Statistical Methods & Applications (2021) 30:1289–1314
https://doi.org/10.1007/s10260-021-00579-1

ORIGINAL PAPER

Structural learning and estimation of joint causal effects among network-dependent variables

Federico Castelletti[1] · Alessandro Mascaro[1,2]

Accepted: 4 July 2021 / Published online: 2 August 2021

Abstract

Bayesian networks in the form of Directed Acyclic Graphs (DAGs) represent an effective tool for modeling and inferring dependence relations among variables, a process known as *structural learning*. In addition, when equipped with the notion of *intervention*, a *causal* DAG model can be adopted to quantify the causal effect on a response due to a hypothetical intervention on some variable. Observational data cannot distinguish between DAGs encoding the same set of conditional independencies (Markov equivalent DAGs), which however can be different from a *causal* perspective. In addition, because causal effects depend on the underlying network structure, uncertainty around the DAG generating model crucially affects the causal estimation results. We propose a Bayesian methodology which combines structural learning of Gaussian DAG models and inference of causal effects as arising from simultaneous interventions on any given set of variables in the system. Our approach fully accounts for the uncertainty around both the network structure and causal relationships through a joint posterior distribution over DAGs, DAG parameters and then causal effects.

Keywords Graphical model · Directed acyclic graph · Structural learning · Causal inference · Bayesian inference

1 Introduction

Graphical models based on Directed Acyclic Graphs (DAGs), also known as *Bayesian networks*, are widely employed to infer dependence relations among variables. Their application to scientific domains abounds, in particular social

✉ Federico Castelletti
federico.castelletti@unicatt.it

Alessandro Mascaro
a.mascaro3@campus.unimib.it

[1] Universita Cattolica del Sacro Cuore, Milan, Italy

[2] Università degli Studi di Milano-Bicocca, Milan, Italy

Table 1 Comparison between causal effects from single and joint interventions

Set	{6,5}		{6,4}	
Node	6	5	6	4
Single	−1.09	−0.14	**−1.09**	−0.51
Joint	−1.09	−0.14	**−0.68**	−0.51

sciences and biology; see for instance Friedman (2004), Yin and Li (2011), Markowetz and Spang (2007) and references therein. A DAG imposes a set of conditional independencies between variables through a DAG-dependent factorization of the joint distribution and provides an effective tool to read-off such relations directly from the graph using graphical criteria.

Real data problems generally suggest that the set of conditional dependencies between variables cannot be postulated in advance, which makes the DAG generating model unknown. In many fields (e.g. genomics) DAG *structural learning* also represents the very ultimate goal of the analysis since it can reveal dependence relations which may help understanding the behavior of biological pathways; see for instance Friedman and Koller (2003) and Sachs et al. (2005). From a statistical viewpoint this issue can be tackled by adopting a *model selection* perspective and several methodologies have been proposed accordingly. On the frequentist side, a distinction can be made between *score* and *constraint*-based methods. The former implement a score function which is maximised over the model space (Chickering 2002) while the latter usually provide a graph estimate by performing a sequence of conditional independence tests; see for instance Kalish and Buhlmann (2007). Moreover, Bayesian methodologies for DAG structural learning estimate a posterior distribution over the space of graphs which in turn provides a coherent quantification of the uncertainty around the data generating model; see for instance Cooper and Herskovits (1992), Ben-David et al. (2015) and the more recent paper Ni et al. (2017) which presents a unified framework for model selection of both directed and undirected graphs. By contrast, Bayesian methods require prior elicitation for each model-dependent parameter which is subject to constraints imposed by the conditional independencies of the underlying DAG structure; see also Geiger and Heckerman (2002).

In their standard formulation, DAGs provide information on the dependence structure between variables in terms of *association*. In many contexts however one might be interested in establishing *causal* relationships, whose precise quantification would require experimental (interventional) data, e.g. produced from randomized controlled experiments; see also Spirtes et al. (2000) and Ellis and Wong (2008). However intervention experiments are not always available since they may be unethical or infeasible. Because causal concepts are relationships that cannot be defined from the (observational) DAG distribution alone (Pearl, 2003) causal inference from non-experimental data requires further assumptions on the DAG generating mechanism. These are encoded in the notions of *intervention* and *post-intervention* distribution, leading to the definition of *causal DAG*; see also Sect. 2.2 for a more detailed discussion. Causal DAGs applied to observational data hence

provide a quantification of the effect of a hypothetical intervention on a specific variable w.r.t. a response of interest. This can be used for instance to predict the effect of a single gene knockout on some other gene or phenotype of interest; see for instance Maathuis et al. (2009). In turn, causal DAGs provide a tool for the design of experiments, because the collection of gene-causal effects can indicate which of them are likely to have a large effect on the response.

In many situations it is impossible to design experiments which act on one specific variable only while keeping all the other fixed. Also, causal effects from simultaneous interventions can significantly deviate from their single-variable counterparts, since the contribution of a given intervened variable in a joint intervention is "adjusted" by the effect of knocking out the others; see also Henckel et al. (2019). The following example aims at illustrating it.

Example 1 Consider DAG $\mathcal{D}$ in Fig. 1 and the parameters associated to its edges, which can be interpreted as the coefficients of a linear Structural Equation Model (SEM); see also Sect. 3. Table 1 provides a comparison between the causal effect of a given variable w.r.t. response node 1 from single and joint (simultaneous) interventions on set of variables. Each column refers to an intervened variable, while the two rows report the causal effect under a single intervention and the causal effect under a joint intervention on a pair of variables. As an instance, it appears that, while the causal effect of (an intervention on) node 6 is the same under single and joint interventions on $\{6, 5\}$ (-1.09), this increases up to -0.68 when node 6 is intervened simultaneously with 4.

DAGs encoding the same conditional independencies are called *Markov equivalent* and cannot be distinguished from observational data. However they can be different from a causal perspective; see also Sect. 2. Accordingly, a frequentist approach would estimate first an *equivalence* class of DAGs using observational data and then a set of DAG-dependent causal effects within the class. This is at the basis of the *IDA* and *joint-IDA* methods developed by Maathuis et al. (2009) and Nandy et al. (2017). However, results can be highly sensitive to the input (estimated) equivalence class, which also depends on the specific structural learning methodology that is adopted; see also Castelletti and Consonni (2021). All of these features suggest the adoption of a unified Bayesian method for DAG structural learning and causal effect estimation which fully accounts for the uncertainty around the underlying network structure. We remark that, within the Guassian setting hereinafter considered, literature on Bayesian causal discovery is, to our

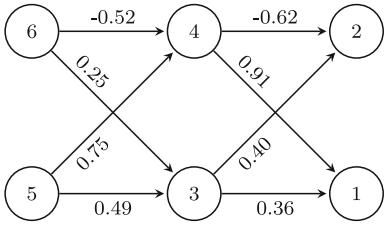

Fig. 1 A DAG with $q = 6$ variables and randomly generated coefficients

knowledge, still narrow, with few recent exceptions such as Castelletti and Consonni (2021).

In this work we present a Bayesian methodology which combines DAG structural learning and causal effect estimation for continuous, Gaussian data. As the result, our method returns an approximated posterior distribution over the space of DAGs and a posterior of the causal effects as arising from simultaneous interventions on any given set of variables. The rest of the paper is organized as follows. In Sect. 2 we introduce some theory and notation on graphical models based on DAGs, causal diagrams and causal discovery. In Sect. 3 we introduce Gaussian DAG models in terms of likelihood factorization and derive the post-intervention distribution and allied causal effect in the general case of simultaneous interventions. Section 4 deals with our Bayesian methodology for structural learning and causal discovery and introduces prior on DAGs and Cholesky parameters. We discuss in Sect. 5 computational details leading to an MCMC scheme for posterior inference. We apply our methodology on simulation settings and real data in Sects. 6 and 7 respectively. Finally, Sect. 8 offers a brief discussion together with possible future developments. All codes are written in R (R Core Team 2017) and are available upon request to the authors.

2 Graphical models, causal diagrams and causal effects

In this section we introduce basic concepts relative to graphical models based on directed acyclic graphs and causal inference. Fur further information on these topics the reader can refer to Pearl (2009) and Lauritzen (1996).

2.1 Graphical models

We briefly introduce the graph notation hereinafter adopted. Let $\mathcal{G} = (V, E)$ be a graph, where $V = \{1, \ldots, q\}$ is a set of nodes (or vertices) and $E \subseteq V \times V$ a set of edges. In what follows, if $(u, v) \in E$ and $(v, u) \notin E$, $\mathcal{G}$ contains a directed edge $u \to v$, while if both $(u, v) \in E$ and $(v, u) \in E$, then $\mathcal{G}$ contains an undirected edge $u - v$. A graph is called *directed* if contains only directed edges. Moreover, a Directed *Acyclic* Graph (DAG) $\mathcal{D}$ is a directed graph which contains no loops, that is sequences of nodes $(u_1, u_2, \ldots, u_k)$ with $u_1 = u_k$, such that there exists a path $u_1 \to u_2 \to \cdots \to u_k$. Moreover, if $(u, v) \in E$ we say that u is a parent of v and denote the set of all parents of v in $\mathcal{D}$ as $\text{pa}_{\mathcal{D}}(v)$. Also, if there exists a directed path from u to v we say that v is a *descendant* of u and let $\text{de}_{\mathcal{D}}(u)$ be the set of all descendants of u in $\mathcal{D}$. Hence, the *non-descendants* of u are all nodes in the set $\text{nd}_{\mathcal{D}}(u) = V \setminus \text{de}_{\mathcal{D}}(u)$.

Let now $(X_1, \ldots, X_q)$ be a random vector. The connection between a graph and probabilistic model $f(x_1, \ldots, x_q)$ for the random vector arises as we associate each variable X_j to a node in the graph. The latter introduces a set of conditional independencies among $X_1, \ldots, X_q$ via the so-called *Markov property* of the graph. As different types of dependence patterns exist, different types of graphs are in general equipped with different Markov properties. A DAG encodes a set of

conditional independencies between variables which can be read-off from the DAG using graphical criteria such as *d-separation* (Pearl, 2009). We then denote with $I(\mathcal{D})$ the set of all conditional independencies implied by $\mathcal{D}$. Let $\mathcal{D}$ be a DAG, $(X_1, \ldots, X_q)$ a collection of random variables. A distribution $f(x_1, \ldots, x_q)$ is said to be *compatible with* the DAG $\mathcal{D}$ or *Markov relative to* $\mathcal{D}$ if it admits the factorization

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

As many distributions may admit the factorization (1), it is possible to define a family of distributions $M(\mathcal{D})$ that are Markov relative to $\mathcal{D}$. For a given $f(x_1, \ldots, x_q) \equiv f$, if we let $I(f)$ be the set of conditional independencies in f, then $f \in M(\mathcal{D})$ if and only if $I(\mathcal{D}) \subseteq I(f)$. Moreover, if $I(\mathcal{D}) = I(f)$, then f is said to be *faithful* to DAG $\mathcal{D}$. This means that the conditional independencies in $\mathcal{D}$ are all and only those embodied in the joint distribution f.

