Statistical Science
2024, Vol. 39, No. 2, 305–321
https://doi.org/10.1214/23-STS905

Bayesian Sample Size Determination for Causal Discovery

Federico Castelletti and Guido Consonni

Abstract. Graphical models based on Directed Acyclic Graphs (DAGs) are widely used to answer causal questions across a variety of scientific and social disciplines. However, observational data alone cannot distinguish in general between DAGs representing the same conditional independence assertions (Markov equivalent DAGs); as a consequence, the orientation of some edges in the graph remains indeterminate. Interventional data, produced by exogenous manipulations of variables in the network, enhance the process of structure learning because they allow to distinguish among equivalent DAGs, thus sharpening causal inference. Starting from an equivalence class of DAGs, a few procedures have been devised to produce a collection of variables to be manipulated in order to identify a causal DAG. Yet, these algorithmic approaches do not determine the sample size of the interventional data required to obtain a desired level of statistical accuracy. We tackle this problem from a Bayesian experimental design perspective, taking as input a sequence of target variables to be manipulated to identify edge orientation. We then propose a method to determine, at each intervention, the optimal sample size to produce an experiment which, with high assurance, will deliver an overall probability of decisive and correct evidence.

Key words and phrases: Active learning, Bayes factor, Bayesian experimental design, directed acyclic graph, intervention.

1. INTRODUCTION

Graphical models based on Directed Acyclic Graphs (DAGs) are widely used to represent dependence relations among random variables [40, 17, 39]. Applications of DAG models in various scientific areas abound, especially in genomics; see, for instance, Friedman [25], Sachs et al. [58], Shojaie and Michailidis [60], Nagarajan, Scutari and Lèbre [46].

It is well known that distinct DAGs can encode the same set of conditional independencies, and their collection is named the Markov equivalence class. Unfortunately, one cannot distinguish between Markov equivalent DAGs using observational data alone [14], without imposing specific assumptions on the statistical model [53]. For each Markov equivalence class, there exists a unique Completed *Partially Directed* Acyclic Graph (CPDAG), also named Essential Graph (EG) [2], which can be taken as representative of the class.

In practice, the structure of a DAG governing the joint distribution of the observations is unknown, and so is the corresponding CPDAG. Learning the structure of a CPDAG has been the subject of several papers. In the frequentist framework, the two most popular methods are the Greedy Equivalence Search (GES) of Chickering [14] and the PC-algorithm of Spirtes, Glymour and Scheines [63], later extended to high-dimensional settings by Kalisch and Bühlmann [37]. From a Bayesian perspective, learning a CPDAG is a model selection problem, which can be approached using the Bayes factor [38] as in Castelletti et al. [8]; see also Castelletti and Peluso [9].

When equipped with assumptions based on do-calculus theory [50], DAGs are also used to answer causal queries in a variety of scientific domains; see Pearl [50] for a scholarly treatment and Pearl [51] for an expository presentation, while Imbens [34] and Dawid [20] offer a more critical view. Because we can only learn a Markov equivalence class using observational data, we will end up with a *collection* of causal effects for the same intervention (each DAG may potentially produce a distinct value). One strategy to handle the resulting multiplicity is to report lower

Federico Castelletti is Assistant Professor, Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy (e-mail: federico.castelletti@unicatt.it). Guido Consonni is Professor, Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy (e-mail: guido.consonni@unicatt.it).

and upper bounds for the causal effect within an equivalence class inferred from the data [43]. Alternatively, one can first summarize multiple effects for each class, and then apply Bayesian Model Averaging (BMA) with respect to the posterior distribution on the space of equivalence classes [7].

For a given CPDAG, a more ambitious line of action is to orient the remaining undirected edges (partly or wholly) in order to sharpen our inference on the causal effect. This can be achieved by applying *interventions* on selected nodes (variables) of the graph, that is, setting exogenously their values (perfect or hard interventions); see Hauser and Bühlmann [31]. More general (nonperfect or soft) interventions are available [72] but are not considered in this paper.

