Bayesian Analysis (2021) **16**, Number 4, pp. 1113–1137

Bayesian Causal Inference in Probit Graphical Models

Federico Castelletti[*,‡] and Guido Consonni[†]

Abstract. We consider a binary response which is potentially affected by a set of continuous variables. Of special interest is the causal effect on the response due to an intervention on a specific variable. The latter can be meaningfully determined on the basis of observational data through suitable assumptions on the data generating mechanism. In particular we assume that the joint distribution obeys the conditional independencies (Markov properties) inherent in a Directed Acyclic Graph (DAG), and the DAG is given a causal interpretation through the notion of interventional distribution. We propose a DAG-probit model where the response is generated by discretization through a random threshold of a continuous latent variable and the latter, jointly with the remaining continuous variables, has a distribution belonging to a zero-mean Gaussian model whose covariance matrix is constrained to satisfy the Markov properties of the DAG; the latter is assigned a DAG-Wishart prior through the corresponding Cholesky parameters. Our model leads to a natural definition of causal effect conditionally on a given DAG. Since the DAG which generates the observations is unknown, we present an efficient MCMC algorithm whose target is the posterior distribution on the space of DAGs, the Cholesky parameters of the concentration matrix, and the threshold linking the response to the latent. Our end result is a Bayesian Model Averaging estimate of the causal effect which incorporates parameter, as well as model, uncertainty. The methodology is assessed using simulation experiments and applied to a gene expression data set originating from breast cancer stem cells.

Keywords: graphical model, directed acyclic graph, DAG-probit, causal inference, DAG-Wishart, modified Cholesky decomposition.

1 Introduction

We consider a system of random quantities, which includes a binary response as well as a collection of continuous variables, and address the problem of determining the causal effect on the response due to an intervention on a given variable. A causal question involves the data generating mechanism after an intervention is applied to the system, and must be carefully distinguished from traditional conditioning of probability theory (Pearl, 2009, Section 2.4). The gold standard for addressing causal questions is represented by randomized controlled experiments; the latter however are not always available because they may be unethical, infeasible, time consuming or expensive (Maathuis and Nandy, 2016). By contrast, *observational* data, that is observations produced without exogenous perturbations of the system, are widely available and often plentiful. The

[*]Università Cattolica del Sacro Cuore, Milan, Italy, federico.castelletti@unicatt.it
[†]Università Cattolica del Sacro Cuore, Milan, Italy, guido.consonni@unicatt.it
[‡]Corresponding author.

 https://doi.org/10.1214/21-BA1260

challenge is then to infer causal effects based on observational data alone. To achieve this goal, it is crucial to set up a suitable conceptual framework which is able to address causal questions; in particular the notion of *joint distribution* for a collection of random variables can only address concepts linked to association, so much so that, by converse, "a causal concept is any relationship that cannot be defined from the distribution alone" (Pearl, 2009, Section 2).

A very useful framework to bridge the gap between the observational and the experimental domains is represented by the Directed Acyclic Graph (DAG), or its allied concept of Structural Equation Model (SEM); see Pearl (1995) and Pearl (2000). DAGs have been extensively used to construct statistical models embodying conditional independence relations (Lauritzen, 1996). Applications are numerous especially in genomics; see for instance Friedman (2004) and Friedman and Koller (2003). With observational data, conditional independence relations will drive inference on DAG and parameter space. On the other hand, the additional syntax and semantics of *causal* DAGs (Pearl, 2000) will be instrumental to define the notion of causal effect.

As in standard probit regression (Albert and Chib, 1993), we assume that the observable binary response is the result of a discretization of a continuous *latent* variable. Next, for a given DAG, we model all continuous random variables, along with the latent, as a multivariate Gaussian family satisfying the corresponding Markov property. We call the resulting setup a *DAG-probit* model, and provide a definition of causal effect on the response which is predicated on a given DAG through the notion of *interventional* distribution (Pearl, 2000). However the structure of the DAG is usually unknown, and this must be taken into account because different DAGs will typically induce distinct causal effects; see the review paper Maathuis and Nandy (2016) and Castelletti and Consonni (2021a) for a Bayesian approach.

In this work we extend the notion of interventional distribution and causal effect (Pearl, 2000; Maathuis et al., 2009) to DAG-probit models. Specifically, we propose a Bayesian method which jointly performs DAG-model determination as well as inference of causal effects in the presence of a binary response. From a computational viewpoint we introduce an MCMC scheme to sample from the joint posterior of models (DAGs) and model-dependent parameters (causal effects) which we implement through an efficient PAS algorithm (Godsill, 2012). The rest of the paper is organized as follows. In Section 2 we review Gaussian DAG-models and define the DAG-probit model. In Section 3 we present the structure of the interventional distribution in its general form, then specialize it to the Gaussian case, and finally extend the definition of causal effect to DAG-probit models. Section 4 presents our Bayesian methodology with particular emphasis on priors for model parameters. An MCMC algorithm for posterior inference on models, parameters and hence causal effects is presented in Section 5. We evaluate the proposed methodology through simulation studies in Section 6, and then apply it to a data set on gene expression measurements derived from breast cancer stem cells (Section 7). Finally a few points for discussion are presented in Section 8. Some theoretical results as well as additional simulation outputs are reported in the Supplementary material (Castelletti and Consonni, 2021b).

2 Model Formulation

In this section we first provide some background material on Gaussian DAG-models with special emphasis on their parameterization (Section 2.1). Next we present our DAG-probit model (Section 2.2). Both sections deal with the likelihood, while choices of prior distributions are discussed in Section 4.

2.1 Gaussian DAG-Models

Let $\mathcal{D} = (V, E)$ be a DAG, where $V = \{1, \ldots, q\}$ is a set of vertices (or nodes) and $E \subseteq V \times V$ a set of edges whose elements are $(u, v) \equiv u \to v$, such that if $(u, v) \in E$ then $(v, u) \notin E$. In addition, $\mathcal{D}$ contains no cycles, that is paths of the form $u_0 \to u_1 \to \cdots \to u_k$ where $u_0 \equiv u_k$. For a given node v, if $u \to v \in E$ we say that u is a *parent* of v (conversely v is a *child* of u). The parent set of v in $\mathcal{D}$ is denoted by $\text{pa}(v)$, the set of children by $\text{ch}(v)$. Moreover, we denote by $\text{fa}(v) = v \cup \text{pa}(v)$ the *family* of v in $\mathcal{D}$. Finally, we say that a DAG is *complete* if all vertices are joined by an edge. For further theory and notation on DAGs we refer to Lauritzen (1996).

We consider a collection of random variables $(X_1, \ldots, X_q)$ and assume that their joint probability density function $f(\boldsymbol{x})$ is Markov w.r.t. $\mathcal{D}$, so that it admits the following factorization

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

In this section, as well as in Section 3, we reason *conditionally* on a given DAG $\mathcal{D}$ without an explicit conditioning in the notation we use. In Section 4 we will instead deal with model (DAG) uncertainty and will reinstate $\mathcal{D}$ in our notation.

If the joint distribution is Gaussian with mean equal to zero, we write

$$X_1, \ldots, X_q \,|\, \boldsymbol{\Omega} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Omega}^{-1}),\ \boldsymbol{\Omega} \in \mathcal{P}_{\mathcal{D}}, \tag{2}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix, and $\mathcal{P}_{\mathcal{D}}$ is the space of symmetric positive definite (s.p.d.) precision matrices Markov w.r.t. $\mathcal{D}$. For a Gaussian DAG-model factorization (1) becomes

$$f(x_1, \ldots, x_q \,|\, \boldsymbol{\Omega}) = \prod_{j=1}^{q} d\mathcal{N}(x_j \,|\, \mu_j(\boldsymbol{x}_{\text{pa}(j)}), \sigma_j^2), \tag{3}$$

where $d\mathcal{N}(\cdot \,|\, \mu, \sigma^2)$ denotes the normal density having mean μ and variance σ^2. The assumption of normality essentially guarantees that the model is *faithful* to the DAG, that is no conditional independence relations other than the ones entailed by the Markov property applied to the DAG are present in the statistical model. The word "essentially" means that faithfulness may fail for particular combinations of parameters; yet this set has Lebesgue measure zero, which translates to zero probability when the parameters are continuous. For an extensive discussion of faithfulness see Sadeghi (2017).

For a stronger version of faithfulness in the Gaussian case which is meant to overcome difficulties in DAG-identification in finite samples see Zhang and Spirtes (2003).

For a given DAG, we assume without loss of generality a *parent ordering* of the nodes which numerically labels the variables so that $i > j$ whenever j is a child of i. A parent ordering always exists, although it is not unique in general. We also remark that a parent ordering is specific to any given DAG under consideration and may change if alternative DAGs are entertained. Importantly, it only represents a convenient device to specify a prior on the parameter space; see Section 4. In the Supplementary material we also show that the prior we employ under any given DAG is invariant to the choice of the parent ordering.

Moreover, we declare node 1, which cannot have children, to be the (latent) outcome variable. Equation (3) can be also written as a *structural equation model*

$$\boldsymbol{L}^\top (X_1, \ldots, X_q)^\top = \boldsymbol{\varepsilon}, \tag{4}$$

where because of the assumed parent ordering $\boldsymbol{L}$ 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 $\boldsymbol{D} = \text{diag}(\boldsymbol{\sigma}^2)$ and $\boldsymbol{\sigma}^2$ is the $(q, 1)$ vector of *conditional* variances whose j-th element is $\sigma_j^2 = \mathbb{V}\text{ar}(X_j \,|\, \boldsymbol{x}_{\text{pa}(j)}, \boldsymbol{\Omega})$. From (4) it follows that

$$\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top. \tag{5}$$

We refer to equation (5) as the *modified Cholesky decomposition* of $\boldsymbol{\Omega}$. Let now $\prec j \succ = \text{pa}(j)$ and $\prec j\,] = \text{pa}(j) \times j$. Representation (5) induces a re-parametrization of $\boldsymbol{\Omega}$ in terms of the Cholesky parameters $\big\{(\sigma_j^2, \boldsymbol{L}_{\prec j\,]}),\ j = 1, \ldots, q\big\}$, where

$$\boldsymbol{L}_{\prec j\,]} = -\boldsymbol{\Sigma}^{-1}_{\prec j \succ}\boldsymbol{\Sigma}_{\prec j\,]}, \quad \sigma_j^2 = \boldsymbol{\Sigma}_{jj\,|\,\text{pa}(j)};$$

see also Cao et al. (2019). Accordingly, equation (3) can be written as

$$f(x_1, \ldots, x_q \,|\, \boldsymbol{D}, \boldsymbol{L}) = \prod_{j=1}^{q} d\mathcal{N}(x_j \mid -\boldsymbol{L}^\top_{\prec j\,]}\boldsymbol{x}_{\text{pa}(j)}, \sigma_j^2), \tag{6}$$

with the understanding that the conditional expectation of X_j in (6) is zero whenever $\text{pa}(j)$ is the empty set.