A further important property of Bayesian networks is *Markov equivalence*. In particular, two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent if and only if $I(\mathcal{D}_1) = I(\mathcal{D}_2)$. It follows that a given set of conditional independencies can be described by several DAGs which are collected into an equivalence class. The latter can be uniquely represented through a *Completed Partially Directed Acyclic graph* (CPDAG) (Chickering 2002), also known as *Essential Graph* (EG) (Andersson et al. 1997) which is obtained as the union (over the edge sets) of all DAGs within the equivalence class.

2.2 Structural causal models and causal diagrams

DAGs are not necessarily carriers of causal information and their common extension to probabilistic graphical models, namely Bayesian networks, only allow to make conditional independence statements. Causal concepts are instead relationships that cannot be deduced from the distribution alone (Pearl, 2009) and accordingly require additional assumptions on the generating mechanism. One possibility is to assume that each parent-child relationship in the network represents a stable and autonomous physical mechanism, which means that it is possible to change one relationship without affecting the others. This assumption leads to the construction of *Structural Causal Models* (SCM) and the corresponding graphical tools, named *causal diagrams*. See also Pearl (2009)[Sect. 1.3.1–1.3.2] for a deep discussion and illustrative examples.

Traditionally, causal concepts are handled in econometrics and social sciences through *linear structural causal models*, that is SCM in which the relationships between variables are assumed to be linear. In general, an SCM can be represented through a system of relations of the form

$$X_j \leftarrow f_j(\mathrm{pa}_{\mathcal{D}}(j), U_j) \quad j = 1, \ldots, q, \tag{2}$$

where $\mathrm{pa}_{\mathcal{D}}(j)$ is now to be interpreted as the set of variables which directly determine the level of X_j. Moreover, U_j is an error term and the left-pointing arrow

indicates a *structural* relation (as opposed to algebraic relations); see also Pearl (2009)[Sect. 1.4].

Given a causal model in the form of (2), drawing an arrow from each variable in $\text{pa}_{\mathcal{D}}(j)$ towards X_j results in a DAG $\mathcal{D}$ called *causal diagram* which is called *Markovian* if it is acyclic and the error terms are jointly independent. It can be proved that every Markovian structural causal model M induces a distribution f which admits the same recursive decomposition (1) that characterizes Bayesian networks. However, causal models in the form (2) are more powerful as the assumptions of stability and autonomous mechanism allow to compute the effect of hypothetical interventions from non-experimental (observational) data.

We now introduce the notion of *intervention*. A *hard* (or *deterministic*) *intervention* on the set of variables $\{X_j, j \in I\}$, $I \subseteq V$, is denoted by $\text{do}\{X_j = \tilde{x}_j\}_{j\in I}$ and is defined as the action of fixing each X_j, $j \in I$, to some chosen value $\tilde{x}_j$. A hard intervention modifies the SCM by replacing each equations $X_j \leftarrow f_j(\text{pa}_{\mathcal{D}}(j), U_j)$ for $j \in I$ with a point mass at $\tilde{x}_j$. From a graphical perspective, the effect of a hard intervention can be represented through the so-called *intervention DAG*. This is obtained from the original DAG $\mathcal{D}$ by removing all edges (u, j) such that $j \in I$ and is denoted by $\mathcal{D}^I$; see also the example in Fig. 2.

A hard intervention $\text{do}\{X_j = \tilde{x}_j\}_{j\in I}$ leads to the definition of *post-intervention distribution* which can be written using the *truncated* factorization

$$f(x_1, \ldots, x_q \,|\, \text{do}\{X_j = \tilde{x}_j\}_{j\in I}) = \begin{cases} \prod_{i\notin I} f(x_i \,|\, \boldsymbol{x}_{\text{pa}_{\mathcal{D}}(\text{i})})\big|_{\{x_j=\tilde{x}_j\}_{j\in I}} & \text{if } x_j = \tilde{x}_j \quad \forall j \in I, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Importantly, the conditional densities in (3) are the same appearing in (1): this means that the post-intervention distribution can be expressed in terms of observational densities.

Moreover, Nandy et al. (2017) define the *total* joint effect of an intervention $\text{do}\{X_j = \tilde{x}_j\}_{j\in I}$ on $X_1 \equiv Y$ as

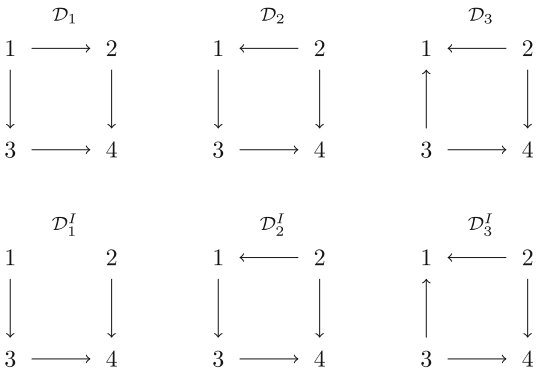

Fig. 2 Three DAGs, $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$, and the corresponding intervention DAGs of an intervention on $I = 2$. Intervention graphs $\mathcal{D}_2^I, \mathcal{D}_3^I$ are equal to the corresponding pre-intervention (observational) DAGs $\mathcal{D}_2, \mathcal{D}_3$, while $\mathcal{D}_1^I$ differs from $\mathcal{D}_1$ for the missing edge $1 \to 2$

$$\boldsymbol{\theta}_Y^I := (\theta_{h,Y}^I)_{h\in I}^\top, \tag{4}$$

where for each $h \in I$

$$\theta_{h,Y}^I := \frac{\partial}{\partial x_h}\mathbb{E}(Y \mid \mathrm{do}\{X_j = \tilde{x}_j\}_{j\in I}) \tag{5}$$

is the causal effect on Y associated to variable X_h in the joint intervention.

2.3 Causal discovery and causal effect estimation

Causal effects can be estimated whenever a causal diagram representing the causal structure of the problem is available. However, often this is not the case and the causal structure must be inferred from the data. Causal discovery methods, that is procedures whose aim is to learn causal DAGs from the data, are traditionally divided into three main classes: constraint-based methods, which estimate equivalence classes of DAGs by testing for conditional independencies between variables; score-based methods, which score DAGs through penalized likelihoods; hybrid methods which combine features of the first two approaches.

The PC algorithm (Spirtes et al. 2000) is one of the most popular algorithms for causal discovery. It is a constraint-based method that assumes acyclicity, causal faithfulness and *causal sufficiency*, where the latter refers to the absence of hidden (latent) variables. The PC algorithm provides an estimate of the CPDAG representing the true causal DAG. Specifically, it first estimates the CPDAG skeleton (that is the undirected graph that would be obtained by removing all the edge orientations from the DAG) and then orients as many edges as possible using various orientation rules; see also Kalish and Buhlmann (2007). For a complete review on causal discovery algorithms the reader can refer to Heinze-Deml et al. (2018).

A slightly different approach has been adopted by Maathuis et al. (2009), who propose a methodology for causal effect estimation from single-node hard interventions in Gaussian models when the DAG is not available. The resulting algorithm is called *IDA* (*Identification when DAG is Absent*). In its basic version, *IDA* first estimates an equivalence class using the PC algorithm (alternatively any other score-based method can be adopted). Next, for each DAG within the input class, the causal effect of X_h on Y is computed using multiple linear regression models. This basic version is slightly modified due to computational reasons. In particular, they propose a faster alternative which only returns the *distinct* causal effects compatible with the input equivalence class, thus avoiding a full enumeration of the DAGs. Their methodology is further extended to *joint* (simultaneous) hard interventions by Nandy et al. (2017), leading to their *joint-IDA* method.

As in the case of single interventions, *joint-IDA* relies on a CPDAG which is estimated up-front, e.g. by using the PC algorithm. Next, three alternative methods for causal estimation from joint interventions are proposed, namely the *path method*,

the *recursive regression for causal effects method* (RRC) and the *modified Cholesky decomposition method* (MCD); see the original paper for details.

3 Causal effects in Gaussian graphical models

In this section we focus on Gaussian DAG models and provide the definition of causal effect under the assumption that the distribution of $X_1, \ldots, X_q$ is jointly normal and Markov w.r.t. a given DAG which is known beforehand. In Sect. 4 we will deal instead under model (DAG) uncertainty.

Let $\mathcal{D} = (V, E)$ be a DAG and assume $(X_1, \ldots, X_q) \,|\, \boldsymbol{\Sigma}, \mathcal{D} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathcal{C}_\mathcal{D}$, the cone of symmetric-positive-definite covariance matrices Markov w.r.t. $\mathcal{D}$. Accordingly, $\boldsymbol{\Sigma}$ reflects the conditional independencies encoded by $\mathcal{D}$. Equation (1) thus becomes

$$f(x_1, \ldots, x_q \,|\, \boldsymbol{\Sigma}, \mathcal{D}) = \prod_{j=1}^{q} f(x_j \,|\, \boldsymbol{x}_{\mathrm{pa}_\mathcal{D}(\mathrm{j})}, \boldsymbol{\Sigma}). \tag{6}$$