The reason why interventions allow to identify the direction of an arrow is that two observationally Markov equivalent DAGs need not be equivalent under intervention, and this fact can be leveraged to split the original equivalence class into smaller interventional equivalence classes [29]. DAG identification through interventions, also named *active learning*, has been the subject of several contributions over the last two decades or so, especially from the computer science community. In particular, He and Geng [32] present graph-based algorithms for optimally selecting intervention variables in order to obtain DAG-identifiability starting from a given Markov equivalence class; we shall return to this issue with greater details in Section 2.2.

The results available on active learning are usually based on a collection of independence tests, which eventually will have to be performed based on sample data. Yet little attention is paid to statistical features of these procedures, and in particular to *sample size determination* (SSD) required to guarantee, before data collection, a reasonably high assurance that desirable inferential properties will be delivered. In this paper, we investigate SSD for active learning in causal discovery from a Bayesian experimental design perspective. Specifically, for a given CPDAG, we frame the problem of edge orientation as a comparison between two competing causal DAG models, and use the Bayes factor as a measure of evidence. Next, based on a sequence of intervention variables leading to causal DAG identification, we determine the corresponding sequence of minimal sample sizes with provable large probability of correct evidence in favor of edge orientation.

The rest of this paper is organized as follows. In Section 2, we present useful background material on DAGs, Markov equivalence classes, interventions and active learning methods for causal discovery. We also present elements of Bayesian experimental design and SSD, which will be required to appreciate the main contribution of our work. In Section 3, we present our strategy for Bayesian SSD, which is then specialized to Gaussian DAG models in Section 4. Section 5 illustrates our method on a simple example and then applies it to a high-dimensional data set. Finally, in Section 6 we discuss some critical points and present new settings of application of the proposed methodology. A few technical results relative to priors for DAG-model parameters and computations of Bayes factors are reported in the Appendix.

2. BACKGROUND

2.1 DAGs, Markov Equivalence and Interventions

Let $\mathcal{D} = (V, E)$ be a Directed Acyclic Graph (DAG) whose vertices $V = \{1, \dots, q\}$ correspond to variables $Y_1, \dots, Y_q$ and $E \subseteq V \times V$ is the set of directed edges. If the joint distribution of the variables factorizes according to $\mathcal{D}$ (see equation (1)), it encodes a set of conditional independence relations, which can be read off from the DAG, for example, by using *d-separation* [50]. We assume that an observational data set $\boldsymbol{Z}$ is available, where

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{z}_1^\top \\ \boldsymbol{z}_2^\top \\ \vdots \\ \boldsymbol{z}_N^\top \end{pmatrix},$$

and $\boldsymbol{z}_i = (z_{i,1}, \dots, z_{i,q})^\top$ for $i = 1, \dots, N$. Let $[\mathcal{D}]$ be the equivalence class of $\mathcal{D}$, which can be identified through $\boldsymbol{Z}$. All DAGs in $[\mathcal{D}]$ are characterized by having the same skeleton (the underlying undirected graph obtained by disregarding edge orientation) and v-structures (subgraphs of the form $u \to v \leftarrow z$ with u and z not connected) [68]. Moreover, each equivalence class can be uniquely represented by a *partially* directed acyclic graph named the Essential Graph (EG) [2] or Completed Partially Directed Acyclic Graph (CPDAG) [14]. Let $\mathcal{E}(\mathcal{D}) \equiv \mathcal{E} = (V, E_{\mathcal{E}})$ be the CPDAG representing $[\mathcal{D}]$. Andersson, Madigan and Perlman [2] show that $\mathcal{E}$ is a chain graph with decomposable chain components. We let $\mathcal{T}$ be the set of chain components of $\mathcal{E}$, with element $\tau \in \mathcal{T}$, and $\mathcal{E}_\tau = (\tau, E_\tau)$ the subgraph of $\mathcal{E}$ induced by τ, where $E_\tau = \{(u, v) \in E_{\mathcal{E}} | u, v \in \tau\}$. Importantly, $\mathcal{T}$ defines a partition of V, and each chain component corresponds to an *undirected* decomposable graph, while edges between nodes belonging to distinct chain components are directed; see also Figure 1 for a simple example.