2.2 DAG-Probit Models

We introduce in this section the general form of a *DAG-probit model*. We assume that the joint distribution of $(X_1, X_2, \ldots, X_q)$ is Gaussian and Markov w.r.t. $\mathcal{D}$ so that its density is as in (6). Recall that X_1 is latent and we are only allowed to observe the binary variable $Y \in \{0, 1\}$. Specifically, for a given threshold $\theta_0 \in (-\infty, +\infty)$, we assume that

$$Y = \begin{cases} 1 & \text{if } X_1 \in [\,\theta_0, +\infty), \\ 0 & \text{if } X_1 \in (-\infty, \theta_0). \end{cases} \tag{7}$$

Combining (6) with (7), the joint density of $(Y, X_1, \ldots, X_q)$ becomes

$$\begin{aligned} f(y, x_1, \ldots, x_q \,|\, \boldsymbol{D}, \boldsymbol{L}, \theta_0) &= f(x_1, \ldots, x_q \,|\, \boldsymbol{D}, \boldsymbol{L}) \cdot \mathbb{1}(\theta_{y-1} < x_1 \leq \theta_y) \\ &= \left\{ \prod_{j=1}^{q} d\mathcal{N}(x_j \,|\, -\boldsymbol{L}_{\prec j]}^{\top} \boldsymbol{x}_{\mathrm{pa}(j)}, \sigma_j^2) \right\} \cdot \mathbb{1}(\theta_{y-1} < x_1 \leq \theta_y), \end{aligned} \tag{8}$$

where $\theta_{-1} = -\infty, \theta_1 = +\infty$. Equation (8) defines a (Gaussian) *DAG-probit model*. A related expression appears in Guo et al. (2015) who model a multivariate distribution of ordered categorical variables through a collection of Gaussian random variables Markov with respect to an undirected graphical model. Now recall from (6) that the conditional distribution of the latent variable X_1 is $\mathcal{N}(-\boldsymbol{L}_{\prec 1]}^{\top} \boldsymbol{x}_{\mathrm{pa}(1)}, \sigma_1^2)$ and, as in standard probit regression, we set $\sigma_1^2 = 1$ for identifiability reasons.

Finally, by considering n independent samples $(y_i, x_{i,2}, \ldots, x_{i,q})$, $i = 1, \ldots, n$, from (8), the *augmented* likelihood can be written as

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{X} \,|\, \boldsymbol{D}, \boldsymbol{L}, \theta_0) &= \prod_{i=1}^{n} f(x_{i,1}, \ldots, x_{i,q} \,|\, \boldsymbol{D}, \boldsymbol{L}) \cdot \mathbb{1}(\theta_{y_i - 1} < x_{i,1} \leq \theta_{y_i}) \\ &= \prod_{j=1}^{q} d\mathcal{N}_n(\boldsymbol{X}_j \,|\, -\boldsymbol{X}_{\mathrm{pa}(j)} \boldsymbol{L}_{\prec j]}, \sigma_j^2 \boldsymbol{I}_n) \cdot \left\{ \prod_{i=1}^{n} \mathbb{1}(\theta_{y_i - 1} < x_{i,1} \leq \theta_{y_i}) \right\}, \end{aligned} \tag{9}$$

where $\boldsymbol{y} = (y_1 \ldots, y_n)^{\top}$, $\boldsymbol{X}$ is the (n, q) augmented data matrix, and $\boldsymbol{X}_A$ is the submatrix of $\boldsymbol{X}$ corresponding to the set A of columns of $\boldsymbol{X}$.

3 Causal Effects

Consider the joint density of the random vector $(X_1, \ldots, X_q)$ Markov w.r.t. a DAG which factorizes as in (1); the latter is referred to as the *observational* (or *pre-intervention*) distribution.

We now introduce the notion of *intervention*. A deterministic intervention on variable X_s, $s \in \{2, \ldots, q\}$ is denoted by $\mathrm{do}(X_s = \tilde{x})$ and consists in setting X_s to the value $\tilde{x}$. The *post-intervention* density of $(X_1, \ldots, X_q)$ is then obtained using the truncated factorization

$$f(x_1, \ldots, x_q \,|\, \mathrm{do}(X_s = \tilde{x})) = \begin{cases} \prod_{j=1, j \neq s}^{q} f(x_j \,|\, \boldsymbol{x}_{\mathrm{pa}(j)})|_{x_s = \tilde{x}} & \text{if } x_s = \tilde{x}, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where, importantly, each term $f(x_j \,|\, \cdot)$ in (10) is the corresponding (pre-intervention) conditional distribution of equation (1); see Pearl (2000). We emphasize that the post-intervention density $f(x_1, \ldots, x_q \,|\, \mathrm{do}(X_s = \tilde{x}))$ is conceptually distinct from the usual *conditional* density $f(x_1, \ldots, x_q \,|\, X_s = \tilde{x})$, which arises out of passive observation of

$X_s = \tilde{x}$. An important feature of equation (10) is that the data generating system is "stable" under exogenous interventions, in the sense that only the local component distribution $f(x_s \,|\, \boldsymbol{x}_{\text{pa}(s)})$ is affected by the intervention and effectively reduces to a point mass on $\tilde{x}$. All the remaining terms are immune to the intervention and thus remain the same. The post-intervention distribution of the (latent) response X_1 is then obtained by integrating (10) w.r.t. $x_2, \ldots, x_q$ which simplifies to

$$f(x_1 \,|\, \text{do}(X_s = \tilde{x})) = \int f(x_1 \,|\, \tilde{x}, \boldsymbol{x}_{\text{pa}(s)}) f(\boldsymbol{x}_{\text{pa}(s)})\, d\boldsymbol{x}_{\text{pa}(s)}; \tag{11}$$

see also Pearl (2000, Theorem 3.2.2).

We now move back to the Gaussian setting of Section 2.1, and assume that $(X_1, X_2, \ldots, X_q) \,|\, \boldsymbol{\Sigma} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ is Markov w.r.t. the underlying DAG. The post-intervention distribution of X_1 can thus be written as

$$\begin{aligned} f(x_1 \,|\, \text{do}(X_s = \tilde{x}), \boldsymbol{\Sigma}) &= \int f(x_1 \,|\, \tilde{x}, \boldsymbol{x}_{\text{pa}(s)}, \boldsymbol{\Sigma}) \cdot f(\boldsymbol{x}_{\text{pa}(s)} \,|\, \boldsymbol{\Sigma})\, d\boldsymbol{x}_{\text{pa}(s)} \\ = \int d\mathcal{N}(x_1 \,|\, \gamma_s \tilde{x} + \boldsymbol{\gamma}^\top \boldsymbol{x}_{\text{pa}(s)}, \delta_1^2) \cdot d\mathcal{N}(\boldsymbol{x}_{\text{pa}(s)} \,|\, \mathbf{0}, \boldsymbol{\Sigma}_{\text{pa}(s),\text{pa}(s)})\, d\boldsymbol{x}_{\text{pa}(s)}, \end{aligned} \tag{12}$$

where $\delta_1^2 = \mathbb{V}\text{ar}(X_1 \,|\, X_s = \tilde{x}, \boldsymbol{x}_{\text{pa}(s)}, \boldsymbol{\Sigma})$. The following Proposition gives the analytic form of the post-intervention distribution of X_1.

Proposition 3.1. *Let* $(X_1, X_2, \ldots, X_q) \,|\, \boldsymbol{\Sigma} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ *and consider the* do *operator* $\text{do}(X_s = \tilde{x})$, $s \in \{2, \ldots, q\}$. *Then the post-intervention distribution of* X_1 *is*

$$f(x_1 \,|\, \text{do}(X_s = \tilde{x}), \boldsymbol{\Sigma}) = d\mathcal{N}\left(x_1 \,|\, \gamma_s \tilde{x}, \frac{\delta_1^2}{1 - (\boldsymbol{\gamma}^\top \boldsymbol{T}^{-1} \boldsymbol{\gamma})/\delta_1^2}\right),$$

where

$$\begin{aligned} \delta_1^2 &= \boldsymbol{\Sigma}_{1 \,|\, \text{fa}(s)}, \\ (\gamma_s, \boldsymbol{\gamma}^\top)^\top &= \boldsymbol{\Sigma}_{1,\text{fa}(s)} \left(\boldsymbol{\Sigma}_{\text{fa}(s),\text{fa}(s)}\right)^{-1}, \\ \boldsymbol{T} &= \left(\boldsymbol{\Sigma}_{\text{pa}(s),\text{pa}(s)}\right)^{-1} + \frac{1}{\delta_1^2} \boldsymbol{\gamma} \boldsymbol{\gamma}^\top, \end{aligned}$$

with the understanding that node s *occupies the first position in the set* $\text{fa}(s)$.

Proof. See Supplementary material (Castelletti and Consonni, 2021b). □

The previous reasoning considered the intervention distribution of the latent response variable X_1 following $\text{do}(X_s = \tilde{x})$. Typically distribution (11) is summarized by its expected value $\mathbb{E}(X_1 \,|\, \text{do}(X_s = \tilde{x}))$. When X_s is continuous, one can define the (total) causal effect as the derivative of $\mathbb{E}(X_1 \,|\, \text{do}(X_s = x))$ w.r.t. x evaluated at $\tilde{x}$: this is especially convenient when the expectation is linear, as in the Gaussian case (12), because the causal effect admits a simple interpretation: it corresponds to the "regression parameter" γ_s associated to variable X_s (Maathuis et al., 2009). Our interest

however lies in the observable response variable Y, and therefore we aim to evaluate $\mathbb{E}(Y \mid \text{do}(X_s = \tilde{x}), \boldsymbol{\Sigma}, \theta_0)$. We thus obtain

$$\begin{aligned} \mathbb{E}(Y \mid \text{do}(X_s = \tilde{x}), \boldsymbol{\Sigma}, \theta_0) &= \Pr(Y = 1 \mid \text{do}(X_s = \tilde{x}), \boldsymbol{\Sigma}, \theta_0) \\ &= \Pr(X_1 \geq \theta_0 \mid \text{do}(X_s = \tilde{x}), \boldsymbol{\Sigma}) \\ &= 1 - \Phi\left(\frac{\theta_0 - \gamma_s \tilde{x}}{\tau_1}\right) \equiv \beta_s(\tilde{x}, \boldsymbol{\Sigma}, \theta_0), \end{aligned} \tag{13}$$

where $\Phi(\cdot)$ denotes the c.d.f. of a standard normal and $\tau_1^2 = \delta_1^2 / \left(1 - (\boldsymbol{\gamma}^\top \boldsymbol{T}^{-1} \boldsymbol{\gamma})/\delta_1^2\right)$. One could then compute the partial derivative of $\mathbb{E}(Y \mid \text{do}(X_s = \tilde{x}), \boldsymbol{\Sigma}, \theta_0)$ w.r.t x evaluated at $\tilde{x}$, and obtain $\phi(\theta_0 - \gamma_s \tilde{x}/\tau_1)\gamma_s/\tau_1$, where $\phi(\cdot)$ is the density function of a standard normal. This however would still depend on $\tilde{x}$. For this reason, and because (13) enjoys an intuitive interpretation being a probability, we will simply denote $\Pr(Y = 1 \mid \text{do}(X_s = \tilde{x}), \boldsymbol{\Sigma}, \theta_0)$ at the causal effect on Y due to an intervention $\text{do}(X_s = \tilde{x})$. Finally, we remark that the causal effect of $\text{do}(X_s = \tilde{x})$ on Y, besides being a function of the value $\tilde{x}$, depends on θ_0 as well as on the covariance matrix $\boldsymbol{\Sigma}$, where the latter is constrained to be Markov w.r.t. the underlying DAG.