Because of the normality assumption, Equation (6) can be equivalently written as a linear SEM. To this end, consider the decomposition of $\boldsymbol{\Sigma} = \boldsymbol{L}^{-\top} \boldsymbol{D} \boldsymbol{L}^{-1}$, where $\boldsymbol{L}$ is a (q, q) matrix of coefficients with diagonal elements equal to 1 and $\boldsymbol{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$ is a diagonal matrix collecting node-conditional variances. From this re-parameterization, the constraints imposed by $\mathcal{D}$ on the model parameters become more apparent since for each (u, v)-element of $\boldsymbol{L}$, $u \neq v$, we have $\boldsymbol{L}_{u,v} \neq 0$ if and only if $u \in \mathrm{pa}_\mathcal{D}(v)$, that is there is an edge $u \to v$ in $\mathcal{D}$. Hence, for $j = 1, \ldots, q$,

$$X_j = -\boldsymbol{L}_{\prec j]}^{\top} \boldsymbol{x}_{\mathrm{pa}_\mathcal{D}(\mathrm{j})} + \varepsilon_j, \quad \varepsilon_j \,|\, \sigma_j^2 \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \sigma_j^2), \tag{7}$$

where $\prec j] = \mathrm{pa}_\mathcal{D}(j) \times j$ and $\boldsymbol{L}_{A \times B}$ denotes the sub-matrix of $\boldsymbol{L}$ with elements belonging to rows and columns indexed by A and B respectively. Therefore,

$$f(x_1, \ldots, x_q \,|\, \boldsymbol{D}, \boldsymbol{L}, \mathcal{D}) = \prod_{j=1}^{q} d\mathcal{N}\big(x_j \,|\, -\boldsymbol{L}_{\prec j]}^{\top} \boldsymbol{x}_{\mathrm{pa}_\mathcal{D}(\mathrm{j})}, \sigma_j^2\big). \tag{8}$$

Let now $I \subseteq \{2, \ldots, q\}$ be an intervention target. The post-intervention distribution of $(X_1, \ldots, X_q)$ given $\mathrm{do}\{X_j = \tilde{x}_j\}_{j \in I}$ becomes

$$f(\boldsymbol{x} \,|\, \mathrm{do}\{X_j = \tilde{x}_j\}_{j \in I}, \boldsymbol{D}, \boldsymbol{L}, \mathcal{D}) = \begin{cases} \prod_{i \notin I} d\mathcal{N}\big(x_i \,|\, -\boldsymbol{L}_{\prec i]}^{\top} \boldsymbol{x}_{\mathrm{pa}_\mathcal{D}(\mathrm{i})}, \sigma_i^2\big)\big|_{\{x_j = \tilde{x}_j\}_{j \in I}} & \text{if } x_j = \tilde{x}_j \quad \forall j \in I, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

An important implication of Equation (9) is that

$$X_1, \ldots, X_q \,|\, \mathrm{do}\{X_j = \tilde{x}_j\}_{j \in I}, \boldsymbol{\Sigma}, \mathcal{D} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma}^I), \tag{10}$$

where

$$\boldsymbol{\Sigma}^I = (\boldsymbol{L}^I)^{-\top}\boldsymbol{D}(\boldsymbol{L}^I)^{-1} \tag{11}$$

and in particular

$$\boldsymbol{L}^I_{u,v} = \begin{cases} 0 & \text{if } v \in I \text{ and } v \neq u \\ \boldsymbol{L}_{u,v} & \text{otherwise.} \end{cases} \tag{12}$$

Equation (10) corresponds to the post-intervention distribution of $(X_1, \ldots, X_q)$ in the Gaussian setting and depends on the covariance matrix $\boldsymbol{\Sigma}^I$. Most importantly, the latter can be reconstructed from the observational parameters $(\boldsymbol{D}, \boldsymbol{L})$ appearing in (7). Finally, the causal effect of X_h on X_1 $(h \in I)$ in a joint intervention on $\{X_j\}_{j\in I}$ is given by

$$\theta^I_{h,1} = \boldsymbol{\Sigma}^I_{h,1}\left(\boldsymbol{\Sigma}^I_{h,h}\right)^{-1}; \tag{13}$$

see also Nandy et al. (2017). It follows that the causal effect $\theta^I_{h,1}$ is a function of the covariance matrix $\boldsymbol{\Sigma}$ which in turn depends on the underlying DAG $\mathcal{D}$. Therefore, inference on DAG $\mathcal{D}$ and its parameter $\boldsymbol{\Sigma}$ will drive inference of causal effects under model uncertainty; see the next section for details.

4 Bayesian inference of causal effects under model uncertainty

Consider n i.i.d. observations $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,q})^\top$, $i = 1, \ldots, n$, from (8) and the (n, q) data matrix $\boldsymbol{X}$, row-binding of the $\boldsymbol{x}_i$'s. The likelihood function relative to $(\boldsymbol{D}, \boldsymbol{L}, \mathcal{D})$ can be written as

$$f(\boldsymbol{X} \mid \boldsymbol{D}, \boldsymbol{L}, \mathcal{D}) = \prod_{j=1}^{q} d\mathcal{N}_n\Big(\boldsymbol{X}_j \mid -\boldsymbol{X}_{\mathrm{pa}_\mathcal{D}(\mathrm{j})}\boldsymbol{L}_{\prec j]}, \sigma_j^2\boldsymbol{I}_n\Big), \tag{14}$$

where $\boldsymbol{X}_A$ denotes the sub-matrix of $\boldsymbol{X}$ with columns indexed by $A \subseteq V$ and $\boldsymbol{I}_n$ is the (n, n) identity matrix. We now proceed by assigning a prior distribution on DAG $\mathcal{D}$ and its Cholesky parameters $(\boldsymbol{D}, \boldsymbol{L})$.

4.1 Prior on DAG $\mathcal{D}$

For a given DAG $\mathcal{D} = (V, E)$, let $\boldsymbol{S}^\mathcal{D}$ be the 0-1 *adjacency matrix* of its skeleton (the underlying undirected graph obtained after removing the orientation of all of its edges), such that for each (u, v)-element in $\boldsymbol{S}^\mathcal{D}$, $\boldsymbol{S}^\mathcal{D}_{u,v} = 1$ if and only if $(u, v) \in E$ or $(v, u) \in E$, 0 otherwise. Conditionally on a prior probability of inclusion $\pi \in (0, 1)$ we then assume $\boldsymbol{S}^\mathcal{D}_{u,v} \mid \pi \overset{\text{iid}}{\sim} \text{Ber}(\pi)$ for each $u > v$. Therefore,

$$p(\boldsymbol{S}^\mathcal{D}) = \pi^{|\boldsymbol{S}^\mathcal{D}|}(1-\pi)^{\frac{q(q-1)}{2} - |\boldsymbol{S}^\mathcal{D}|}, \tag{15}$$

where $|\boldsymbol{S}^\mathcal{D}|$ is the number of edges in $\mathcal{D}$ (equivalently in its skeleton) and $q(q-1)/2$ corresponds to the maximum number of edges in a DAG on q nodes. Finally we set

$$p(\mathcal{D}) \propto p(\boldsymbol{S}^{\mathcal{D}}), \tag{16}$$

for any $\mathcal{D} \in \mathcal{S}_q$, where $\mathcal{S}_q$ is the space of all DAGs on q nodes. The resulting prior thus depends on the DAG skeleton only and assigns equal prior weights to DAGs having the same number of edges. Hyperparameter π can be tuned to reflect some prior knowledge of sparsity in the DAG space, when this information is available. Moreover, one can adopt distinct hyperparameters $\pi_{u,v}$, to include edge-specific prior information on the edge inclusions.

4.2 Prior on Cholesky parameters $(\boldsymbol{D},\boldsymbol{L})$

Consider now the DAG-constrained covariance matrix $\boldsymbol{\Sigma} \in \mathcal{C}_{\mathcal{D}}$ and its modified Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{L}^{-\top}\boldsymbol{D}\boldsymbol{L}^{-1}$; see also Sect. 3. Moreover, without loss of generality assume a parent ordering of the nodes such that $u > v$ whenever u is a parent of v. We assign a prior to $\boldsymbol{\Sigma}$ through a *DAG-Wishart* prior on $(\boldsymbol{D}, \boldsymbol{L})$ with hyperparameter $\boldsymbol{U}$ (a $q \times q$ s.p.d. matrix) and shape hyperparameter $\boldsymbol{a}^{\mathcal{D}} = (a_1^{\mathcal{D}}, \ldots, a_q^{\mathcal{D}})^{\top}$ (Ben-David et al., 2015). The DAG-Wishart distribution has been introduced as a conjugate prior for covariance matrices Markov w.r.t. a DAG $\mathcal{D}$ and therefore provides an extension of the standard Wishart distribution which can be adopted for unconstrained covariance matrices (equivalently, complete DAG models); see also Ben-David et al. (2015) for details.

An interesting feature of the DAG-Wishart distribution is that it induces a re-parameterization of $\boldsymbol{\Sigma}$ in terms of node-parameters $(\boldsymbol{L}_{\prec j]}, \sigma_j^2)$ that are a priori independent with distribution

$$\begin{aligned} \sigma_j^2 \mid \mathcal{D} &\sim \text{ I-Ga}\left(a_j^{\mathcal{D}}, \frac{1}{2}\boldsymbol{U}_{jj|\prec j\succ}\right), \\ \boldsymbol{L}_{\prec j]} \mid \sigma_j^2, \mathcal{D} &\sim \mathcal{N}_{|\text{pa}_{\mathcal{D}}(j)|}\left(-\boldsymbol{U}_{\prec j\succ}^{-1}\boldsymbol{U}_{\prec j]}, \sigma_j^2\boldsymbol{U}_{\prec j\succ}^{-1}\right), \end{aligned} \tag{17}$$

where $\boldsymbol{U}_{jj|\prec j\succ} = \boldsymbol{U}_{jj} - \boldsymbol{U}_{[j\succ}\boldsymbol{U}_{\prec j\succ}^{-1}\boldsymbol{U}_{\prec j]}$, $\prec j] = \text{pa}_{\mathcal{D}}(j) \times j$, $[j\succ = j \times \text{pa}_{\mathcal{D}}(j)$ and $\prec j\succ = \text{pa}_{\mathcal{D}}(j) \times \text{pa}_{\mathcal{D}}(j)$. Hyperparameters $a_j^{\mathcal{D}}$ and $\boldsymbol{U}$ are specific to each DAG model under consideration. However, it can be shown that the default choice $a_j^{\mathcal{D}} = \frac{1}{2}(a + |\text{pa}_{\mathcal{D}}(j)| - q + 1)$ $(a > q - 1)$ guarantees compatibility among prior distributions for Markov equivalent DAGs; see also Castelletti and Consonni (2020). Also, a standard choice for $\boldsymbol{U}$, hereinafter adopted, is $\boldsymbol{U} = g\boldsymbol{I}_q$ $(g > 0)$ which reflects a prior belief of (marginal) independence among variables. Hyperparameter g regulates the strength of our prior statement: lower values of g correspond to less informative prior distributions on $\boldsymbol{\Omega}$. Under this choice, Equation (17) becomes

$$\begin{aligned} \sigma_j^2 \mid \mathcal{D} &\sim \text{I-Ga}\left(a_j^{\mathcal{D}}, \frac{1}{2}\, g\right), \\ \boldsymbol{L}_{\prec j]} \mid \sigma_j^2, \mathcal{D} &\sim \mathcal{N}_{|\text{pa}_{\mathcal{D}}(j)|}\left(\boldsymbol{0}, \sigma_j^2 \left(g \boldsymbol{I}_{|\text{pa}_{\mathcal{D}}(j)|}\right)^{-1}\right). \end{aligned} \tag{18}$$