Under DAG $\mathcal{D}$, the joint density of $\boldsymbol{Y} = (Y_1, \dots, Y_q)$ is assumed to factorize as follows:

$$f(\boldsymbol{y}|\mathcal{D}) = \prod_{j=1}^{q} f(y_j | \boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)}), \tag{1}$$

where $\mathrm{pa}_{\mathcal{D}}(j)$ is the set of parents of node j in $\mathcal{D}$ and $\boldsymbol{y}_A$ is the vector of variables representing nodes in $A \subseteq V$. Consider now an intervention on Y_u, $u \in V$, which replaces Y_u with a new r.v. $\widetilde{Y}_u$ having density $\widetilde{f}_u(\cdot)$; this is

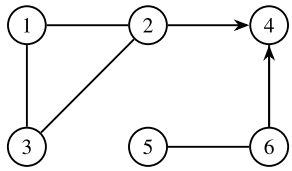

FIG. 1. *A CPDAG with three chain components* $\tau_1 = \{1,2,3\}, \tau_2 = \{4\}, \tau_3 = \{5,6\}$. *Edges between chain components are directed, while edges linking nodes belonging to the same chain component are undirected.*

slightly more general than the simple hard intervention because the assigned value is allowed to be generated according to a random mechanism. We call Y_u the manipulated variable (also named *intervention target*) [29], and the *do-operator* [50] is used to denote such an intervention. The *post-intervention* joint distribution of $\boldsymbol{Y}$ is defined as

$$(2)\quad f\big(\boldsymbol{y}|\mathrm{do}(Y_u = \widetilde{Y}_u), \mathcal{D}\big) = \tilde{f}_u(y_u) \prod_{j \neq u} f(y_j|\boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)}),$$

where the $f(y_j|\boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)})$'s are the preintervention densities appearing in (1), and y_u is the value taken on by $\widetilde{Y}_u$. We emphasize that the assumption embedded in equation (2) equips the DAG with a causal meaning. It states that the Markovian structure (1) of the joint distribution is stable under exogenous perturbations of the system save for its uth component, which is the only term affected by the intervention. For a critical appraisal of this assumption, see Dawid [20].

As mentioned in Section 1, interventional data can be used to identify the orientation of an undirected edge in $\mathcal{E}$. Specifically, let $\mathcal{E}$ be a CPDAG and suppose that the undirected edge $u - v$ occurs in $\mathcal{E}$. This implies that there are two DAGs in the Markov equivalence class of $\mathcal{E}$, which are identical save for the orientation of the edge between u and v: $\mathcal{D}_0$, which contains $u \leftarrow v$, and $\mathcal{D}_1$ containing $u \rightarrow v$; see Chickering [13] and Andersson, Madigan and Perlman [2], Lemma 3.2. From (2), one can show, assuming faithfulness [63], that following an intervention on Y_u,

$$(3)\quad Y_u \perp\!\!\!\perp Y_v \quad \text{under } \mathcal{D}_0, \qquad Y_u \not\perp\!\!\!\perp Y_v \quad \text{under } \mathcal{D}_1.$$

Result (3) is a consequence of *d-separation* [17], Section 5.3. For instance, independence under $\mathcal{D}_0$ stems from the fact that u and v are *separated* in the moral graph of the ancestral set of $\{u, v\}$.

2.2 Active Learning for Causal Discovery

Because of faithfulness, by manipulating a sufficiently large number of nodes, we can in principle identify a DAG structure through independence tests between each intervened variable and its neighbors. He and Geng [32] propose an optimal design strategy, first estimate a Markov equivalence class and then orient which minimizes the number of manipulated variables that are needed to guarantee that an equivalence class is progressively partitioned into smaller classes, eventually comprising a single DAG. Specifically, a sequence of manipulated variables $S = (u_1, \ldots, u_K)$ is *sufficient* for the CPDAG $\mathcal{G}$ if one can identify a single DAG from all possible DAGs in $\mathcal{G}$ after variables in S are manipulated; additionally, the minimal (sufficient) sequence is defined as follows.

