Received: 16 October 2019 | Accepted: 30 March 2020

DOI: 10.1111/biom.13281

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC METHODOLOGY

Bayesian inference of causal effects from observational data in Gaussian graphical models

Federico Castelletti | **Guido Consonni**

Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy

Correspondence
Federico Castelletti, Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy.
Email: federico.castelletti@unicatt.it

Funding information
Università Cattolica del Sacro Cuore, Grant/Award Numbers: D.3.2, D1

Abstract

We assume that multivariate observational data are generated from a distribution whose conditional independencies are encoded in a Directed Acyclic Graph (DAG). For any given DAG, the causal effect of a variable onto another one can be evaluated through intervention calculus. A DAG is typically not identifiable from observational data alone. However, its Markov equivalence class (a collection of DAGs) can be estimated from the data. As a consequence, for the same intervention a set of causal effects, one for each DAG in the equivalence class, can be evaluated. In this paper, we propose a fully Bayesian methodology to make inference on the causal effects of any intervention in the system. Main features of our method are: (a) both uncertainty on the equivalence class and the causal effects are jointly modeled; (b) priors on the parameters of the modified Cholesky decomposition of the precision matrices across all DAG models are constructively assigned starting from a unique prior on the complete (unrestricted) DAG; (c) an efficient algorithm to sample from the posterior distribution on graph space is adopted; (d) an objective Bayes approach, requiring virtually no user specification, is used throughout. We demonstrate the merits of our methodology in simulation studies, wherein comparisons with current state-of-the-art procedures turn out to be highly satisfactory. Finally we examine a real data set of gene expressions for *Arabidopsis thaliana*.

KEYWORDS

causal inference, directed acyclic graph, graphical model, Markov equivalence class, objective Bayes, observational data

1 | INTRODUCTION

Discovering causal relationships between variables represents a well-established area in statistical science (Pearl, 2000; Imbens and Rubin, 2015). Applications are ubiquitous, with genomics representing a fast-growing field (Sachs *et al.*, 2005; Pingault *et al.*, 2018). In *knock-out* experiments, one or more genes are made inoperative through some external *intervention*, and the goal is inferring the effect of such an intervention on a response variable of interest. Measurements produced from gene perturbation (called *interventional data*) should be contrasted with *observational data*, which are instead collected from an idle system that has not been subjected to any intervention; see, for instance, Sachs *et al.* (2005). If both observational and interventional data are available, one can establish causal relationships among genes by comparing *pre-* and *postintervention* measurements on the response variable; see, for instance, Pearl (2000) or the more recent paper by Hauser and Bühlmann (2015).

In this work, we address the issue of inferring causal effects in a more challenging setting, namely when *only* observational data are available, a more common and realistic framework in many scientific investigations as intervention experiments are typically very expensive or even unfeasible. In order to tackle the problem effectively, we assume that the conditional independencies inherent in the multivariate *observational distribution* of measurement variables can be read off from a graph (where nodes represent variables) through the *Markov property* associated to the graph; see, for instance, Friedman (2004) and Shojaie and Michailidis (2009). In particular, we use Directed Acyclic Graphs (DAGs) that are especially suited for causal reasoning based on the notion of *interventional distribution* (Pearl, 2000).

In the following, we assume that the observations are generated according to a Gaussian distribution satisfying the factorization inherited from a DAG, also called Markov property; see Lauritzen (1996). In applications the assumed generating DAG is unknown and thus needs to be estimated. A difficulty we face is that the true generating DAG is not identifiable from observational data because its conditional independencies can be encoded in different DAGs that can be collected into a (Markov) *equivalence class*. For each equivalence class, there also exists a unique representative chain graph, called Essential Graph (EG; Andersson *et al.* 1997) or Completed Partially Directed Acyclic Graph (CPDAG: Chickering 2002). Given that observational data can only identify a (potentially large) equivalence class of Gaussian DAGs, and that the causal effect on Y of an intervention on X is tied to each specific DAG (see Section 2), the best we can hope for is to identify a *collection* of causal effects (one for each DAG in the equivalence class). Note that identifiability can be guaranteed under additional assumptions within the Gaussian model (Peters and Bühlmann, 2014) or under alternative data-generating models; see, for instance, Mahmoudi and Wit (2018), Hoyer *et al.* (2009), and Shimizu *et al.* (2006).

An important contribution to the estimation of causal effects from observational data when the generating DAG is unknown is contained in Maathuis *et al.* (2009), which combines graphical model selection and causal effect estimation in high-dimensional settings. Specifically, they first learn an equivalence class of DAGs using the PC algorithm (alternatively, any other score-based method can be used). Next, they propose two different strategies to estimate the causal effect of a variable on a response. The first one enumerates all DAGs in the equivalence class and for each one estimates the causal effect. As this first strategy is computationally expensive, a more efficient algorithm that returns only the *distinct* causal effects within a given equivalence class is implemented. However, both approaches are predicated on the choice of a single EG estimate, with no associated measure of uncertainty. We improve on this by presenting a Bayesian method that combines structural learning of the equivalence class as well as inference on causal effects. In both cases, a fully Bayes posterior distribution is obtained using an objective approach requiring virtually no prior elicitation from the user.

The rest of the paper is organized as follows. In Section 2, we review relevant notation for graphical models, in particular DAGs, together with the notions of interventional distribution and causal effect. Section 4 presents our methodology for Bayesian inference on causal effects. This requires inference on a covariance matrix Markov with respect to a DAG that is pursued adopting an objective Bayesian prior-to-posterior analysis on the modified Cholesky parameterizaton (Section 3). We then propose two different Bayesian strategies to obtain a summary estimate of the causal effect. We evaluate the performance of our proposal through simulation scenarios in Section 5, while Section 6 applies our method to the analysis of gene expression data for *Arabidopsis thaliana*. Finally, Section 7 offers a brief discussion together with possible future developments.

2 | DAGS, INTERVENTIONS, AND CAUSAL EFFECTS

Let $\mathcal{D} = (V, E)$ be a DAG, where $V = \{1, \dots, q\}$ denotes a set of nodes (or vertices) and $E = V \times V$ a set of directed edges, so that if $(u, v) \in E$, then $(v, u) \notin E$. For a given DAG $\mathcal{D}$, if there is an edge $u \to v$ we say that u is a *parent* of v (conversely, v is a child of u) and denote the parent set of v in $\mathcal{D}$ as $\mathrm{pa}_{\mathcal{D}}(v)$. The *family* of v in $\mathcal{D}$ is $\mathrm{fa}_{\mathcal{D}}(v) = v \cup \mathrm{pa}_{\mathcal{D}}(v)$. Further notions and details may be found, for instance, in the book of Lauritzen (1996).

Consider a collection of q random variables $(X_1, \dots, X_q)$; for convenience we will sometimes use the convention $X_1 = Y$, where Y denotes the response. Given a DAG $\mathcal{D} = (V, E)$, we associate each variable to a node in V. To simplify the notation, for the remainder of the section we consider a fixed DAG and omit the subscript $\mathcal{D}$ (eg, by writing $\mathrm{pa}(v)$ instead of $\mathrm{pa}_{\mathcal{D}}(v)$). Let the joint probability density function of $(X_1, \dots, X_q)$ be denoted by $f(\cdot)$. We assume that $f(\cdot)$ obeys the Markov property of the DAG so that it factorizes according to

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

We refer to (1) as the *observational* (or *preintervention*) distribution.

To define formally a causal effect, we require the notion of intervention and the allied *do-operator* (Pearl, 2000). Let $\mathrm{do}(X_i = \tilde{x}_i)$ denote a (deterministic) intervention consisting in setting X_i to the value $\tilde{x}_i$. The *postintervention* distribution is obtained using the truncated factorization, namely

$$\begin{aligned} f(x_1, \dots, x_q \mid \mathrm{do}(X_i = \tilde{x}_i)) \\ = \begin{cases} \prod_{j=1, j \neq i}^{q} f(x_j \mid \boldsymbol{x}_{\mathrm{pa}(j)})|_{x_i = \tilde{x}_i} & \text{if } x_i = \tilde{x}_i, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

Note that the $f(x_j \mid \cdot)$-terms appearing in (2) are precisely the (preintervention) conditional distributions of Equation (1). We can interpret this as a form of "stability" of the data-generating mechanism with respect to an intervention. The postintervention distribution of Y is then obtained by integrating (2) with respect to $x_2, \dots, x_q$,

$$f(y \mid \mathrm{do}(X_i = \tilde{x}_i)) = \int f(y \mid \tilde{x}_i, \boldsymbol{x}_{\mathrm{pa}(i)}) f(\boldsymbol{x}_{\mathrm{pa}(i)})\, d\boldsymbol{x}_{\mathrm{pa}(i)}; \quad (3)$$

see Pearl (2000, Theorem 3.2.2). It is common to summarize the postintervention distribution (3) through its expected value, and to define the *causal effect* of $\mathrm{do}(X_i = \tilde{x}_i)$ on Y, which we denote by γ_i, as

$$\frac{\partial}{\partial x} \mathbb{E}\big(Y \mid \mathrm{do}(X_i = x)\big)\big|_{x=\tilde{x}_i}; \quad (4)$$

see Maathuis *et al.* (2009). Consider now the case in which $(X_1, \dots, X_q)^\top = \boldsymbol{x}$ are distributed according to a Gaussian DAG-model,

$$\boldsymbol{x} \mid \boldsymbol{\Sigma} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} \in \mathcal{C}_{\mathcal{D}}$, the space of symmetric positive definite (s.p.d.) covariance matrices Markov with respect to $\mathcal{D}$. For later purposes, we will also use the parameterization $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ with $\boldsymbol{\Omega} \in \mathcal{P}_{\mathcal{D}}$, the space of s.p.d. precision matrices Markov with respect to $\mathcal{D}$. For a Gaussian DAG-model, the factorization (1) becomes

$$f(x_1, \dots, x_q \mid \boldsymbol{\Sigma}) = \prod_{j=1}^{q} f(x_j \mid \boldsymbol{x}_{\mathrm{pa}(j)}, \boldsymbol{\Sigma})$$

and the associated marginal postintervention distribution of Y can be written as

$$\begin{aligned} &f(y \mid \mathrm{do}(X_i = \tilde{x}_i), \boldsymbol{\Sigma}) \\ &\quad = \int f(y \mid \tilde{x}_i, \boldsymbol{x}_{\mathrm{pa}(i)}, \boldsymbol{\Sigma}) f(\boldsymbol{x}_{\mathrm{pa}(i)} \mid \boldsymbol{\Sigma})\, d\boldsymbol{x}_{\mathrm{pa}(i)}. \end{aligned}$$