Therefore, a prior on the DAG Cholesky parameters $(\boldsymbol{D}, \boldsymbol{L})$ is given by

$$p(\boldsymbol{D}, \boldsymbol{L} \mid \mathcal{D}) = \prod_{j=1}^{q} p(\boldsymbol{L}_{\prec j]} \mid \sigma_j^2)\, p(\sigma_j^2). \tag{19}$$

Moreover, because of conjugacy of (18) with the likelihood (14), the posterior distribution of $(\boldsymbol{D}, \boldsymbol{L})$ given $\boldsymbol{X}$ is such that

$$\begin{aligned} \sigma_j^2 \mid \boldsymbol{X}, \mathcal{D} &\sim \text{I-Ga}\left(a_j^{\mathcal{D}} + \frac{n}{2}, \frac{1}{2}\,(g + b_j)\right), \\ \boldsymbol{L}_{\prec j]} \mid \sigma_j^2 \boldsymbol{X}, \mathcal{D} &\sim \mathcal{N}_{|\text{pa}_{\mathcal{D}}(j)|}\left(-\hat{\boldsymbol{L}}_{\prec j]}, \sigma_j^2 \left(g \boldsymbol{I}_{|\text{pa}_{\mathcal{D}}(j)|} + \boldsymbol{S}_{\text{pa}_{\mathcal{D}}(j)}\right)^{-1}\right), \end{aligned} \tag{20}$$

where

$$\begin{aligned} \boldsymbol{S}_{\text{pa}_{\mathcal{D}}(j)} &= \boldsymbol{X}_{\text{pa}_{\mathcal{D}}(j)}^{\top} \boldsymbol{X}_{\text{pa}_{\mathcal{D}}(j)}, \\ \hat{\boldsymbol{L}}_{\prec j]} &= \left(g \boldsymbol{I}_{|\text{pa}_{\mathcal{D}}(j)|} + \boldsymbol{S}_{\text{pa}_{\mathcal{D}}(j)}\right)^{-1} \boldsymbol{X}_{\text{pa}_{\mathcal{D}}(j)}^{\top} \boldsymbol{X}_j, \\ b_j &= \boldsymbol{X}_j^{\top} \boldsymbol{X}_j - \boldsymbol{X}_j^{\top} \boldsymbol{X}_{\text{pa}_{\mathcal{D}}(j)} \left(g \boldsymbol{I}_{|\text{pa}_{\mathcal{D}}(j)|} + \boldsymbol{S}_{\text{pa}_{\mathcal{D}}(j)}\right)^{-1} \boldsymbol{X}_{\text{pa}_{\mathcal{D}}(j)}^{\top} \boldsymbol{X}_j. \end{aligned}$$

5 Computational details

In this section we detail the MCMC scheme that we adopt to sample from the target distribution

$$p(\boldsymbol{D}, \boldsymbol{L}, \mathcal{D} \mid \boldsymbol{X}) \propto f(\boldsymbol{X} \mid \boldsymbol{D}, \boldsymbol{L}, \mathcal{D})\, p(\boldsymbol{D}, \boldsymbol{L} \mid \mathcal{D})\, p(\mathcal{D}), \tag{21}$$

and therefore to approximate the posterior distribution of the causal effect in (13) which is a function of the Cholesky parameters $(\boldsymbol{D}, \boldsymbol{L})$; see Equations (12) and (13). An efficient sampler can be implemented by resorting to a *Partial Analytic Structure* (PAS) algorithm (Godsill, 2012) whereas the update of DAG $\mathcal{D}$ and model parameters $(\boldsymbol{D}, \boldsymbol{L})$ is performed in two steps.

5.1 Update of DAG $\mathcal{D}$

In the first step, given the current DAG $\mathcal{D}$, a new DAG $\mathcal{D}^*$ is drawn from a suitable proposal distribution $q(\mathcal{D}^* \mid \mathcal{D})$ which is defined as follows. We consider three types of operators that locally modify a DAG: insert a directed edge (InsertD $u \to v$ for short), delete a directed edge (DeleteD $u \to v$) and reverse a directed edge (ReverseD $u \to v$). For a given $\mathcal{D} \in \mathcal{S}_q$, being $\mathcal{S}_q$ the set of all DAGs on q nodes, we then construct the set of *valid* operators $\mathcal{O}_{\mathcal{D}}$, that is operators whose resulting graph

is a DAG. A DAG $\mathcal{D}^*$ is then called a *direct successor* of $\mathcal{D}$ if it can be reached by applying an operator in $\mathcal{O}_\mathcal{D}$ to $\mathcal{D}$. Therefore, given the current $\mathcal{D}$ we propose $\mathcal{D}^*$ by uniformly sampling an element in $\mathcal{O}_\mathcal{D}$ and applying it to $\mathcal{D}$. Since there is a one-to-one correspondence between each operator and resulting DAG, the probability of transition is $q(\mathcal{D}^* \mid \mathcal{D}) = 1/|\mathcal{O}_\mathcal{D}|$, for each $\mathcal{D}^*$ direct successor of $\mathcal{D}$.

Consider now two DAGs $\mathcal{D}$ and $\mathcal{D}^*$ which differ by one edge $(u, v) \in \mathcal{D}$, $(u, v) \notin \mathcal{D}^*$ and let $(\boldsymbol{D}^\mathcal{D}, \boldsymbol{L}^\mathcal{D})$ and $(\boldsymbol{D}^{\mathcal{D}^*}, \boldsymbol{L}^{\mathcal{D}^*})$ be the corresponding Cholesky parameters. Notice that these differ only with regard to their v-th component $((\sigma_v^\mathcal{D})^2, \boldsymbol{L}^\mathcal{D}_{\prec v]})$ and $((\sigma_v^{\mathcal{D}^*})^2, \boldsymbol{L}^{\mathcal{D}^*}_{\prec v]})$. Moreover, the remaining parameters $\{(\sigma_r^\mathcal{D})^2, \boldsymbol{L}^\mathcal{D}_{\prec r]};\, r \neq v\}$ and $\{(\sigma_r^{\mathcal{D}^*})^2, \boldsymbol{L}^{\mathcal{D}^*}_{\prec r]};\, r \neq v\}$ are componentwise *equivalent* between the two graphs because they refer to structurally equivalent conditional models. This feature is crucial for the correct application of the PAS algorithm; see also Wang and Li (2012)[Sect. 5.2].

Under a PAS algorithm the acceptance probability for a DAG $\mathcal{D}^*$ (defined as above) is given by $\alpha_{\mathcal{D}^*} = \min\{1; r_{\mathcal{D}^*}\}$ where

$$r_{\mathcal{D}^*} = \frac{p(\boldsymbol{X}, \boldsymbol{D}^{\mathcal{D}^*} \setminus (\sigma_v^{\mathcal{D}^*})^2, \boldsymbol{L}^{\mathcal{D}^*} \setminus \boldsymbol{L}^{\mathcal{D}^*}_{\prec v]} \mid \mathcal{D}^*)}{p(\boldsymbol{X}, \boldsymbol{D}^\mathcal{D} \setminus (\sigma_v^\mathcal{D})^2, \boldsymbol{L}^\mathcal{D} \setminus \boldsymbol{L}^\mathcal{D}_{\prec v]} \mid \mathcal{D})} \cdot \frac{p(\mathcal{D}^*)}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D} \mid \mathcal{D}^*)}{q(\mathcal{D}^* \mid \mathcal{D})}, \tag{22}$$

and

$$p(\boldsymbol{X}, \boldsymbol{D}^\mathcal{D} \setminus (\sigma_v^\mathcal{D})^2, \boldsymbol{L}^\mathcal{D} \setminus \boldsymbol{L}^\mathcal{D}_{\prec v]} \mid \mathcal{D}) = \int\int f(\boldsymbol{X} \mid \boldsymbol{D}^\mathcal{D}, \boldsymbol{L}^\mathcal{D}, \mathcal{D})\, p(\boldsymbol{D}^\mathcal{D}, \boldsymbol{L}^\mathcal{D} \mid \mathcal{D})\, d\boldsymbol{L}^\mathcal{D}_{\prec v]}\, d(\sigma_v^\mathcal{D})^2$$

(similarly under $\mathcal{D}^*$). In addition, because of the likelihood and prior factorizations in (14) and (19), it can be shown that (22) reduces to

$$r_{\mathcal{D}^*} = \frac{m(\boldsymbol{X}_v \mid \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}^*}(v)}, \mathcal{D}^*)}{m(\boldsymbol{X}_v \mid \boldsymbol{X}_{\mathrm{pa}_\mathcal{D}(v)}, \mathcal{D})} \cdot \frac{p(\mathcal{D}^*)}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D} \mid \mathcal{D}^*)}{q(\mathcal{D}^* \mid \mathcal{D})}, \tag{23}$$

where, because of conjugacy of the Normal-Inverse-Gamma prior with the Normal density,

$$m(\boldsymbol{X}_v \mid \boldsymbol{X}_{\mathrm{pa}_\mathcal{D}(v)}, \mathcal{D}) = (2\pi)^{-\frac{n}{2}} \frac{\left|g\boldsymbol{I}_{|\mathrm{pa}_\mathcal{D}(v)|}\right|^{1/2}}{\left|g\boldsymbol{I}_{|\mathrm{pa}_\mathcal{D}(v)|} + \boldsymbol{S}_{\mathrm{pa}_\mathcal{D}(v)}\right|^{1/2}} \cdot \frac{\Gamma\left(a_v^\mathcal{D} + \frac{n}{2}\right)}{\Gamma\left(a_v^\mathcal{D}\right)} \frac{\left(\frac{g}{2}\right)^{a_v^\mathcal{D}}}{\left(\frac{g+b_v}{2}\right)^{a_v^\mathcal{D}+n/2}}, \tag{24}$$

and $\boldsymbol{S}_{\mathrm{pa}_\mathcal{D}(v)}$, b_v are defined as in Sect. 4.2.