DEFINITION 2.1 (He and Geng [32]). Let $S = (u_1, \ldots, u_K)$ be a sequence of manipulated variables. Then S is minimal if $|S| = \min\{|S_l| : S_l \in \mathcal{S}\}$, where $|S_l|$ is the number of elements in S_l, and $\mathcal{S}$ is the set of all sufficient sequences.

Note that Definition 2.1 represents the basis for defining an optimal design in He and Geng [32]. An important feature of their method is that it can be implemented *locally*, that is, separately for each chain component of the CPDAG, thus greatly expediting the learning process. They also propose two kinds of intervention experiments: *batch*, which determines upfront the minimum set of variables to be manipulated so that undirected edges are all oriented after the interventions, and *sequential*, wherein at each step one chooses a target variable, which reduces the current Markov equivalence class into the smallest possible, until reaching a singleton set containing a unique DAG. We will return to the approach of He and Geng [32] in Section 3.2, when we present our method for SSD.

Hauser and Bühlmann [30] propose two methods for active learning based on sequential intervention experiments. The first one is a greedy approach for single-vertex interventions that maximizes the number of edges that can be oriented after each intervention. The second one yields in polynomial time a minimum set of targets of arbitrary size that guarantees full identifiability. There are two noteworthy features of their approach. First, it overcomes some computational inefficiencies related to the enumeration of all DAGs within each chain component inherent in [32]. In addition, it considers *all* data (observational and interventional) collected up to the current point, which are jointly modeled as in the GIES method [29]. Importantly, for the motivation of our paper, their analysis emphasizes the role played by the sample size of the collected data at each stage, although their investigation is only conducted through simulations across a few scenarios; for similar considerations, see also Castelletti and Consonni [6], Section 7. Further relevant papers on active learning are Tong and Koller [67], Meganck, Leray and Manderick [44], Eberhardt [23], Hyttinen, Eberhardt and Hoyer [33], and more recently von Kügelgen et al. [69], Squires et al. [64], Peng, Shen and Pan [52].

2.3 Bayesian Experimental Design and Sample Size Determination

The Bayesian approach to experimental design has a long tradition. Lindley was a pioneer and supported a

decision-theoretic approach; see, for instance, Lindley [41]. Following that approach, Chaloner and Verdinelli [12] presented a unified perspective on the topic with an excellent review up to the mid-1990s. Another almost contemporary review was provided in DasGupta [18].

In this paper, we focus on a specific aspect of design, namely SSD. This was conceptualized in the influential book Raiffa and Schlaifer [55] and has been the subject of several papers in the years to follow. In the 1997 issue of the *Journal of the Royal Statistical Society Series D* entirely devoted to SSD, several papers adopted the Bayesian viewpoint; in particular, we single out Lindley [42], which is based on the principle of the maximization of expected utility, Weiss [71] which deals with hypothesis testing and Adcock [1] which presents a review. Because of its more pragmatic content, Bayesian SSD has been widely discussed in a variety of applied contexts, notably clinical trials, an early instance being Spiegelhalter and Freedman [62]; see also the comprehensive book by Spiegelhalter, Abrams and Myles [61] and references therein. O'Hagan and Stevens [48] carefully distinguished two objectives, analysis and design, leading to the use of two distinct priors for SSD: the *analysis* and the *design* prior. The simultaneous use of two different priors for the same parameter is actually not new: in a different context, it was advocated in an earlier paper by Etzioni and Kadane [24].

Any approach to SSD is predicated on the type of statistical inference one wishes to perform. This is often the test of an hypothesis on a parameter of interest, which typically reduces to comparing a simple null hypothesis against a two-sided alternative, or two composite hypotheses, each being one-sided. Spiegelhalter, Abrams and Myles ([61], Section 6.5) discuss a hybrid, as well as a full, Bayesian approach to the problem. In the *hybrid* case, a standard frequentist size-α null-rejection region is considered. Next, a prior is assigned to the parameter, and the classical power function is integrated with respect to the prior, leading to an *unconditional*, or expected, "classical" power. Equivalently, one evaluates the (prior)-predictive probability that the test statistic falls in the rejection region of the null hypothesis. Clearly, classical *conditional* power used in SSD can be recovered as a special case by assigning a degenerate prior on a fixed value of the parameter. The optimal sample size is finally derived so that the unconditional power is equal to a prespecified value, 80% say. The *full* Bayesian approach instead first needs to specify when the null hypothesis should be rejected, a sort of "Bayesian significance." One option is to require that the posterior probability of the null falls below a fixed threshold. In a preposterior analysis, when the observations are yet to be collected and should be regarded as random variables, the above is an uncertain event and implicitly defines a sampling rejection region for the null [61].

