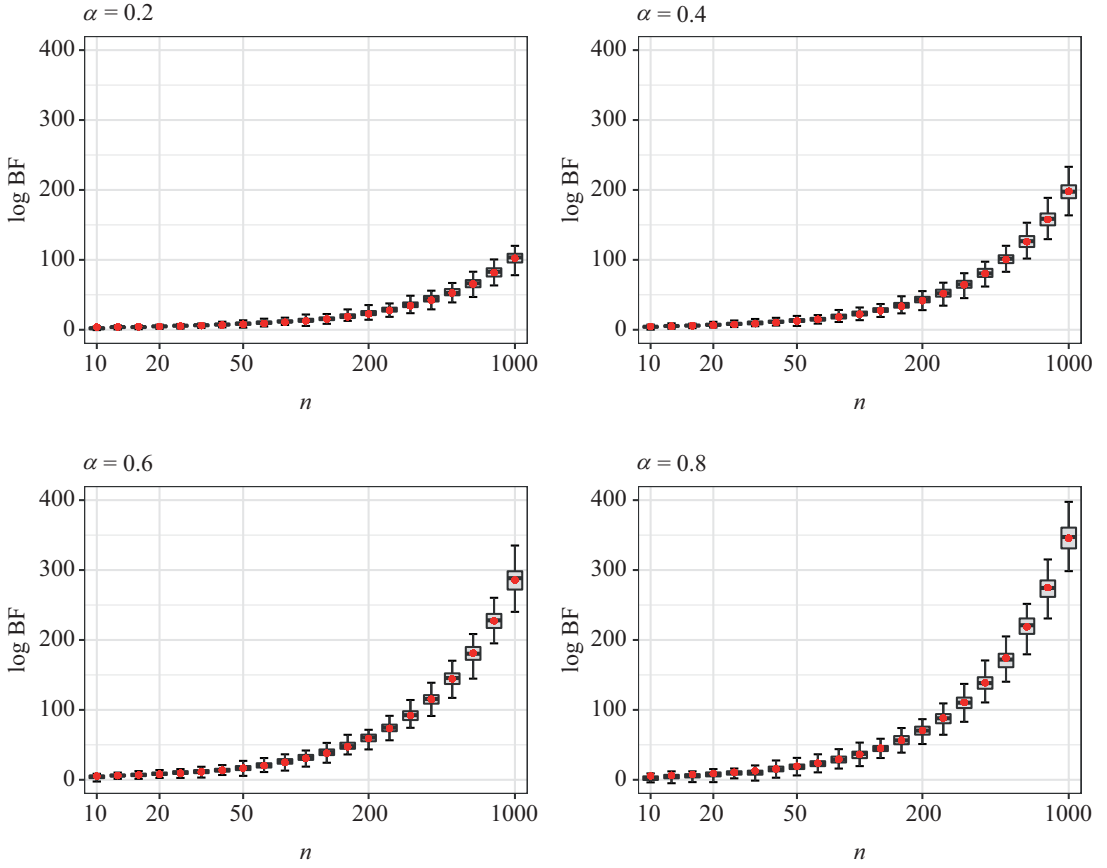

Fig. 2. Simulated data in the balanced setting. Distribution, over $N = 100$ simulated datasets, of the log Bayes factor of $\mathcal{D}_0$ against $\mathcal{D}_1$ for increasing sample sizes n and values of $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$, corresponding to increasing balanced proportions of observational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

adopts the Bayes factor in (8) to compute the acceptance ratio between any two competing DAGs $\mathcal{D}$, $\tilde{\mathcal{D}}$. More specifically, we consider as a target distribution the marginal posterior $p(\mathcal{D} \mid X) \propto m(X \mid \mathcal{D})\, p(\mathcal{D}), \mathcal{D} \in \mathcal{S}_q$, 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 $\tilde{\mathcal{D}}$ is proposed from a suitable proposal distribution $q(\tilde{\mathcal{D}} \mid \mathcal{D})$ 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 $\tilde{\mathcal{D}}$ under a Metropolis–Hastings algorithm is given by

$$\alpha_{\tilde{\mathcal{D}}} = \min\left\{1; \mathrm{BF}(\tilde{\mathcal{D}}, \mathcal{D}) \cdot \frac{p(\tilde{\mathcal{D}})}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D} \mid \tilde{\mathcal{D}})}{q(\tilde{\mathcal{D}} \mid \mathcal{D})}\right\}$$

with $\mathrm{BF}(\tilde{\mathcal{D}}, \mathcal{D})$ as in (7). The 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. Now let $\boldsymbol{\mathcal{D}}$ be the set of distinct DAGs visited by the MCMC chain. The posterior probability of $\mathcal{D} \in \boldsymbol{\mathcal{D}}$ can be