5.2 Update of Cholesky parameters $(\boldsymbol{D},\boldsymbol{L})$

In the second step we then sample the model-dependent parameters $(\boldsymbol{D}, \boldsymbol{L})$ conditionally on the accepted DAG from their full conditional distribution in (20). In addition, samples from the Cholesky parameters $(\boldsymbol{D}, \boldsymbol{L})$ can be used to recover $\boldsymbol{\Sigma}^I$

and then compute the causal effect for any given target of intervention nodes $I \subseteq \{2, \ldots, q\}$ as in (13).

5.3 Posterior inference

Our MCMC scheme is summarized in Algorithm 1. Its output consists of a collection of DAGs and Cholesky parameters $\left\{\mathcal{D}^{(s)}, (\boldsymbol{D}^{(s)}, \boldsymbol{L}^{(s)})\right\}_{s=1}^{S}$ sampled from the posterior (21) and a collection of causal effect coefficients $\left\{\theta_{h,1}^{I(s)}\right\}_{s=1}^{S}$ for an input intervention target I and each $h \in I$.

Algorithm 1: MCMC scheme for posterior inference

Input: An (n, q) data matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q)$, a target of intervention I
Output: S samples from the posterior distribution of $(\boldsymbol{D}, \boldsymbol{L}, \mathcal{D})$ and $\theta_{h,1}^{I}$, for $h \in I$

1 Initialize $\mathcal{D}^{(0)}$, e.g. the empty DAG;
2 **for** $s = 1, \ldots, S$ **do**
3 Sample $\mathcal{D}^{*}$ from $q(\mathcal{D}^{*} \mid \mathcal{D}^{(s-1)})$ and set $\mathcal{D}^{(s)} = \mathcal{D}^{*}$ with probability $\alpha_{\mathcal{D}}$, otherwise $\mathcal{D}^{(s)} = \mathcal{D}^{(s-1)}$;
4 Sample $(\boldsymbol{D}^{(s)}, \boldsymbol{L}^{(s)})$ from the posterior distribution $p(\boldsymbol{D}, \boldsymbol{L} \mid \boldsymbol{X}, \mathcal{D}^{(s)})$ using (20);
5 Construct $\boldsymbol{L}^{I^{(s)}}$ as in (12) and recover $\boldsymbol{\Sigma}^{I^{(s)}}$ using (11);
6 For each $h \in I$ obtain $\theta_{h,1}^{I(s)}$ as in (13).
7 **end**

Moreover, the posterior probability of a DAG $\mathcal{D}$ can be approximated from the MCMC output as

$$p(\mathcal{D} \mid \boldsymbol{X}) \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\left\{\mathcal{D}^{(s)} = \mathcal{D}\right\}, \tag{25}$$

which corresponds to the DAG frequency of visits in the chain. Alternatively, approximations of posterior model probabilities can be obtained from re-normalized marginal likelihoods; see also García-Donato and Martínez-Beneito (2013) for a discussion.

From the same output, summaries of interest can be also obtained. In particular, for each pair of nodes (u, v) we can compute the (estimated) posterior probability of edge inclusion

$$\hat{p}(u \to v \mid \boldsymbol{X}) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\left\{u \to v \in \mathcal{D}^{(s)}\right\}, \tag{26}$$

where $\mathbb{1}\left\{u \to v \in \mathcal{D}^{(s)}\right\} = 1$ if $\mathcal{D}^{(s)}$ contains $u \to v$, 0 otherwise. Moreover, an overall summary of the posterior distribution of each causal effect $\theta_{h,1}^{I}$ can be computed as

$$\hat{\theta}^{I}_{h,1} = \frac{1}{S}\sum_{s=1}^{S} \theta^{I(s)}_{h,1}, \tag{27}$$

which corresponds to a Bayesian Model Averaging (BMA) estimate wherein posterior model probabilities are approximated through their MCMC frequencies of visits. Equation (27) naturally incorporates the uncertainty around both the underlying causal DAG model and the allied DAG-dependent parameters.

6 Simulation study

6.1 Settings

In this section we evaluate our methodology through simulation experiments. Specifically, we construct different scenarios by varying the sample size $n \in \{50, 100, 200, 500\}$, the number of intervened nodes (size of the target) $s \in \{2, 4\}$ and the number of nodes $q \in \{10, 20\}$. Under each simulation scenario defined by n we generate $N = 30$ datasets, each obtained as follows. We first randomly sample a *sparse* DAG $\mathcal{D}$ by fixing a probability of edge inclusion equal to 0.2. We then generate the corresponding (true) Cholesky parameters $(\boldsymbol{D}, \boldsymbol{L})$ by drawing the non-zero elements of $\boldsymbol{L}$ in $[-1, -0.1] \cup [0.1, 1]$ while fixing $\boldsymbol{D} = \boldsymbol{I}_q$. We finally construct the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{L}^{-\top}\boldsymbol{D}\boldsymbol{L}^{-1}$ and generate n multivariate i.i.d. observations, representing an (n, q) dataset $\boldsymbol{X}$, from the Gaussian DAG-model $\mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Moreover, for a given s we randomly choose a target I consisting of s nodes randomly sampled from $\{2, \ldots, q\}$. We then recover the post-intervention parameters $\boldsymbol{L}^I$ and $\boldsymbol{\Sigma}^I$ using (11) and (12); the true set of causal effects $\{\theta^I_{h,1}\}_{h \in I}$ follows from (13). For each simulated dataset we run $S = 15000$ iterations of Algorithm 1 to approximate the posterior distribution over DAGs, Cholesky parameters and causal effects.

6.2 Graph selection

To assess the accuracy of our method in recovering the underlying causal structure we compare DAG point estimates that can be retrieved from our MCMC output with the corresponding true DAGs. Similarly, we evaluate the performance of the frequentist PC algorithm, the structural learning method underlying the IDA approach of Maathuis et al. (2009). Specifically, with regard to our method, we consider both the Median Probability (DAG) Model (MPM) and the Maximum A Posteriori (MAP) DAG estimates, where the former corresponds to the DAG obtained by including those edges whose posterior probability of inclusion exceeds 0.5, while the latter corresponds to the DAG with highest MCMC frequency of visits. We implement PC algorithm at significance level $\alpha \in \{0.01, 0.05, 0.10\}$. In addition, because PC outputs a CPDAG, starting from each of our DAG estimates (MPM and MAP) we construct the representative CPDAG, that is the CPDAG

representing the equivalence class of the estimated DAG. We compare each CPDAG estimate with the CPDAG representing the equivalence class of the true DAG in terms of Structural Hamming Distance (SHD) and Structural Intervention Distance (SID). SHD corresponds to the number of edge insertions, deletions or flips needed to transform the estimated graph into the true graph. SID was instead introduced by Peters and Bühlmann (2015) and is based on a graphical criterion quantifying the closeness between two graphs in terms of the corresponding sets of compatible intervention distributions; see also the original paper for details. Lower values of SHD and SID correspond to better performances. Our results are summarized in the box-plots of Figs. 3 and 4 which report the distribution of the two indexes over the $N = 30$ simulations. It appears that our MPM and MAP estimates

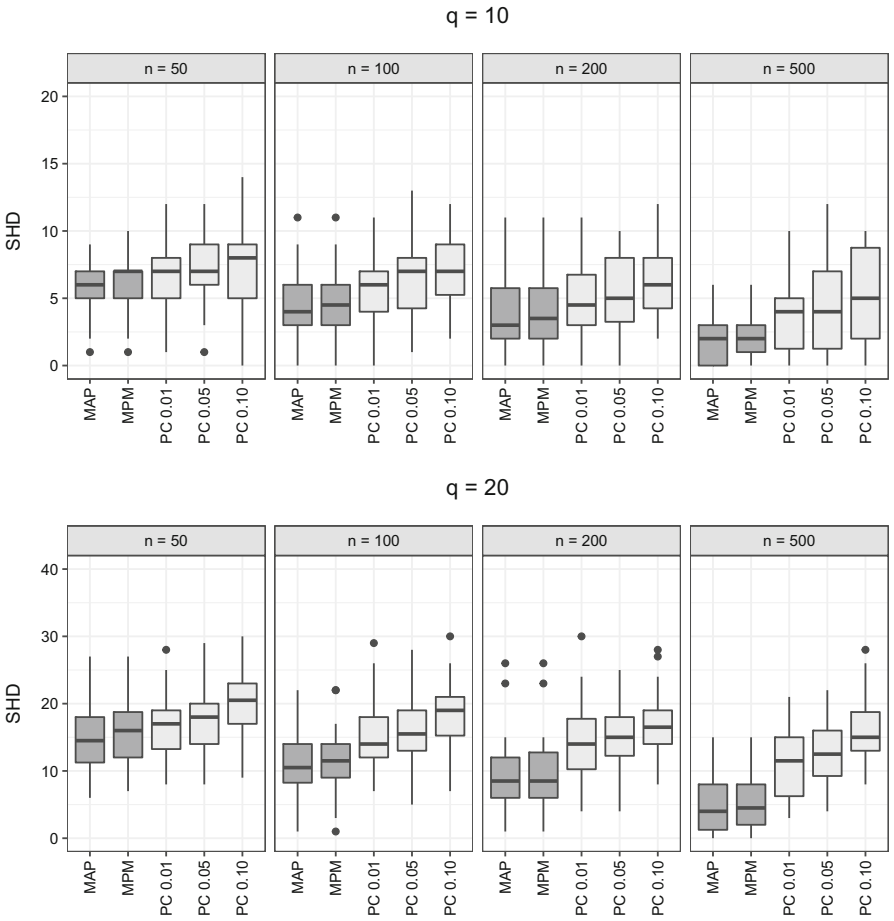

Fig. 3 Simulation study. Distribution over $N = 30$ simulated datasets of the Structural Hamming Distance (SHD) between estimated and true CPDAGs. Methods under comparison are: our Bayesian method with output the Median Probability Model (MPM) and Maximum A Posteriori (MAP) graph estimates and the PC algorithm implemented at significance level $\alpha \in \{0.01, 0.05, 0.10\}$, respectively PC 0.01, PC 0.05, PC 0.10