Moreover, because of the normality assumption, the expectation of Y conditionally on $(\tilde{x}_i, \boldsymbol{x}_{\mathrm{pa}(i)})$ is

$$\mathbb{E}\big(Y \mid \tilde{x}_i, \boldsymbol{x}_{\mathrm{pa}(i)}, \boldsymbol{\Sigma}\big) = \gamma_i \tilde{x}_i + \boldsymbol{\gamma}_{\mathrm{pa}(i)}^\top \boldsymbol{x}_{\mathrm{pa}(i)}, \quad (5)$$

whence

$$\mathbb{E}(Y \mid \mathrm{do}(X_i = \tilde{x}_i), \boldsymbol{\Sigma}) = \gamma_i \tilde{x}_i + \int \boldsymbol{\gamma}_{\mathrm{pa}(i)}^\top \boldsymbol{x}_{\mathrm{pa}(i)}\, d\boldsymbol{x}_{\mathrm{pa}(i)}, \quad (6)$$

so that, applying (4), the causal effect of $\mathrm{do}(X_i = \tilde{x}_i)$ on Y is γ_i. It follows that γ_i corresponds to the coefficient of X_i in the conditional expectation of Y on $\boldsymbol{x}_{\mathrm{fa}(i)}$. If instead $Y \in \mathrm{pa}(i)$, then the causal effect of X_i on Y is zero by definition.

The relationship between γ_i and $\boldsymbol{\Sigma}$ can be easily explicated as follows. Let $\boldsymbol{\Sigma}_{Y,\mathrm{fa}(i)}$ be the marginal covariance matrix of $(Y, \boldsymbol{x}_{\mathrm{fa}(i)})$ that we partition as

$$\big[\boldsymbol{\Sigma}_{Y,\mathrm{fa}(i)}\big] = \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{Y,\mathrm{fa}(i)} \\ \boldsymbol{\Sigma}_{\mathrm{fa}(i),Y} & \boldsymbol{\Sigma}_{\mathrm{fa}(i),\mathrm{fa}(i)} \end{bmatrix}.$$

Then, the conditional expectation of Y is

$$\mathbb{E}\big(Y \mid \boldsymbol{x}_{\mathrm{fa}(i)}, \boldsymbol{\Sigma}\big) = \big[\boldsymbol{\Sigma}_{Y,\mathrm{fa}(i)}\big] \big(\boldsymbol{\Sigma}_{\mathrm{fa}(i),\mathrm{fa}(i)}\big)^{-1} \boldsymbol{x}_{\mathrm{fa}(i)}.$$

Recalling now that $\mathrm{fa}(i) = i \cup \mathrm{pa}(i)$, we obtain

$$\gamma_i = \Big[\big[\boldsymbol{\Sigma}_{Y,\mathrm{fa}(i)}\big] \big(\boldsymbol{\Sigma}_{\mathrm{fa}(i),\mathrm{fa}(i)}\big)^{-1}\Big]_1 \quad (7)$$

where subscript 1 refers to the first element of the vector.

3 | BAYESIAN INFERENCE OF DAG MODEL PARAMETERS

Recall from the previous section that the causal effect as defined in Equation (4) is a function of the covariance matrix $\boldsymbol{\Sigma}$, which is Markov with respect to the underlying DAG $\mathcal{D}$. Accordingly, we will proceed by making inference on $\boldsymbol{\Sigma}$ and then derive inference on γ_i. Under a Bayesian framework, one requires a suitable prior for $\boldsymbol{\Sigma}$ that we construct based on the Cholesky parameterization of a DAG model. This parameterization is also used in Ni *et al.* (2017) who provide a unified framework for the analysis of different types of graphical models (undirected, directed, and hybrid).

3.1 | Cholesky parameterizations and DAG-Wishart priors

Consider a DAG $\mathcal{D} = (V, E)$ and assume a *parent ordering* of its vertices. This means that if there is an arrow $u \to v$, equivalently $(u, v) \in E$, then $u > v$. A parent ordering always exists but it is not unique in general. We assume that $\boldsymbol{x} \mid \boldsymbol{\Sigma} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathcal{C}_{\mathcal{D}}$, the space of s.p.d. covariance matrices Markov with respect to $\mathcal{D}$ (again omitting the subscript $\mathcal{D}$ for simplicity). An alternative parameterization of the Gaussian DAG model can be obtained through the corresponding *structural equation model*

$$\boldsymbol{L}^\top \boldsymbol{x} = \boldsymbol{\varepsilon}, \quad (8)$$

where $\boldsymbol{L}$ is a (q, q) lower triangular matrix of coefficients, $\boldsymbol{L} = \{\boldsymbol{L}_{ij}, i \geq j\}$, such that $\boldsymbol{L}_{ij} \neq 0$ if and only if $i \to j$ and $\boldsymbol{L}_{ii} = 1$. Moreover, $\boldsymbol{\varepsilon}$ is a $(q, 1)$ vector of error terms, $\boldsymbol{\varepsilon} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{\sigma}^2)$ and $\boldsymbol{\sigma}^2$ is the $(q, 1)$ vector of *conditional* variances whose jth element is $\sigma_j^2 = \mathbb{V}\mathrm{ar}(X_j \mid \boldsymbol{x}_{\mathrm{pa}(j)}, \boldsymbol{\Sigma}) = \boldsymbol{\Sigma}_{jj \mid \mathrm{pa}(j)}$. It follows that $\boldsymbol{\Sigma} = \boldsymbol{L}^{-\top} \boldsymbol{D} \boldsymbol{L}^{-1}$, where $A^{-\top} = (A^{-1})^\top$. Alternatively, if we let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ be the

(q, q) precision matrix, we obtain

$$\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^{\top}. \tag{9}$$

We refer to Equation (9) as the *modified Cholesky decomposition* of $\boldsymbol{\Omega}$. Let now $\prec j \succ = \mathrm{pa}(j)$ and $\prec j\,] = \mathrm{pa}(j) \times j$. The DAG Cholesky parameterization of $\boldsymbol{\Omega}$ is given by the node-parameters $\boldsymbol{\theta}_j = (\boldsymbol{D}_{jj}, \boldsymbol{L}_{\prec j\,]})$, $j = 1, \dots, q$, where

$$\boldsymbol{D}_{jj} = \boldsymbol{\Sigma}_{jj\,|\,\mathrm{pa}(j)}, \quad \boldsymbol{L}_{\prec j\,]} = -\boldsymbol{\Sigma}_{\prec j \succ}\boldsymbol{\Sigma}_{\prec j\,]}$$

and $\boldsymbol{\Sigma}_{\prec j \succ}$ is the sub-matrix of $\boldsymbol{\Sigma}$, whose rows and columns refer to $\mathrm{pa}(j)$. A general *DAG-Wishart* distribution on $(\boldsymbol{D}, \boldsymbol{L})$ has density

$$p(\boldsymbol{D}, \boldsymbol{L}) \propto \exp\left[-\frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^{\top})\boldsymbol{U}\right\}\right] \prod_{j=1}^{q} \boldsymbol{D}_{jj}^{-\frac{a_j}{2}}, \tag{10}$$

with $\boldsymbol{U}$ a s.p.d. matrix and $a_j > |\mathrm{pa}(j)| + 2$; see Ben-David *et al.* (2015) and Cao *et al.* (2019). A more useful representation of (10) is given in terms of the marginal and conditional distributions of $\boldsymbol{D}_{jj}$ and $\boldsymbol{L}_{\prec j\,]}$, respectively,

$$\begin{aligned} \boldsymbol{D}_{jj} &\sim \text{I-Ga}\left(\frac{a_j}{2} - \frac{|\mathrm{pa}(j)|}{2} - 1, \frac{1}{2}\boldsymbol{U}_{jj|\prec j \succ}\right), \\ \boldsymbol{L}_{\prec j\,]}|\boldsymbol{D}_{jj} &\sim \mathcal{N}_{|\mathrm{pa}(j)|}\left(-\boldsymbol{U}_{\prec j \succ}^{-1}\boldsymbol{U}_{\prec j\,]}, \boldsymbol{D}_{jj}\boldsymbol{U}_{\prec j \succ}^{-1}\right), \end{aligned} \tag{11}$$

where I-Ga(a, b) stands for an Inverse-Gamma distribution with shape $a > 0$ and rate $b > 0$ having expectation $b/(a-1)$ $(a > 1)$. Assume now that the DAG is *complete*, that is, all pairs of nodes are joined by an edge. The corresponding Markov distribution is any *unrestricted* joint distribution on the q variables. In the Gaussian setting, this means that the precision matrix $\boldsymbol{\Omega}$ is unconstrained, that is $\boldsymbol{\Omega} \in \mathcal{P}$, the space of all s.p.d precision matrices. If $\boldsymbol{\Omega}$ is assigned a standard Wishart prior, written $\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{U})$, with density

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

where $a > q - 1$ and $\boldsymbol{U}$ a s.p.d. matrix, the *induced* distribution on $(\boldsymbol{L}, \boldsymbol{D})$ becomes a special case of (11), which we name a *complete DAG-Wishart*, wherein $a_j = a + q - 2j + 3$. This result appears in Ben-David *et al.* (2015, Formula (6)); note, however, that their Jacobian is not correctly written; it should be $\prod_{j=1}^{q} \boldsymbol{D}_{jj}^{-(q-j+2)}$, where we have replaced their p (the number of variables) with our symbol q.