4 Bayesian Inference

In this section we introduce priors for $(\boldsymbol{\Omega}, \theta_0, \mathcal{D})$, which we structure as $p(\boldsymbol{\Omega}, \theta_0, \mathcal{D}) = p(\boldsymbol{\Omega} \mid \mathcal{D}) p(\mathcal{D}) p(\theta_0)$. Further distributional results useful for our MCMC scheme of Section 5 are also presented. We briefly preview here the essential features.

To start with, consider $p(\boldsymbol{\Omega} \mid \mathcal{D})$, $\boldsymbol{\Omega} \in \mathcal{P}_{\mathcal{D}}$. We first proceed to the re-parameterization $\boldsymbol{\Omega} \mapsto (\boldsymbol{D}, \boldsymbol{L})$ presented in Subsection 2.1, and specify a DAG-Wishart prior (Cao et al., 2019) on the Cholesky parameters $(\boldsymbol{D}, \boldsymbol{L})$. We achieve this goal using a highly parsimonious elicitation procedure, which we briefly detail in Section 4.1; see also our Supplementary material for more information. For the unknown threshold $\theta_0 \in (-\infty, +\infty)$, we assume a uniform prior, so that $p(\theta_0) \propto 1$ (Section 4.3). Finally, a prior on DAG $\mathcal{D}$ is assigned through independent Bernoulli distributions on the elements of the skeleton of $\mathcal{D}$ (Section 4.2).

4.1 Prior on the Cholesky Parameters

Consider first a DAG $\mathcal{D} = (V, E)$ which is *complete*, so that the precision matrix $\boldsymbol{\Omega}$ is unconstrained. A standard conjugate prior is the Wishart distribution, $\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{U})$ having expectation $a\boldsymbol{U}^{-1}$, where $a > q - 1$ and $\boldsymbol{U}$ is a s.p.d. matrix. In absence of substantive prior information, a standard choice for the hyperparameter $\boldsymbol{U}$, hereinafter adopted, is $\boldsymbol{U} = g\,\boldsymbol{I}_q$, where $g > 0$ and $\boldsymbol{I}_q$ is the (q, q) identity matrix. The induced prior on the Cholesky parameters consistent with the DAG parent ordering is such that the node parameters $(\sigma_j^2, \boldsymbol{L}_{\prec j\,]})$, $j = 1, \ldots, q$, are independent with distribution

$$\sigma_j^2 \quad \sim \quad \text{I-Ga}\left(\frac{a_j}{2} - \frac{|\text{pa}(j)|}{2} - 1, \frac{g}{2}\right),$$

$$\boldsymbol{L}_{\prec j]} \,|\, \sigma_j^2 \;\sim\; \mathcal{N}_{|\mathrm{pa}(j)|}\left(\mathbf{0}, \frac{1}{g}\sigma_j^2\, \boldsymbol{I}_{|\mathrm{pa}(j)|}\right), \tag{14}$$

where $|A|$ is the number of elements in the set A, $a_j = a + q - 2j + 3$; see Ben-David et al. (2015, Supplemental B). The symbol I-Ga(a, b) stands for an Inverse-Gamma distribution with shape $a > 0$ and rate $b > 0$ having expectation $b/(a-1)$ $(a > 1)$. In addition, to guarantee the identifiability of the DAG-probit model, we fix $\sigma_1^2 = 1$, so that instead of $p(\sigma_1^2, \boldsymbol{L}_{\prec 1]})$ we need only to consider $p(\boldsymbol{L}_{\prec 1]})$ with $\boldsymbol{L}_{\prec 1]} \sim \mathcal{N}_{|\mathrm{pa}(1)|}(\mathbf{0}, g^{-1}\boldsymbol{I}_{|\mathrm{pa}(1)|})$; see also Section 2.2. Recall that (14) applies only to a complete DAG $\mathcal{D}$.

Consider now the case in which $\mathcal{D}$ is not complete. The idea is to leverage (14) to set up a general method to construct a prior on $(\boldsymbol{D}, \boldsymbol{L})$, the Cholesky parameters of $\boldsymbol{\Omega} \in \mathcal{P}_{\mathcal{D}}$, which can then be applied to *any* given DAG $\mathcal{D}$. To this end we follow the procedure of Geiger and Heckerman (2002), which we detail in the Supplementary material. For a given DAG $\mathcal{D}$ we show that the prior assigned to its Cholesky parameters is

$$\begin{aligned} \sigma_j^2 \;&\sim\; \text{I-Ga}\left(\frac{a_j^{\mathcal{D}}}{2}, \frac{g}{2}\right), \\ \boldsymbol{L}_{\prec j]} \,|\, \sigma_j^2 \;&\sim\; \mathcal{N}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\left(\mathbf{0}, \frac{1}{g}\sigma_j^2\, \boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\right), \end{aligned} \tag{15}$$

independently for $j =, \ldots, q$, where $a_j^{\mathcal{D}} = a + |\mathrm{pa}_{\mathcal{D}}(j)| - q + 1$. Finally we can write

$$p(\boldsymbol{D}, \boldsymbol{L}) = \prod_{j=1}^{q} p(\sigma_j^2, \boldsymbol{L}_{\prec j]}), \quad (\boldsymbol{L}, \boldsymbol{D}) \in \Theta_{\mathcal{D}}, \tag{16}$$

where $\Theta_{\mathcal{D}}$ is the image of the space $\mathcal{P}_{\mathcal{D}}$ under the mapping $\boldsymbol{\Omega} \mapsto (\boldsymbol{D}, \boldsymbol{L})$.

4.2 Prior on DAG Space

For a given DAG $\mathcal{D}$, let $\boldsymbol{A}^{\mathcal{D}}$ be the (symmetric) 0-1 adjacency matrix of the skeleton of $\mathcal{D}$ whose (u, v) element is denoted by $\boldsymbol{A}^{\mathcal{D}}_{(u,v)}$. Conditionally on the edge inclusion probability π, we first assign a Bernoulli prior independently to each element $\boldsymbol{A}^{\mathcal{D}}_{(u,v)}$ belonging to the lower-triangular part, that is: $\boldsymbol{A}^{\mathcal{D}}_{(u,v)} \,|\, \pi \overset{iid}{\sim} \mathrm{Ber}(\pi), u > v$. As a consequence we get

$$p(\boldsymbol{A}^{\mathcal{D}}) = \pi^{|\boldsymbol{A}^{\mathcal{D}}|}(1-\pi)^{\frac{q(q-1)}{2} - |\boldsymbol{A}^{\mathcal{D}}|}, \tag{17}$$

where $|\boldsymbol{A}^{\mathcal{D}}|$ denotes the number of edges in the skeleton, equivalently the number of entries equal to one in the lower-triangular part of $\boldsymbol{A}^{\mathcal{D}}$. Finally, we set $p(\mathcal{D}) \propto p(\boldsymbol{A}^{\mathcal{D}})$, for $\mathcal{D} \in \mathcal{S}_q$, where $\mathcal{S}_q$ is the set of all DAGs on q nodes.

4.3 Posterior Distribution of θ_0

As mentioned, in absence of substantive prior information, we assign a flat improper prior to the threshold $\theta_0 \in (-\infty, \infty)$, $p(\theta_0) \propto 1$. Accordingly, we need to prove that the

posterior of θ_0 is proper. The next proposition details under which conditions propriety is guaranteed.

Proposition 4.1. *Under the prior (14) for $(\boldsymbol{D}, \boldsymbol{L})$, $p(\mathcal{D})$ as in Section 4.2 for DAG $\mathcal{D}$ and the improper prior $p(\theta_0) \propto 1$ for θ_0, the posterior of θ_0 is proper provided the sample contains at least two observations with distinct values for Y, that is $y_i = 1$, $y_{i'} = 0$ $(i \neq i')$.*

Proof. See Supplementary material (Castelletti and Consonni, 2021b). □

Additionally, we prove in the Supplementary material that under the conditions of Proposition 4.1 the joint posterior of $(\boldsymbol{D}, \boldsymbol{L}, \mathcal{D}, \theta_0, \boldsymbol{X}_1)$ is proper. Clearly, alternative priors for θ_0 might have been used; yet the full conditional of θ_0 would not be amenable to direct sampling. As a consequence, posterior inference on θ_0 is performed through a Metropolis Hastings step inside our MCMC scheme; see Section 5 for details.

5 MCMC Scheme

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

$$p(\boldsymbol{D}, \boldsymbol{L}, \mathcal{D}, \theta_0, \boldsymbol{X}_1 \,|\, \boldsymbol{y}, \boldsymbol{X}_{-1}) \propto f(\boldsymbol{y}, \boldsymbol{X} \,|\, \boldsymbol{D}, \boldsymbol{L}, \mathcal{D}, \theta_0)\, p(\boldsymbol{D}, \boldsymbol{L} \,|\, \mathcal{D})\, p(\mathcal{D}), \tag{18}$$

now emphasizing the dependence on DAG $\mathcal{D}$, where $\boldsymbol{X}_{-1} = (\boldsymbol{X}_2, \ldots, \boldsymbol{X}_q)$, and the term $p(\theta_0)$ has been omitted because it is proportional to one.

