Network structure learning under uncertain interventions

Federico Castelletti

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

and

Stefano Peluso

Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, Milan

January 28, 2022

Abstract

Gaussian Directed Acyclic Graphs (DAGs) represent a powerful tool for learning the network of dependencies among variables, a task which is of primary interest in many fields and specifically in biology. Different DAGs may encode equivalent conditional independence structures, implying limited ability, with observational data, to identify causal relations. In many contexts however, measurements are collected under heterogeneous settings where variables are subject to exogenous interventions. Interventional data can improve the structure learning process whenever the *targets* of an intervention are known. However, these are often uncertain or completely unknown, as in the context of drug target discovery. We propose a Bayesian method for learning dependence structures and intervention targets from data subject to interventions on unknown variables of the system. Selected features of our approach

include a DAG-Wishart prior on the DAG parameters, and the use of variable selction priors to express uncertainty on the targets. We provide theoretical results on the correct asymptotic identification of intervention targets and derive sufficient conditions for Bayes factor and posterior ratio consistency of the graph structure. Our method is applied in simulations and real-data world settings, to analyze perturbed protein data and assess antiepileptic drug therapies. Details of the MCMC algorithm and proofs of propositions are provided in the Supplementary Material, together with more extensive results on simulations and applied studies.

Keywords: Directed acyclic graph; DAG-Wishart prior; Interventional data; Target discovery.

1 Introduction

1.1 Motivation and framework

Graphical models based on directed networks have been widely employed to understand dependence relations between variables, a crucial problem in many scientific areas, especially in biology (Friedman, 2004; Shojaie and Michailidis, 2009). Typically, the network structure is inferred under the assumption that multivariate data have been generated by a stable system. More realistically however, measurements can be heterogeneous, meaning that modifications in the generating mechanism, e.g. due to exogenous interventions, have occurred.

An instance is genomic medicine, where interactions between genes provide insights on the genesis and progression of diseases, whose occurrence is reflected by aberrations in the gene-network functioning. In this setting, drug therapies capable of gene-inhibition can be applied to regulate and restore dependencies in the gene-network structure. However, the effect of drug treatments at gene level can be uncertain or completely unknown (Paananen

and Fortino, 2019; Marton et al., 1998), so that discovering the targets of an intervention or therapy becomes of interest in itself. Drug target discovery is also essential for the development of personalized treatments, to identify genes that are affected by drugs and in turn evaluate patients' response to therapies; see Rawat et al. (2020) for a recent discussion.

In this paper we consider multivariate data generated from a system subject to unknown interventions, and we propose a novel method for learning their dependence structure and the effects of interventions. We represent the data generating mechanism through a Directed Acyclic Graph (DAG) which allows for a factorization of the joint distribution in terms of "parent-child" relations between nodes (variables). An intervention modifies the original DAG structure by dropping the dependence of each intervened node (the *intervention target*) from its parents. Deterministic interventions assume that each intervened variable is set equal to a constant level, an assumption reasonable in some contexts, such as gene-knockout experiments; by converse, stochastic interventions (Korb et al., 2004), that we adopt in the current paper, are more general and replace the conditional distribution of the intervened node with that of a new random variable, independent from all parent nodes.

1.2 Related works

The problem of learning DAGs from interventional data has received some interest over the last years. In particular, some methodologies for structure learning of DAGs given interventions with known targets have been developed; see for instance Hauser and Bühlmann (2015) and Castelletti and Consonni (2019) for a frequentist and Bayesian approach respectively. When unknown targets are allowed, Eaton and Murphy (2007) apply the dynamic programming algorithm of Koivisto and Sood (2004) on a graph augmented with interven-

tional nodes, to estimate edge inclusion probabilities and interventions. Their method is implemented for categorical data, using the the Bayesian-Dirichlet score of Heckerman and Geiger (1995), although it could be adapted to the Gaussian case. The authors show in simulations that their method can correctly recover both the intervention targets and the graph structure under specific settings. Importantly however, the augmentation with interventional nodes increases the graph size, and causes the method to be practically feasible only up to 20 nodes. Higher dimensions compel to severely restrictive assumptions on the interventions: (i) in terms of the number of intervened nodes, thus ruling out interventions with a diffuse impact, as in our application of Section 6.1, or (ii) in terms of interventions forced to act on distinct nodes, thus excluding cases where different but related drug therapies can partially share effects on the same node, as in the analysis of antiepileptic therapies *Valporate* and *Carbamazepine*, discussed in Section 6.2.

A similar idea of graph augmentation to include hidden nodes representing intervention targets has been proposed in Zhang et al. (2017): they apply constraint-based methods, as the PC algorithm of Spirtes et al. (2000), on the augmented graph to identify the network skeleton and v-structures, and then recover arrow directions through invariance considerations. Specifically, the distribution of a target node, conditional on its causal parents, should not change when interventions affect other nodes; this idea was first adopted in Peters et al. (2016), who propose a method that estimates causal effects, and that can be iterated with the purpose of network learning under uncertain interventions. The PC algorithm is also used by He and Geng (2016) who first recover from a collection of datasets group-specific network structures, then pooled together to infer the causal graph. The methods of Zhang et al. (2017) and He and Geng (2016) are both substantially different from our framework as they are based on multiple hypothesis tests and do not

provide any uncertainty of the estimated graph. In addition, He and Geng (2016) do not perform target estimation, while Zhang et al. (2017) recover the manipulated variables, but without identifying the interventional settings they refer to. More recently, Ke et al. (2019) propose an optimization-based Bayesian method for network learning, limited to categorical interventional and observational data. Finally, Squires et al. (2020) suggest a greedy search algorithm for causal structure learning with unknown interventions and target estimation, in the presence of multiple datasets, one of which has to be purely observational.

1.3 Contribution and structure of the paper

In this paper we propose a Bayesian methodology for joint structure learning of Gaussian DAGs and intervention targets that extend the literature along the following directions: (i) we build a new modelling framework where observational data are not strictly required, the excessive reliance on multiple tests is avoided, and unknown interventions are represented as indicator vector parameters, rather than auxiliary nodes that increase the graph dimension; (ii) we demonstrate theoretically, and validate empirically, the correct asymptotic identification of the targets and of the equivalence class of the true DAG; (iii) we propose a novel MCMC algorithm for joint posterior analysis over the space of graphs and interventions, without resorting to optimization routines. In addition, we emphasize that our method is practically feasible on graphs of dimension larger than those studied so far in the Bayesian literature, without imposing restrictive assumptions on the structure of the interventions. Finally, differently from other Bayesian approaches for DAG structure learning, our method revolves around *arbitrary* DAGs, i.e. with completely unknown ordering of the nodes; see Ni et al. (2017, 2019) for a comparison.

The rest of the paper is organized as follows. In Section 2 we briefly summarize the main

concepts about DAGs and interventions, introduce our Gaussian DAG-model, and priors on DAGs, intervention targets and DAG-parameters. Asymptotic theoretical properties of target identification and graph learning are described in Section 3, whilst in Section 4 we develop the MCMC scheme for posterior inference on DAGs and targets. Simulation studies to assess the performance of our method are conducted in Section 5, while Section 6 presents applications to real data and comparisons with alternative methods available in the literature. Finally, Section 7 contains a brief discussion together with possible extensions of our methodology. Further details on our MCMC algorithm, proofs of propositions, and more in-depth simulation and real-world studies are provided in the Supplementary Material.

2 Model formulation

2.1 Directed acyclic graphs and interventions

Let $\mathcal{D} = (V, E)$ be a Directed Acyclic Graph (DAG), where $V = \{1, \dots, q\}$ is a set of nodes and $E \subseteq V \times V$ a set of edges. If $(u, v) \in E$, then $(v, u) \notin E$ and we say that u is a *parent* of node v and v is a *child* of u. The set of all parents of v in $\mathcal{D}$ is denoted by $\mathrm{pa}_{\mathcal{D}}(v)$. Consider a collection of q random variables, $X_1, \dots, X_q$. We assume that the joint distribution f factorizes according to $\mathcal{D}$ as

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

If (1) holds, $f(x_1, \dots, x_q \,|\, \mathcal{D})$ is said to obey the *Markov property* of $\mathcal{D}$. In our context of interventional data, Equation (1) is also called *observational* or *pre-interventional* distribution.

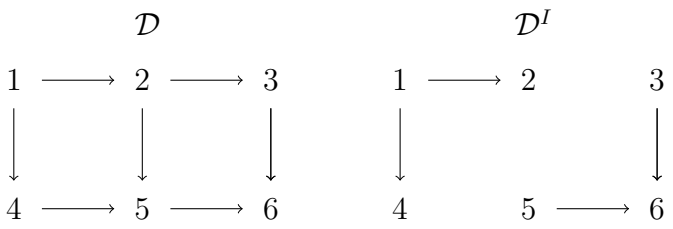

Figure 1: A DAG $\mathcal{D}$ and the corresponding intervention DAG $\mathcal{D}^I$ for the target $I = \{3, 5\}$.