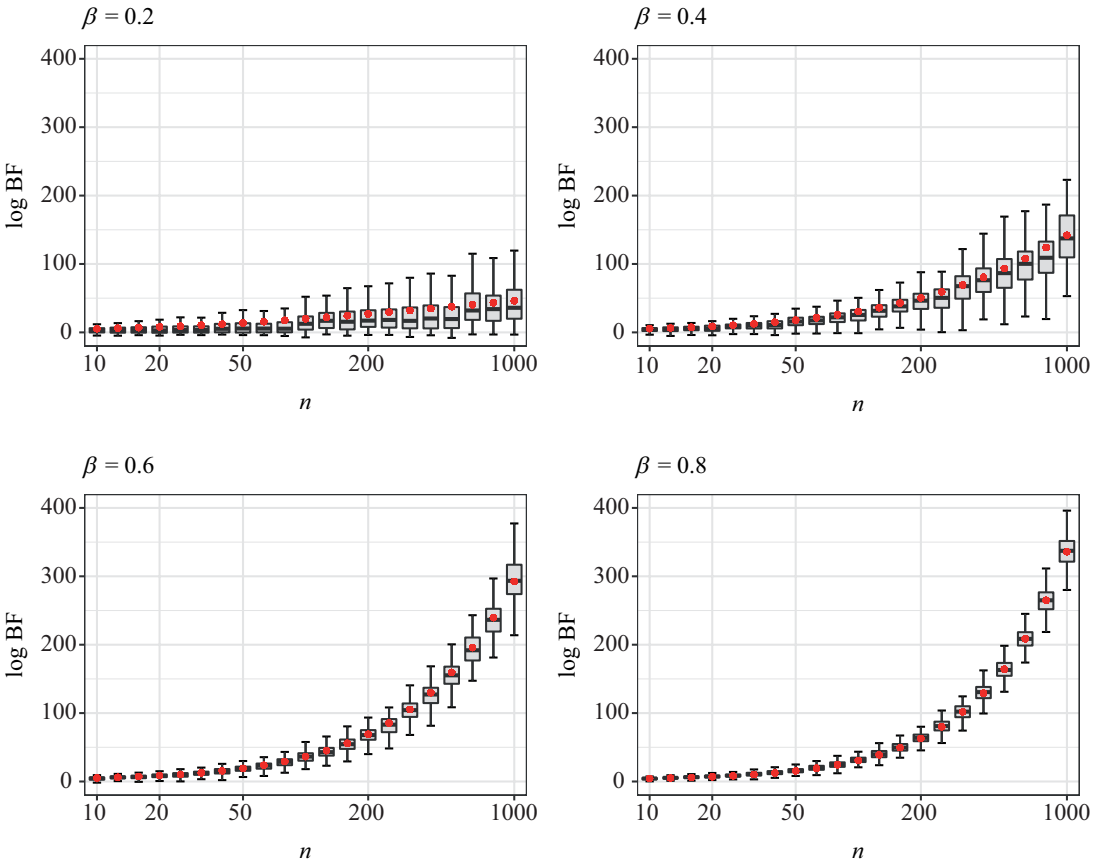

Fig. 3. Simulated data in the observationally dominant setting. Distribution, over $N = 100$ simulated datasets, of the log Bayes factor of $\mathcal{D}_0$ against $\mathcal{D}_1$ for increasing sample sizes n and values of $\beta \in \{0.2, 0.4, 0.6, 0.8\}$, corresponding to increasing proportions of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