If one does not want to use prior probabilities of the hypotheses for SSD, an alternative is to adopt the Bayes factor [38] (BF) directly as a measure of evidence. This is the approach taken in Weiss [71], which considers testing a point null against a general two-sided alternative under a normal model with known variance. A useful feature of this early paper is that it produces the plots of the prior-predictive distribution of the BF under the null and the alternative (represented by a normal prior for the mean parameter). It is apparent that, for a variety of reasonable sample sizes, the BF is likely to reach convincing evidence according to traditional scales (e.g., Table 1) when the alternative is assumed to hold; while this is hardly ever the case when the null is assumed to be true; see Weiss [71] for a numerical illustration of this phenomenon. This imbalance in the learning rate happens because the null hypothesis is nested into the alternative, so that the BF grows essentially as the square root of the sample size under the null, whereas the rate of growth is exponential under the alternative; for a theoretical justification, see Dawid [4]. This fact suggests that treating the symmetrically two-nested hypothesis for SSD can be problematic. One possible solution to this problem is setting distinct evidential thresholds for the acceptance of the two hypotheses. An alternative is to use type I error rate to fix the threshold for rejecting H_0, and then determine the sample size required to have a high Bayesian power; these are discussed in Weiss [71].

Gelfand and Wang [70] present a simulation-based framework for Bayesian SSD capable of handling more complex settings such as generalized linear models and hierarchical models, as well as planning an experiment for model separation (choice between two models). Their framework makes a repeated use of the *fitting* and *sampling* priors, which play the same role of the analysis and design priors of O'Hagan and Stevens [48].

De Santis [22] extends the evidential approach of Royall [57, 56] to Bayesian SSD. Since his work introduces important concepts useful also for this paper, we provide below a short summary.

Consider two hypotheses H_0 and H_1, and let y^n be a sample of observations of size n. Let $\mathrm{BF}_{01}(y^n)$ be the Bayes factor in favor of H_0 against H_1, and let $p(H_i)$ be the prior probability associated with H_i, $i = 0, 1$. For a fixed value γ_0, we say that the data provide *decisive* evidence in favor of H_0 at level γ_0 if $\Pr(H_0|y^n) > \gamma_0$, equivalently if $\mathrm{BF}_{01}(y^n) > \omega\frac{\gamma_0}{1-\gamma_0} := k_0$, where $\omega = p(H_1)/p(H_0)$ is the prior odds. Similarly, for a fixed value γ_1, the data provide decisive evidence in favor of H_1 at level γ_1 if $\Pr(H_1|y^n) > \gamma_1$, equivalently if $\mathrm{BF}_{01}(y^n) < \omega\frac{1-\gamma_1}{\gamma_1} := 1/k_1$. Once the data have been gathered, the BF will be computed and evaluated against k_0 and k_1. For a suitably large value k_0, $\mathrm{BF}_{01}(y^n) > k_0$ will be considered *decisive* evidence in favor of H_0, and similarly, for

TABLE 1
Classification scheme for the interpretation of Bayes factor BF_{01} *(from Schönbrodt and Wagenmakers* [59] *adapted from Jeffreys* [35])