3.2 | Parameter priors under model uncertainty

When several models are entertained (model uncertainty), the construction of priors on the parameters of each model requires specific care; see Consonni *et al.* (2018) for a recent review. The key issue is *compatibility* of priors across models, a notion that is especially critical for DAG-graphical models, as we clarify below. Geiger and Heckerman (2002) (G&H) propose a method to construct parameter priors for the comparison of DAG-models that ensures identical marginal likelihoods for DAGs belonging to the same equivalence class: a basic compatibility requirement because, under the Gaussian assumption, DAGs within the same equivalence class are not distinguishable using observational data alone (more on this in Section 4). Their method assumes some regularity conditions on the likelihood (*complete model equivalence, regularity, likelihood modularity*) that are satisfied by any Gaussian model. The distinctive feature of their approach regards, however, the construction of the prior. This can be split into two steps. The first one (*prior modularity*) states that, given two distinct DAG models with the *same* set of parents for vertex j, the prior for the node-parameter $\boldsymbol{\theta}_j$ must be the same under both models, namely

$$p(\boldsymbol{\theta}_j \,|\, \mathcal{D}_h) = p(\boldsymbol{\theta}_j \,|\, \mathcal{D}_k)$$

for any pair of distinct DAGs $\mathcal{D}_h$ and $\mathcal{D}_k$ such that $\mathrm{pa}_{\mathcal{D}_h}(j) = \mathrm{pa}_{\mathcal{D}_k}(j)$. The second one (*global parameter independence*) states that for every DAG model $\mathcal{D}$, the parameters $\{\boldsymbol{\theta}_j;\, j = 1, \dots, q\}$ should be a priori independent, that is

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \prod_{j=1}^{q} p(\boldsymbol{\theta}_j \,|\, \mathcal{D}).$$

If one follows the above path, it can be shown that *all* parameter priors are completely determined by a *unique* prior on the parameter of *any* of the (equivalent) complete DAGs (Geiger and Heckerman, 2002, Theorem 1); additionally all DAG-models within the same equivalence class will be scored equally (same marginal likelihood); see Theorem 3 in G&H. Finally, with regard to zero-mean Gaussian DAG models for a set of variables of dimension at least three, the Wishart distribution on the unconstrained precision matrix of a complete DAG-model is characterized as the *only* prior that guarantees global parameter independence (section 5 of G&H). The bottom line is that the construction of all priors across DAG-models is driven by a *single* Wishart distribution on an unconstrained precision matrix.

We follow the procedure of G&H to induce a prior on the Cholesky parameterization of $\boldsymbol{\Omega} \in \mathcal{P}_{\mathcal{D}}$, namely $(\boldsymbol{D}, \boldsymbol{L})$, starting from a prior on the unrestricted precision matrix for a complete DAG. Specifically, let $\mathcal{D}$ be an arbitrary DAG and assume a parent ordering of its nodes. For each node $j \in \{1, \dots, q\}$, let $\{\boldsymbol{D}_{jj}, \boldsymbol{L}_{\prec j\,]}\}$ be the Cholesky parameters associated to node j, and identify a complete DAG $\mathcal{D}^{C(j)}$ such that $\mathrm{pa}_{\mathcal{D}^{C(j)}}(j) = \mathrm{pa}_{\mathcal{D}}(j)$. Let $\{\boldsymbol{D}_{jj}^{C(j)}, \boldsymbol{L}_{\prec j\,]}^{C(j)}\}$ be the Cholesky parameters of node j under the complete DAG $\mathcal{D}^{C(j)}$. We then assign to $\{\boldsymbol{D}_{jj}, \boldsymbol{L}_{\prec j\,]}\}$ the same prior of $\{\boldsymbol{D}^{C(j)}_{jj}, \boldsymbol{L}^{C(j)}_{\prec j\,]}\}$ that can be gathered from Equation (11) in the complete DAG-Wishart version.

If interest centers on obtaining the posterior on the DAG Cholesky parameters $(\boldsymbol{D}, \boldsymbol{L})$, as opposed to computing marginal likelihoods of DAG models, it is rather expedient to compute first the posterior on the unconstrained $\boldsymbol{\Omega}$, which by conjugacy is still Wishart, and then recover, through the procedure of G&H, the posterior on $(\boldsymbol{D}, \boldsymbol{L})$. This is the method we follow in the next subsection.

3.3 | Objective Bayes analysis

Because of our discussion in Section 3.2, to assign a prior on the precision matrix under any DAG $\mathcal{D}$ we only require a single Wishart prior on an unconstrained precision matrix. Consider a random sample of size n of multivariate observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, where $\boldsymbol{x}_i = (x_{i,1}, \dots, x_{i,q})^\top$, and $\boldsymbol{x}_i \mid \boldsymbol{\Omega} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$, $i = 1, \dots, n$, where $\boldsymbol{\Omega}$ is unconstrained. Let the (n, q) data matrix $\boldsymbol{X}$ be obtained by row-binding the individual $\boldsymbol{x}_i^\top$'s. In the absence of substantive prior information, we assume a default noninformative (N) prior

$$p^N(\boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{-1}; \tag{12}$$

see, for instance, Press (1982). The sampling density of $\boldsymbol{X}$, equivalently the likelihood function for $\boldsymbol{\Omega}$, is

$$f(\boldsymbol{X} \mid \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Omega}\, \boldsymbol{X}^\top \boldsymbol{X}\right)\right\};$$

whence

$$\boldsymbol{\Omega} \mid \boldsymbol{X} \sim \mathcal{W}_q\left(n + q - 1, \boldsymbol{X}^\top \boldsymbol{X}\right). \tag{13}$$

Equation (13) induces a complete DAG Wishart posterior on $(\boldsymbol{D}, \boldsymbol{L})$—Equation (11)—with $a_j = n + 2q - 2j + 2$, which can be easily sampled from. Next one can generate posterior draws of the Cholesky parameters of $\boldsymbol{\Sigma} \in \mathcal{C}_\mathcal{D}$, $(\boldsymbol{D}_{jj}, \boldsymbol{L}_{\prec j\,]})$, $j = 1, \dots, q$, which provide posterior draws from $\boldsymbol{\Sigma} = \boldsymbol{L}^{-\top}\boldsymbol{D}\boldsymbol{L}^{-1}$. From the latter, one can finally obtain posterior draws of the causal effect parameter γ_i for any *given* DAG $\mathcal{D}$, using (7).

4 | BAYESIAN INFERENCE ON CAUSAL EFFECTS

Recall from Section 1 that Markov equivalent DAGs can be collected into equivalence classes, each one being represented by an EG (Andersson *et al.*, 1997). Bayesian methods for EG model selection generate a posterior distribution over the EG space, which is generally approximated *via* Markov chain Monte Carlo (MCMC) techniques because an exhaustive exploration of the EG space is not feasible even in moderately sized problems (Gillispie and Perlman, 2002). As an EG $\mathcal{G}$ collects together a *set* of DAGs, *several* distinct causal effects of X_i on Y are typically associated with $\mathcal{G}$, because the causal effect γ_i is DAG-dependent through the set $\mathrm{fa}_\mathcal{D}(i)$; see Equation (4). When making inference on causal effects using observational data we are thus confronted with two sources of uncertainty: (a) one concerns the *structure* (EG) of the data-generating mechanism; (b) the other refers to the *size* of the actual causal effect, as distinct causal effects are consistent with any given EG.