An *intervention* on the node $j \in V$ is defined as the action of setting X_j to the value of a random variable U_j having density $\widetilde{f}(u_j)$. A *joint intervention* on $I \subseteq V$ fix, for each $j \in I$, X_j to U_j, with $\{U_j\}_{j \in I}$ mutually independent. I is called an *intervention target* and the *do-operator* $\text{do}\{X_j = U_j\}_{j \in I}$ (Pearl, 2000) is used to denote such an intervention. As a consequence of the intervention, the original dependence in $\mathcal{D}$ of node j from its parents $\text{pa}_{\mathcal{D}}(j)$ is dropped, and the *intervention DAG* of $\mathcal{D}$ is defined, following Hauser and Bühlmann (2012), as $\mathcal{D}^I = (V, E^I)$, where $E^I = \{(u, v) \in E \,|\, v \notin I\}$. See also Figure 1 for an example of DAG and related intervention DAG. The intervention on I replaces the conditional density of each node $j \in I$ in (1), that is $f(x_j \,|\, \boldsymbol{x}_{\text{pa}_{\mathcal{D}}(j)})$, with $\widetilde{f}(x_j)$. Therefore, the *post-intervention* distribution of $X_1, \ldots, X_q$ given the operator $\text{do}\{X_j = U_j\}_{j \in I}$ is obtained from (1) using the truncated factorization

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

When there are no interventions, Equation (2) reduces to Equation (1). Multiple independent interventions form a *family* of intervention targets $\mathcal{I} = \{I_1, \ldots, I_K\}$, where each $I_k \subseteq V$ and index k refers to the *k-th* intervention.

2.2 Gaussian DAG-models

We assume $(X_1, \ldots, X_q) \,|\, \boldsymbol{\Sigma}, \mathcal{D} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathcal{C}_\mathcal{D}$, the space of all covariance matrices Markov w.r.t. $\mathcal{D}$. A Gaussian DAG-model can be equivalently written as a linear Structural Equation Model (SEM) of the form $\boldsymbol{L}^\top (X_1, \ldots, X_q)^\top = \boldsymbol{\varepsilon}$, where $\boldsymbol{L}$ is a $q \times q$ matrix with unit diagonal elements and $\boldsymbol{\varepsilon} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{D})$, $\boldsymbol{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$; see Kaplan (2009). Accordingly, the decomposition $\boldsymbol{\Sigma} = \boldsymbol{L}^{-\top} \boldsymbol{D} \boldsymbol{L}^{-1}$ holds. For each (u, v)-element of $\boldsymbol{L}$ and $u \neq v$ we have $\boldsymbol{L}_{u,v} \neq 0$ if and only if $u \in \mathrm{pa}_\mathcal{D}(v)$, and it holds that

$$f(x_1, \ldots, x_q \,|\, \boldsymbol{D}, \boldsymbol{L}, \mathcal{D}) = \prod_{j=1}^{q} \varphi\big(x_j \,|\, -\boldsymbol{L}_{\prec j]}^\top \boldsymbol{x}_{\mathrm{pa}_\mathcal{D}(j)}, \sigma_j^2\big), \tag{3}$$

where φ is the Gaussian density function, $\prec j\,] = \mathrm{pa}_\mathcal{D}(j) \times j$ and $\boldsymbol{L}_{A \times B}$ denotes the sub-matrix of $\boldsymbol{L}$ with elements belonging to rows and columns indexed by A and B respectively. The set $\big\{\big(\boldsymbol{L}_{\prec j]}, \sigma_j^2\big)\big\}_{j=1}^q$ represents the collection of *observational* parameters (because they index the observational model).

For any intervention target $I \subseteq V$, we further assume that each interventional density $\widetilde{f}$ in (2) is zero-mean Gaussian $\widetilde{f}(u_j) = \varphi(u_j \,|\, 0, \phi_j), j \in I$, with $U_j \perp\!\!\!\perp U_{j'}$ for each $j \neq j'$. With this specific choice of interventional density, we can now replace the conditioning event $\mathrm{do}\{X_j = U_j\}_{j \in I}$ with the parameters $\{I, \boldsymbol{\Phi} = \{\phi_j\}_{j \in I}\}$, where $\boldsymbol{\Phi}$ is the collection of *interventional* parameters. The post-intervention distribution of $(X_1, \ldots, X_q)$ thus becomes

$$f\big(x_1, \ldots, x_q \,|\, \boldsymbol{D}, \boldsymbol{L}, \boldsymbol{\Phi}, I, \mathcal{D}\big) = \prod_{j \notin I} \varphi\big(x_j \,|\, -\boldsymbol{L}_{\prec j]}^\top \boldsymbol{x}_{\mathrm{pa}_\mathcal{D}(j)}, \sigma_j^2\big) \prod_{j \in I} \varphi(x_j \,|\, 0, \phi_j). \tag{4}$$

We now split the available observations of $X_1, \ldots, X_q$ among K datasets, as arising from the family of K independent interventions: each $n^{(k)} \times q$ dataset $\boldsymbol{X}^{(k)}$ consists of a collection of $n^{(k)}$ i.i.d. multivariate observations $\boldsymbol{x}_i^{(k)} = \big(x_{i,1}^{(k)}, \ldots, x_{i,q}^{(k)}\big)^\top$ (rows of the data matrix $\boldsymbol{X}^{(k)}$) associated to intervention target I_k, for $i = 1, \ldots, n^{(k)}$. Accordingly, the

post intervention distribution related to intervention k is $f\big(\boldsymbol{x}_i^{(k)} \,|\, \boldsymbol{D}, \boldsymbol{L}, \boldsymbol{\Phi}^{(k)}, I_k, \mathcal{D}\big)$, where $\boldsymbol{\Phi}^{(k)} = \{\phi_j^{(k)}\}_{j\in I}$ are the interventional parameters associated to the k-th intervention. Given the collection of datasets $\boldsymbol{X} = \big(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}\big)^\top$, the likelihood function is finally

$$f\big(\boldsymbol{X} \,|\, \boldsymbol{\theta}, I_1, \ldots, I_K, \mathcal{D}\big) = \prod_{k=1}^{K} \prod_{i=1}^{n^{(k)}} f\big(x_{i,1}^{(k)}, \ldots, x_{i,q}^{(k)} \,|\, \boldsymbol{D}, \boldsymbol{L}, \boldsymbol{\Phi}^{(k)}, I_k, \mathcal{D}\big), \tag{5}$$

where $\boldsymbol{\theta} = \big\{\boldsymbol{D}, \boldsymbol{L}, \boldsymbol{\Phi}^{(1)}, \ldots, \boldsymbol{\Phi}^{(K)}\big\}$ is the collection of all DAG-dependent (observational and interventional) parameters.

2.3 Prior on DAG parameter $\boldsymbol{\theta}$

Conditionally on DAG $\mathcal{D}$ and the collection of targets $I_1, \ldots, I_K$, we first assign a prior to the observational parameters $(\boldsymbol{D}, \boldsymbol{L})$. Recall that $\boldsymbol{\Sigma} = \boldsymbol{L}^{-\top}\boldsymbol{D}\boldsymbol{L}^{-1}$, where $\boldsymbol{\Sigma}$ is the covariance matrix of a multivariate Gaussian random variable Markov w.r.t. DAG $\mathcal{D}$. We assign $(\boldsymbol{D}, \boldsymbol{L})$ a DAG-Wishart prior with hyperparameter $\boldsymbol{U}$ (a $q\times q$ positive definite matrix) and shape hyperparameter $\boldsymbol{a}^{\mathcal{D}} = (a_1^{\mathcal{D}}, \ldots, a_q^{\mathcal{D}})^\top$; see Ben-David et al. (2015) and Cao et al. (2019). Also, a standard choice, hereinafter adopted, is $\boldsymbol{U} = g\boldsymbol{I}_q$ ($g > 0$). The DAG-Wishart distribution induces a re-parameterization of $\boldsymbol{\Sigma}$ in terms of the node-parameters $\big\{(\boldsymbol{L}_{\prec j\,]}, \sigma_j^2)\big\}_{j=1}^q$, independent across $j = 1, \ldots, q$, and with distribution

$$\sigma_j^2 \sim \text{ I-Ga}\left(\frac{1}{2}a_j^{\mathcal{D}}, \frac{1}{2}\,g\right), \quad \boldsymbol{L}_{\prec j\,]} \,|\, \sigma_j^2 \sim \ \mathcal{N}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\left(\boldsymbol{0}, \sigma_j^2\left(g\boldsymbol{I}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\right)^{-1}\right), \tag{6}$$

where $\prec j\,] = \mathrm{pa}_{\mathcal{D}}(j) \times j$ and I-Ga(a, b) stands for an Inverse-Gamma distribution with shape $a > 1$ and rate $b > 0$ having expectation $b/(a-1)$. From (6), the prior on the observational parameters $(\boldsymbol{D}, \boldsymbol{L})$ is given by $p(\boldsymbol{D}, \boldsymbol{L}) = \prod_{j=1}^q p\left(\boldsymbol{L}_{\prec j\,]} \,|\, \sigma_j^2\right) p\left(\sigma_j^2\right)$. Hyper-parameters $a_j^{\mathcal{D}}$ are specific to each DAG model, and it can be shown that the default choice

(hereinafter adopted) $a_j^{\mathcal{D}} = a + |\text{pa}_{\mathcal{D}}(j)| - q + 1$ $(a > q - 1)$ guarantees compatibility among prior distributions for Markov equivalent DAGs; see Peluso and Consonni (2020). In particular, we set $a = q$, the minimum integer value that guarantees a proper prior distribution, regardless of the specific $\mathcal{D}$.