Bayes factor	Evidence category
> 100	Extreme evidence for H_0
$30 - 100$	Very strong evidence for H_0
$10 - 30$	Strong evidence for H_0
$3 - 10$	Moderate evidence for H_0
$1 - 3$	Anecdotal evidence for H_0
1	No evidence
$1/3 - 1$	Anecdotal evidence for H_1
$1/10 - 1/3$	Moderate evidence for H_1
$1/30 - 1/10$	Strong evidence for H_1
$1/100 - 1/30$	Very strong evidence for H_1
$< 1/100$	Extreme evidence for H_1

a large enough k_1, $\mathrm{BF}_{01}(y^n) < 1/k_1$ will be considered *decisive* evidence in favor of H_1. While in the exposition so far the values of k_i depend on the threshold probabilities γ_i and the prior probabilities $p(H_i)$, one can fix k_i directly having in mind a classification of evidence based on the BF, such as that provided by Schönbrodt and Wagenmakers [59], which is an adjustment of the original table presented in Jeffreys [35]; see Table 1. The latter provides broad indications for researchers working in the social sciences, especially experimental psychology. Finally, we declare that $1/k_1 \leq \mathrm{BF}_{01}(y^n) \leq k_0$ corresponds to *inconclusive* evidence.

It is instructive to consider the probability of evidential support provided by the Bayes factor *conditionally* on each H_i. Thus, we obtain for $i, j = 0, 1$ and $i \neq j$:

- $p_i^I(k_0, k_1, n)$: the probability of *Inconclusive evidence*, namely $\frac{1}{k_1} \leq \mathrm{BF}_{01}(y^n) \leq k_0$, conditionally on H_i;
- $p_i^{DC}(k_i, n)$: the probability of *Decisive and Correct evidence*, namely $\mathrm{BF}_{ij}(y^n) > k_i$, conditionally on H_i;
- $p_i^M(k_j, n)$: the probability of *Misleading evidence*, namely $\mathrm{BF}_{ij}(y^n) < \frac{1}{k_j}$, conditionally on H_i.

The *unconditional* probability for any of the above three types of evidential support can be obtained by averaging the corresponding conditional probability w.r.t. the prior probabilities $p(H_i)$. In particular, we have

$$p^{DC}(k_0, k_1, n) = p(H_0)\, p_0^{DC}(k_0, n) + p(H_1)\, p_1^{DC}(k_1, n),$$

which represents the overall preexperimental evaluation of the potential success of the experiment. Hence, it is proposed to choose the optimal sample size n^* based on $p^{DC}(k_0, k_1, n)$. Specifically, for $\zeta \in (0, 1)$,

$$n^* = \min\{n \in \mathbb{N} : p^{DC}(k_0, k_1, n) \geq \zeta\}. \tag{4}$$

Of course, besides guaranteeing *ex ante* a fairly high level for $p^{DC}(k_0, k_1, n)$, it would be also useful to control that the unconditional probabilities of inconclusive and misleading evidence are fairly low.

Recall that $p^{DC}(k_0, k_1, n)$ is a weighted mixture of two components. Accordingly, criterion (4) is not suitable if the aim is to control one of the two probabilities of correct and decisive evidence rather than the average. This can be the case in clinical trials, where interest centers on one hypothesis, H_i say. In this case, it seems more appropriate to select the optimal sample size n_i^* by controlling directly p_i^{DC}.

Schönbrodt and Wagenmakers [59] also rely on the BF to plan a design to detect with high probability an effect when it exists. In our setting, this corresponds to decisive and correct evidence in favor of the alternative hypothesis when the null represents absence of an effect. Similar to Weiss [71], they demonstrate the usefulness of plotting the distribution of the BF under the null, as well as under the alternative hypothesis. Computations are performed based on simulations in a fixed-n design, although an open-ended sequential design as well as a sequential design with maximal n are considered.

More recently, Pan and Banerjee [49] attempt to provide a simulation-based framework for Bayesian SSD making explicit use of design and analysis priors. Working primarily in the setting of conjugate Bayesian linear regression models, the required computational power for SSD is relatively modest. They also show that several frequentist results can be obtained as special cases of their general Bayesian approach.

3. BAYESIAN SAMPLE SIZE DETERMINATION FOR ACTIVE LEARNING

3.1 Analysis Prior and Bayes Factor Computation

In this section, we first consider a Bayesian model for the observations conditionally on an input CPDAG $\mathcal{E}$. Next, we derive the Bayes Factor (BF) between two specific DAG models belonging to the equivalence class represented by $\mathcal{E}$. The resulting BF is used in the testing procedure (3), which underlies the approach to SSD we describe in Section 3.2.