4.1 | EG model selection

Issue (a) is a structural learning problem that we address using the objective Bayes (OB) methodology developed in Castelletti *et al.* (2018). The core of their method consists in an MCMC algorithm that approximates the posterior distribution over the EG space. Specifically, starting from a chain-graph representation of an EG $\mathcal{G}$ (Andersson *et al.*, 1997), they derive a closed-form expression for the marginal likelihood $m_\mathcal{G}(\boldsymbol{X})$ of $\mathcal{G}$. This is based on a noninformative prior coupled with a fractional Bayes factor methodology (O'Hagan, 1995) and the G&H method described in Section 3.2. Let $\mathcal{S}_q$ be the set of all EGs on q nodes. They first assign a prior $p(\mathcal{G})$ to $\mathcal{G}, \mathcal{G} \in \mathcal{S}_q$, by imposing a Bernoulli-beta distribution independently to each element of the adjacency matrix of the skeleton (underlying undirected graph) of $\mathcal{G}$, denoted by $\mathcal{G}^u$,

$$\begin{aligned}\mathcal{G}^u_{(j)} \mid \pi &\sim \mathrm{Ber}(\pi), \quad j = 1, \dots, \frac{q(q-1)}{2},\\ \pi &\sim \mathrm{Beta}(a, b),\end{aligned} \tag{14}$$

where $\mathcal{G}^u_{(j)}$ is the jth element of the vectorized lower triangular part of the adjacency matrix of $\mathcal{G}^u$ and $q(q-1)/2$ corresponds to the maximum number of edges in a graph on q nodes. Different choices of the hyperparameters a and b are possible and allow to regulate the prior probability of inclusion of each edge; see the original paper for details. Therefore, the posterior distribution of $\mathcal{G}$ given the data $\boldsymbol{X}$ is obtained as

$$p(\mathcal{G} \mid \boldsymbol{X}) = \frac{m_\mathcal{G}(\boldsymbol{X})p(\mathcal{G})}{\sum_{\mathcal{G} \in \mathcal{S}_q} m_\mathcal{G}(\boldsymbol{X})p(\mathcal{G})} \tag{15}$$

and is approximated *via* MCMC methods. The same output can be used to approximate any feature thereof, such as the probability of inclusion of a specific edge that plays a crucial role in the first strategy for causal effect estimation proposed above. In addition, OBES can be adapted to account for *sparsity* in the EG space by limiting the model space to those graphs having fewer edges than a specified threshold M. This represents a reasonable condition if some knowledge of sparsity is available, like in gene regulatory networks. In addition, such a constraint can ease computations especially in a high-dimensional setting (large q) that is, however, far from our context. Therefore, we can easily relax it, without paying substantial computational costs.

4.2 | Causal effect estimation

We now address issue (b), namely the evaluation of the causal effect of X_i on Y. Let $\mathcal{G}$ be an EG. To simplify our notation, we also denote with $\mathcal{G}$ the corresponding Markov equivalence class. In principle, each DAG $\mathcal{D} \in \mathcal{G}$ returns a distinct causal effect $\gamma_i^{\mathcal{D}}$, which we write as $\gamma^{\mathcal{D}}$ to streamline notation. One option would be to enumerate all DAGs within the equivalence class $\mathcal{G}$ and then estimate each $\gamma^{\mathcal{D}}$. However, such enumeration can be quite demanding and possibly unfeasible even for small problems with $q > 8$. A solution is offered by noticing that different DAGs within the same equivalence class may lead to the *same* causal effect. This happens for any two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ such that $\mathrm{pa}_{\mathcal{D}_1}(i) = \mathrm{pa}_{\mathcal{D}_2}(i)$ because then $\mathrm{fa}_{\mathcal{D}_1}(i) = \mathrm{fa}_{\mathcal{D}_2}(i)$ too; see Equation (7). Maathuis *et al.* (2009) propose a fast and localized algorithm (Algorithm 3 in their original paper) to estimate causal effects starting from an (estimated) equivalence class without a full enumeration. An intermediate output of their algorithm that is of special interest to us identifies all the distinct sets of parents of X_i compatible with the given essential graph.

We now present two alternative proposals for the Bayesian estimation of causal effects building on the methodology developed in Section 3 as well as in the current section. Given an equivalence class of DAGs represented by the EG $\mathcal{G}$, let $\{\gamma_l(\mathcal{G});\, l = 1, \dots, L_{\mathcal{G}}\}$ denote the collection of $L_{\mathcal{G}}$ distinct causal effects of X_i on Y (again dropping subscript i for simplicity). A "natural" choice for an overall measure of causal effect conditionally on $\mathcal{G}$ is given by the *average conditional causal effect*

$$\gamma_{\mathrm{avg}}(\mathcal{G}) = \frac{1}{L_{\mathcal{G}}} \sum_{l=1}^{L_{\mathcal{G}}} \gamma_l(\mathcal{G}). \tag{16}$$

We now further elaborate on the rationale behind (16). Let $\{\mathcal{D}_l(\mathcal{G});\, l = 1, \dots, L_{\mathcal{G}}\}$ denote the set of distinct DAGs in the equivalence class represented by $\mathcal{G}$. The parameter $\gamma_l(\mathcal{G})$ associated to $\mathcal{D}_l(\mathcal{G})$ is a "regression" coefficient in the *observational* conditional expectation of Y in Equation (5), where the expectation is implicitly conditioned on $\mathcal{D}_l(\mathcal{G})$. For two distinct DAGs $\mathcal{D}_l(\mathcal{G})$ and $\mathcal{D}_{l'}(\mathcal{G})$ in the same equivalence class, the meaning of $\gamma_l(\mathcal{G})$ and $\gamma_{l'}(\mathcal{G})$ would be *different* because the set of regressors $\{x_i, \boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}_l}(i)}\}$ can be different from $\{x_i, \boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}_{l'}}(i)}\}$. As a consequence, in the observational setting of (5), the average in (16) could be questioned because it takes averages of intrinsically distinct quantities. On the other hand, $\gamma_l(\mathcal{G})$ is also the coefficient associated to X_i in the *interventional* expectation in (6), so that it represents the difference in the (interventional) expectation of Y between $\mathrm{do}(X_i = \widetilde{x}_i + 1)$ and $\mathrm{do}(X_i = \widetilde{x}_i)$. Although the estimate of this quantity will depend on $\mathcal{D}_l(\mathcal{G})$, its meaning does not, because $\mathcal{D}_l(\mathcal{G})$ enters through the parent set variables $\boldsymbol{x}_{\mathrm{pa}_{\mathcal{D}_l}(i)}$ that are integrated out in (6). This justifies taking a summary measure of the various $\gamma_l(\mathcal{G})$'s as, for instance, the average. An *estimate* of $\gamma_{\mathrm{avg}}(\mathcal{G})$ is provided by the corresponding conditional expectation

$$\overline{\gamma}_{\mathrm{avg}}(\mathcal{G}; \boldsymbol{X}) = \frac{1}{L_{\mathcal{G}}} \sum_{l=1}^{L_{\mathcal{G}}} \mathbb{E}\{\gamma_l(\mathcal{G}) \mid \boldsymbol{X}, \mathcal{G}\}. \tag{17}$$

Clearly, a frequentist analogue of (17) could be computed upon replacing $\mathbb{E}\{\gamma_l(\mathcal{G}) \mid \boldsymbol{X}, \mathcal{G}\}$ with a suitable estimate $\widehat{\gamma}_l(\mathcal{G}; \boldsymbol{X})$. Note that in our Bayesian setting $\overline{\gamma}_{\mathrm{avg}}(\mathcal{G}; \boldsymbol{X})$ is readily available using draws from the posterior distribution of each $\gamma_l(\mathcal{G})$, as described in Section 3.

The estimate $\overline{\gamma}_{\mathrm{avg}}(\mathcal{G}; \boldsymbol{X})$ is predicated on an unknown EG $\mathcal{G}$. A frequentist approach solves the problem by estimating $\mathcal{G}$ upfront, as in Maathuis *et al.* (2009) who employ the sample version of the PC-algorithm (Spirtes *et al.*, 2000; Kalisch and Bühlmann, 2007). This has the disadvantage of disregarding *model uncertainty*, which instead is fully accounted for in our Bayesian setting through the posterior distribution on the space of the essential graphs. We propose to employ the latter in two distinct ways, which in turn give rise to two strategies for causal effect estimation.

The first one relies on a single summary of the posterior distribution on graph space. This is achieved by constructing a graph that includes only those edges whose marginal posterior probability of inclusion exceeds a given threshold. A popular choice for such threshold is 0.5, as in the original median probability model of Barbieri and Berger (2004). With a suitable adaptation, this leads to the *projected median probability graph model* (Castelletti **et al.**, 2018). An interesting alternative for fixing the threshold relies on the Bayesian False Discovery Rate (FDR; Müller *et al.* 2007; Peterson *et al.* 2015). Either way we denote with $\mathcal{G}^*$ the resulting graph. Given $\mathcal{G}^*$, we then construct the corresponding set of distinct causal effects $\{\gamma_1(\mathcal{G}^*), \dots, \gamma_{L_{\mathcal{G}^*}}(\mathcal{G}^*)\}$, and obtain $\overline{\gamma}_{\mathrm{avg}}(\mathcal{G}^*; \boldsymbol{X})$ as in (17): we refer to this estimate as OB-MED.

Our second proposal instead employs the full posterior distribution $p(\mathcal{G} \mid \boldsymbol{X})$ in (15). Recall that the OBES output consists of a collection of EGs $\{\mathcal{G}_k, k = 1, \dots, K\}$ visited by the MCMC chain. For each $\mathcal{G}_k$, we can approximate its posterior probability as

$$p(\mathcal{G}_k \mid \boldsymbol{X}) \approx \frac{m_{\mathcal{G}_k}(\boldsymbol{X})p(\mathcal{G}_k)}{\sum_{k=1}^{K} m_{\mathcal{G}_k}(\boldsymbol{X})p(\mathcal{G}_k)}, \tag{18}$$

TABLE 1 Simulation study

n	OB-MED	PC 0.01	PC 0.05	PC 0.10	GES 0	GES 0.5	GES 1
50	6.22	13.78	13.45	13.43	17.88	7.42	7.08
100	5.40	12.30	12.15	13.95	17.15	6.60	5.78
200	2.50	9.16	8.78	9.84	7.62	1.56	2.03

Note. Average structural Hamming distances between estimated EGs and true EGs, over 40 data sets, for sample size $n \in \{50, 100, 200\}$. Performances are measured for the OBES projected median probability graph model (OB-MED), the PC algorithm at significance levels $1\%, 5\%, 10\%$ (respectively, PC 0.01, PC 0.05, PC 0.10) and the GES algorithm with tuning parameter equal to 0, 0.5, and 1 (respectively, GES 0, GES 0.5, GES 1).

that is by re-normalizing each $m_{\mathcal{G}_k}(\boldsymbol{X})p(\mathcal{G}_k)$ term. A natural alternative to (18) is to use the MCMC frequency of visits of graph $\mathcal{G}_k$; for a comparison of these two methods, see García-Donato and Martínez-Beneito (2013).