5.1 Update of $(\boldsymbol{D}, \boldsymbol{L}, \mathcal{D})$

From (18) the full conditional distribution of $(\boldsymbol{D}, \boldsymbol{L}, \mathcal{D})$ is

$$p(\boldsymbol{D}, \boldsymbol{L}, \mathcal{D} \,|\, \boldsymbol{y}, \boldsymbol{X}, \theta_0) \propto p(\boldsymbol{X} \,|\, \boldsymbol{D}, \boldsymbol{L}, \mathcal{D}) p(\boldsymbol{D}, \boldsymbol{L} \,|\, \mathcal{D}) p(\mathcal{D})$$

using (9), where $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_q)$ is the (n, q) augmented data matrix.

To sample from $p(\boldsymbol{D}, \boldsymbol{L}, \mathcal{D} \,|\, \boldsymbol{X})$ we adopt a reversible jump MCMC algorithm which takes into account the partial analytic structure (PAS, Godsill 2012) of the DAG-Wishart distribution to sample DAG $\mathcal{D}$ and the Cholesky parameters $(\boldsymbol{D}, \boldsymbol{L})$ from their full conditional. A similar approach was implemented in Wang and Li (2012) for Gaussian undirected graphical models using G-Wishart priors. Details about the PAS algorithm and its implementation in our DAG setting are reported in the Supplementary material (Castelletti and Consonni, 2021b).

Specifically, at each iteration of the MCMC scheme, we first propose a new DAG $\mathcal{D}'$ from a suitable proposal distribution $q(\mathcal{D}' \,|\, \mathcal{D})$; see again our Supplementary material. In particular, it is shown that when proposing a DAG $\mathcal{D}'$ which differs from the current graph $\mathcal{D}$ by one edge (h, j), the acceptance probability for $\mathcal{D}'$ is given by

$$\alpha_{\mathcal{D}'} = \min\left\{1, \frac{m(\boldsymbol{X}_j \,|\, \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}'}(j)}, \mathcal{D}')}{m(\boldsymbol{X}_j \,|\, \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}, \mathcal{D})} \cdot \frac{p(\mathcal{D}')}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D} \,|\, \mathcal{D}')}{q(\mathcal{D}' \,|\, \mathcal{D})}\right\} \tag{19}$$

where, for $j \in \{2, \dots, q\}$,

$$m(\boldsymbol{X}_j \,|\, \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}, \mathcal{D}) = (2\pi)^{-\frac{n}{2}} \frac{\left|\boldsymbol{T}_j\right|^{1/2}}{\left|\bar{\boldsymbol{T}}_j\right|^{1/2}} \cdot \frac{\Gamma\left(\frac{a_j^{\mathcal{D}}}{2} + \frac{n}{2}\right)}{\Gamma\left(a_j^{\mathcal{D}}/2\right)} \left[\frac{1}{2}g\right]^{a_j^{\mathcal{D}}/2} \left[\frac{1}{2}\big(g + \boldsymbol{X}_j^\top \boldsymbol{X}_j - \hat{\boldsymbol{L}}_j^\top \bar{\boldsymbol{T}}_j \hat{\boldsymbol{L}}_j\big)\right]^{-(a_j^{\mathcal{D}}+n)/2} \tag{20}$$

with

$$\begin{aligned} \boldsymbol{T}_j &= g\boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|} \\ \bar{\boldsymbol{T}}_j &= g\boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|} + \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}^\top \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)} \\ \hat{\boldsymbol{L}}_j &= \big(g\boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|} + \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}^\top \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}\big)^{-1} \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}^\top \boldsymbol{X}_j, \end{aligned}$$

$a_j^{\mathcal{D}} = a + |\mathrm{pa}_{\mathcal{D}}(j)| - q + 1$ and $\boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}$ denotes the $(n, |\mathrm{pa}_{\mathcal{D}}(j)|)$ sub-matrix of $\boldsymbol{X}$ whose columns belong to the set $\mathrm{pa}_{\mathcal{D}}(j)$. For $j = 1$, because we fixed $\sigma_1^2 = 1$, we have instead

$$m(\boldsymbol{X}_1 \,|\, \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(1)}, \mathcal{D}) = (2\pi)^{-\frac{n}{2}} \frac{\left|\boldsymbol{T}_1\right|^{1/2}}{\left|\bar{\boldsymbol{T}}_1\right|^{1/2}} \cdot \exp\left\{-\frac{1}{2}\big(\boldsymbol{X}_1^\top \boldsymbol{X}_1 - \hat{\boldsymbol{L}}_1^\top \bar{\boldsymbol{T}}_1 \hat{\boldsymbol{L}}_1\big)\right\}, \tag{21}$$

with $\boldsymbol{T}_1, \bar{\boldsymbol{T}}_1, \hat{\boldsymbol{L}}_1$ defined in analogy with the expressions appearing after (20); see the Supplementary material (Castelletti and Consonni, 2021b) for details. Moreover, given DAG $\mathcal{D}$ and $\boldsymbol{X}_1$, the full conditional of $(\boldsymbol{D}, \boldsymbol{L})$ reduces to the augmented posterior $p(\boldsymbol{D}, \boldsymbol{L} \,|\, \boldsymbol{X})$, which is conditional on the actual data $(\boldsymbol{X}_2, \dots, \boldsymbol{X}_q)$ as well as the latent values $\boldsymbol{X}_1$ and can be easily sampled from. Specifically, since

$$f(\boldsymbol{X} \,|\, \boldsymbol{D}, \boldsymbol{L}) = \prod_{j=1}^{q} d\mathcal{N}_n(\boldsymbol{X}_j \,|\, -\boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)} \boldsymbol{L}_{\prec j]}, \sigma_j^2 \boldsymbol{I}_n) \tag{22}$$

and because of (16) and conjugacy of the Normal-Inverse-Gamma prior in (14) with the Normal density, the posterior distribution of the Cholesky parameters given $\boldsymbol{X}$ is, for $j = 2, \dots, q$,

$$\begin{aligned} \sigma_j^2 \,|\, \boldsymbol{X} &\sim \text{I-Ga}\left(\frac{a_j^{\mathcal{D}}}{2} + \frac{n}{2}, \frac{1}{2}\big(g + \boldsymbol{X}_j^\top \boldsymbol{X}_j - \hat{\boldsymbol{L}}_j^\top \bar{\boldsymbol{T}}_j \hat{\boldsymbol{L}}_j\big)\right), \\ \boldsymbol{L}_{\prec j]} \,|\, \sigma_j^2, \boldsymbol{X} &\sim \mathcal{N}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\big(-\hat{\boldsymbol{L}}_j, \sigma_j^2 \bar{\boldsymbol{T}}_j^{-1}\big). \end{aligned} \tag{23}$$

Moreover, for node 1 where $\sigma_1^2 = 1$, we have

$$\boldsymbol{L}_{\prec 1]} \,|\, \boldsymbol{X} \sim \mathcal{N}_{|\mathrm{pa}_{\mathcal{D}}(1)|}\big(-\hat{\boldsymbol{L}}_1, \bar{\boldsymbol{T}}_1^{-1}\big). \tag{24}$$

5.2 Update of $\boldsymbol{X}_1$ and $\boldsymbol{\theta}_0$

Updating of $\boldsymbol{X}_1 = (x_{1,1}, \dots, x_{n,1})^\top$ can be performed by direct sampling from the full conditional distribution of each latent observation $x_{i,1}$,

$$f(x_{i,1} \,|\, y_i, x_{i,2}, \dots, x_{i,q}, \boldsymbol{D}, \boldsymbol{L}, \mathcal{D}, \theta_0) \propto f(x_{i,1} \,|\, \boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(1)}, \boldsymbol{L}_{\prec j]}) \cdot \mathbb{1}(\theta_{y_i - 1} < x_{i,1} \leq \theta_{y_i}),$$

which corresponds to a $\mathcal{N}(-\boldsymbol{L}_{\prec 1]}^{\top}\boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(1)},1)$ truncated at the interval $(\theta_{y_i-1},\theta_{y_i}]$.

Finally, the cut-off θ_0 is updated through a Metropolis Hastings step where, given the current value θ_0, a candidate value g_0 is proposed from $q(g_0\,|\,\theta_0)=d\mathcal{N}(g_0\,|\,\theta_0,\sigma_0^2)$. We then set $\theta_0=g_0$ with probability

$$\alpha_\theta=\min\{1;r_\theta\}\,, \tag{25}$$

where

$$r_\theta=\frac{\prod_{i=1}^{n}\left[\Phi\big(g_{y_i}\,|\,-\boldsymbol{L}_{\prec 1]}^{\top}\boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(1)},1\big)-\Phi\big(g_{y_i-1}\,|\,-\boldsymbol{L}_{\prec 1]}^{\top}\boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(1)},1\big)\right]}{\prod_{i=1}^{n}\left[\Phi\big(\theta_{y_i}\,|\,-\boldsymbol{L}_{\prec 1]}^{\top}\boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(1)},1\big)-\Phi\big(\theta_{y_i-1}\,|\,-\boldsymbol{L}_{\prec 1]}^{\top}\boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}}(1)},1\big)\right]}\cdot\frac{d\mathcal{N}\left(\theta_0\,|\,g_0,\sigma_0^2\right)}{d\mathcal{N}\left(g_0\,|\,\theta_0,\sigma_0^2\right)},$$

and $g_{-1}=\infty, g_1=+\infty$.

5.3 Algorithm

Algorithm 1 summarizes our MCMC scheme. The output is a collection of DAGs $\big\{\mathcal{D}^{(t)}\big\}_{t=1}^{T}$ and Cholesky parameters $\big\{\big(\boldsymbol{D}^{\mathcal{D}^{(t)}},\boldsymbol{L}^{\mathcal{D}^{(t)}}\big)\big\}_{t=1}^{T}$ approximatively sampled from the target distribution (18). In particular we can compute posterior summaries of interest such as the posterior probabilities of edge inclusion, namely

$$\hat{p}_{u\to v}(\boldsymbol{y},\boldsymbol{X}_2,\ldots,\boldsymbol{X}_q)\equiv\hat{p}_{u\to v}=\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{u\to v}\big\{\mathcal{D}^{(t)}\big\}, \tag{26}$$

where $\mathbb{1}_{u\to v}\big\{\mathcal{D}^{(t)}\big\}$ takes value 1 if and only if $\mathcal{D}^{(t)}$ contains the edge $u\to v$. Moreover, we can reconstruct the covariance matrices $\big\{\boldsymbol{\Sigma}^{\mathcal{D}^{(t)}}\big\}_{t=1}^{T}$ using (5). The latter can be subsequently retrieved to obtain for selected $s\in\{2,\ldots,q\}$ and intervention value $\tilde{x}$ the collection of causal effects $\big\{\beta_s^{(t)}(\tilde{x})\big\}_{t=1}^{T}$ defined in (13), where we set $\beta_s^{(t)}(\tilde{x})\equiv\beta_s\big(\tilde{x},\boldsymbol{\Sigma}^{\mathcal{D}^{(t)}},\theta_0^{(t)}\big)$. An overall summary of the causal effect of $\mathrm{do}(X_s=\tilde{x})$ on Y can be computed as

$$\hat{\beta}_s^{BMA}(\tilde{x})=\frac{1}{T}\sum_{t=1}^{T}\beta_s^{(t)}(\tilde{x}), \tag{27}$$

which corresponds to a Bayesian Model Averaging (BMA) estimate where posterior (DAG) model probabilities are approximated through the MCMC frequencies of visits; see García-Donato and Martínez-Beneito (2013) for a discussion of the merits of frequency based estimators in large model spaces.