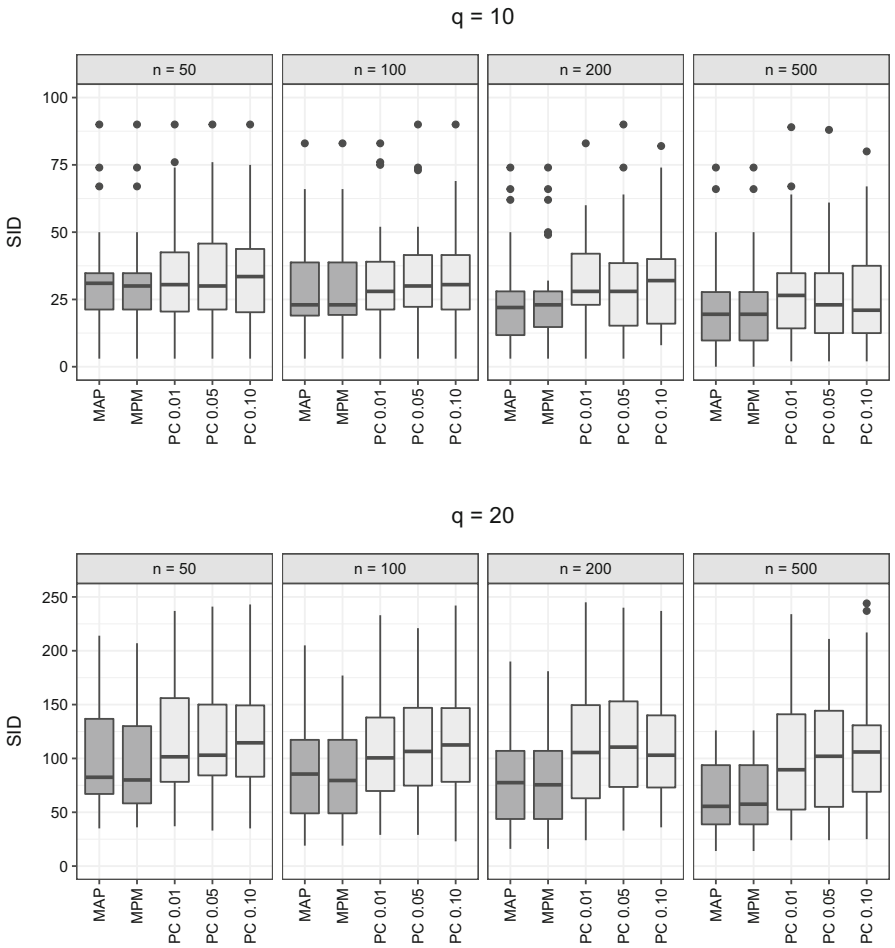

Fig. 4 Simulation study. Distribution over $N = 30$ simulated datasets of the Structural Intervention Distance (SID) between estimated and true CPDAGs. Methods under comparison are: our Bayesian method with output the Median Probability Model (MPM) and Maximum A Posteriori (MAP) graph estimates and the PC algorithm implemented at significance level $\alpha \in \{0.01, 0.05, 0.10\}$, respectively PC 0.01, PC 0.05, PC 0.10

are competitive with PC for moderate sample sizes and perform slightly better than PC as n increases both in terms of SHD and SID.

6.3 Causal effect estimation

We now evaluate the performance of our method in causal effect estimation. To this end, under each simulation we compute the BMA estimate (27) for each intervened node $h \in I$. Each estimated causal effect $\hat{\theta}^I_{h,1}$, $h \in I$, is compared with the corresponding true causal effect $\theta^I_{h,1}$ by computing the absolute-value distance

$$d_h^{BMA} = \left| \hat{\theta}_{h,1}^I - \theta_{h,1}^I \right|. \tag{28}$$

We also include in our analysis the *joint-IDA* method of Nandy et al. (2017) (see also Sect. 2.3). In particular, for the graph selection step we implement PC algorithm at significance level $\alpha = 0.01$ which has also been shown to perform better in sparse settings Kalish and Buhlmann 2007. *Joint-IDA* recovers for each intervened node $h \in I$ the set of distinct causal effects compatible with the input CPDAG. This is then summarized through the arithmetic mean which provides an estimate of $\theta_{h,1}^I$, $h \in I$. The *joint-IDA* estimate is compared with the true causal effect by computing the absolute-value distance d_h^{IDA} following (28).

Results are summarized in the box-plots of Fig. 5 which reports the distribution of d_h^{BMA} and d_h^{IDA} across the 30 simulated datasets and intervened nodes, for increasing values of the sample size n, different number of variables q and different size of the target $s \in \{2, 4\}$. Clearly, lower values of the distance correspond to better performances.

It appears that both methods improve their performances as the sample size increases. However, our BMA-based method outperforms *joint-IDA* under all scenarios and in particular in the setting $s = 4$. One possible reason is that,

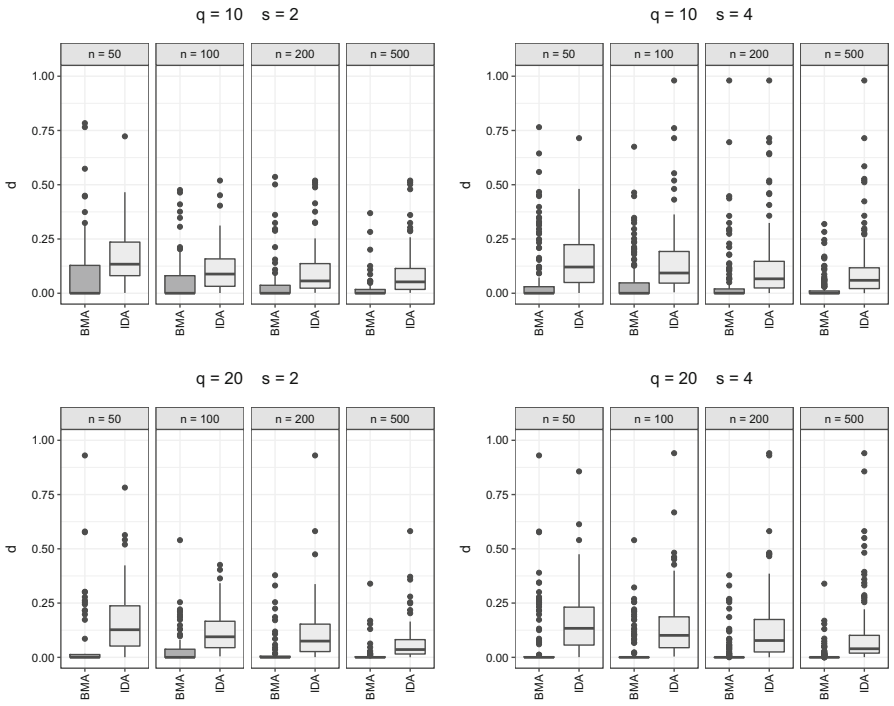

Fig. 5 Simulation study. Distribution of the absolute-value distance d between estimated and true causal effects for size of the target $s \in \{2, 4\}$, number of variables $q \in \{10, 20\}$ and sample size $n \in \{50, 100, 200, 500\}$. Methods under comparison are: our BMA-based approach (BMA) and the *Joint-IDA* method (IDA)

differently from our Bayesian method, *joint-IDA* relies on a given (estimated) equivalence class of DAGs. Indeed, causal inference results strongly depend on the input CPDAG estimate and therefore on the accuracy in the graph selection. Anyway, results obtained under different CPDAG estimates (e.g. the PC algorithm for different significance levels) did not lead to improvements in the causal effect estimation. By contrast, our MCMC-based method relies on a posterior distribution over a collection of DAGs some of which, although lying outside the true-DAG equivalence class, might be "structurally similar" to the true causal DAG and still result in a causal effect which is close to the true one. This result is also consistent with the behavior observed in the Structural Intervention Distance (Fig. 5).

We finally investigate the computational time required by our method. Figure 6 summarizes the behaviour of the running time (averaged over 12 replicates) *per* iteration, as a function of $q \in \{10, \ldots, 100\}$ for $n = 500$ (upper panel), and as a function of $n \in \{50, \ldots, 1000\}$ for $q = 50$ (lower panel). We also highlight that the computational burden of our method might increase with the number of variables also because the size of the DAG space grows more than exponentially in q and accordingly a larger number of MCMC iterations are in general required to reach convergence. By converse, the computational burden of IDA is generally lower.

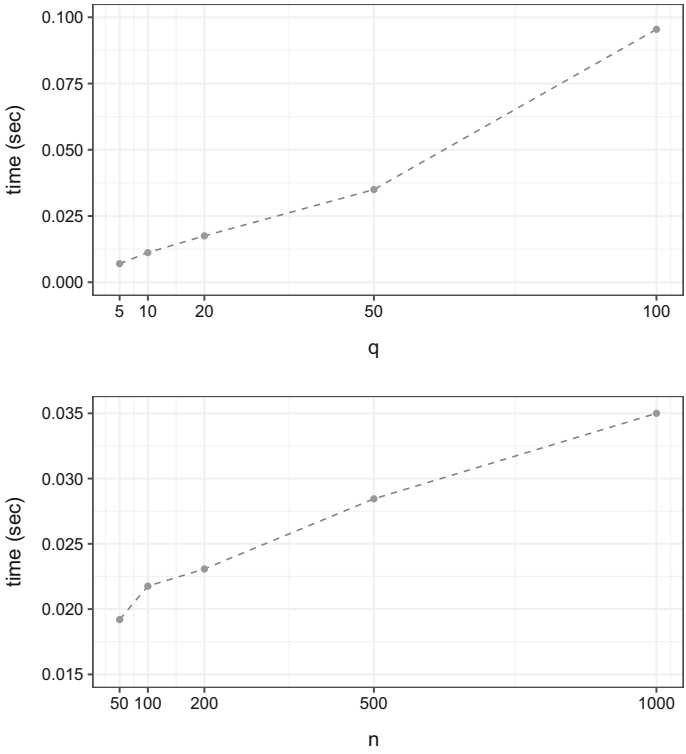

Fig. 6 Computational time (in seconds) *per* iteration, as a function of the number of variables q for fixed $n = 500$ (upper plot) and as a function of the sample size n for fixed $q = 50$ (lower plot), averaged over 12 simulated datasets

However, we remark that our method provides not only a point estimate of DAGs and causal effects, but also an approximation of the whole posterior distribution over the joint space of DAGs and parameters.