Because of our discussion on the meaning of the causal effect parameters $\gamma_l(\mathcal{G})$'s across DAGs following Equation (16), we can take advantage of the posterior distribution on the space of graphs and use Bayesian Model Averaging (BMA; Hoeting *et al.* 1999). Accordingly an alternative to (17) is represented by the BMA estimate of the causal effect

$$\overline{\gamma}_{BMA}(\boldsymbol{X}) = \sum_{\mathcal{G}_k} \mathbb{E}\{\gamma_{\text{avg}}(\mathcal{G}_k) \mid \boldsymbol{X}, \mathcal{G}_k\} p(\mathcal{G}_k \mid \boldsymbol{X}), \qquad (19)$$

which will be referred to as OB-MA in the sequel. The proposed strategies are summarized in two algorithms presented as Supporting Information to the online version of this paper.

5 | SIMULATIONS

In this section, we evaluate the performance of our method through simulation studies. Specifically, we fix the number of nodes $q = 20$ and consider different sample sizes $n \in \{50, 100, 200\}$. Under each scenario defined by n, we randomly generate 40 DAGs with probability of edge inclusion $p_{\text{edge}} = 0.1$ using the `randomDAG` function in the R package `pcalg`. Next, under each $\mathcal{D}$, n independent and identically distributed observations are generated from the system of linear equations

$$X_{i,j} = \mu_j + \sum_{k \in \text{pa}_{\mathcal{D}}(j)} \beta_{k,j} X_{i,k} + \varepsilon_{i,j}, \qquad (20)$$

where $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ independently with $\mu_j = 0$ and $\sigma_j^2 = 1$, while regression coefficients $\beta_{k,j}$ are uniformly chosen in the interval [1,2].

This results in a collection of 40 (observational) data sets. For each $\mathcal{D}$, we determine its representative EG $\mathcal{G}$. Next, for each $\mathcal{G}$, we randomly select a target node $i \in \{2, \dots, 20\}$ and compute the (true) set of distinct causal effects of do$(X_i = \widetilde{x}_i)$ on $X_1 \equiv Y$, which we summarize through the average $\overline{\gamma}_{\text{true}}$. The resulting collection of 40 true average causal effects, $\overline{\gamma}_{\text{true}}^{(1)}, \dots, \overline{\gamma}_{\text{true}}^{(40)}$, will be the benchmark of our comparison. For each data set, we then proceed as follows.

We first implement the OBES method with $T = 25\,000$ iterations to approximate the posterior over the EG space as in Equation (18). Next, for each EG $\mathcal{G}$, we construct the set of distinct causal effects and produce a sample of size $S = 5000$ from the posterior of each coefficient $\gamma_l^{(k)}$. The overall average causal effect is computed as in Equation (19). We denote the resulting collection of estimates as $(\overline{\gamma}_{OB-MA}^{(1)}, \dots, \overline{\gamma}_{OB-MA}^{(40)})$.

We also consider the OB-MED method. Accordingly, from the EG estimate $\mathcal{G}^*$ (projected median probability graph model), we compute the conditional average causal effect $\overline{\gamma}_{\text{avg}}(\mathcal{G}^*)$ as in (16). The collection of estimated parameters is denoted with $(\overline{\gamma}_{OB-MED}^{(1)}, \dots, \overline{\gamma}_{OB-MED}^{(40)})$.

Finally, we consider the method of Maathuis *et al.* (2009) that first uses the PC algorithm to infer an EG, and then estimates the set of distinct causal effects through maximum likelihood estimation of regression parameters in suitably defined linear regression models; see the original paper for details. The PC algorithm is based on a sequence of conditional independence tests that we implement for significance level $\alpha \in \{1\%, 5\%, 10\%\}$. In addition, for graph estimation we employ the Greedy Equivalence Search (GES) algorithm of Chickering (2002); see also Hauser and Bühlmann (2015). GES is computed for three different optimization criteria: the Bayesian Information Criterion (GES 0) and the Extended Bayesian Information Criterion with tuning coefficient 0.5 and 1 (GES 0.5 and GES 1 respectively), corresponding to increasing sparsity in the estimation procedure.

We start by evaluating the performances of the various methods in recovering the true EG structure as measured by the Structural Hamming Distance (SHD) between estimated and true graph. SHD represents the number of edge insertions, deletions, or flips needed to transform a graph (the estimated EG) into another (the true EG). The average SHD computed over the 40 replicates is reported in Table 1. As it is apparent, the PC algorithm and the GES 0 method present the worst performances in recovering the structure of the true EG. On the other hand, GES is much more sensitive to the choice of its tuning parameter, and improves the EG estimation as sparsity is encouraged (GES 0.5 and GES 1). In general, results from GES 0.5 and GES 1 are comparable with the performance of OB-MED that remains highly competitive under all scenarios. Moreover, all methods improve their performance as the sample size increases.

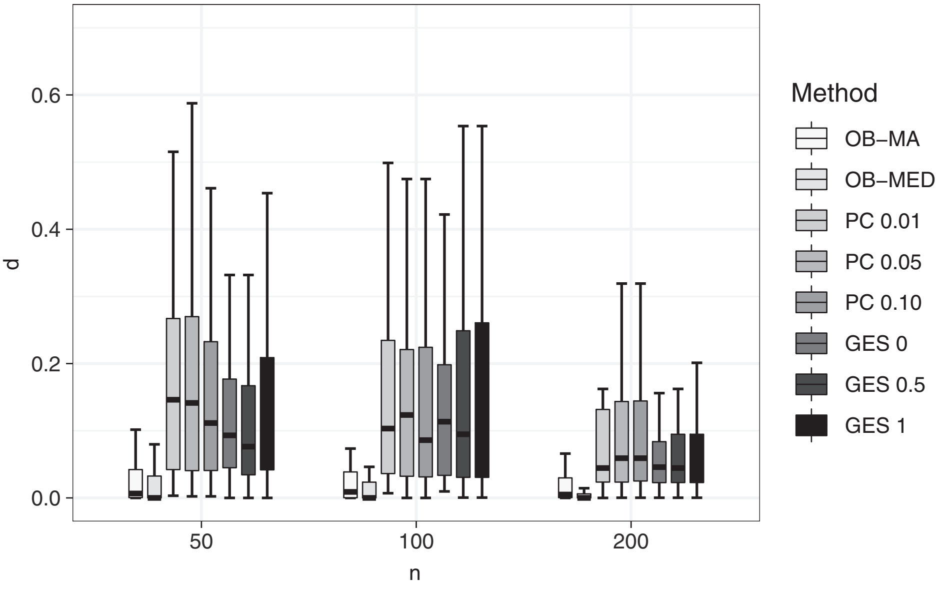

FIGURE 1 Simulation study. Absolute-value distance between estimated and true causal effect over 40 data sets, for sample size $n \in \{50, 100, 200\}$. The comparison is between our OB method with model averaging of causal effects (OB-MA), OB with OBES projected median probability model (OB-MED), IDA with EG estimation using the PC algorithm at significance levels 1%, 5%, 10% (respectively PC 0.01, PC 0.05, PC 0.10) and the GES algorithm with tuning parameter equal to 0, 0.5, and 1 (respectively GES 0, GES 0.5, GES 1)

Next, we compare the results regarding the causal effect estimation. To this end, let $\overline{\gamma}_M^{(i)}$, $i = 1, \dots, 40$, be the estimated (average) causal effect under the ith simulation for the generic method M under comparison, $\overline{\gamma}_{\text{true}}^{(i)}$ the corresponding (average) true causal effect. For each simulation and method, we measure the absolute-value distance between the estimated and true causal effect, $d_M(i) = |\overline{\gamma}_M^{(i)} - \overline{\gamma}_{\text{true}}^{(i)}|$. Results, for sample sizes $n \in \{50, 100, 200\}$, are summarized in the box-plots of Figure 1, where we report the distribution of the distance $d_M(i)$, $i = 1, \dots, 40$, for each method included in the study. Results show that both OB-MED and OB-MA outperform all the other methods under each scenario. Intuitively, the poor performance of the PC algorithm depends on a poor recovery of the underlying EG (as revealed in Table 1), which consequently affects the correct identification of the set of distinct causal effects. Moreover, GES generally presents slightly better performances than the PC, especially for $n = 200$. With regard to our method, it is interesting to observe that OB-MED, even if based on a single model estimate, outperforms OB-MA that instead relies on a collection of estimated EGs. The same behavior is observed under each scenario, and becomes more evident as the sample size increases.