Consider now the interventional parameters $\left\{\boldsymbol{\Phi}^{(k)}\right\}_{k=1}^{K}$, where each $\boldsymbol{\Phi}^{(k)}$ is a collection of node-parameters $\left\{\phi_j^{(k)}\right\}_{j\in I_k}$. Because each $\phi_j^{(k)}$ corresponds to an unconditional variance in a post-intervention distribution where each node $j \in I_k$ has no parents, we can set

$$\phi_j^{(k)} \sim \text{I-Ga}\left(a_j^{(k)}, b_j^{(k)}\right), \tag{7}$$

independently, where $a_j^{(k)} = (a-q+1)/2$ and $b_j^{(k)} = g/2$, following the same elicitation procedure leading to (6). The prior on the collection of interventional parameters is therefore $p(\boldsymbol{\Phi}^{(1)}, \ldots, \boldsymbol{\Phi}^{(K)}) = \prod_{k=1}^{K} \prod_{j\in I_k} p(\phi_j^{(k)})$, leading to a conditional prior on $\boldsymbol{\theta}$ of the form

$$p(\boldsymbol{\theta} \,|\, I_1, \ldots, I_K, \mathcal{D}) = p(\boldsymbol{D}, \boldsymbol{L}) \cdot p(\boldsymbol{\Phi}^{(1)}, \ldots, \boldsymbol{\Phi}^{(K)}) = \prod_{j=1}^{q} \left\{ p\big(\boldsymbol{L}_{\prec j]} \,|\, \sigma_j^2\big) p(\sigma_j^2) \prod_{k: j\in I_k} p(\phi_j^{(k)}) \right\}. \tag{8}$$

2.4 Prior on targets $I_1, \ldots, I_K$

Consider now the collection of targets $I_1, \ldots, I_K$, where $I_k \subseteq \{1, \ldots, q\}$, $k = 1, \ldots, K$. For convenience, we represent each target I_k as an indicator vector $\boldsymbol{h}_k = (h_k(1), \ldots, h_k(q))^\top$ such that for each $j = 1, \ldots, q$, $h_k(j) = 1$ if $j \in I_k$, and 0 otherwise. Conditionally on a prior probability $\pi_k(j) \in (0, 1)$, we can assign a prior to I_k through q independent Bernoulli distributions on $\boldsymbol{h}_k$,

$$p\big(I_k \,|\, \boldsymbol{\pi}_k\big) = p\big(\boldsymbol{h}_k \,|\, \boldsymbol{\pi}_k\big) = \prod_{j=1}^{q} \pi_k(j)^{h_k(j)} \big(1 - \pi_k(j)\big)^{1-h_k(j)}, \tag{9}$$

where $\boldsymbol{\pi}_k = (\pi_k(1), \ldots, \pi_k(q))^\top$. Assuming prior independence among intervention targets, we then set $p(I_1, \ldots, I_K \,|\, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K) = \prod_{k=1}^{K} p(\boldsymbol{h}_k \,|\, \boldsymbol{\pi}_k)$. In addition we assign, for $j =$

$1, \ldots, q$ and $k = 1, \ldots, K$, $\pi_k(j) \overset{\text{iid}}{\sim} \text{Beta}(a_k, b_k)$, which leads to the integrated prior for I_k

$$p(I_k) = \frac{\Gamma(a_k + b_k)}{\Gamma(a_k) + \Gamma(b_k)} \cdot \frac{\Gamma\big(a_k + |I_k|\big)\Gamma\big(q - |I_k| + b_k\big)}{\Gamma\big(a_k + b_k + q\big)}, \tag{10}$$

where $|I_k| = \sum_{j=1}^{q} h_k(j)$ corresponds to the number of intervened nodes under intervention k. Expression (10) resembles the multiplicity correction prior introduced in Scott and Berger (2010) for variable selection.

2.5 Prior on DAG $\mathcal{D}$

For a given DAG $\mathcal{D} = (V, E)$, let $\boldsymbol{S}^{\mathcal{D}}$ be the 0-1 *adjacency matrix* of its skeleton (the underlying undirected graph obtained after removing the orientation of its edges), such that for each (u, v)-element in $\boldsymbol{S}^{\mathcal{D}}$, $\boldsymbol{S}^{\mathcal{D}}_{u,v} = 1$ if and only if $(u, v) \in E$ or $(v, u) \in E$, and 0 otherwise. Given some prior probability of inclusion $\eta \in (0, 1)$, we assume $\boldsymbol{S}^{\mathcal{D}}_{u,v} \overset{\text{iid}}{\sim} \text{Ber}(\eta)$ for each $u > v$, so that $p(\boldsymbol{S}^{\mathcal{D}}) = \eta^{|\boldsymbol{S}^{\mathcal{D}}|}(1 - \eta)^{\frac{q(q-1)}{2} - |\boldsymbol{S}^{\mathcal{D}}|}$, where $|\boldsymbol{S}^{\mathcal{D}}|$ is the number of edges in $\mathcal{D}$ (equivalently in its skeleton) and $q(q-1)/2$ corresponds to the maximum number of edges in a DAG with q nodes. Finally we set $p(\mathcal{D}) \propto p(\boldsymbol{S}^{\mathcal{D}})$, for any $\mathcal{D} \in \mathcal{S}_q$, where $\mathcal{S}_q$ is the space of all DAGs on q nodes. Such a prior only depends on the number of edges in the graph and can easily reflect prior knowledge of sparsity (Castelletti et al., 2018). Other priors, specific for DAGs and based on the number of compatible perfect orderings of the vertices, are also present in the literature (Friedman and Koller, 2003; Kuipers and Moffa, 2017).

3 Theoretical properties of target and graph learning

In the present section we investigate the asymptotic behaviour of the false positive and false negative rates associated to the estimation of the intervention targets, and the correct asymptotic identification of the graphical structure. The dependence on a given graph $\mathcal{D}$ is assumed and omitted in the first part of this section, and later reinstated when we discuss graph learning. Let $I_{01}, \dots, I_{0k}$ be the true unknown intervention targets, $\boldsymbol{\Phi}_0^{(1)}, \dots, \boldsymbol{\Phi}_0^{(K)}$ the true interventional parameters, and $(\boldsymbol{D}_0, \boldsymbol{L}_0)$ the true observational parameters corresponding to the Cholesky decomposition of the variance (precision) matrix $\boldsymbol{\Sigma}_0$ ($\boldsymbol{\Omega}_0$). For a given node j and dataset k, and with $\boldsymbol{\theta} = \left\{\boldsymbol{D}, \boldsymbol{L}, \boldsymbol{\Phi}^{(1)}, \dots, \boldsymbol{\Phi}^{(K)}\right\}$, we first define the posterior log-odds of an intervention on node j in dataset k as

$$\tilde{\gamma}_j^{(k)}(\boldsymbol{X}, \boldsymbol{\theta}) := logit\left(h_k(j) = 1 \,|\, \boldsymbol{X}, \boldsymbol{\theta}\right) = logit\left(h_k(j) = 1 \,|\, \boldsymbol{X}^{(k)}, \boldsymbol{D}, \boldsymbol{L}, \boldsymbol{\Phi}^{(k)}\right),$$

where $logit(A) = \ln(\mathbb{P}(A)/\mathbb{P}(\bar{A}))$ for some event A and its complement $\bar{A}$. The posterior conditional probability of an intervention on node j in dataset k is

$$\mathbb{P}(h_k(j) = 1 \,|\, \cdot) \propto \pi_k(j) \varphi_{n^{(k)}}\left(\boldsymbol{X}_j^{(k)} \,|\, \boldsymbol{0}, \phi_j^{(k)} \boldsymbol{I}_{n_j^{(k)}}\right),$$

where φ_p is the density function of a p-variate Gaussian r.v. and $\boldsymbol{X}_j^{(k)}$ denotes column indexed by j in dataset $\boldsymbol{X}^{(k)}$. For the complement event we have instead

$$\mathbb{P}(h_k(j) = 0 \,|\, \cdot) \propto (1 - \pi_k(j)) \varphi_{n^{(k)}}\left(\boldsymbol{X}_j^{(k)} | - \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}^{(k)} \boldsymbol{L}_{\prec j\,]}, \sigma_j^2 \boldsymbol{I}_{n_j^{(k)}}\right).$$

We further remove the dependence from $\boldsymbol{\theta}$ by considering the conditional expectation

$$\gamma_j^{(k)}(\boldsymbol{X} \,|\, \mathcal{A}_j^k) := \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X}, \mathcal{A}_j^k}\left[\tilde{\gamma}_j^{(k)}(\boldsymbol{X}, \boldsymbol{\theta})\right],$$

where the conditioning event is $\mathcal{A}_j^k = \{j \in I_{0k}\}$, therefore interpreted as the posterior expected log-odds of an intervention on node j in dataset k, given that j is indeed a true

intervened node (target) in the dataset. In the following proposition we show that the posterior expected log-odds $\gamma_j^{(k)}(\boldsymbol{X} \mid \mathcal{A}_j^k)$ correctly diverges when the intervention has truly occurred, and the asymptotic normality of the scaled log-odds.

Proposition 3.1. *For* $\mathcal{K}_j(\mathcal{D}) := a_j^{\mathcal{D}}/\boldsymbol{U}_{jj|\prec j\succ}$ *we have*

$$\gamma_j^{(k)}(\boldsymbol{X} \mid \mathcal{A}_j^k)/n^{(k)} \xrightarrow{d} \mathcal{N}\left(\frac{1}{2}\left(\mathcal{K}_j(\mathcal{D})\phi_{0j}^{(k)} - 1 + \mathrm{tr}\boldsymbol{\Sigma}_{0\prec j\succ}/g - \ln\left(\mathcal{K}_j(\mathcal{D})\phi_{0j}^{(k)}\right) + C_1\right),\right.$$
$$\left.\left(\phi_{0j}^{(k)}\mathcal{K}_j(\mathcal{D}) - 1\right)^2/(2n^{(k)}) + \mathrm{tr}\boldsymbol{\Sigma}_{0\prec j\succ}^2/(2n^{(k)})\right),$$

where $C_1 = \ln(a_j^{\mathcal{D}}/2) - \psi\left(a_j^{\mathcal{D}}/2\right)$ *and* ψ *is the digamma function. Also,* $\gamma_j^{(k)}(\boldsymbol{X} \mid \mathcal{A}_j^k) \xrightarrow{a.s.} +\infty$.