Under a chain graph $\mathcal{E}$, the joint density of $\boldsymbol{Y}$ factorizes [3] as

$$f(\boldsymbol{y}|\boldsymbol{\theta}_{\mathcal{E}}) = \prod_{\tau \in \mathcal{T}} f_\tau(\boldsymbol{y}_\tau | \boldsymbol{y}_{\mathrm{pa}_{\mathcal{E}}(\tau)}, \boldsymbol{\theta}_\tau), \tag{5}$$

where $\boldsymbol{\theta}_{\mathcal{E}} = \{\boldsymbol{\theta}_\tau, \tau \in \mathcal{T}\}$ is a parameter indexing the graphical model $\mathcal{E}$. A specific feature of $\mathcal{E}$ is that *all* nodes in τ share the same parents $\mathrm{pa}_{\mathcal{E}}(\tau)$ [2], Theorem 4.1(iii). We will further assume that the prior on $\boldsymbol{\theta}_{\mathcal{E}}$ factorizes as

$$p(\boldsymbol{\theta}_{\mathcal{E}}) = \prod_{\tau \in \mathcal{T}} p(\boldsymbol{\theta}_\tau), \tag{6}$$

a condition, which can be named *global* parameter independence following Castelo and Perlman [11].

To recover a DAG structure from $\mathcal{E}$, we need to determine the orientation of all the undirected edges in $\mathcal{E}$. Since each undirected edge $u - v$ belongs to one chain component only, say τ, we can restrict our attention to $\mathcal{E}_\tau$, the undirected decomposable graph of chain component τ, and work separately on each chain component because of factorizations (5) and (6). A further useful feature, highlighted in He and Geng [32], Theorem 4, is the following: if neither cycles nor v-structures are created during the process of edge orientation in a given chain component, then neither cycles nor v-structures are introduced in the whole graph, too. Moreover, because a CPDAG is uniquely characterized by its skeleton and v-structures [2], any DAG obtained by orienting the original CPDAG $\mathcal{E}$ as described above still belongs to the equivalence class of $\mathcal{E}$.

Consider the orientation of edge $u - v$ with $u, v \in \tau$ and write for simplicity $\mathcal{E}_\tau \equiv \mathcal{G}$. According to (3), if Y_u is subjected to an intervention, Y_u is becomes independent of Y_v if $u \leftarrow v$; on the other hand, there is dependence between Y_u and Y_v if $u \rightarrow v$. To determine edge orientation, we first write explicitly the general term $f_\tau(\cdot)$ in (5) using the standard factorization of the joint distribution for decomposable graphical models [40]. For better clarity, we use X for the variables in chain component τ, and $\{X_1, \ldots, X_T\}$ to denote $\{Y_j, j \in \tau\}$, where $T = |\tau|$. Let also $C_1, \ldots, C_K$ be a sequence of cliques of the decomposable graph $\mathcal{G}$. Consider now, for $k = 2, \ldots, K$, the three types of sets

$$H_k = C_1 \cup \cdots \cup C_k,$$
$$S_k = C_k \cap H_{k-1},$$
$$R_k = C_k \setminus H_{k-1},$$

which are called *history*, *separators* and *residuals*, respectively, and set $R_1 = H_1 = C_1, S_1 = \varnothing$. Note that $C_1 \cup R_2 \cup \cdots \cup R_K = V$ and also $R_k \cap R_{k'} = \varnothing$. Additionally, let the sequence be (weakly) perfect, that is: (i) for all $i > 1$ there is a $k < i$ such that $S_i \subseteq C_k$; (ii) the sets S_k are all complete [40], page 14. It is then possible to number the vertices of a decomposable graph starting from those in C_1, then those in R_1, R_2 and so on. In this way, we obtain a *perfect numbering of vertices*, and a *perfect directed version* $\mathcal{G}^<$ of $\mathcal{G}$ by directing its edges from lower to higher numbered vertices. Hence, we can write

$$f(\boldsymbol{x}|\boldsymbol{\theta}_{\mathcal{G}^<}) = \prod_{k=1}^{K} f(\boldsymbol{x}_{R_k}|\boldsymbol{x}_{S_k}, \boldsymbol{\theta}_{R_k}); \tag{7}$$