6 | REAL DATA ANALYSIS: GENE EXPRESSIONS IN *ARABIDOPSIS THALIANA*

In this section, we apply our methodology to the analysis of gene expressions in *Arabidopsis thaliana* (Wille *et al.*, 2004). The data set is publicly available as Supporting Information to the online version of the paper. The complete data set consists of gene expression measurements taken in *Arabidopsis thaliana* grown under $n = 118$ different conditions (such as light or darkness, or growth hormones). Note that we treat all units as exchangeable, although in principle one could set up a multiple graph framework accounting for differences between observations generated under different settings; see Peterson *et al.* (2015). From a graphical model perspective, each gene corresponds to a node in the graph, while chemical reactions between them are represented by edges. As described in the original paper, interactions between 39 genes generate a pathway that can be partitioned into two components: the mevalonate and nonmevalonate (MEV) pathways.

In the following, we focus on $q = 13$ genes involved in the MEV pathway, which also received particular attention in Wille *et al.* (2004). Besides structural learning, we are also interested in estimating the causal effect of an intervention on a specific gene on the remaining ones according to the framework discussed in Section 2.

We run $T = 25\,000$ iterations of OBES by fixing the maximum number of edges to $M = 3q$ and setting $a = 1$, $b = (2q - 2)/3 - 1$ in the EG prior (14), which corresponds to a prior probability of inclusion of 0.125 for each edge. We then focus on causal effect estimation by applying the OB-MA strategy. Accordingly, we use the OBES output to approximate the posterior distribution over the visited EGs as in (18). Next, under each EG $\mathcal{G}_k$, we construct for each pair of nodes (u, v) $(u \neq v)$ the set of distinct causal effects and provide for each gamma parameter an approximation to its posterior

$\mathcal{G}$	$p(\mathcal{G} \mid \boldsymbol{X})$	$\mathcal{G}_S$	$\mathrm{pa}_{\mathcal{G}}(\mathrm{PPDS1})$	Average causal effects
$\mathcal{G}_1$	4.23%	$PPDS1 \longrightarrow PPDS2$	$\emptyset$	$\overline{\gamma}_{1,1} = 0.91$
$\mathcal{G}_2$	3.69%	$PPDS1 \longrightarrow PPDS2$	$\emptyset$	$\overline{\gamma}_{2,1} = 0.91$
$\mathcal{G}_3$	3.25%	$PPDS1 \longrightarrow PPDS2$	$\emptyset$	$\overline{\gamma}_{3,1} = 0.91$
$\mathcal{G}_4$	2.47%	$MECPS$ $\vert$ $PPDS1 \longrightarrow PPDS2$ $\vert$ HDR	$\emptyset$ $\{MECPS\}$ $\{HDR\}$	$\overline{\gamma}_{4,1} = 0.91$ $\overline{\gamma}_{4,2} = 0.92$ $\overline{\gamma}_{4,3} = 0.85$

FIGURE 2 *Arabidopsis thaliana*. Four top EGs $\mathcal{G}_1, \dots, \mathcal{G}_4$ with associated posterior probabilities $p(\mathcal{G} \mid \boldsymbol{X})$. Intervened node is $PPDS1$, response node is $PPDS2$. Sub-graphs $\mathcal{G}_S$, $S = \mathrm{cl}_{\mathcal{G}}(PPDS1) \cup PPDS2$, help identify the set of distinct causal effects. Column $\mathrm{pa}_{\mathcal{G}}(PPDS1)$ specifies the parents of $PPDS1$ compatible with the structure of $\mathcal{G}$, leading to different estimates of causal effect coefficients (fifth column)

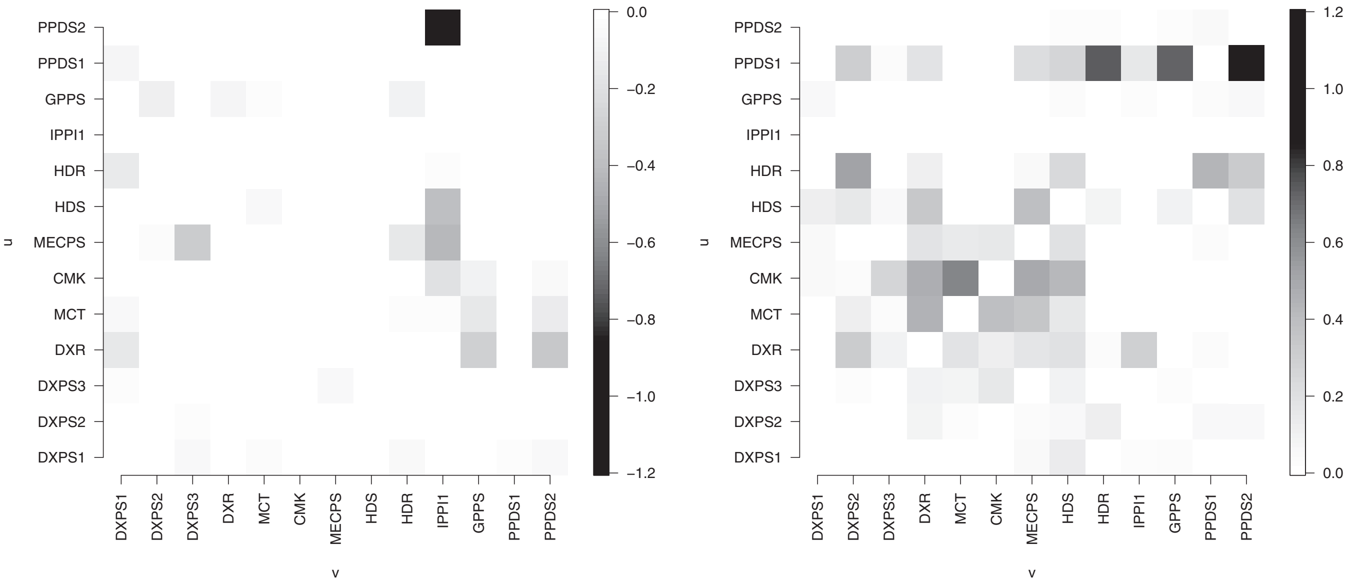

FIGURE 3 *Arabidopsis thaliana*. Heat maps of average causal effects of $\mathrm{do}(X_u = \widetilde{x}_u)$ on X_v estimated by the OB-MA method for each pair of nodes (u, v). Left panel: negative causal effects; right panel: positive causal effects

distribution. As an instance of the output, we provide in Figure 2 information about the (four) EGs with highest estimated posterior probability and focus on the causal effect of an intervention on variable $PPDS1$ with respect to $PPDS2$. Let $\mathrm{ne}_{\mathcal{G}}(u)$ be the *neighbor* set of u in $\mathcal{G}$, that is the set of all nodes v such that there exists $v - u$ and $\mathrm{bd}_{\mathcal{G}}(u) = \mathrm{ne}_{\mathcal{G}}(u) \cup \mathrm{pa}_{\mathcal{G}}(u)$ be the *boundary* of u in $\mathcal{G}$. Let also $\mathrm{cl}_{\mathcal{G}}(u) = u \cup \mathrm{bd}_{\mathcal{G}}(u)$ be the *closure* of u in $\mathcal{G}$. Accordingly, we report for each EG $\mathcal{G}$ the sub-graph $\mathcal{G}_S$, $S = \mathrm{cl}_{\mathcal{G}}(PPDS1) \cup PPDS2$, which help identify the set of distinct causal effects. Each causal effect indeed corresponds to one possible set $\mathrm{pa}_{\mathcal{G}}(PPDS1)$ compatible with the structure of $\mathcal{G}$ (fourth column of the table), leading to a distinct estimate $\overline{\gamma}$ obtained as in (17) (fifth column). From the same output, we can observe that the posterior probability $p(\mathcal{G} \mid \boldsymbol{X})$ of each graph never exceeds the 5% threshold. Therefore, there exists a high degree of uncertainty around graph estimation, and accordingly results might be significantly biased if causal effect estimation is performed by relying on a single model estimate. Our results are summarized in the two heat maps of Figure 3 where