Proof. See Supplementary Material.

The proposition states that, if $j \in I_k$, i.e. node j is a target under intervention k, this will be detected with sample size large enough and for any given graph $\mathcal{D}$, therefore with a false negative rate eventually zero. The scaled log-odds of the (correct) target classification has asymptotic Gaussian distribution. Note that the mean of the asymptotic distribution increases when $a_j^{\mathcal{D}}$ is large. This is typical of a node with many parents in a large graph, which makes easier the identification of an intervention, because the latter will suppress many dependence relations. We refer the reader to the Supplementary Material for a more extensive discussion of the proposition.

In the opposite case of no intervention, we analyse the behaviour of

$$\bar{\gamma}_j^{(k)}(\boldsymbol{X} \mid \bar{\mathcal{A}}_j^k) := \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X},\bar{\mathcal{A}}_j^k}\left[-\tilde{\gamma}_j^{(k)}(\boldsymbol{X}, \boldsymbol{\theta})\right],$$

where the conditioning event is $\bar{\mathcal{A}}_j^k = \{j \notin I_{0k}\}$, therefore interpreted as the posterior expected log-odds of no intervention on node j in dataset k, given that j is not an unknown intervened node. With the following result we show conditions for which this quantity diverges when there is no intervention, and its asymptotic normality.

Proposition 3.2. *For* $\delta_{jk} := \frac{a_j^{(k)}}{b_j^{(k)}}\sigma_{0j}^2 - 1$ *we have*

$$\bar{\gamma}_j^{(k)}(\boldsymbol{X} \mid \bar{\mathcal{A}}_j^k)/n^{(k)} \xrightarrow{d} \mathcal{N}\left(\frac{1}{2}\left(\frac{a_j^{(k)}}{b_j^{(k)}}\boldsymbol{\Sigma}_{0jj} - 1 - \ln\frac{a_j^{(k)}}{b_j^{(k)}}\sigma_{0j}^2 + C_2\right),\right.$$
$$\left.\left(\frac{\boldsymbol{\Sigma}_{0jj}}{\sigma_{0j}^2}\delta_{jk}\right)^2/(2n^{(k)}) + \frac{\boldsymbol{\Sigma}_{0jj}^2 - \sigma_{0j}^4}{\sigma_{0j}^4}\left[\delta_{jk}/n^{(k)} + \left(\frac{5\boldsymbol{\Sigma}_{0jj} - 3\sigma_{0j}^2}{\boldsymbol{\Sigma}_{0jj} + \sigma_{0j}^2}\right)/(2n^{(k)})\right]\right),$$

where $C_2 = \ln\left(a_j^{(k)}\right) - \psi\left(a_j^{(k)}\right)$ *and* ψ *is the digamma function. Also,* $\gamma_j^{(k)}(\boldsymbol{X} \mid \bar{\mathcal{A}}_j^k) \xrightarrow{a.s.} +\infty$.

Proof. See Supplementary Material.

Proposition 3.2 tells that a true negative case of no intervention, i.e. $j \notin I_k$, will be eventually detected with sample size large enough. Note that the mean of the asymptotic distribution is closer to zero when node j is independent from any other node. Intuitively, it is more difficult to understand the absence of an intervention since there are no parent-child relations that are removed by the intervention on node j, which makes the intervention effect less apparent. We refer the reader to the Supplementary Material for a more extensive discussion. In the following sections we develop and implement an MCMC algorithm that empirically confirm the correct detection of nodes which are targeted by interventions.

The above discussion focuses on the correct identification of interventions for a given DAG. We now prove model selection consistency of the true DAG observational equivalence class; the latter, combined with consistent estimation of targets, allows to identify the group-specific intervention graphs. Data within group k is used to find interventions specific to that group, and observational data from all groups are combined to make inference on the underlying observational DAG structure. In Section 3 of Supplementary material, we first extend the conjugacy result on the DAG-Wishart prior of Ben-David et al. (2015) to interventional Gaussian multivariate data from multiple groups; then we

prove, following Cao et al. (2019) and Peluso and Consonni (2020), its Bayes factor and posterior ratio consistency outside $[\mathcal{D}_0]$, the equivalence class of the true DAG, and its asymptotic compatibility within $[\mathcal{D}_0]$.

We have *Bayes factor consistency* if, for all $\mathcal{D} \neq \mathcal{D}_0$, the Bayes factor

$$BF_{\mathcal{D},\mathcal{D}_0} = \frac{m(\boldsymbol{X} \mid \mathcal{D}, I_1, \ldots, I_K)}{m(\boldsymbol{X} \mid \mathcal{D}_0, I_1, \ldots, I_K)} \xrightarrow{\bar{P}} 0,$$

whenever $\mathcal{D}_0$ is the true DAG generating $\boldsymbol{X}$, where $\xrightarrow{\bar{P}}$ denotes convergence in probability, $\bar{P}$ is the probability measure under the true DAG $\mathcal{D}_0$, and $m(\boldsymbol{X} \mid \mathcal{D}, I_1, \ldots, I_k)$ is the marginal (or integrated) likelihood. We have *posterior ratio consistency* if, with $\mathcal{D}_0$ being the true DAG, it holds that

$$\max_{\mathcal{D} \neq \mathcal{D}_0} \frac{p(\mathcal{D} \mid \boldsymbol{X}, I_1, \ldots, I_K)}{p(\mathcal{D}_0 \mid \boldsymbol{X}, I_1, \ldots, I_K)} = \max_{\mathcal{D} \neq \mathcal{D}_0} BF_{\mathcal{D},\mathcal{D}_0}(\boldsymbol{X} \mid I_1, \ldots, I_k) \frac{p(\mathcal{D})}{p(\mathcal{D}_0)} \xrightarrow{\bar{P}} 0. \tag{11}$$

For each $j \in V$, let $\tilde{n}_j = \sum_{k: j \in I_k} n^{(k)}$ and $n_j^* = \sum_{k: j \notin I_k} n^{(k)}$ be the number of observations among groups $k = 1, \ldots, K$ such that node j is respectively intervened and not.

Proposition 3.3. *Let $\mathcal{D}_0$ be the true DAG. Assume $(\boldsymbol{D}, \boldsymbol{L}) \mid \mathcal{D}$ follows a DAG-Wishart distribution with hyperparameters $\boldsymbol{U}$ and $\boldsymbol{a}^{\mathcal{D}}$ as in Equation* (8)*, and consider the likelihood function in Equation* (5)*. If (a) $a_j^{\mathcal{D}} = a + |\mathrm{pa}_{\mathcal{D}}(j)| - q + 1$, (b) $\tilde{n}_j = o(n_j^*)$ for all $j \in V$, and (c) for all $j \neq l \in V$ there exists a k such that $j \notin I_k$ and $l \notin I_k$ hold, then as $n \to \infty$,*

$$\begin{aligned} &i) \quad \max_{\mathcal{D} \notin [\mathcal{D}_0]} \frac{p(\mathcal{D} \mid \boldsymbol{X}, I_1, \ldots, I_K)}{p(\mathcal{D}_0 \mid \boldsymbol{X}, I_1, \ldots, I_K)} \xrightarrow{\bar{P}} 0, \\ &ii) \quad \frac{p(\mathcal{D} \mid \boldsymbol{X}, I_1, \ldots, I_K)}{p(\mathcal{D}_0 \mid \boldsymbol{X}, I_1, \ldots, I_K)} \xrightarrow{\bar{P}} \frac{p(\mathcal{D})}{p(\mathcal{D}_0)} \quad \textit{for all } \mathcal{D} \in [\mathcal{D}_0]. \end{aligned}$$

Proof. See Supplementary Material. □

Proposition 3.3 shows that posterior ratio and Bayes factor consistency under DAG-Wishart prior holds *outside* the Markov equivalence class of the true generating DAG $\mathcal{D}_0$. On the other hand, the posterior ratio tends to the prior ratio (Bayes factor equal to one) within the true equivalence class. This result is coherent with Peluso and Consonni (2020), in the context of a single-group observational dataset. We refer the reader to the Supplementary Material for a discussion of the assumptions underlying the result.