6.4 Robustness to model misspecification

To evaluate the performance of our method under non-Gaussian data we implement a simulation study where for a given DAG $\mathcal{D}$ data are generated under a Structural Equation Model of the form

$$X_j = -\boldsymbol{L}_{\prec j]}\boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}}(\mathrm{j})} + \varepsilon_j$$

for $j \in \{1\ldots, q\}$. With regard to the error term distribution, we consider three different scenarios:

1. $\varepsilon_j \overset{iid}{\sim} \mathcal{N}(0,1)$ (Gaussian),
2. $\varepsilon_j \overset{iid}{\sim} \bar{t}_6$ (Student-t)
3. $\varepsilon_j \overset{iid}{\sim} \mathrm{Unif}(-\sqrt{3}, \sqrt{3})$ (Uniform),

where $\bar{t}_6$ denotes a *scaled t* distribution with 6 degrees of freedom. The specific parameter choice in 2-3 guarantees that $\mathbb{V}\mathrm{ar}(\varepsilon_j) = 1$ which is therefore coherent with setting 1. We fix $q = 10$ and, for each sample size $n \in \{50, 100, 200, 500\}$ and scenario 1-2-3, we perform $N = 30$ independent simulations using the same generating DAGs and SEMs used in our previous simulation studies. We apply our method to evaluate the causal effect on X_1 of a joint intervention on $s = 2$ randomly selected nodes among $\{2, \ldots, 10\}$. The resulting estimates are compared with the

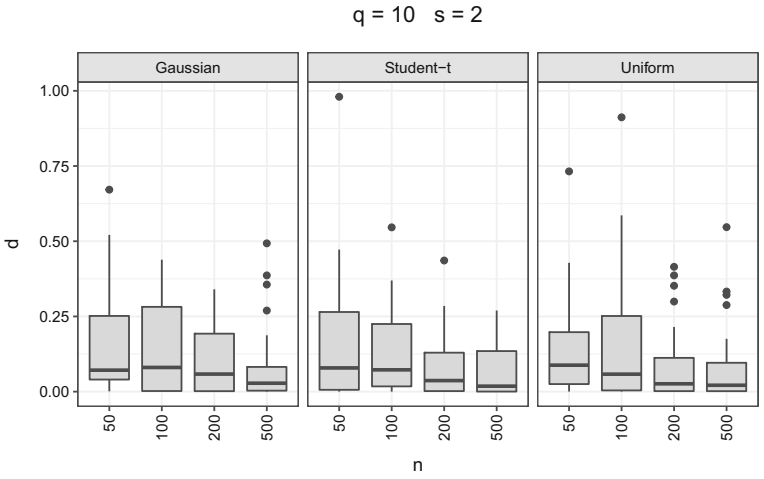

Fig. 7 Simulation study. Distribution of the absolute-value distance d between estimated and true causal effects for Gaussian and non-Gaussian distributed data (Student-t and Uniform), for size of the target $s = 2$, number of variables $q = 10$ and sample size $n \in \{50, 100, 200, 500\}$

true causal effects in terms of the absolute-value distance d as in (28). Our results are summarized in the box-plots of Fig. 7. Results compared across scenarios 1-2-3 are quite similar, showing that our method, when applied to causal effect estimation, is somewhat robust with respect to different distributions of the error term in the SEM model.

7 Real data analysis

In this section we apply our methodology and *joint-IDA* to the "Wine quality" dataset of Cortez et al. (2009); the dataset is publicly available at `https://archive.ics.uci.edu/`. In our analysis we include observations of seven continuous variables measuring physicochemical properties of a Portoguese wine called *Vinho verde*, and a response variable representing a sensory score of the wine quality (ranging in 0–10) given by $n = 1593$ independent assessors.

This dataset has been often used for prediction tasks, i.e. to evaluate the quality of wine on the basis of its physicochemical properties only. However, one might be also interested in causal questions, such as whether intervening on one (or more) physicochemical property may change the wine sensory score. As a consequence, this can lead to identify the target of intervention which produces the largest increase in the score.

We run Algorithm 1 to approximate the posterior distribution of DAGs, DAG-parameters and causal effects for any variable in the system and the *joint-IDA* method based on a CPDAG estimated obtained from PC algorithm. Because one can reasonably assume that the quality score does not affect any of the physicochemical properties (but rather the opposite is argued), we restrict the space of DAGs by imposing that node 1 (the sensory score) cannot have descendant nodes. Such a constraint introduces prior information on the causal structure which is suggested by the concrete problem. In our MCMC algorithm this is achieved by limiting the set of valid operators of type *Insert* involving node 1 to those of the form $u \rightarrow 1$; see also Sect. 5. In the PC algorithm instead, this background information is included with the following procedure: we first estimate the skeleton between variables $X_1, \ldots, X_q$ as in the standard first step of PC. Next, we orient undirected edges between variables Y and covariates $X_2, \ldots, X_q$ as $X_j \rightarrow Y$, while apply Meek's orientation rules to orient the sub-graph of $X_2, \ldots, X_q$; see also Kalish and Buhlmann (2007) for details.

We first assess the convergence of the MCMC algorithm by running two independent chains of length $S = 30000$. Figure 8 summarizes the estimated posterior probabilities of edge inclusion (Equation 26) computed from each MCMC chain. The two resulting heatmaps suggest a highly satisfactory agreement between the two chains.

Starting from our MCMC output we consider both the Maximum a Posteriori (MAP) and the Median Probability Model (MPM) as DAG estimates. However, we stress that our final BMA estimate does not rely on a single DAG but rather on a full posterior of DAGs and accordingly a single DAG estimate is only constructed as an overall graph summary. The two graphs are reported in Fig. 9, together with the

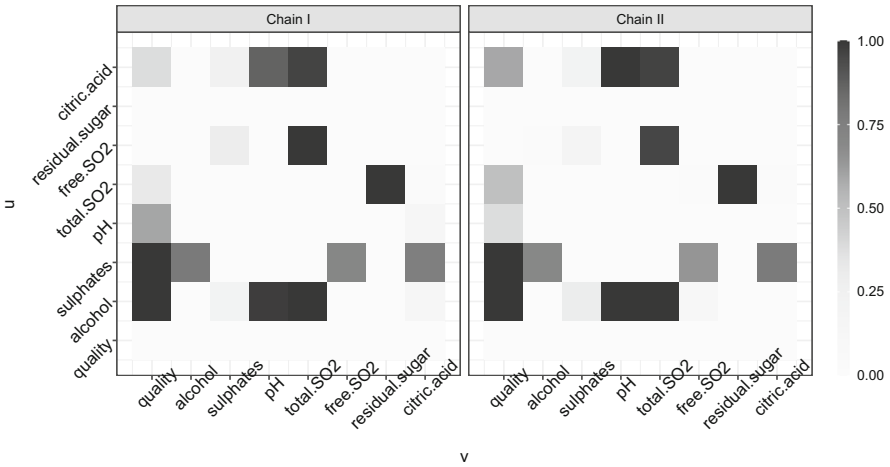

Fig. 8 Real data analysis. Heat maps with estimated posterior probabilities of edge inclusion obtained under two independent MCMC chains

DAG estimate obtained from the modified version of PC (implemented at significance level $\alpha = 0.01$). There are only few differences between the three estimates, the most notable being the presence of an additional edge from *total.SO2* to *quality* in the PC estimate.

We now present our results on causal effect estimation. Specifically, we first consider single-node interventions and compute the BMA and *joint-IDA* estimates of the causal effect on the response for each node (physicochemical property). Moreover, for each pair of nodes, $\{h, k\}$ we obtain the corresponding BMA and *joint-IDA* causal effect estimates under a joint intervention on $\{X_h, X_k\}$. Results are summarized in the left-side heatmaps of Fig. 10. Each (h, k)-element $(h \neq k)$ represents the BMA (upper panel) and *joint-IDA* (lower panel) causal effect estimate of X_k on $Y = X_1$ in a joint intervention on $\{X_h, X_k\}$; main diagonal-elements correspond to the causal effects as obtained from single-node interventions. It appears that an increase in variables *alcohol* and *sulphates* may result in an increase in wine *quality*. By converse, a similar effect can be achieved by reducing the level of *pH* and *total.SO2* since the two covariates exhibit negative causal effects.

The right-side heatmaps of Fig. 10 reports for each pair (h, k) the sum of the corresponding two (absolute value) BMA (upper panel) and *joint-IDA* (lower panel) causal effect estimates obtained under the joint intervention on $\{X_h, X_k\}$, that is $|\hat{\theta}^{\{h,k\}}_{h,1}| + |\hat{\theta}^{\{h,k\}}_{k,1}|$. Each of these terms provides an overall measure of the "strength" of the causal effect that a joint intervention on the two variables might produce on the response. As a consequence, this collection of coefficients allows to identify which pair of variables is associated to the largest potential increase in quality sensory score. In particular, it appears that a joint intervention on variables *alcohol* and *sulphates* has the largest effect on the response variable. This result is invariant with respect to the method used, as it can be observed by comparing the upper and

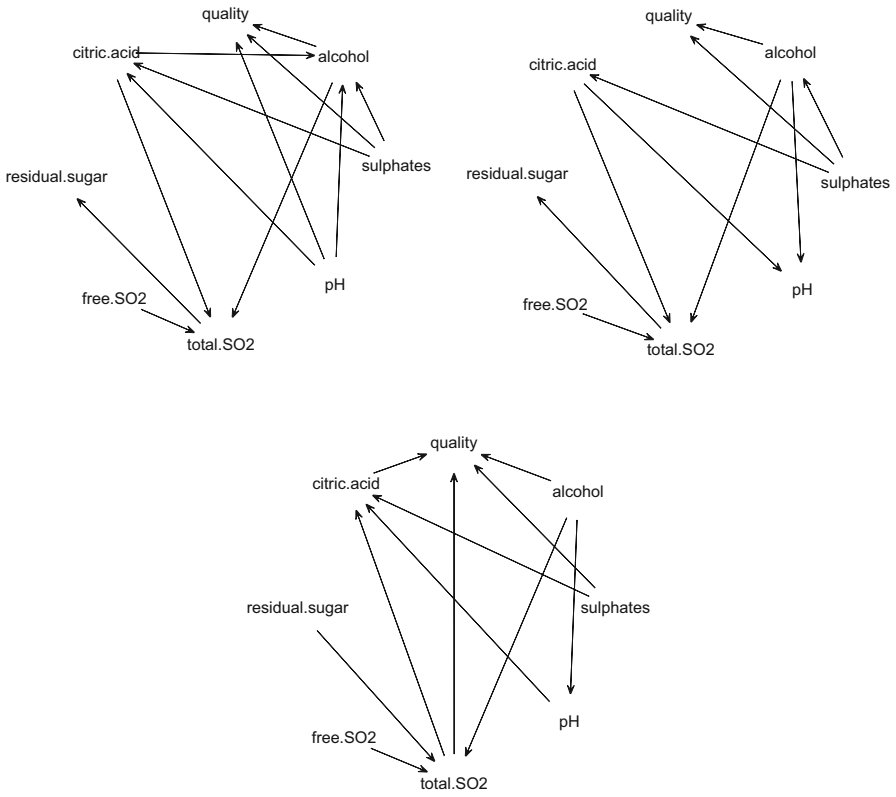

Fig. 9 Real data analysis. Comparison between estimated graphs. From upper-left to bottom: maximum a posteriori, median probability and modified-PC DAG estimates

lower heatmaps of Fig. 10. Substantial differences between the two methods appear, instead, for variable *total.SO2*, which under *joint-IDA* is associated with a (negative) causal effect on *quality*. In addition, *joint-IDA* causal effect estimates are somewhat higher than those obtained under our BMA method. We remark that the effect of joint interventions on more than two variables can be evaluated in a similar way. However, for simplicity of exposition we have limited our analysis to the case of pair-nodes interventions.