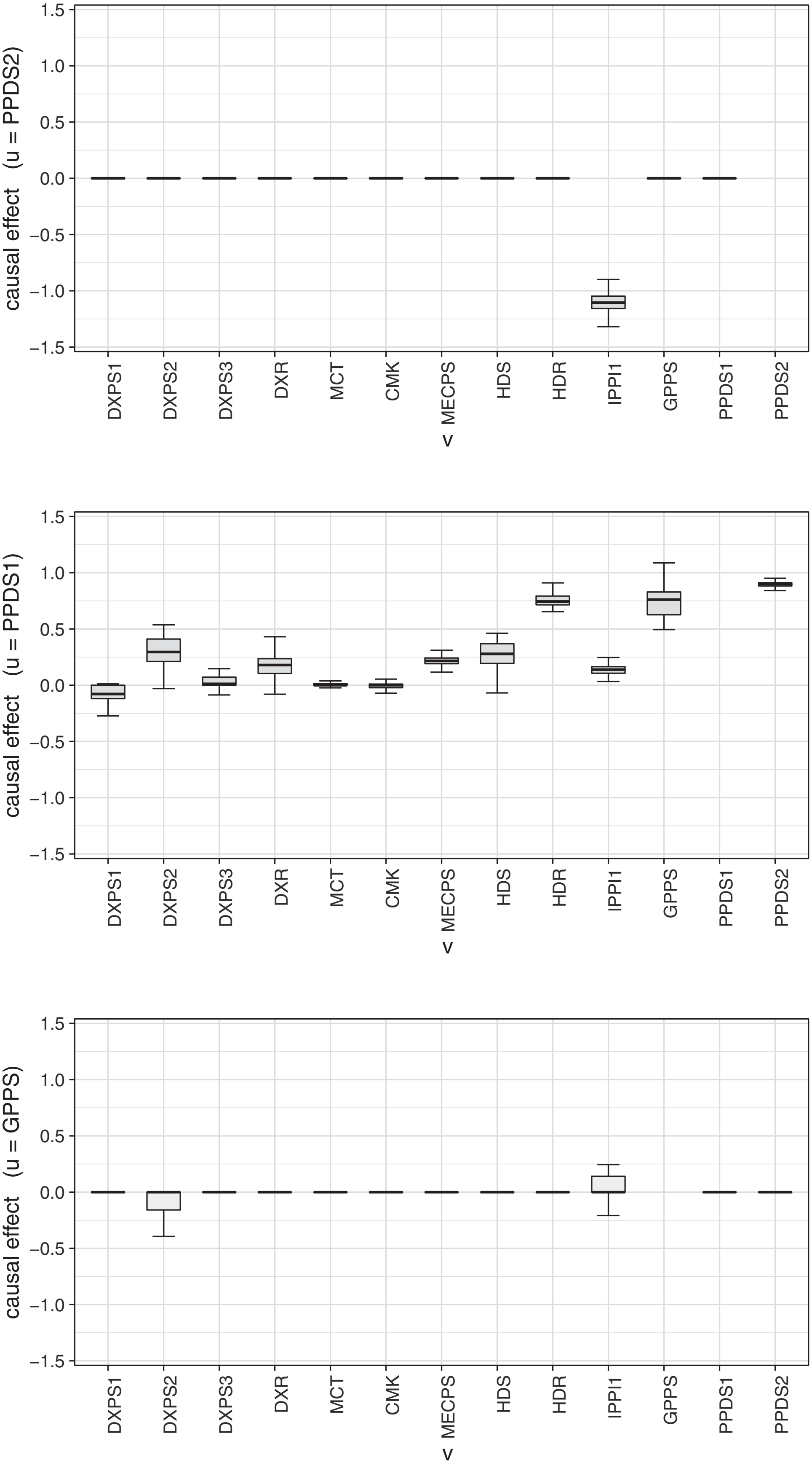

FIGURE 4 *Arabidopsis thaliana*. Intervened nodes (u) are $PPDS2, PPDS1, GPPS$. Posterior distribution of the causal effect parameter of do($X_u = \widetilde{x}_u$) on each potential response X_v based on the OB-MA strategy

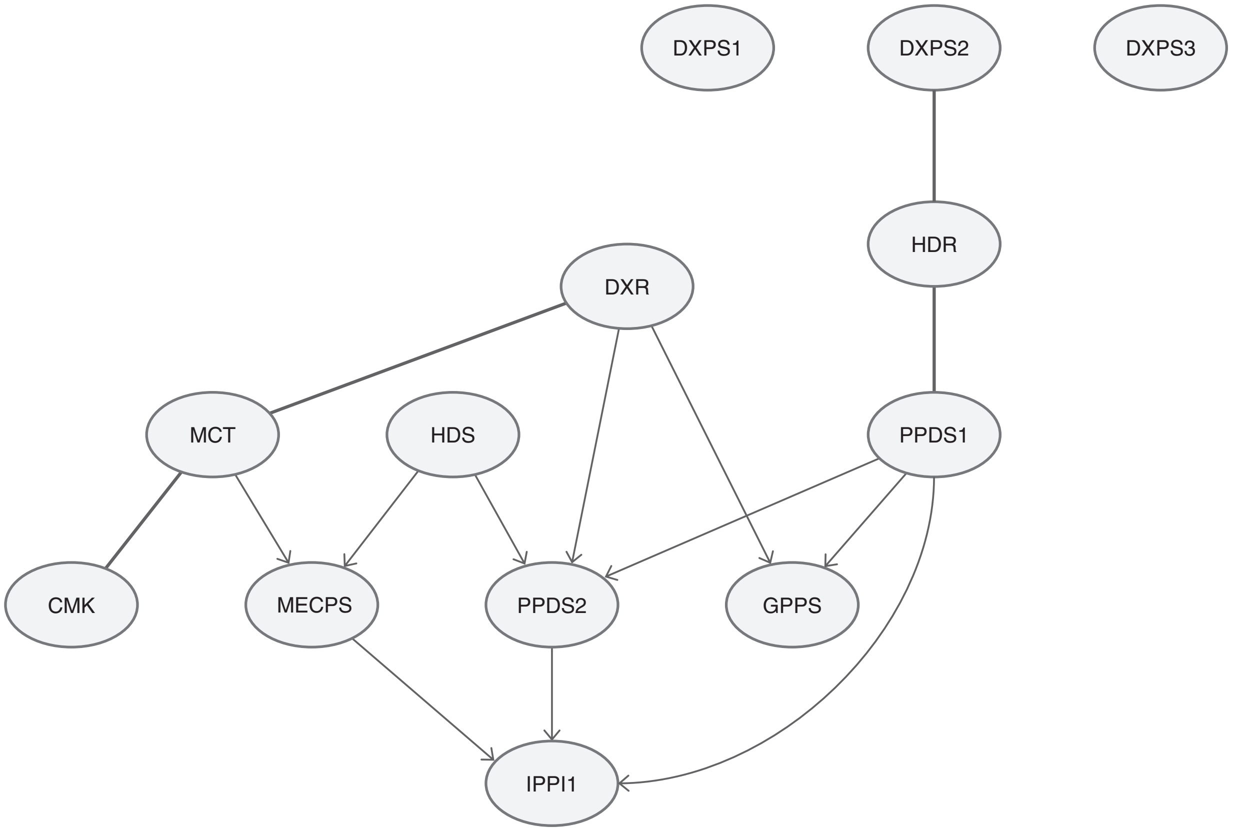

FIGURE 5 *Arabidopsis thaliana*. Estimated graph obtained from the OBES method

we report, for each pair of nodes (u, v), the OB-MA causal effect estimates of $\text{do}(X_u = \tilde{x}_u)$ on X_v. In addition, for three selected intervened nodes $u \in \{PPDS2, PPDS1, GPPS\}$, we report in the box-plots of Figure 4 the posterior distributions of the causal effect parameter of $\text{do}(X_u = \tilde{x}_u)$ on each potential response X_v obtained from the OB-MA strategy.

As a comparison, we also adopt the OB-MED method. Specifically, we start using the OBES output to estimate the posterior probabilities of edge inclusion. Next, we compute $FDR(k)$ for a grid of thresholds $k \in [0, 1]$ and select the maximum value of k such that $FDR(k) < 0.10$, which we denote by $k^* = 0.75$; see also Peterson *et al.* (2015). The resulting projected median probability graph is reported in Figure 5. Using this EG estimate we then identify, for each pair of nodes (u, v) $(u \neq v)$, the set of distinct causal effects and provide for each one an approximation to its posterior distribution. Again, the results are summarized in the two heat maps of Figure 6 where we report, for each pair of nodes (u, v), the average causal effect of $\text{do}(X_u = \tilde{x}_u)$ on X_v.

Results obtained from the two strategies look similar. In particular, OB-MED, which relies on a single EG estimate, clearly distinguishes between zero and nonzero causal effects. Differently, OB-MA, which is based on a (large) collection of EGs, returns further estimated causal effect coefficients that are (slightly) far from zero, in addition to those identified by the previous strategy. Moreover, as a comparison with existing results for the structural learning issue (Wille *et al.*, 2004), our EG estimate (median projected probability graph model) also captures the path DXR → MCT → CMK and the link between PPDS1 and PPDS2, with associated positive causal effects as it emerges from both the two strategies. On the other hand, our results reveal strong negative causal effects of genes PPDS2, HDS, MECPS onto IPPI1. Comparisons with the IDA method of Maathuis *et al.* (2009) is provided in the Supporting Information to the online version of the paper.

7 | DISCUSSION

In this paper, we present a Bayesian methodology for causal effects estimation from observational data. Specifically, we assume that the observations are generated from a Gaussian model Markov with respect to a DAG that describes conditional independencies as well as causal relationships between variables. As only observational data are available, the underlying DAG is identifiable only up to an equivalence class (or its representative essential graph), and accordingly causal effects are not uniquely determined. For a given equivalence class of DAGs, we first introduce an objective Bayes methodology to estimate causal effects; next we describe an MCMC