6 Simulations

In this section we evaluate the performance of our method through simulation studies. Specifically, for each combination of number of nodes $q\in\{20,40\}$ and sample size

Algorithm 1: MCMC scheme to sample from (18).

Input: A dataset $(\boldsymbol{y}, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_q)$
Output: T samples from the posterior (18)

Initialize $\mathcal{D}^{(0)}$, e.g. the empty DAG, the cut-offs $\theta_{-1}^{(0)} = -\infty, \theta_0^{(0)} = 0, \theta_1^{(0)} = +\infty$ and the latent variables $\boldsymbol{x}_1^{(0)}$, e.g. $x_{i,1}^{(0)} \overset{ind}{\sim} \mathcal{N}(0,1)$ truncated at $(\theta_{y_i-1}^{(0)}, \theta_{y_i}^{(0)}]$;
for $t = 1, \ldots, T$ **do**
  Sample $\mathcal{D}'$ from $q(\mathcal{D}' \,|\, \mathcal{D}^{(t-1)})$ and set $\mathcal{D}^{(t)} = \mathcal{D}'$ with probability $\alpha_{\mathcal{D}}$ (19), otherwise $\mathcal{D}^{(t)} = \mathcal{D}^{(t-1)}$;
  Sample $\left(\boldsymbol{D}^{\mathcal{D}^{(t)}}, \boldsymbol{L}^{\mathcal{D}^{(t)}}\right)$ from its augmented posterior distribution (23);
  For $i = 1, \ldots, n$, independently sample $x_{i,1}$ from $\mathcal{N}\big(-\boldsymbol{L}_{\prec 1]}^{\mathcal{D}^{(t)}\top} \boldsymbol{x}_{i,\mathrm{pa}_{\mathcal{D}^{(t)}}(1)}, 1\big)$ truncated at $(\theta_{y_i-1}^{(t)}, \theta_{y_i}^{(t)}]$;
  Propose a cut-off g_0 from $q(g_0 \,|\, \theta_0^{(t)})$ and set $\theta_0^{(t)} = g_0$ with probability α_θ (25), otherwise $\theta_0^{(t)} = \theta_0^{(t-1)}$; set $\theta_{-1}^{(t)} = -\infty, \theta_1^{(t)} = +\infty$.
end

$n \in \{100, 200, 500\}$, which we call simulation *scenario*, we generate 40 DAGs using a probability of edge inclusion equal to $p = 3/(2q-2)$ to induce sparsity; see Peters and Bühlmann (2014). For each DAG $\mathcal{D}$ we then proceed as follows. We identify a parent ordering and fix $\boldsymbol{D}^{\mathcal{D}} = \boldsymbol{I}_q$ and then randomly sample the entries of $\boldsymbol{L}^{\mathcal{D}}$ in the interval $[-2,-1] \cup [1,2]$; next we generate a dataset consisting of n i.i.d. q-dimensional observations from the structural equation model (4) which also includes the $(n,1)$ vector of latent observations; we finally fix the threshold $\theta_0 = 0$ and obtain the 0-1 vector of responses $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ as in (7).

We apply Algorithm 1 to approximate the target distribution in (18) by setting the number of MCMC iterations $T = 25000$ for $q = 20$, and $T = 50000$ for $q = 40$. We also set $g = 1/n$ and $a = q$ in the prior on the Cholesky parameters of $\boldsymbol{\Omega}$ (14) and $\sigma_0^2 = 0.25$ in the proposal density for the cut-off θ_0.

We begin by evaluating the global performance of our method in learning the graph structure. To this end, we first estimate the posterior probabilities of edge inclusion by computing $\hat{p}_{u \to v}(\cdot)$ in (26) for each pair of distinct nodes u, v. Next, we consider a threshold for edge inclusion $k \in [0,1]$ and for a given k construct a graph estimate by including those edges (u,v) whose posterior probability exceeds k. The resulting graph is compared with the true DAG through the sensitivity (SEN) and specificity (SPE) indexes, respectively defined as

$$SEN = \frac{TP}{TP + FN}, \quad SPE = \frac{TN}{TN + FP},$$

where TP, TN, FP, FN are the numbers of true positives, true negatives, false positives and false negatives, based on the adjacency matrix of the estimated graph.

	$n = 100$	$n = 200$	$n = 500$
$q = 20$	93.89	94.19	95.12
$q = 40$	90.94	94.91	97.19

Table 1: Simulations. Area under the curve (percentage values) computed from the average ROC curves in Figure 1 for number of nodes $q \in \{20, 40\}$ and sample sizes $n \in \{100, 200, 500\}$.

The two indexes are used to construct a receiver operating characteristic (ROC) curve. Specifically, for each scenario defined by q and n, we present a ROC curve constructed as follows. For each threshold k, we compute SEN and $(1 - SPE)$ under each of the 40 DAGs used in the simulation. The point whose coordinates are the mean of each of the two measures corresponds to one dot in Figure 1. The collection of dots connected by lines represents an average ROC curve. We proceed similarly to compute the 5th and 95th percentile and obtain the grey band.

To better appreciate Figure 1, we also compute, for each simulation scenario (q, n), the area under the (average) ROC curve (AUC) whose values are reported in Table 1. They are close or above 94% under the three sample sizes considered when $q = 20$. When $q = 40$ AUC exceeds 90% for $n = 100$ and rises to over 97% for $n = 500$.

A more specific check on the ability of our method in recovering the structure of the underlying DAG can be considered. Since Y is the response, interest centers on the causal effect on Y following an intervention on a variable in the system. A natural group of intervention variables is represented by the set of parents of the latent node X_1 either because they directly influence X_1 (and hence Y) or because they act as intermediate nodes along a pathway originating from a variable upstream in the graph. To this end, under each simulation scenario, we fix the threshold for edge inclusion $k^* = 0.5$ and include those edges $u \to 1$ such that $\hat{p}_{u \to 1}(\cdot) \geq 0.5$ in analogy with the median probability model of Barbieri and Berger (2004). The resulting 0-1 vector of indicators for edge inclusion is $\boldsymbol{a} = (a_{1,1}, \ldots, a_{q,1})^\top$, where $a_{1,1} = 0$ while, for $u = 2, \ldots, q$, $a_{u,1} = 1$ if $u \to 1$ is included, 0 otherwise. Next we compute the proportion of predictors that are correctly classified,

$$p^* = \frac{1}{q-1} \sum_{u=2}^{q} \mathbb{1}\{a_{u,1} = \boldsymbol{A}^{\mathcal{D}}_{(u,1)}\},$$

where $\boldsymbol{A}^{\mathcal{D}}_{(u,v)}$ denotes the (u, v) element of the adjacency matrix of $\mathcal{D}$. The results are summarized in the box-plots of Figure 2 where we report the frequency distribution of p^* computed over the 40 true DAGs. While for $n = 100$ the proportion of correctly classified edges presents some variability with a median which is nevertheless around 80% ($q = 40$) and 90% ($q = 20$), the performance greatly improves as the sample size increases with practically all values being close to 1.

We now focus on causal effect estimation. Under each simulated DAG $\mathcal{D}$ and parameters $(\boldsymbol{D}^{\mathcal{D}}, \boldsymbol{L}^{\mathcal{D}})$ we first compute the (true) covariance matrix $\boldsymbol{\Sigma}^{\mathcal{D}}$ using (5). Now recall from (13) that the causal effect on Y is a probability which also depends on the

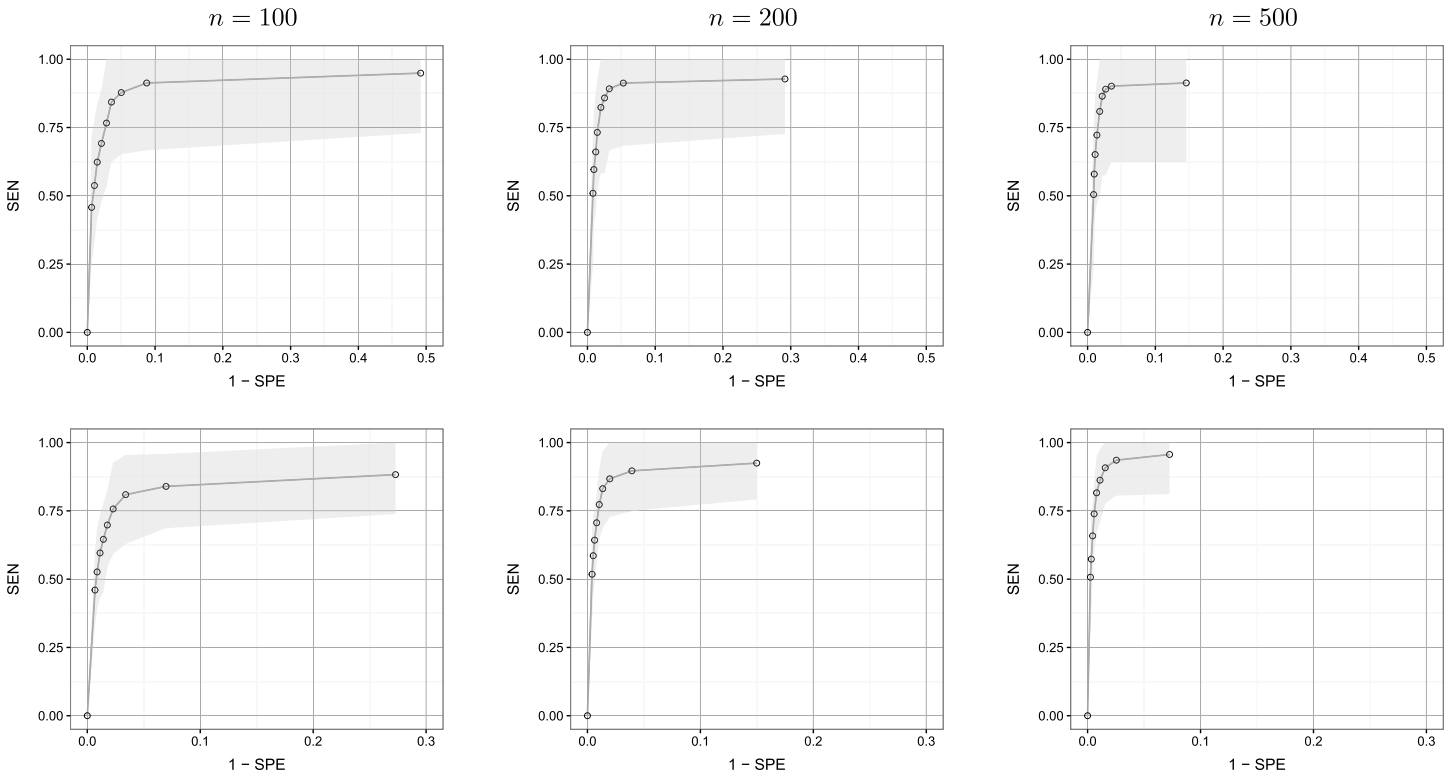

Figure 1: Simulations. Receiver operating characteristic (ROC) curve obtained under varying thresholds for the posterior probabilities of edge inclusion for each combination of number of nodes $q = \{20, 40\}$ (first and second row respectively) and sample size $n \in \{100, 200, 500\}$. Dots and connecting line describe the (average over the 40 simulated DAGs) ROC curve, while the grey area represents the 5th–95th percentile band.