8 Discussion

In this paper we present a Bayesian methodology for structural learning of dependence and causal relations among variables. In particular, we assume that multivariate *observational* data have been generated by a Directed Acyclic Graph (DAG) model which is unknown. Of special interest is also the causal effect of a specific variable on a response arising from a joint intervention on several variables in the system. The latter depends on the underlying network structure which

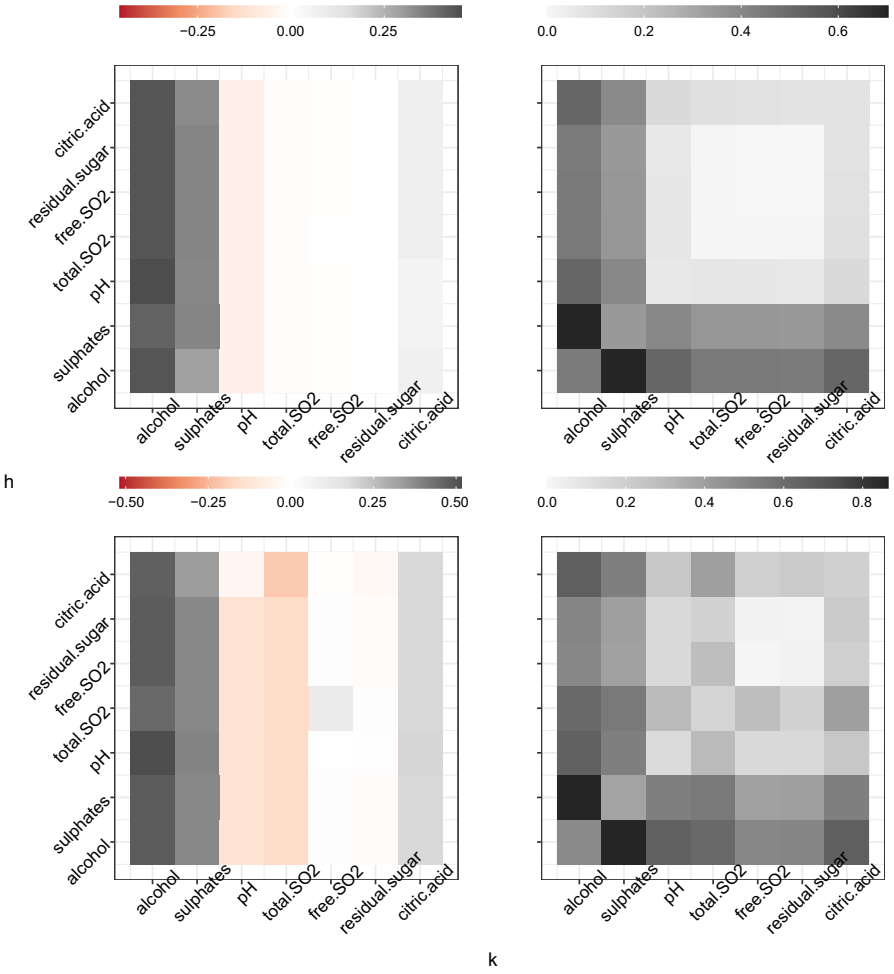

Fig. 10 Real data analysis. Left panel: BMA (*top*) and *joint-IDA* (*bottom*) estimates of the causal effect of X_k on Y in a joint intervention on $\{X_h, X_k\}$. Right panel: sum of absolute-value BMA (*top*) and *joint-IDA* (*bottom*) estimates obtained from joint interventions on $\{X_h, X_k\}$

therefore needs to be estimated. Accordingly, our method combines DAG structural learning and causal effect estimation, leading to a posterior distribution over the space of DAGs, DAG parameters and causal effects. Simulation results show that our method is highly competitive with the frequentist benchmark *joint-IDA* and leads to improved estimates of joint causal effects especially in scenarios characterized by a moderate sample size. On the other hand, our (score-based) methodology requires an approximated posterior distribution over the space of DAGs and parameters, which might become computationally expensive as the number of variables increases. Differently, *joint-IDA* has been specifically developed for high dimensional settings and therefore can efficiently perform even when thousands of variables are involved. However, its output relies on a single

estimated equivalence class of DAGs whose identification may affect the causal estimation results.

Joint interventions lead to causal effects that can significantly deviate from their single-node counterparts. Accordingly, a desired effect on the response can be obtained through a unique intervention involving several variables simultaneously, rather than a sequence of single-node interventions. Since the number of possible joint interventions grows exponentially in the number of variables, the investigation of an optimization strategy which identifies the optimal intervention target producing the desired level of the response could be of interest.

In general, a DAG cannot be uniquely identified from observational data and accordingly a possibly large collection of causal effects is estimated. Randomized intervention experiments producing interventional data can be used to improve the identifiability of the data generating model which consequently reduces the uncertainty around the causal effect estimate; see also Castelletti and Consonni (2020). In principle, one could then perform sequential simultaneous intervention leading to the identification of the true causal effect. This issue can be tackled from an optimal design of experiment perspective implementing an objective function whose optimization reduces the uncertainty related to each BMA causal effect estimate of interest.

In this paper we consider causal effect estimation from joint *hard* interventions. A more general framework, named *soft* interventions, assumes that parent-child dependencies are "modified" but yet preserved after intervention. In this setting, Correa and Bareinboim (2020) introduce a set of rules (named σ-calculus) for the identifiability of causal effects arising from soft interventions. They then show how these rules can be applied to identify the causal effect of an interventions from a combination of observational and *interventional* data, the latter arising from exogenous perturbations of selected variables in the system. A Bayesian framework for causal effect estimation under soft interventions is however still lacking to our knowledge and is currently under investigation by the authors.

Funding Open access funding provided by Università Cattolica del Sacro Cuore within the CRUI-CARE Agreement.

References

Andersson SA, Madigan D, Perlman MD et al (1997) A characterization of Markov equivalence classes for acyclic digraphs. Ann Stat 25(2):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 arxiv:1109.4371v5

Castelletti F, Consonni G (2021) Bayesian causal inference in probit graphical models. Bayesian Anal. Advance publication

Castelletti F, Consonni G (2020) Discovering causal structures in Bayesian Gaussian directed acyclic graph models. J R Stat Soc Ser A Stat Soc 183(4):1727–1745

Castelletti F, Consonni G (2021) Bayesian inference of causal effects from observational data in Gaussian graphical models. Biometrics 77(1):136–149

Chickering DM (2002) Learning equivalence classes of Bayesian-network structures. J Mach Learn Res 2:445–498

Cooper G, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Mach Learn 9:309–347

Correa J, Bareinboim E (2020) A calculus for stochastic interventions: causal effect identification and surrogate experiments. In: Proceedings of the AAAI conference on artificial intelligence 34(06):10093–10100

Cortez P, Teixeira J, Cerdeira A, Almeida F, Matos T, Reis J (2009) Using data mining for wine quality assessment. In: Gama J, Costa VS, Jorge AM, Brazdil PB (eds) Discovery Science. Springer, Berlin, pp 66–79

Ellis B, Wong WH (2008) Learning causal Bayesian network structures from experimental data. J Am Stat Assoc 103(482):778–789

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303:799–805

Friedman N, Koller D (2003) Being Bayesian about network structure. a Bayesian approach to structure discovery in Bayesian networks. Mach Learn 50(2):95–125

García-Donato G, Martínez-Beneito MA (2013) On sampling strategies in Bayesian variable selection problems with large model spaces. J Am Stat Assoc 108(501):340–352

Geiger D, Heckerman D (2002) Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. Annals Stat 30(5):1412–1440

Godsill SJ (2012) On the relationship between Markov chain Monte Carlo methods for model uncertainty. J Comput Graph Stat 10(2):230–248

Heinze-Deml C, Maathuis MH, Meinshausen N (2018) Causal structure learning. Ann Rev Stat Appl 5:371–391

Henckel L, Perković E, Maathuis MH (2019) Graphical criteria for efficient total effect estimation via adjustment in causal linear models. arXiv pre-print https://arxiv.org/abs/1907.02435

Kalish M, Buhlmann P (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J Mach Learn Res 8:613–36

Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford

Maathuis MH, Kalisch M, Buhlmann P (2009) Estimating high-dimensional intervention effects from observational data. Ann Stat 37(6A):3133–3164

Markowetz F, Spang R (2007) Inferring cellular networks—a review. BMC Bioinformatics 8(S5)

Nandy P, Maathuis MH, Richardson TS et al (2017) Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. Ann Stat 45(2):647–674

Ni Y, Stingo FC, Baladandayuthapani V (2017) Sparse multi-dimensional graphical models: a unified Bayesian framework. J Am Stat Assoc 112(518):779–793

Pearl J (2003) Statistics and causal inference: a review. Test 12:281–345

Pearl J (2009) Causality. Cambridge University Press, Cambridge

Peters J, Bühlmann P (2015) Structural intervention distance for evaluating causal graphs. Neural Comput 27(3):771–799

R Core Team (2017) R: A Language and Environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

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

Spirtes P, Glymour CN, Scheines R, Heckerman D (2000) Causation, prediction, and search. MIT Press, Cambridge

Wang H, Li SZ (2012) Efficient Gaussian graphical model determination under G—Wishart prior distributions. Electron J Stat 6:168–198

Yin J, Li H (2011) A sparse conditional Gaussian graphical model for analysis of genetical genomics data. Ann Appl Stat 5(4):2630–2650