approximated as

$$\hat{p}(\mathcal{D} \mid X) = \frac{m(X \mid \mathcal{D})p(\mathcal{D})}{\sum_{\mathcal{D} \in \boldsymbol{\mathcal{D}}} m(X \mid \mathcal{D})p(\mathcal{D})} = \left\{1 + \sum_{\mathcal{D}^* \neq \mathcal{D}} \frac{p(\mathcal{D}^*)}{p(\mathcal{D})} \mathrm{BF}(\mathcal{D}^*, \mathcal{D})\right\}^{-1}, \tag{15}$$

while it is $\hat{p}(\mathcal{D} \mid X) = 0$ 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 DAG, $\hat{\mathcal{D}}$, 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

$$\hat{p}(u \to v \mid X) = \sum_{\mathcal{D} \in \boldsymbol{\mathcal{D}}} \hat{p}(\mathcal{D} \mid X) \mathbb{1}(u \to v \in \mathcal{D}), \tag{16}$$

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

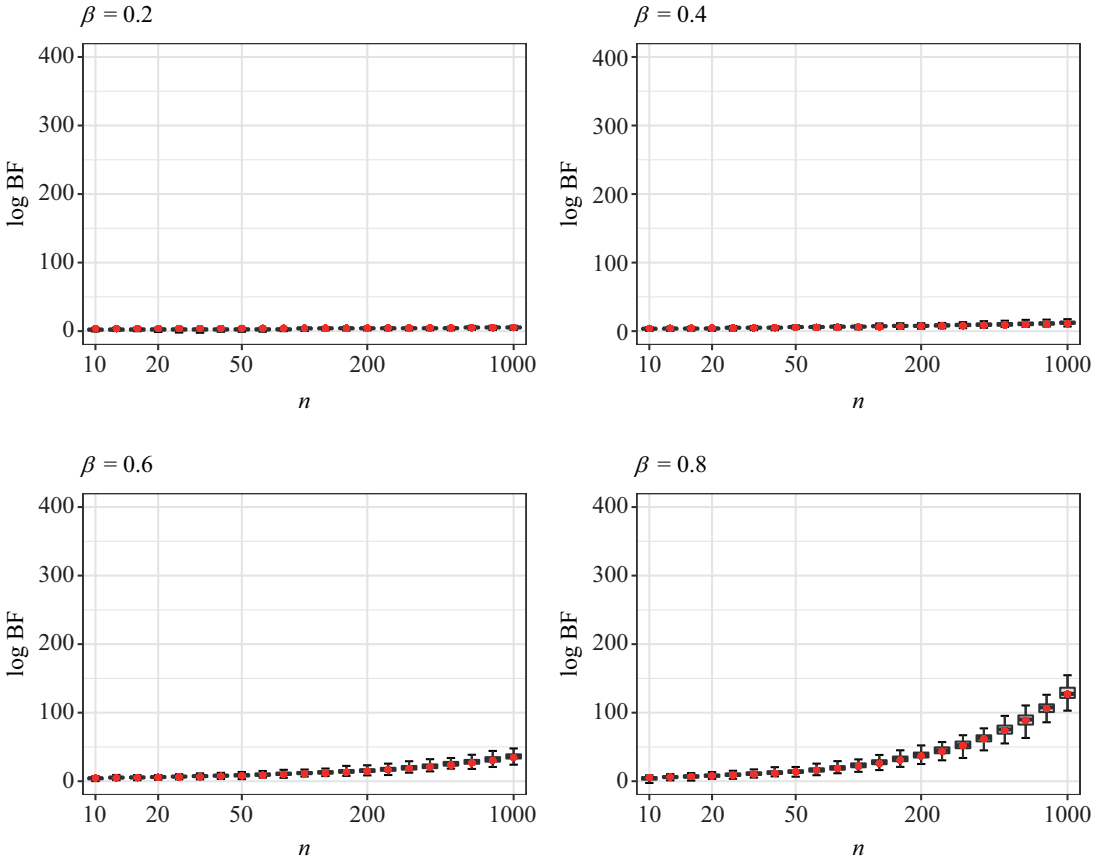

Fig. 4. Simulated data in the interventionally dominant setting. Distribution, over $N = 100$ simulated datasets, of the log Bayes factor of $\mathcal{D}_0$ against $\mathcal{D}_1$ for increasing sample sizes n and values of $\beta \in \{0.2, 0.4, 0.6, 0.8\}$, corresponding to increasing proportions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