4 MCMC scheme and posterior inference

We construct a collapsed Metropolis-Hastings sampler (Metropolis et al., 1953) on the space of DAGs and intervention targets to approximate the marginal posterior

$$p(I_1, \dots, I_K, \mathcal{D} \,|\, \boldsymbol{X}) \propto m\big(\boldsymbol{X} \,|\, I_1, \dots, I_K, \mathcal{D}\big) \cdot p(I_1, \dots, I_K)\, p(\mathcal{D}). \tag{12}$$

The full conditional distribution of $\mathcal{D}$ is $p(\mathcal{D} \,|\, I_1, \dots, I_K, \boldsymbol{X}) \propto m(\boldsymbol{X} \,|\, I_1, \dots I_K, \mathcal{D})\, p(\mathcal{D})$. To update DAG $\mathcal{D}$ we implement a Metropolis-Hastings step where, given the current DAG, a new DAG $\widetilde{\mathcal{D}}$ is proposed from as suitable proposal $q(\widetilde{\mathcal{D}} \,|\, \mathcal{D})$ and accepted with probability

$$\alpha = \min\left\{1; \frac{m\big(\boldsymbol{X} \,|\, I_1, \dots, I_K, \widetilde{\mathcal{D}}\big)}{m\big(\boldsymbol{X} \,|\, I_1, \dots, I_K, \mathcal{D}\big)} \cdot \frac{p(\widetilde{\mathcal{D}})}{p(\mathcal{D})} \cdot \frac{q(\mathcal{D} \,|\, \widetilde{\mathcal{D}})}{q(\widetilde{\mathcal{D}} \,|\, \mathcal{D})}\right\}. \tag{13}$$

Conditionally on DAG $\mathcal{D}$ we update the K targets $I_1, \dots, I_K$ (equivalently, the indicator vectors $\boldsymbol{h}_1, \dots, \boldsymbol{h}_K$ introduced in Section 2.4) sequentially. For a given k, the full conditional of I_k is $p(I_k \,|\, \{I_s\}_{s \neq k}, \mathcal{D}, \boldsymbol{X}) \propto m(\boldsymbol{X} \,|\, I_1, \dots, I_K, \mathcal{D}) p(I_1, \dots, I_K)$. Update of I_k conditionally on $\{I_s\}_{s \neq k}$ and DAG $\mathcal{D}$ is again performed through a Metropolis Hastings step, where a new target $\widetilde{I}_k$ proposed from $q(\widetilde{I}_k \,|\, I_k)$ is accepted with probability

$$\beta_k = \min\left\{1; \frac{m\big(\boldsymbol{X} \,|\, \widetilde{I}_k, \{I_s\}_{s \neq k}, \mathcal{D}\big)}{m\big(\boldsymbol{X} \,|\, I_k, \{I_s\}_{s \neq k}, \mathcal{D}\big)} \cdot \frac{p(\widetilde{I}_k)}{p(I_k)} \cdot \frac{q(I_k \,|\, \widetilde{I}_k)}{q(\widetilde{I}_k \,|\, I_k)}\right\}. \tag{14}$$

We refer the reader to Section 4 of Supplementary Material for full details. The output is a collection of DAGs $\left\{\mathcal{D}^{(s)}\right\}_{s=1}^{S}$ and targets $\left\{I_1^{(s)}, \ldots, I_K^{(s)}\right\}_{s=1}^{S}$ approximately sampled from the posterior (12), where S is the number of finally kept MCMC iterations. Given this output, we can estimate, for each node j and target I_k, the posterior probability of intervention

$$\widehat{p}(j \in I_k \,|\, \boldsymbol{X}) \equiv \widehat{p}_{j \in I_k} = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\left\{j \in I_k^{(s)}\right\}, \tag{15}$$

a measure of evidence that intervention k acts on node j. In addition, an approximated marginal posterior distribution over the DAG space is available, with each DAG posterior probability estimated as $\widehat{p}(\mathcal{D} \,|\, \boldsymbol{X}) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\left\{\mathcal{D}^{(s)} = \mathcal{D}\right\}$, for $\mathcal{D} \in \mathcal{S}_q$, the set of all DAGs on q nodes. Furthermore, we can estimate with

$$\widehat{p}(u \to v \,|\, \boldsymbol{X}) \equiv \widehat{p}_{u \to v} = \frac{1}{S} \sum_{t=1}^{S} \mathbb{1}_{u \to v}\left\{\mathcal{D}^{(s)}\right\} \tag{16}$$

the posterior probability of inclusion of each directed edge $u \to v$, where $\mathbb{1}_{u \to v}\left\{\mathcal{D}^{(s)}\right\} = 1$ if $\mathcal{D}^{(s)}$ contains $u \to v$, 0 otherwise.

5 Simulation study

5.1 Simulated settings

To assess the performance of our method, we construct various simulated settings by varying the number of variables $q \in \{20, 40\}$, the sample size of each dataset $\boldsymbol{X}^{(k)}$ with increasing values $n^{(k)} \in \{10, 20, 50, 100, 200, 500\}$, for a number of interventions (datasets) $K = 4$. A family of intervention targets $I_1, \ldots, I_K$ is generated under two scenarios resembling different degrees of "sparsity" in the targets. Scenario *Sparse* is characterized by a moderate

number of interventions, with each target I_k obtained by drawing without replacement $s \in \{2, 4\}$ nodes, respectively for $q = 20$ or $q = 40$. On the other hand, in Scenario *Diffuse* we assign each node to one of the K targets. As a consequence, each variable is involved in one of the K interventions, with an overall larger number of simultaneous interventions (sizes of the targets I_k). Under each scenario defined by $(q, n^{(k)})$ and settings *Sparse* and *Diffuse*, we perform 40 simulations, each corresponding to a true DAG $\mathcal{D}$, family of targets $\{I_1, \ldots, I_K\}$ and resulting in a (multiple with $K = 4$ groups) dataset. For each simulation we first generate a topologically ordered DAG $\mathcal{D}$ with probability of edge inclusion $p_{edge} = 2/q$. Given DAG $\mathcal{D}$ and family targets $I_1, \ldots, I_K$, we then generate parameters $\boldsymbol{D}$, $\boldsymbol{L}$ and $\boldsymbol{\Phi}^{(k)} = \left\{\phi_j^{(k)}, j \in I_k\right\}$, by fixing $\boldsymbol{D} = \boldsymbol{I}_q$, and $\phi_j^{(k)} = 0.1$ for each $j \in I_k$ and $k = 1, \ldots, K$; non-zero elements of $\boldsymbol{L}$ are uniformly chosen in the interval $[-1, -0.1] \cup [0.1, 1]$; see also Equation (4). Next, for each $k = 1 \ldots, K$, $n^{(k)}$ i.i.d. interventional data collected in the $n^{(k)} \times q$ data matrix $\boldsymbol{X}^{(k)}$ are generated as in (4). Each dataset is therefore a collection of K interventional data matrices $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$.

For comparison we include the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm of Squires et al. (2020), with significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1% and IGSP 0.001% respectively), as recommended in the original paper. We also include Algorithm 1 of He and Geng (2016), implemented at significance level 0.05. As a further benchmark, we also construct a baseline node-wise regression approach, by adapting to our interventional setting the two-stage adaptive lasso method of Han et al. (2016); we call this benchmark *Node-wise*. Finally, we include the Greedy Interventional Equivalence Search (GIES) method of Hauser and Bühlmann (2012), implemented using the Extended Bayesian Information Criterion (EBIC, Foygel and Drton 2010) with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5 and GIES 1 respectively) as also recommended in the

original paper. The method of GIES was developed for known intervention targets that we input, as in an *oracle* setting, using the *true* intervened nodes. All methods can be adopted for DAG structure learning, whilst only IGSP and *Node-wise* can perform target estimation. Finally, notice that IGSP requires $n^{(k)} > q$, so that results of IGPS for $n^{(k)} \in \{10, 20\}$ are missing. Further details on benchmarks, together with convergence diagnostics and more extensive simulation settings are provided in Supplementary Material.

5.2 Results

We fix the number of MCMC iterations $S \in \{25000, 50000\}$ for $q \in \{20, 40\}$ respectively. In addition we set $g = 1/n$ in the Inverse Gamma priors (6) and (7), and $\eta = 1/q$ in the prior on DAG (Section 2.5) which corresponds to a prior probability of edge inclusion smaller than the expected level of sparsity, as commonly recommended; see for instance Barbieri and Berger (2004). Finally we fix $a_k = 1/q$, $b_k = 1$ for each $k = 1, \dots, K$ in the Beta prior leading to (10). Other scenarios not reported for brevity show that the results are quite insensitive to these hyperparameter choices, expecially for large sample sizes; for instance when we fix $\pi = 2/q$ there is a negligible change for $n^{(k)} = \{10, 20\}$ and no change is observed for higher sample sizes.

We start by evaluating the performance of the methods in identifying the intervention targets. With regard to our Bayesian method, we first compute for each simulation, the posterior probabilities $\widehat{p}_{j \in I_k}$ in (15). Next, by fixing a threshold of inclusion $z = 0.5$ we provide a target estimate $\widehat{I}_k$, $k = 1, \dots, K$, by including in $\widehat{I}_k$ all nodes j such that $\widehat{p}_{j \in I_k} > z$. Estimated targets $\widehat{I}_k$'s are compared with the true targets by computing the false positive

and false negative rates, respectively defined as

$$\text{FPR} = \frac{\sum_{j,k} \mathbb{1}\{j \in \widehat{I}_k, j \notin I_k\}}{\sum_{j,k} \mathbb{1}\{j \notin I_k\}}, \quad \text{FNR} = \frac{\sum_{j,k} \mathbb{1}\{j \notin \widehat{I}_k, j \in I_k\}}{\sum_{j,k} \mathbb{1}\{j \in I_k\}}; \tag{17}$$

similarly for the other frequentist methods under evaluation. Results for $q = 40$ setting are summarized in the box-plots of Figure 2, whilst results for $q = 20$ are reported in the Supplementary Material. This reports the distribution of FPR and FNR constructed across the $N = 40$ simulations for three methods under evaluation and increasing sample sizes $n^{(k)}$ under scenarios *Sparse* and *Diffuse*. With regard to our method (*Bayes*) we notice that coherently with the theoretical results of Section 3, both sources of error vanish as sample size increases. This tendency is more evident for FNR that rapidly goes to zero already at moderate sample sizes, e.g. $n^{(k)} = 20$. It is clear the outperformance of our proposal, relative to the benchmarks, with *Node-wise* performing equally well only in terms of FNR.