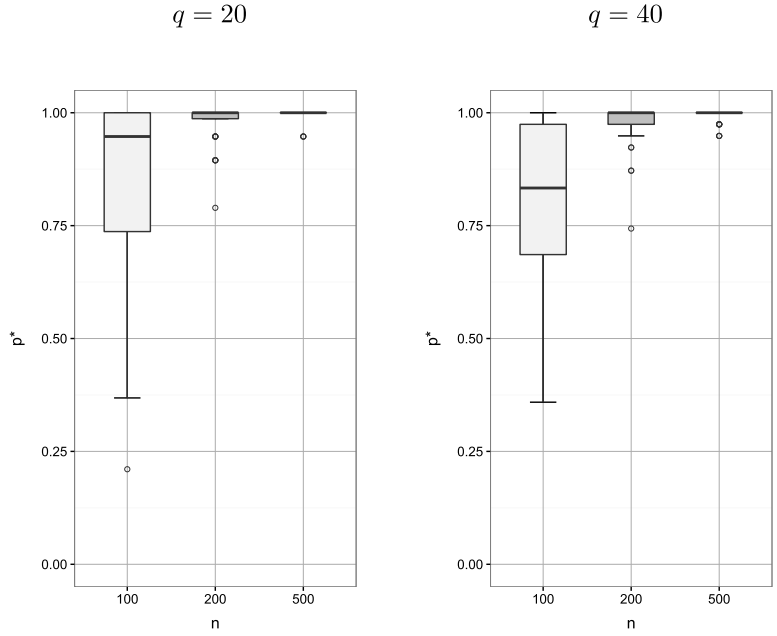

Figure 2: Simulations. Distribution across 40 simulated datasets of the proportion of predictors p^* that are correctly classified given a threshold for edge inclusion $k^* = 0.5$ for each combination of number of nodes $q \in \{20, 40\}$ and sample size $n \in \{100, 200, 500\}$.

level $\tilde{x}$ assigned to the intervened variable X_s. For each intervened node $s \in \{2, \ldots, q\}$ we evaluate $\beta_s(\tilde{x}, \boldsymbol{\Sigma}^{\mathcal{D}}, \theta_0) \equiv \beta_s^{true}(\tilde{x})$ at each observed value of X_s in the simulation scenario, $(x_{1,s}, \ldots, x_{n,s})$, and obtain the $(n, 1)$ vector of causal effects $\big(\beta_s^{true}(x_{1,s}), \ldots, \beta_s^{true}(x_{n,s})\big)^\top$. Next we produce the collection of BMA estimates $\hat{\beta}_s^{BMA}(x_{1,s}), \ldots, \hat{\beta}_s^{BMA}(x_{n,s})$ according to equation (27). To evaluate the performance of our method in estimating the causal effect we consider the differences $\big(\beta_s^{true}(x_{i,s}) - \hat{\beta}_s^{BMA}(x_{i,s})\big)$ and compute the mean absolute error (MAE)

$$MAE_s = \frac{1}{n} \sum_{i=1}^{n} \big|\beta_s^{true}(x_{i,s}) - \hat{\beta}_s^{BMA}(x_{i,s})\big|,$$

for each intervened node $s = 2, \ldots, q$. Results are summarized in the box-plots of Figure 3, where we report the distribution of the MAE (constructed across the 40 DAGs and nodes $s = 2, \ldots, q$) as a function of n. As expected, MAE decreases and approaches 0 as the sample size n grows for both values of q. Notice that the median value of MAE in the worst case scenario ($q = 20, n = 100$) is about half of one percent.

Finally, we also explore settings where $n \leq q$: in particular we include simulation results for $q = 40$ and $n \in \{10, 20, 40\}$. Again, we generate 40 DAGs and the allied parameters as in our first simulation study. Results are summarized in the box-plots

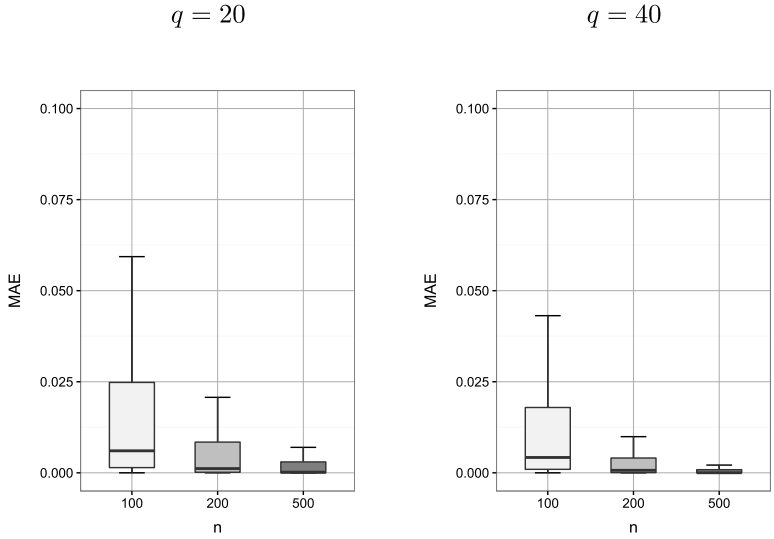

Figure 3: Simulations. Distribution over 40 datasets and nodes $s \in \{2, \ldots, q\}$ of the mean absolute error (MAE) of BMA estimates of true causal effects. Results are presented for each combination of number of nodes $q \in \{20, 40\}$ and sample size $n \in \{100, 200, 500\}$.

of Figure 4, where we report the distribution of the MAE, constructed across the 40 DAGs and nodes $s \in \{2, \ldots, q\}$ as a function of n. It appears that, even if sample sizes are moderate, MAE decreases as n grows.

7 Analysis of Gene Expressions from Breast Cancer Cells

In this section we apply our method to a gene expression dataset presented in Yin et al. (2014). The aim of the original study was to evaluate the ability of a gene signature derived from breast cancer stem cells to predict the risk of metastasis and survival in breast cancer patients. To this end, a collection of genes which are believed to be the main responsible for tumor initiation, progression, and response to therapy was considered. The study was motivated by recent literature establishing the existence of a rare population of cells, called *stem-like cells*, which supposedly represent the cellular origin of cancer; see for instance O'Brien et al. (2006). Gene-expression levels were measured on $n = 198$ breast cancer patients of which 62 manifested distant metastasis.

Evaluating the causal effect on Y due to an hypothetical intervention on a specific gene which sets its expression level may help understand which genes are particularly relevant for determining distant metastasis. This in turn can be useful to identify epigenetic modifications capable of setting genes "on" or "off"; see for instance Abdul et al. (2017) and Campbell et al. (2017).

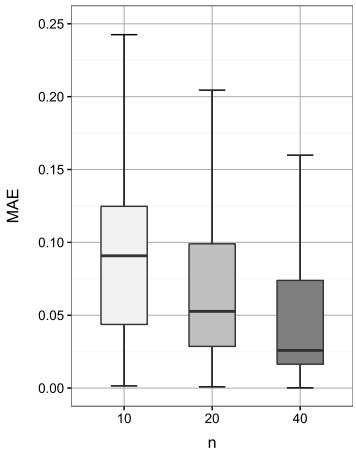

Figure 4: Simulations. Distribution over 40 datasets and nodes $s \in \{2, \ldots, q\}$ of the mean absolute error (MAE) of BMA estimates of true causal effects. Results are presented for number of nodes $q = 40$ and sample size $n \in \{10, 20, 40\}$.

We first implemented our method on the complete dataset which includes 41 genes and a binary response variable Y indicating the occurrence (absence or presence, respectively $Y = 0$ and $Y = 1$) of distant metastasis. For simplicity of exposition, we disregarded those genes which appeared to be irrelevant for causal effect estimation, because not related to the response, nor to other genes and therefore we finally included in our analysis $q = 28$ genes.

We apply Algorithm 1 by fixing the number of MCMC iterations $T = 120000$ and after standardizing observations from the continuous variables $X_2, \ldots, X_q$. We also set $g = 1/n$ and $a = q$ in the prior on the Cholesky parameters of $\mathbf{\Omega}$ as in the simulation scenarios of Section 6. We first use the MCMC output to estimate the posterior probability of inclusion of each directed edge $u \to v$, that we report in the heat map of Figure 5. Results show a substantial degree of sparsity in the underlying graph structure and only 48 edges have a posterior probability of inclusion exceeding 0.5. Moreover, among the 28 genes, only gene IL8, for which $\hat{p}_{\text{IL8}\to Y}(\cdot) = 0.70$, seems to directly affect the response variable.

To evaluate the incidence of each gene on the probability of recurrence we compute the causal effect (13) on the response due to an intervention on a specific gene. To this end, starting from the MCMC output we produce a BMA estimate $\hat{\beta}_s^{BMA}(\tilde{x})$ for each gene $s = 2, \ldots, q$ according to (27). Since the causal effect depends on the level $\tilde{x}$ assigned to the intervened variable X_s, we evaluate $\hat{\beta}_s^{BMA}(\tilde{x})$ at each observed value of X_s, that is $x_{1,s}, \ldots, x_{n,s}$. The results are reported in Figure 6, where each box-plot refers to a gene $s \in \{2, \ldots, q\}$ and summarizes the distribution of $n = 198$ BMA estimates, $\{\hat{\beta}_s^{BMA}(x_{i,s})\}_{i=1}^n$. Because the data were standardized, the ranges of $X_2, \ldots, X_q$ are similar and we can meaningfully compare results across genes.