5.3. *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 eight-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 $S = 10\,000$. Given the output, we first recover the maximum a posteriori DAG estimate, whose representative $\mathcal{I}$-EG coincides with the true $\mathcal{I}$-EG in Fig. 5(b). The true DAG is reported in Fig. 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 (16) for each possible edge (u, v), $u, v \in \{1, \ldots, q\}$. The results are summarized in the heat map of Fig. 5(c). It appears that the posterior probability of inclusion is approximately one for all directed edges that are included in the estimated $\mathcal{I}$-essential graph and zero otherwise, with the exception of the two directed edges Author $\rightarrow$ Bar and Author $\leftarrow$ Bar whose posterior probabilities are approximately equal to 0.5. Indeed, the posterior distribution of DAGs computed from (15) is concentrated around

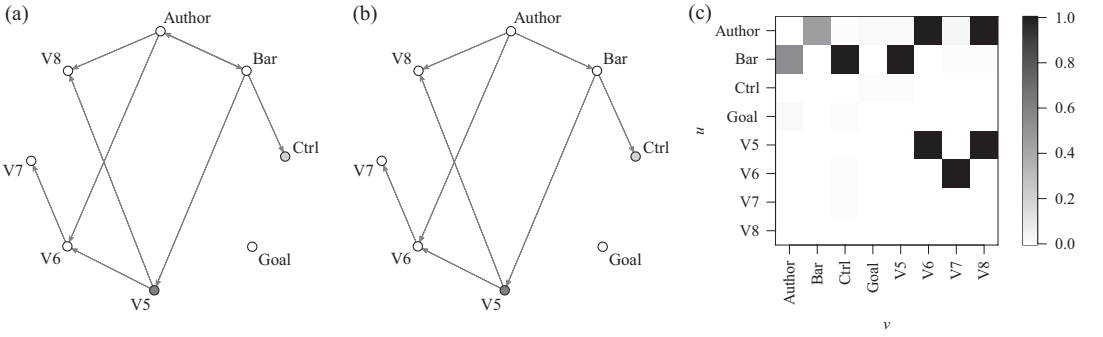

Fig. 5. gmInt data. (a) True DAG, (b) true and estimated $\mathcal{I}$-essential graph and (c) heat map collecting the estimated posterior probabilities of edge inclusion $\hat{p}(u \to v \mid X)$.

the two DAGs belonging to the true $\mathcal{I}$-Markov equivalence class, which divides the overall posterior over the DAG space into two equal parts.

5.4. *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, § 5.3.1) for the construction of the datasets analysed here. 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 $S = 10\,000$ on each of the five datasets to approximate the posterior (15) from which we recover, as a summary of the entire output, the maximum a posteriori 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 method of Hauser & Bühlmann (2012), a search-and-score method based on penalized maximum likelihood estimation that jointly models 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 expressions 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. The 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,

Table 1. *DREAM4 data. Structural Hamming distance between true and estimated DAGs for each of the five datasets*

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

Bayes MAP, our Bayesian approach with the maximum a posteriori DAG estimate; GIES 0, the greedy interventional equivalence search method implementing the Bayesian information criterion; GIES 0.5, the extended Bayesian information criterion with tuning coefficient $\gamma \in \{0.5\}$; GIES 1, the extended Bayesian information criterion with tuning coefficient $\gamma \in \{1\}$.

with no further adaptations, for 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 special case, the whole prior elicitation construction is trivially respected, since, for all $k = 1, \ldots, K$, there are no interventional parameters $\theta^{(k)}$. 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) that 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 equivalent and nonequivalent graphs, respectively, 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 lines as 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 end, He & Geng (2008) proposed 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 that 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) proposed 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 adopted a Bayes factor computed

from observational data, and is therefore limited to targets designed by a batch strategy. On the other hand, our methodology is based on a Bayes factor 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

Castelletti was partially supported by the Università Cattolica del Sacro Cuore. Peluso acknowledges partial support from the Swiss National Science Foundation. We thank the editor, the associate editor and two reviewers for their comments and suggestions that improved the paper.