We then evaluate the overall performance of each methodology in recovering the DAG structure. With regard to our method, we provide a DAG estimate by computing first $\widehat{p}_{u\to v}$, the (estimated) posterior probability of inclusion, for each edge (u, v) as in (16); then we fix a threshold for edge inclusion $z = 0.5$ and obtain an estimate $\widehat{\mathcal{D}}$ by including all edges such that $\widehat{p}_{u\to v} > 0.5$, as in the median probability model proposed by Barbieri and Berger (2004) in a linear regression framework. We compare $\widehat{\mathcal{D}}$ with the true DAG by measuring the Structural Hamming Distance (SHD, Tsamardinos et al. 2006) between the two graphs; similarly for each DAG estimate directly outputted by the other methods; lower values of SHD correspond to better performances. Results for $q = 40$ are summarized in the box-plots of Figure 3, where each plot reports the distribution of SHD across the $N = 40$ simulated datasets for the various methods and increasing sample sizes $n^{(k)} \in \{10, \ldots, 500\}$ under Scenarios *Sparse* and *Diffuse*. It is clear the tendency of a better and better recovery of the true graphical structure as we increase the amount of available

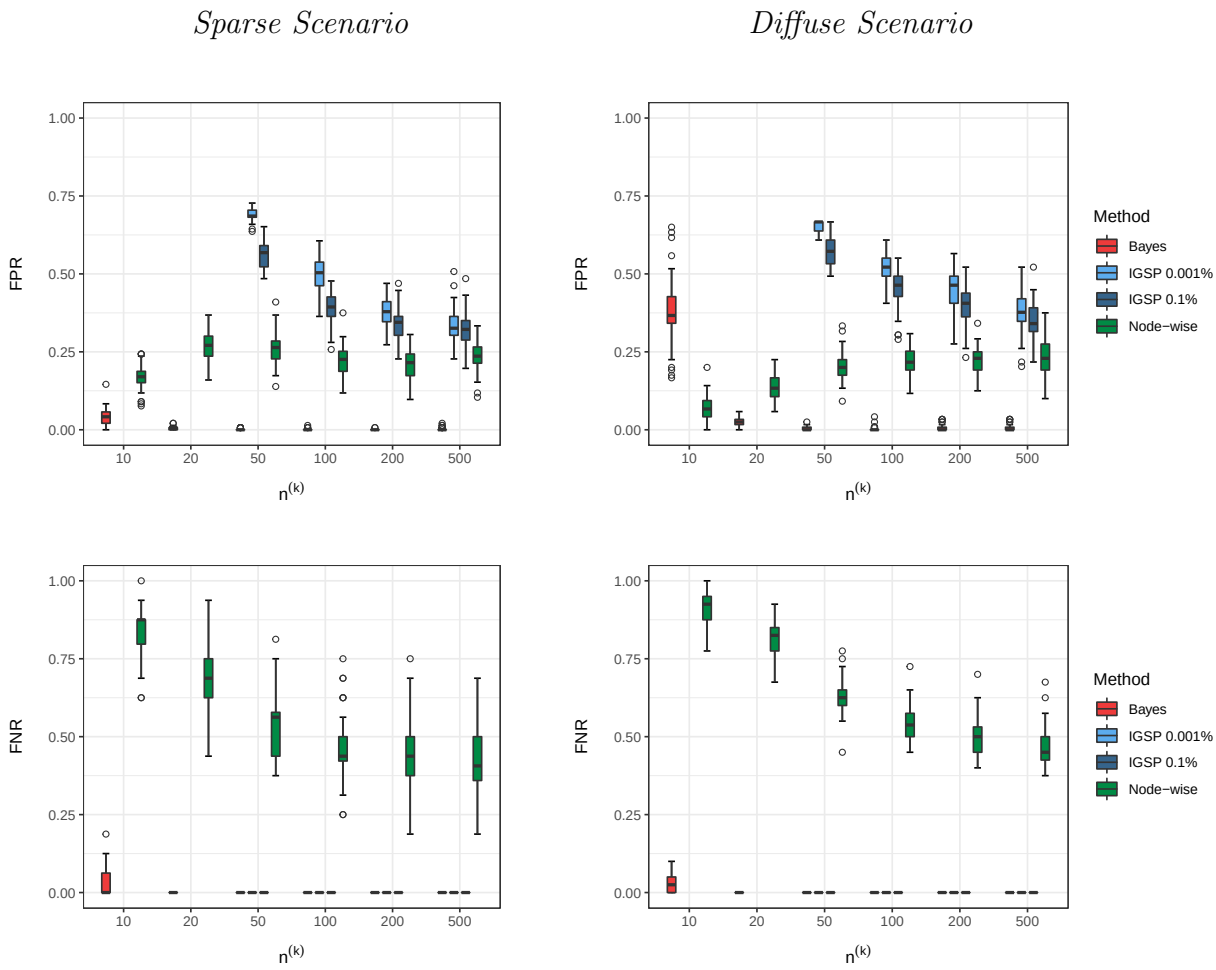

Figure 2: Simulations. Distribution of the False Positive Rate (FPR, first row) and False Negative Rate (FNR, second row) across $N = 40$ simulated datasets under Scenarios *Sparse* (first column) and *Diffuse* (second column) for number of nodes $q = 40$ and increasing sample sizes $n^{(k)}$. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1% and IGSP 0.001%) and node-wise regression (Node-wise).

data, and an overall better performance of *Bayes* relative to all the benchmarks. The only exception is GIES 0.5, however implemented with knowledge of the true targets, that outperforms our Bayesian method in few settings characterized by small sample sizes, where indeed target identification was more difficult for our method. However, it performs worse than *Bayes* as $n^{(k)}$ increases, especially under Scenario *Sparse*.

6 Real data analyses

6.1 Protein signalling data

In the current section we apply our methodology to the protein signalling data presented in Sachs et al. (2005). The dataset, provided as a supplement to the original paper, collects simultaneous measurements of multiple phosphorylated proteins and phospholipid components in individual primary human immune system cells. Measurements of $q = 11$ phosphorylated proteins and phospholipids were collected after a series of stimulatory cues and inhibitory interventions obtained from the administration of distinct reagents. Each reagent induces a perturbation of the proteins' pathway since it affects either one of the signalling molecules directly or some (unmeasured) receptor enzymes. More specifically, seven datasets are associated to known interventions, while other two (reagents CD3/CD28 and ICAM-2) refer to general (unknown) perturbations; see also Table 1 in Sachs et al. (2005) and our Table 1. The same dataset was analysed by Castelletti and Consonni (2019) who implemented their OBIES method on the collection of measurements associated to known targets to learn the structure of an interventional essential graph.

We include in our analysis all nine datasets, by assuming known intervention targets for the first seven, whilst considering the last two as characterized by uncertain interventions

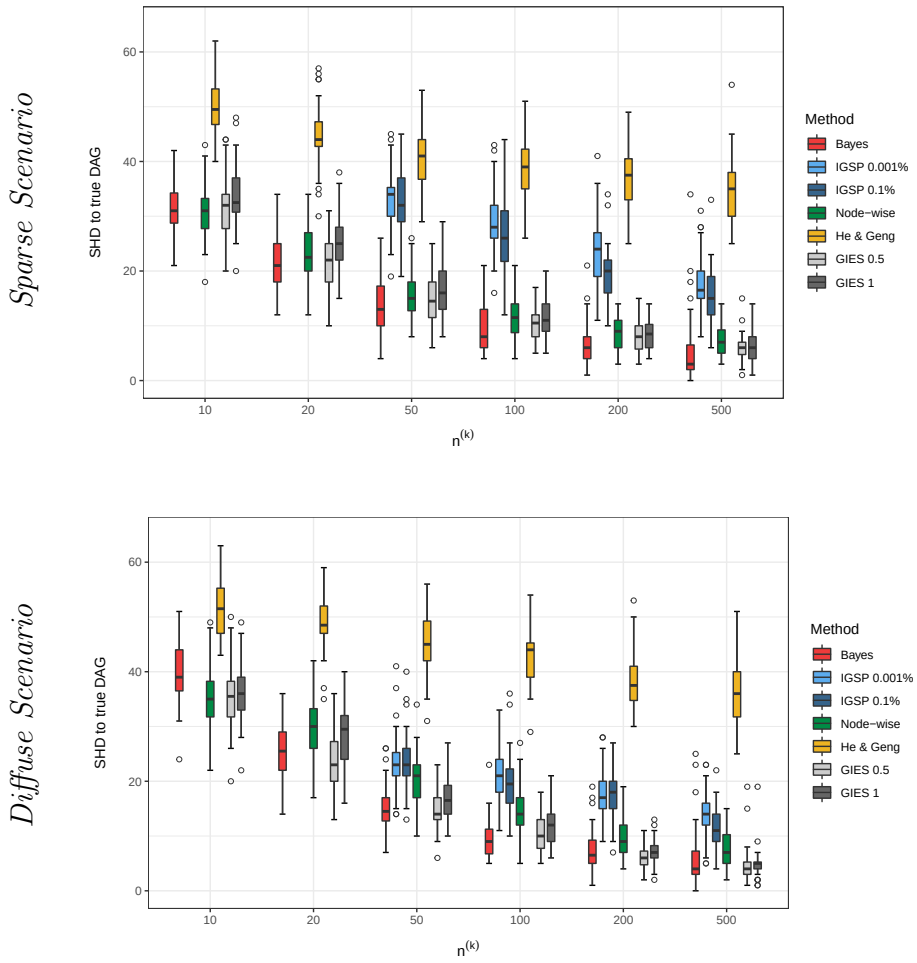

Figure 3: Simulations. Distribution across $N = 40$ simulated datasets of the Structural Hamming Distance (SHD) between estimated and true DAG for number of nodes $q = 40$ and increasing sample sizes $n^{(k)}$ under *Sparse* and *Diffuse* Scenarios. Methods under comparison are: our Bayesian methodology (Bayes), the *Unknown Target Interventional Greedy Sparsest Permutation* algorithm implemented at significance level $\alpha \in \{0.1\%, 0.001\%\}$ (IGSP 0.1%, IGSP 0.001%), node-wise regression (Node-wise), Algorithm 1 of He and Geng (2016) (He & Geng) and the *Greedy Interventional Equivalence Search* method with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5, GIES 1).