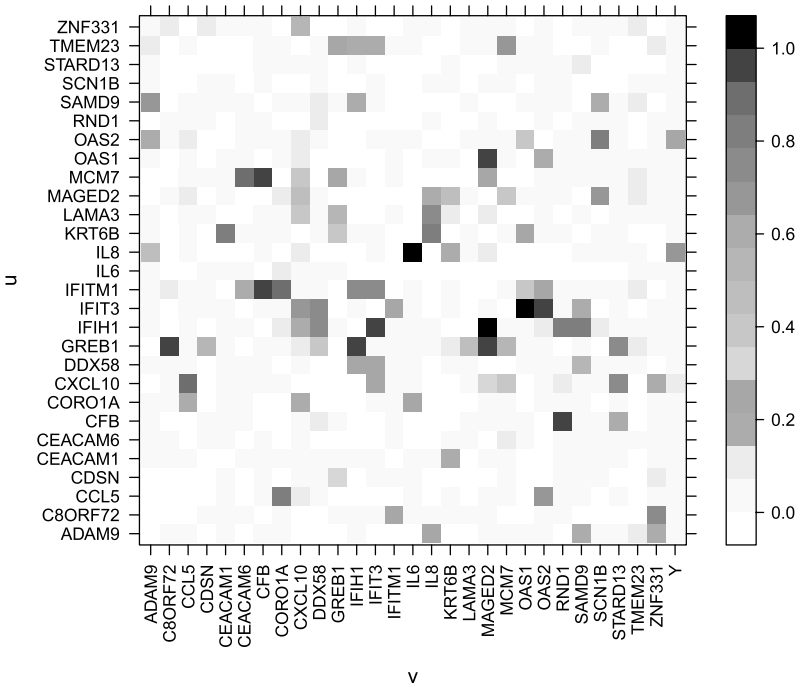

Figure 5: Gene expression data. Heat map with estimated marginal posterior probabilities of edge inclusion for each edge $u \to v$.

Recall from Proposition 3.1 that γ_s is the covariance between X_s and X_1. If $\gamma_s = 0$, equation (13) shows that the causal effect on Y due to an intervention on X_s does not vary with $\tilde{x}$. If prior information is weak in relation to the sample size, the estimate of the causal effect will be close to the overall frequency of distant metastasis in the sample (0.31). This is the situation exhibited by most genes in Figure 6. On the other hand if γ_s is not zero, the collection of causal effects evaluated at x_{is}, $i = 1, \ldots, n$ will vary. Since the observations are centred, their average is zero and the causal effects will be spread around the value corresponding to the average $\bar{x}_s = 0$ whose estimate is 0.31 as indicated above. This is what happens for a few genes such as IL8, OAS2 and KRT6B, which exhibit a much greater variability of the causal effect across their measurements, implying that their regulation can influence the occurrence of distant metastasis. In particular, gene IL8 has also been identified as having a potential impact on cancer cells in several studies (Waugh and Wilson, 2008). Other genes which stand out in terms of variability are OAS2 and KRT6B, with the latter not directly linked to Y (as one can see from the heat map of Figure 5) and exhibiting a moderate causal effect on Y which is likely due to the strong association of KRT6B with IL8 (as it emerges from the posterior probabilities of edge inclusion in Figure 5).

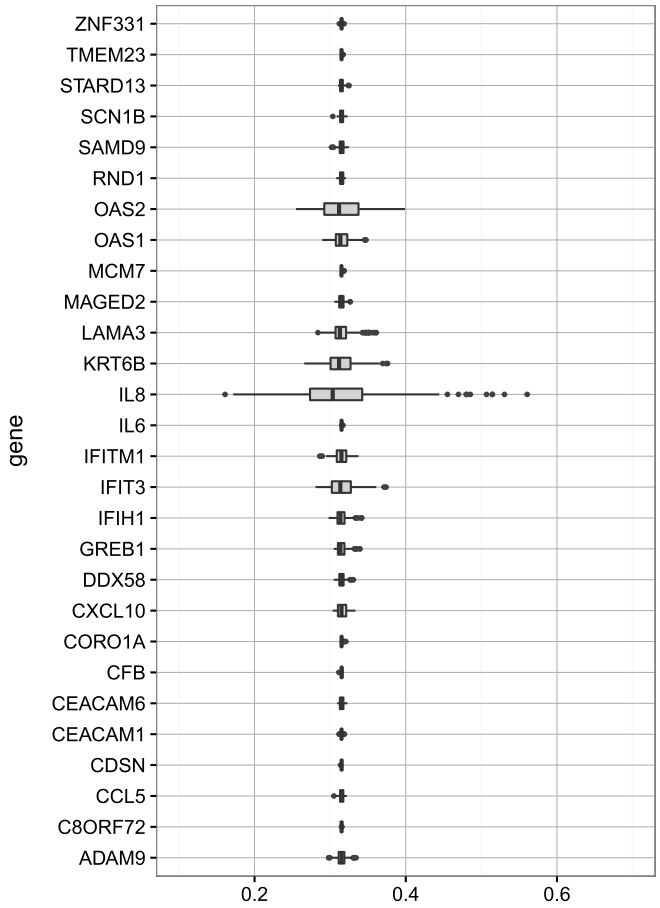

Figure 6: Gene expression data. Box-plots of BMA estimate of causal effect. Each box-plot refers to one of the 28 genes s, and represents the $n = 198$ BMA estimates computed at each observed value $(x_{1,s}, \ldots, x_{n,s})$ of expression for gene s.

For genes IL8 and OAS2 we also report in Figure 7 more detailed results for causal effect estimation. In particular, each plot reports the BMA estimates $\{\hat{\beta}_s^{BMA}(x_{i,s})\}_{i=1}^n$ (represented by $n = 198$ dots), and the corresponding credible regions at level 95% represented by the grey area. Results show that increasing expression levels of IL8 are likely to increase the presence of distant metastasis, with BMA estimates of the probability of recurrence ranging in the interval $[0.18; 0.54]$. This is consistent with results we have obtained showing that most of the mass of the distribution of the coefficient γ_s for these genes is assigned to the positive half-line; see also the discussion after (13). Moreover, more extreme levels of IL8 are associated with larger credible regions. A similar behavior, although less pronounced, is observed for gene OAS2 with BMA estimates ranging between $[0.25; 0.41]$.

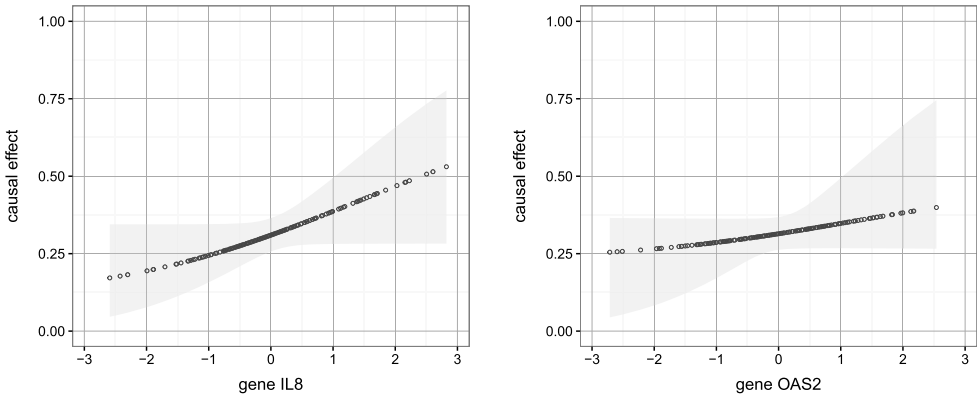

Figure 7: Gene expression data. BMA estimates (dots) and credible regions at level 95% (grey area) for two selected genes, IL8 and OAS2.

8 Discussion

We consider a system of real-valued variables together with a binary response; interest lies in the evaluation of the causal effect on the response due to an external intervention on a variable. Given observational data, this goal can be pursued by introducing some *causal* assumptions on the data generating mechanism; see for instance Maathuis and Nandy (2016). In particular, we assume that multivariate observations are generated from an unknown DAG without latent variables (*causal sufficiency*) and that the observational distribution is *faithful* to the DAG. These assumptions, coupled with the notion of intervention and the *do*-calculus, lead to a post-intervention DAG factorization that, being based on observational node-conditional distributions, allows to estimate causal effects from nonexperimental (observational) data; see also Pearl (2009).

Our *DAG-probit* model assumes that the binary response is generated by standard thresholding applied to a continuous latent variable, and that the joint distribution of all continuous variables belongs to a zero-mean Gaussian Directed Acyclic Graph (DAG) model. We then proceed by assigning a prior to the DAG-constrained covariance matrix through a DAG-Wishart distribution on the corresponding Cholesky parameters. Our elicitation procedure only requires to set the hyperparameters of a single unique Wishart distribution and guarantees score equivalence, meaning that marginal likelihoods of Markov equivalent DAG models are all equal (Geiger and Heckerman, 2002). This feature will have a useful implication, as we discuss at the end of this section.

Because the structure of the data-generating DAG is unknown, we construct an MCMC sampler whose target is the joint posterior distribution of the DAG and the allied Cholesky parameters. This is achieved by carefully tailoring a Partial Analytic Structure (PAS) algorithm to our DAG setting. As a by-product, we recover the MCMC sequence of causal effects corresponding to each visited DAG; this represents the input

to our final Bayesian Model Averaging (BMA) estimate, which naturally accounts for model uncertainty on the underlying graph structure.

The assumption of jointly normally distributed random variables can be a source of concern whenever one faces a concrete data analysis. With regard to our application the Gaussian assumption has been often used to analyse gene expression data; see for instance Dobra et al. (2004) and Markowetz and Spang (2007). In addition, it allows to easily incorporate the binary outcome through a latent component, and results in an efficient algorithm, because of closed-form expressions both for the posterior distribution of parameters, as well as for the marginal likelihood of models.

Besides the assumption of normality, our model posits a unique, yet unknown, graphical structure as the generating mechanism of all observations. Nevertheless, some problems may suggest to partition the units into groups each having a specific graphical structure which can be however related to the other ones, as in gene expressions collected on multiple tissues from the same individual (Xie et al., 2017). In this setting a multiple graphical model setup would be more appropriate to encourage similarities between group graphical structures; see for instance Peterson et al. (2015) for a Bayesian analysis of multiple Gaussian undirected graphical models. The latter could be a useful starting point for an extension of our DAG-probit model to multiple groups.