Supplementary material

The Supplementary Material includes (i) additional material on the DAG marginal likelihood in the general and Gaussian cases, (ii) proofs of the 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. The 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. *Ann. Statist.* **25**, 505–41.

Ben-David, E., Li, T., Massam, H. & Rajaratnam, B. (2015). High dimensional Bayesian inference for Gaussian directed acyclic graph models. *arXiv:* 1109.4371v5.

Cao, X., Khare, K. & Ghosh, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Ann. Statist.* **47**, 319–48.

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. *Int. Statist. Rev.* **88**, 752–75.

Castelletti, F. & Consonni, G. (2022). Bayesian sample size determination for causal discovery. *arXiv:* 2206.00755v1.

Castelletti, F. & Peluso, S. (2022). Network structure learning under uncertain interventions. *J. Am. Statist. Assoc.*, doi: 10.1080/01621459.2022.2037430.

Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proc. 11th Conf. Uncertainty Artif. Intel.*, pp. 87–98. San Francisco, CA: Morgan Kaufmann.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2**, 445–98.

Consonni, G., La Rocca, L. & Peluso, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. *Scand. J. Statist.* **44**, 741–64.

Correa, J. & Bareinboim, E. (2020). A calculus for stochastic interventions: causal effect identification and surrogate experiments. In *Proc. 34th AAAI Conf. Artif. Intel.*, pp. 10093–100. New York: AAAI Press.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–317.

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**, 4079–86.

Eberhardt, F. (2008). Almost optimal intervention sets for causal discovery. In *Proc. 24th Conf. Uncertainty Artif. Intel.*, pp. 161–8. Arlington, VA: AUAI Press.

Foygel, R. & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, vol. 23, Lafferty, J., Williams, C., Shawe-Taylor, J. Zemel, R. & Culotta, A. eds. New York: Curran Associates.

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

GARCÍA-DONATO, G. & MARTÍNEZ-BENEITO, M. A. (2013). On sampling strategies in Bayesian variable selection problems with large model spaces. *J. Am. Statist. Assoc.* **108**, 340–52.

GEIGER, D. & HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30**, 1412–40.

HAUSER, A. & BÜHLMANN, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13**, 2409–64.

HAUSER, A. & BÜHLMANN, P. (2015). Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Statist. Soc.* B **77**, 291–318.

HE, Y. & GENG, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.* **9**, 2523–47.

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

HÜNERMUND, P. & BAREINBOIM, E. (2023). Causal inference and data fusion in econometrics. *arXiv:* 1912.09104v4.

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*, vol. 33, Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H. eds. New York: Curran Associates, pp. 9551–61.

KALISCH, M. & BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8**, 613–36.

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

KASS, R. E. & RAFTERY, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.

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

KORB, K. B., HOPE, L. R., NICHOLSON, A. E. & AXNICK, K. (2004). Varieties of causal intervention. In *PRICAI 2004: Trends in Artificial Intelligence*, Zhang, C. W., Guesgen, H. & Yeap, W. K. Berlin: Springer, pp. 322–31.

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

MARBACH, D., PRILL, R. J., SCHAFFTER, T., MATTIUSSI, C., FLOREANO, D. & STOLOVITZKY, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Nat. Acad. Sci.* **107**, 6286–91.

MARBACH, D., SCHAFFTER, T., MATTIUSSI, C. & FLOREANO, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J. Comp. Biol.* **16**, 229–39.

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

PELUSO, S. & CONSONNI, G. (2020). Compatible priors for model selection of high-dimensional Gaussian DAGs. *Electron. J. Statist.* **14**, 4110–32.

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 Rev. Genet.* **19**, 566–80.

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 Syst. Biol.* **7**, 1–19.

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

ROVERATO, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.* **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**, 523–9.

VERMA, T. & PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proc. 6th Ann. Conf. Uncertainty Artif. Intel.*, pp. 255–70. New York: Elsevier Science.

YANG, K., KATCOFF, A. & UHLER, C. (2018). Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proc. 35th Int. Conf. Machine Learning*, vol. 80, J. Dy & A. Kraus. PMLR, pp. 5541–50.

[*Received on* 5 *December* 2022. *Editorial decision on* 26 *April* 2023]