Reagent	Akt in.	G06976	Psitect	U0126	LY294002	PMA	B2cAMP	CD3/C28	ICAM-2
Target	Akt	PKC	PIP2	Mek	Akt	PKC	PKA	?	?
Sample size	911	723	810	799	848	913	707	853	902

Table 1: Sachs data. Intervention targets and sample sizes for each of the nine administrated reagents giving rise to the collection of nine datasets. Symbol ? indicates an unknown target for the corresponding reagent.

that we learn with our methodology. We run $S = 25000$ iterations of our MCMC algorithm to approximate the posterior distribution over the space of DAGs and intervention targets; we fix $g = 1/n$ in the priors on DAG parameters (6) and (7), a prior probability of edge inclusion $\eta = 1/q$ (Section 2.5) and $a_k = 1/q$, $b_k = 1$ $(k = 1, \ldots, K)$ in the Beta prior leading to (10).

Results are summarized in Figure 4, with the heat map of posterior probabilities of intervention for each node and each of the two uncertain intervention groups CD3/CD28 and ICAM-2. Black dots for the first seven interventions represent the targets that were assumed to be known. Our findings are coherent with biological literature establishing that both reagents ICAM-2 and CD3/C28 are capable of activating enzyme ZAP70 and in turn signalling nodes Mek, PLC and PKC among others. These are indeed selected as promising targets under at least one of the two interventions. We also report in the right panel of Figure 4 a DAG estimate obtained by including those edges whose posterior probability of inclusion exceeds 0.5: dark and light grey circles identify proteins whose posterior probability of intervention computed under CD3/C28 and ICAM-2 respectively is larger than 0.5, and therefore represent plausible intervention targets for the two reagents. We stress that a higher cardinality of the estimated intervention set in the two datasets with

unknown targets is not unexpected, since Sachs et al. (2005) refer to *general* perturbations that overall stimulated the cell, against *specific* perturbations acting on defined set of molecules in the remaining datasets with known targets. We refer to Sections 6 and 7 of Supplementary Material for sensitivity analyses and detailed comparisons with alternative methods. In particular, *He & Geng* does not provide target estimates, whilst we share 3 out of 4 targets with *Node-wise* in the dataset CD3/C28 and 1 out of 2 with *IGPS* in the dataset ICAM-2. Also, *Node-wise* and *He & Geng* estimate the same graph skeleton of our method, whilst *IGPS*, that wrongly assumes one dataset to be purely observational, misses two links. The benchmark *Node-wise* also share many edge orientations; on the other hand, with *He & Geng* many edges remain unoriented.

6.2 Gene expression profiles under antiepileptic drug therapies

In this section we consider a gene expression dataset relative to patients affected by epilepsy. The original dataset (publicly available at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143272`) includes measurements of about $13,000$ genes collected on subjects divided into: healthy patients, epilepsy patients drug-naïve treated and subjects treated with one antiepileptic monotherapy among *Valporate*, *Phenytoin* and *Carbamazepine*. The aim of the original study was to identify mRNA expression biomarkers associated with the disease and the antiepileptic drug response. Results in Rawat et al. (2020) revealed that patients showing differential response to antiepileptic monotherapies were also characterized by differential blood gene expression profiles.

In the following we consider four groups of subjects (*Drug-naïve*, *Valporate*, *Phenytoin* and *Carbamazepine*) to evaluate the effect of each drug therapy, relative to the untreated patients (*Drug-naïve* group). We include in our analysis 100 genes that were most differen-

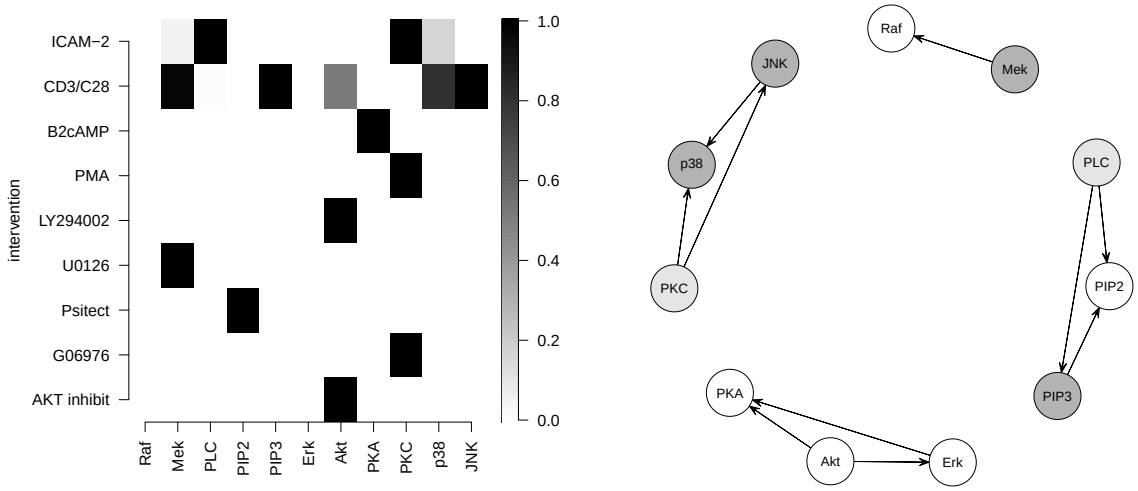

Figure 4: Sachs data. Heat map with estimated posterior probabilities of intervention computed under each of the nine interventions for each node u ($u = \text{Raf}, \ldots, \text{JNK}$), (left side). Estimated DAG with dark (light) grey circles representing nodes whose posterior probability of intervention under reagent CD3/C28 (ICAM-2) exceeds 0.5 (right side).

tially expressed between healthy and unhealthy drug-naïve patients as resulting from *limma* (*Linear Models for Microarray and RNA-Seq Data*) tests (Smyth, 2005) and therefore represent suitable candidate targets to evaluate patients' response to each drug therapy. Our algorithm is run for number of iterations $S = 1200000$ by fixing prior hyperparameters as in Section 6.1. Because our interest lies in discovering drug-induced effects (interventions) on patients who received one of the drug therapies as opposed to unhealty, yet untreated, patients, we consider drug-naïve individuals as a ground (reference) group and fix the corresponding intervention target as the empty set.

We first compute, for each node/gene $v \in \{1, \ldots, 100\}$ and each intervention target I_2, I_3, I_4 corresponding to one of the three drug therapies, the posterior probabilities of intervention; see Equation (15). Results reported in the Supplementary Material show that there are few genes exhibiting a high posterior probability of intervention under some of the treatments. Specifically, only six genes, that are reported in the sub-map of Figure 5, are associated with probabilities of intervention exceeding 0.5.

In addition, we show in the right panel of Figure 5 the estimated (median probability) sub-graph of these genes, including parent and child nodes. This DAG estimate can help understanding how gene dependencies modify in force of an intervention after one of the drug therapies is administrated. The implementation of alternative methods (Section 5.1) specifically designed for DAG and target learning under uncertain interventions revealed several difficulties that are specific to this kind of data. In particular, while IGSP cannot be applied since $n^{(k)} < q$, the baseline *Node-wise* method analyses separately the K datasets and does not identify any dependence relation between genes, even when implemented for different values of the tuning parameter . Finally, the approach of He and Geng (2016), shown to underperform in simulation, does not provide target estimates and outputs a DAG

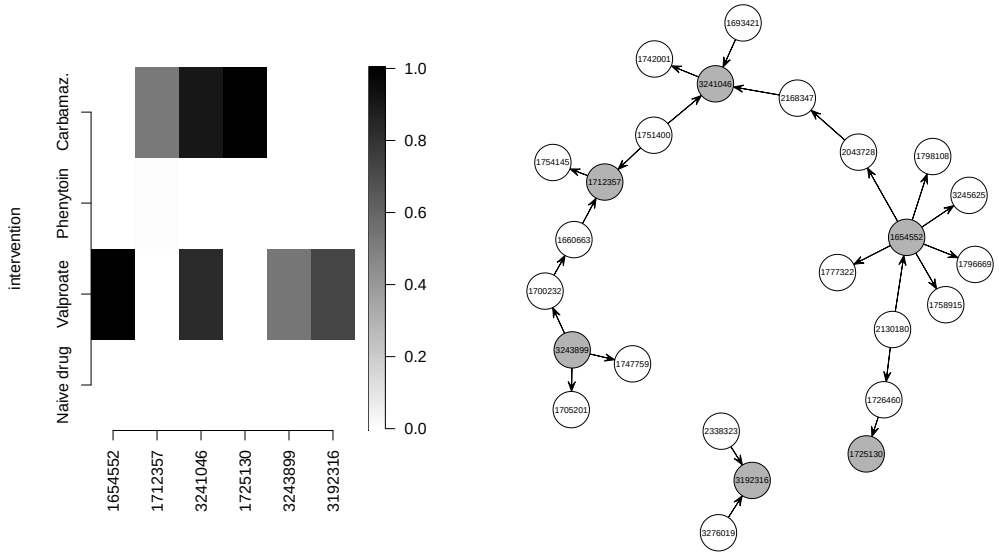

Figure 5: Epilepsy data. Left panel: Heat map with estimated posterior probabilities of intervention computed under each intervention (drug treatment) for selected nodes (genes $1654552, \ldots, 3192316$). Right panel: Estimated (pre-intervention) sub-graph of six selected genes, parent and child nodes, with grey circles representing intervention targets.

with a huge number of links, even for values of tuning parameter α encouraging sparsity. We refer the reader to Sections 6 and 7 of Supplementary Material for sensitivity analyses and further illustrations of the results.