In this work we consider causal effects as obtained from interventions on single nodes. However in practice an exogenous intervention may affect many variables (genes) simultaneously and accordingly one may want to predict for instance the effect of a double or triple gene knockout on the response. Causal effect estimation from joint interventions is carried out in a Gaussian setting by Nandy et al. (2017) using a frequentist approach. Their results show that the causal effect of X_s on the response in a joint intervention on a given set of variables can be still expressed as a function of the covariance matrix Markov w.r.t. $\mathcal{D}$. The same problem can be tackled by adopting a Bayesian methodology which combines DAG structural learning and causal effect estimation and is currently under investigation by ourselves. In addition, an extension to DAG-probit models should be feasible along the lines of this paper.

The methodology adopted in this work revolves around DAGs. However, it is known that in the Gaussian setting DAGs encoding the same conditional independencies (Markov equivalent DAGs) are not distinguishable using observational data (Verma and Pearl, 1990) and can be collected into Markov equivalence classes (MECs). Accordingly, when the goal of the analysis is structural learning (model selection) MECs represent the appropriate inferential object (Andersson et al., 1997). However, if the objective is causal effect estimation, this is no longer so, because Markov equivalent DAGs may return distinct causal effects. An inspection of (13) reveals the reason: a causal effect depends on the parent set of the intervened node, and this may differ among DAGs within the same MEC. Yet MECs can be exploited also for causal inference, as we now detail. In a frequentist setting, Maathuis et al. (2009) first estimate a MEC using the classic PC algorithm (Spirtes et al., 2000), and then provide an estimate of the causal effect under each DAG within the estimated equivalence class. Alternatively, a Bayesian methodology would first determine the posterior distribution on MEC space, and then, conditionally on a given MEC, compute the posterior of each causal effect within the

class (one for each DAG). A single MEC causal effect estimate can be obtained by averaging effects across DAGs, using uniform weights on equivalent DAGs. Finally, an overall Bayesian Model Averaging (BMA) estimate can be obtained by averaging MEC-conditional estimates using posterior probabilities of MECs as weights; for details see (Castelletti and Consonni, 2021a). We remark that the above strategies require an exhaustive enumeration of all DAGs belonging to a MEC, which is not feasible even for a moderate number of nodes. Accordingly one considers only the distinct causal effects within a given MEC, because these values can be efficiently recovered (Maathuis et al., 2009, Algorithm 3) even in high-dimensional settings. In this work we seemingly ignore the issue of DAG Markov equivalence, and propose a causal inference procedure which directly focuses on DAG space, rather than MEC space. However, as already remarked at the beginning of this section, our method for parameter prior construction across DAG models guarantees score equivalence for DAGs within the same MEC. This, together with a uniform prior on DAGs within the same MEC, ensures that causal effects associated to Markov equivalent DAGs will be assigned equal weights in the resulting BMA estimate.

Supplementary Material

Supplement to Bayesian causal inference in probit graphical models (DOI: 10.1214/21-BA1260SUPP; .pdf). The Supplementary material contains the proof of Propositions 3.1 and 4.1, a detailed exposition of the PAS algorithm adopted in Section 5, additional simulation results and comparisons and an investigation of the computational time of our algorithm.

References

Abdul, Q., Yu, B., Chung, H., Jung, H., and J. S., C. (2017). "Epigenetic modifications of gene expression by lifestyle and environment." *Archives of Pharmacal Research*, 40: 1219–1237. doi: https://doi.org/10.1007/s12272-017-0973-3. 1128

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88(422): 669–679. URL http://www.jstor.org/stable/2290350 MR1224394. 1114

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). "A characterization of Markov equivalence classes for acyclic digraphs." *The Annals of Statistics*, 25(2): 505–541. MR1439312. doi: https://doi.org/10.1214/aos/1031833662. 1133

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32: 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 1125

Ben-David, E., Li, T., Massam, H., and Rajaratnam, B. (2015). "High dimensional Bayesian inference for Gaussian directed acyclic graph models." *arXiv preprint*. arXiv:1109.4371 1120

Campbell, K. L., Landells, C. E., Fan, J., and Brenner, D. R. (2017). "A systematic review of the effect of lifestyle interventions on adipose tissue gene expression: Implications for carcinogenesis." *Obesity*, 25(S2): S40–S51. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/oby.22010 1128

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(1): 319–348. MR3909935. doi: https://doi.org/10.1214/18-AOS1689. 1116, 1119

Castelletti, F. and Consonni, G. (2021a). "Bayesian inference of causal effects from observational data in Gaussian graphical models." *Biometrics*, 77(1): 136–149. MR4229727. doi: https://doi.org/10.1111/biom.13281. 1114, 1134

Castelletti, F. and Consonni, G. (2021b). "Supplementary Material of "Bayesian Causal Inference in Probit Graphical Models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1260SUPP. 1114, 1118, 1121, 1122

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). "Sparse graphical models for exploring gene expression data." *Journal of Multivariate Analysis*, 90(1): 196–212. URL http://www.sciencedirect.com/science/article/pii/S0047259X04000259 MR2064941. doi: https://doi.org/10.1016/j.jmva.2004.02.009. 1133

Friedman, N. (2004). "Inferring cellular networks using probabilistic graphical models." *Science*, 303(5659): 799–805. URL https://science.sciencemag.org/content/303/5659/799 1114

Friedman, N. and Koller, D. (2003). "Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks." *Machine Learning*, 50(1-2): 95–125. URL https://doi.org/10.1023/A:1020249912095 1114

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(501): 340–352. MR3174624. doi: https://doi.org/10.1080/01621459.2012.742443. 1123

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(5): 1412–1440. MR1936324. doi: https://doi.org/10.1214/aos/1035844981. 1120, 1132

Godsill, S. J. (2012). "On the relationship between Markov chain Monte Carlo methods for model uncertainty." *Journal of Computational and Graphical Statistics*, 10(2): 230–248. MR1939699. doi: https://doi.org/10.1198/10618600152627924. 1114, 1121

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2015). "Graphical models for ordinal data." *Journal of Computational and Graphical Statistics*, 24(1): 183–204. MR3328253. doi: https://doi.org/10.1080/10618600.2014.889023. 1117

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press. MR1419991. 1114, 1115

Maathuis, M. and Nandy, P. (2016). "A review of some recent advances in causal inference." In Bühlmann, P., Drineas, P., Kane, M., and van der Laan, M. (eds.), *Handbook of Big Data*, 387–408. Chapman and Hall/CRC. MR3674827. 1113, 1114, 1132

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). "Estimating high-dimensional intervention effects from observational data." *The Annals of Statistics*, 37(6A): 3133–3164. MR2549555. doi: https://doi.org/10.1214/09-AOS685. 1114, 1118, 1133, 1134

Markowetz, F. and Spang, R. (2007). "Inferring cellular networks – a review." *BMC Bioinformatics*, 8(S5). doi: https://doi.org/10.1186/1471-2105-8-S6-S5. 1133

Nandy, P., Maathuis, M. H., and Richardson, T. S. (2017). "Estimating the effect of joint interventions from observational data in sparse high-dimensional settings." *Ann. Statist.*, 45(2): 647–674. MR3650396. doi: https://doi.org/10.1214/16-AOS1462. 1133

O'Brien, C. A., Pollett, A., Gallinger, S., and Dick, J. E. (2006). "A human colon cancer cell capable of initiating tumour growth in immunodeficient mice." *Nature*, 445: 106–110. URL https://doi.org/10.1038/nature05372 1128

Pearl, J. (1995). "Causal diagrams for empirical research." *Biometrika*, 82(4): 669–688. URL http://www.jstor.org/stable/2337329 MR1380809. doi: https://doi.org/10.1093/biomet/82.4.669. 1114

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge. MR1744773. 1114, 1117, 1118

Pearl, J. (2009). "Causal inference in statistics: An overview." *Statistics Surveys*, 3: 96–146. MR2545291. doi: https://doi.org/10.1214/09-SS057. 1113, 1114, 1132

Peters, J. and Bühlmann, P. (2014). "Identifiability of Gaussian structural equation models with equal error variances." *Biometrika*, 101(1): 219–228. URL http://www.jstor.org/stable/43305605 MR3180667. doi: https://doi.org/10.1093/biomet/ast043. 1124

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). "Bayesian inference of multiple Gaussian graphical models." *Journal of the American Statistical Association*, 110(509): 159–174. PMID: 26078481. MR3338494. doi: https://doi.org/10.1080/01621459.2014.896806. 1133

Sadeghi, K. (2017). "Faithfulness of probability distributions and graphs." *Journal of Machine Learning Research*, 18(148): 1–29. URL http://jmlr.org/papers/v18/17-275.html MR3763782. 1115

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search (2nd edition)*. Cambridge, MA: The MIT Press. MR1815675. 1133

Verma, T. and Pearl, J. (1990). "Equivalence and Synthesis of Causal Models." In

Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI 90, 255–270. New York, NY, USA: Elsevier Science Inc. 1133

Wang, H. and Li, S. Z. (2012). "Efficient Gaussian graphical model determination under G-Wishart prior distributions." *Electronic Journal of Statistics*, 6: 168–198. MR2879676. doi: https://doi.org/10.1214/12-EJS669. 1121

Waugh, D. J. and Wilson, C. (2008). "The interleukin-8 pathway in cancer." *Clinical Cancer Research*, 14(21): 6735–6741. doi: https://doi.org/10.1158/1078-0432.CCR-07-4843. 1130

Xie, Y., Liu, Y., and Valdar, W. (2017). "Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics." *Biometrika*, 103(3): 493–511. MR3551780. doi: https://doi.org/10.1093/biomet/asw035. 1133

Yin, Z.-Q., Liu, J.-J., Xu, Y.-C., Yu, J., Ding, G.-H., Yang, F., Tang, L., Liu, B.-H., Ma, Y., Xia, Y.-W., Lin, X.-L., and Wang, H.-X. (2014). "A 41-gene signature derived from breast cancer stem cells as a predictor of survival." *Journal of Experimental & Clinical Cancer Research*, 33(49). doi: https://doi.org/10.1186/1756-9966-33-49. 1128

Zhang, J. and Spirtes, P. (2003). "Strong Faithfulness and Uniform Consistency in Causal Inference." In Meek, C. and Kjærulff, U. (eds.), *UAI '03, Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence, Acapulco, Mexico, August 7–10 2003*, 632–639. Morgan Kaufmann. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=983&proceeding_id=19 1116

Acknowledgments

We thank three reviewers, an Associate Editor and the Editor for constructive comments that helped improve the paper. Support from COSTNET (COST Action CA15109) and UCSC (D1 and 2019-D.3.2 research grant) is gratefully acknowledged.