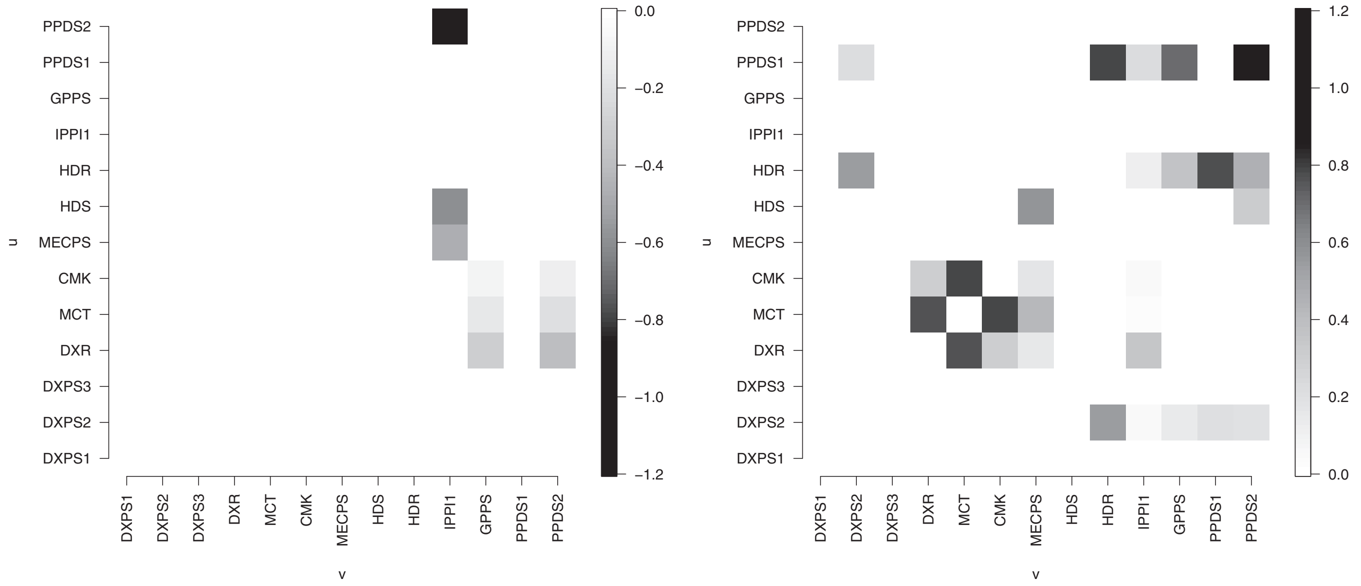

FIGURE 6 *Arabidopsis thaliana*. Heat maps of average causal effects of do($X_u = \widetilde{x}_u$) on X_v estimated by the OB-MED method for each pair of nodes (u, v). Left panel: negative causal effects; right panel: positive causal effects

algorithm to approximate the posterior distribution on graph space. At this stage, we present two alternative strategies. One summarizes the posterior distribution through a single graph and computes the estimate of the causal effect conditionally on that graph (OB-MED). Alternatively, an overall estimate of the causal effect can be obtained through Bayesian model averaging (OB-MA). The performance of our method is evaluated on simulation scenarios, and shown to behave appreciably better than existing benchmark methods under both strategies.

A technical feature that deserves some thought regards the estimation of the causal effect γ_i in (6). Our approach is to express γ_i as a function of the covariance matrix $\boldsymbol{\Sigma}_{Y,\mathrm{fa}(i)}$ as in (7); accordingly inference on γ_i is induced from the posterior of $\boldsymbol{\Sigma}_{Y,\mathrm{fa}(i)}$. On the other hand, based on (5), Maathuis *et al.* (2009) suggest to estimate γ_i using least squares in the "ordinary" regression model $Y \sim X_i + \boldsymbol{x}_{pa(i)}$. In so doing, however, the Markov constraints inherent in $\boldsymbol{\Sigma} \in \mathcal{C}_{\mathcal{D}}$ through the DAG $\mathcal{D}$ are lost, and this may produce inadequate estimates of causal effects. We further investigated the issue through simulation experiments, replacing the ordinary least squares estimate of γ_i with that induced by a Markov-DAG-constrained MLE of $\boldsymbol{\Sigma} \in \mathcal{C}_{\mathcal{D}}$ (Rütimann and Bühlmann, 2009) followed by (7) and obtained appreciably better results.

Throughout this work, we assume that there are no unmeasured confounders. However, this assumption may be unreasonable in a variety of contexts and can potentially produce biased results both in the graph and in the causal effect estimation. An important contribution on this topic is presented in Frot *et al.* (2019), who introduce a novel methodology to estimate Markov equivalence classes of DAGs in the presence of hidden variables using frequentist concepts.

Our model assumes that the observations are jointly normally distributed. This appears quite reasonable in our application (after a log-transformation of the variables) but may not be realistic in other contexts. A few recent papers that address the problem of extending the scope of causal DAGs beyond the Gaussian setting are Mahmoudi and Wit (2018) who consider nonparanormal models, while Hoyer *et al.* (2009) and Shimizu *et al.* (2006) use *nonlinear* models and *linear-non-Gaussian* models, respectively, for causal discovery.

The definition of causal effect that we employ following Pearl (2000) requires the notion of a DAG because of the crucial role played by the parents of the intervened node. However, DAGs cannot capture feedback loops, a feature that may be useful to model gene regulatory networks, as demonstrated in current research. In particular, the recent paper Ni *et al.* (2018) adopts Reciprocal Graphical Models (RGM) (Koster, 1997) to investigate gene regulatory relationships, thus allowing for the presence of loops. This is made possible by including additional variables (ie, copy numbers of genes) that act as parents of corresponding gene nodes. A natural question that arises is how to meaningfully relate the notion of causal effect to an RGM.

ACKNOWLEDGMENTS

We thank Hélène Massam for helpful communications on the DAG-Wishart. We also thank an associate editor and a reviewer for useful comments that led to an improved version of the paper. Support from COSTNET (COST Action CA15109) and UCSC (D1 and D.3.2 research grant) is gratefully acknowledged.

DATA AVAILABILITY STATEMENT

The *Arabidopsis thaliana* dataset of Wille *et al.* (2004) is provided as Supporting Information available with this paper at the Biometrics website on Wiley Online Library.

ORCID

Federico Castelletti https://orcid.org/0000-0001-7911-2942
Guido Consonni https://orcid.org/0000-0002-1252-5926

REFERENCES

Andersson, S.A., Madigan, D. and Perlman, M.D. (1997) A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25, 505–541.

Barbieri, M. and Berger, J. (2004) Optimal predictive model selection. *The Annals of Statistics*, 32, 870–897.

Ben-David, E., Li, T., Massam, H. and Rajaratnam, B. (2015) High dimensional Bayesian inference for Gaussian directed acyclic graph models. [Preprint] Available at: https://arxiv.org/abs/1109.4371v5.

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

Castelletti, F., Consonni, G., Della Vedova, M. and Peluso, S. (2018) Learning Markov equivalence classes of directed acyclic graphs: an objective Bayes approach. *Bayesian Analysis*, 13, 1231–1256.

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

Consonni, G., Fouskakis, D., Liseo, B. and Ntzoufras, I. (2018) Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13, 627–679.

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

Frot, B., Nandy, P. and Maathuis, M.H. (2019) Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 459–487.

García-Donato, G. and Martínez-Beneito, M.A. (2013) On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, 108, 340–352.

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

Gillispie, S.B. and Perlman, M.D. (2002) The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence*, 141, 137–155.

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

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999) Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors). *Statistical Science*, 14, 382–417.

Hoyer, P.O., Janzing, D., Mooij, J.M., Peters, J. and Schölkopf, B. (2009) Nonlinear causal discovery with additive noise models. In: Koller, D., Schuurmans, D., Bengio, Y. and Bottou, L. (Eds.) *Advances in Neural Information Processing Systems 21*. Red Hook, NY: Curran Associates, Inc, pp. 689–696.

Imbens, G.W. and Rubin, D.B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Kalisch, M. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8, 613–36.

Koster, J.T. (1997) Markov properties of nonrecursive causal models. *The Annals of Statistics*, 24, 2148–2177.

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

Maathuis, M.H., Kalisch, M. and Bühlmann, P. (2009) Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37, 3133–3164.

Mahmoudi, M.S. and Wit, E. (2018) Estimating causal effects from nonparanormal observational data. *The International Journal of Biostatistics*, 14, forthcoming.

Müller, P., Parmigiani, G. and Rice, K. (2007) FDR and Bayesian multiple comparisons rules. In: Bernardo, J.M., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A. and West, M. (Eds.) *Bayesian Statistics 8*. Oxford: Oxford University Press.

Ni, Y., Ji, Y. and Müller, P. (2018) Reciprocal graphical models for integrative gene regulatory network analysis. *Bayesian Analysis*, 13, 1095–1110.

Ni, Y., Stingo, F.C. and Baladandayuthapani, V. (2017) Sparse multi-dimensional graphical models: a unified Bayesian framework. *Journal of the American Statistical Association*, 112, 779–793.

O'Hagan, A. (1995) Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 99–138.

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

Peters, J. and Bühlmann, P. (2014) Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101, 219–228.

Peterson, C., Stingo, F.C. and Vannucci, M. (2015) Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110, 159–174.

Pingault, J.B., O'Reilly, P.F., Schoeler, T., Ploubidis, G.B., Rijsdijk, F. and Dudbridge, F. (2018) Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics*, 19, 566–580.

Press, S.J. (1982) *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Malabar, FL: Krieger Publishing Company, Inc.

Rütimann, P. and Bühlmann, P. (2009) High dimensional sparse covariance estimation via directed acyclic graphs. *Electronic Journal of Statistics*, 3, 1133–1160.

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

Shimizu, S., Hoyer, P.O., Hyvärinen, A. and Kerminen, A.J. (2006) A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7, 2003–2030.

Shojaie, A. and Michailidis, G. (2009) Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*, 16, 407–26.

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction and Search*, 2nd edition. Cambridge, MA: MIT Press.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelić, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., Bühlmann, P. (2004) Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology*, 5, A92.

SUPPORTING INFORMATION

Web Appendices, Tables and Figures referenced in Sections 5 and 6 are available with this paper at the Biometrics website on Wiley Online Library. The material contains two algorithms summarizing the proposed strategies, additional results from our simulation study, and application to *Arabidopsis thaliana* as well as the R code implementing our method. The latter includes the two algorithms presented in the paper together with an illustrative example.

How to cite this article: Castelletti F, Consonni G. Bayesian inference of causal effects from observational data in Gaussian graphical models. *Biometrics*. 2021;77:136–149. https://doi.org/10.1111/biom.13281