7 Discussion

We have developed a Bayesian statistical methodology for simultaneous learning of network-based structure dependencies and of intervention targets. Our proposal is useful in those contexts where the data are subject to interventions, but each intervention affects unknown

variables in the network, as in diffuse protein perturbations or drug target discovery. We implement a novel MCMC sampler to approximate the posterior distribution on the space of DAGs and intervention targets. When applied to genomic data collected under various drug treatments, our method provides insights on the dependence relations between genes and the effect of distinct drug therapies. We also provide theoretical guarantees of the methodology, by looking at the consistency in recovering the true intervention targets and graph. Our theoretical results are supported by rigorous simulation studies, showing an overall outperformance of our method with respect to alternative approaches, in terms of target and structure learning.

Since changes in brain networks are known to be involved in stimulus-response associations (Boettiger and D'Esposito, 2005), a potential field of application of the developed methodology is on the joint learning of dependent changes in functional Magnetic Resonance Imaging (fMRI) activations within brain regions and functional connectivities between regions; see for instance Warnick et al. (2018). The human brain is an oriented network system of brain regions involving directional connectivity, and the mainstream statistical approach relies on the theory of random networks (Simpson et al., 2013). Still, many statistical issues remain unaddressed , with the study of complex dependencies among brain regions a fertile area of methodological development, often based on simplistic inferential frameworks. For instance, in the network construction process from raw fMRI data up to the adjacency matrix, methods for estimating functional connectivities between network nodes typically rely on association measures, whilst modeling methods remain relatively limited for brain network estimation. After estimating a functional brain network, the following step often involves various methods of crude thresholding of the connectivity matrix (Telesford et al., 2011), to remove weak connections and produce an adjacency matrix

which notes the presence or absence of a functional connection between any two nodes. We conjecture that an implementation of our methodology directly on fMRI data can lead to an alternative reliable estimation of the adjacency matrix characterizing the brain network and of the activated target areas under stimuli, avoiding the arbitrary steps involved in the process of network construction.

Our approach to structure learning of DAGs and targets relies on the *do-calculus* theory and the allied notion of intervention DAG (Pearl, 2000). Accordingly, we are able to recover interventions which modify the DAG Markov property, by destroying parent-child dependencies for each intervened variable. Alternative definitions of intervention are available in the literature. Among these, Kocaoglu et al. (2019) consider from a theoretical perspective the case of *soft interventions*, where parent-child dependencies are "modified" but yet preserved after intervention, and develop graphical criteria to represent the post-intervention DAG Markov property and characterize DAGs Markov equivalence. A Bayesian methodology for structure learning under soft interventions is possible, following the lines of Castelletti and Consonni (2019), and an extension to *uncertain soft interventions* is of interest, based on the premises developed in the current work.

Acknowledgments Research of F.C. was partially supported by UCSC (D1 and 2019-D.3.2 research grants). We gratefully acknowledge a helpful discussion with Guido Consonni (UCSC) during the drafting of this paper.

SUPPLEMENTARY MATERIAL

The file **supplementary_material.pdf** provides supplemental information to our paper, and is organized as follows. Sections 1 and 2 contain proofs and discussions of Propositions

3.1 and 3.2. In Section 3 we show and discuss results on graph consistency. Sections 4 and 5 provide details about the proposed MCMC scheme and additional simulated results. Finally, Sections 6 and 7 contain sensitivity analyses to hyperparameter choices and further results for the two real data applications.

References

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

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

Boettiger, C. A. and M. D'Esposito (2005). Frontal networks for learning and executing arbitrary stimulus-response associations. *Journal of Neuroscience 25*(10), 2723–2732.

Cao, X., K. Khare, and M. Ghosh (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *The Annals of Statistics 47*(1), 319–348.

Castelletti, F. and G. Consonni (2019). Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways. *The Annals of Applied Statistics 13*(4), 2289–2311.

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

Eaton, D. and K. Murphy (2007). Belief net structure learning from uncertain interventions. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 107–114.

Foygel, R. and M. Drton (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems 23*, pp. 2020–2028.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science 303*(5659), 799–805.

Friedman, N. and D. Koller (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning 50*(1), 95–125.

Han, S. W., G. Chen, M.-S. Cheon, and H. Zhong (2016). Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. *Journal of the American Statistical Association 111*(515), 1004–1019.

Hauser, A. and P. Bühlmann (2015). Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(1), 291–318.

Hauser, A. and P. Bühlmann (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research 13*(79), 2409–2464.

He, Y. and Z. Geng (2016). Causal network learning from multiple interventions of unknown manipulated targets. *arXiv preprint arXiv:1610.08611*.

Heckerman, D. and D. Geiger (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning 20*, 197–243.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions (2nd ed.)*. SAGE Publications.

Ke, N. R., O. Bilaniuk, A. Goyal, S. Bauer, H. Larochelle, B. Schölkopf, M. C. Mozer, C. Pal, and Y. Bengio (2019). Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*.

Kocaoglu, M., A. Jaber, K. Shanmugam, and E. Bareinboim (2019). Characterization and learning of causal graphs with latent variables from soft interventions. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32, pp. 14369–14379. Curran Associates, Inc.

Koivisto, M. and K. Sood (2004). Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research 5*, 549–573.

Korb, K. B., L. R. Hope, A. E. Nicholson, and K. Axnick (2004). Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*, pp. 322–331. Springer.

Kuipers, J. and G. Moffa (2017). Partition MCMC for inference on acyclic digraphs. *Journal of the American Statistical Association 112*(517), 282–299.

Marton, M. J., J. L. DeRisi, H. A. Bennett, V. R. Iyer, M. R. Meyer, C. J. Roberts, R. Stoughton, J. Burchard, D. Slade, H. Dai, D. E. Bassett, L. H. Hartwell, P. O.

Brown, and S. H. Friend (1998). Drug target validation and identification of secondary drug target effects using dna microarrays. *Nature Medicine 4*, 1293–1301.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of Chemical Physics 21*(6), 1087–1092.

Ni, Y., F. C. Stingo, and V. Baladandayuthapani (2017). Sparse multi-dimensional graphical models: A Unified bayesian framework. *Journal of the American Statistical Association 112*(518), 779–793.

Ni, Y., F. C. Stingo, and V. Baladandayuthapani (2019). Bayesian graphical regression. *Journal of the American Statistical Association 114*(525), 184–197.

Paananen, J. and V. Fortino (2019). An omics perspective on drug target discovery platforms. *Briefings in Bioinformatics 21*(6), 1937–1953.

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

Peluso, S. and G. Consonni (2020). Compatible priors for model selection of high-dimensional Gaussian DAGs. *Electronic Journal of Statistics 14*(2), 4110–4132.

Peters, J., P. Bühlmann, and N. Meinshausen (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 78*(5), 947–1012.

Rawat, C., S. Kushwaha, A. K. Srivastava, and R. Kukreti (2020). Peripheral blood

gene expression signatures associated with epilepsy and its etiologic classification. *Genomics 112*(1), 218 – 224.

Rawat, C., R. Kutum, S. Kukal, A. Srivastava, U. R. Dahiya, S. Kushwaha, S. Sharma, D. Dash, L. Saso, A. K. Srivastava, and R. Kukreti (2020). Downregulation of peripheral PTGS2/COX-2 in response to valproate treatment in patients with epilepsy. *Scientific Reports 10*(2546).

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

Scott, J. G. and J. O. Berger (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics 38*(5), 2587–2619.

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

Simpson, S. L., F. D. Bowman, and P. J. Laurienti (2013). Analyzing complex functional brain networks: fusing statistics and network science to understand the brain. *Statistics surveys 7*, 1.

Smyth, G. K. (2005). *limma: Linear Models for Microarray Data*, pp. 397–420. New York, NY: Springer New York.

Spirtes, P., C. N. Glymour, R. Scheines, and D. Heckerman (2000). *Causation, prediction, and search*. MIT press.

Squires, C., Y. Wang, and C. Uhler (2020). Permutation-based causal structure learning

with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR.

Telesford, Q. K., S. L. Simpson, J. H. Burdette, S. Hayasaka, and P. J. Laurienti (2011). The brain as a complex system: using network science as a tool for understanding the brain. *Brain connectivity 1*(4), 295–308.

Tsamardinos, I., L. E. Brown, and C. F. Aliferis (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning 65*(1), 31–78.

Warnick, R., M. Guindani, E. Erhardt, E. Allen, V. Calhoun, and M. Vannucci (2018). A Bayesian approach for estimating dynamic functional network connectivity in fMRI data. *Journal of the American Statistical Association 113*(521), 134–151.

Zhang, K., B. Huang, J. Zhang, C. Glymour, and B. Schölkopf (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, Volume 2017, pp. 1347. NIH Public Access.