see Dawid and Lauritzen [21], equation (35). The kth term in (7) can be further written (omitting subscript k to ease notation) as

$$f(\boldsymbol{x}_R|\boldsymbol{x}_S, \boldsymbol{\theta}_R) = \prod_{l=1}^{|R|} f(x_{R,l}|x_{R,1}, \ldots, x_{R,l-1}, \boldsymbol{x}_S, \theta_{R,l}), \tag{8}$$

where $x_{R,l}$ is the lth term of $\boldsymbol{x}_R$. Importantly, the previous decomposition holds for any ordering $(x_{R,1}, \ldots, x_{R,|R|})$ of $\boldsymbol{x}_R$. Also, we can always choose clique $C_1 = R_1$ to be that which contains edge $u - v$ [40], Lemma 2.18. Now consider two perfect directed versions of $\mathcal{G}$:

- $\mathcal{G}_0^< \equiv \mathcal{D}_0$, containing $u \leftarrow v$,
- $\mathcal{G}_1^< \equiv \mathcal{D}_1$, containing $u \rightarrow v$,

such that $\mathcal{D}_0$ and $\mathcal{D}_1$ are identical except for the edges $u \leftarrow v$ and $u \rightarrow v$.

Consider now the assignment of a prior distribution on the parameter indexing the model based on $\mathcal{D}_i$, $i = 0, 1$. If the DAG-model satisfies certain regularity assumptions, we can follow the general procedure of Geiger and Heckerman [27] for eliciting its parameter prior. This is the idea: one starts from a *unique* prior on the parameter of a *complete* DAG-model, wherein all vertices are linked so that no conditional independencies are implied. Next, parameters indexing the same conditional distributions are given identical priors under *any* DAG, which in turn are derived from the unique prior assigned at the beginning under a complete DAG. An advantage of this method is that it represents an effective way to build compatible priors [19, 16] across models. An important consequence of compatibility is that *marginal* data distributions (marginal likelihoods) will involve the distributions of vertices and neighbor variables derived from a single prior, and this dramatically simplifies the elicitation procedure. More details on prior assignments are provided in Appendix 6.

Using (7) and (8) together with independence of the parameters $\{\boldsymbol{\theta}_{R_k}, k = 1, \ldots, K\}$ as well as independence of $\{\theta_{R,l}, l = 1, \ldots, |R|\}$, where R stands for a generic residual R_k, the marginal data distributions under the two DAG models following an intervention on X_u are, respectively,

$$\begin{aligned} &f\big(\boldsymbol{x}|\mathrm{do}(X_u = \widetilde{X}_u), \mathcal{D}_0\big) \\ &\quad = \tilde{f}_u(x_u) m(x_v) \\ &\qquad \times m(\boldsymbol{x}_{R_1 \setminus \{u,v\}}|x_u, x_v) \prod_{k=2}^{K} m(\boldsymbol{x}_{R_k}|\boldsymbol{x}_{S_k}), \end{aligned} \tag{9}$$

$$\begin{aligned} &f\big(\boldsymbol{x}|\mathrm{do}(X_u = \widetilde{X}_u), \mathcal{D}_1\big) \\ &\quad = \tilde{f}_u(x_u) m(x_v|x_u) \\ &\qquad \times m(\boldsymbol{x}_{R_1 \setminus \{u,v\}}|x_u, x_v) \prod_{k=2}^{K} m(\boldsymbol{x}_{R_k}|\boldsymbol{x}_{S_k}), \end{aligned} \tag{10}$$

where $\widetilde{X}_u \sim \tilde{f}_u(\cdot)$. Recall that the conditional distributions $m(\cdot|\cdot)$ as well as the marginal $m(\cdot)$ appearing in the right-hand side of (9) and (10) are derived from the same prior: hence, terms involving the same arguments are identical. Let now

$$\begin{aligned} &H_0 : \text{the interventional distribution is (9)}, \\ &H_1 : \text{the interventional distribution is (10)}. \end{aligned} \tag{11}$$

Based on a sample of size n,

$$\boldsymbol{X}^n = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix},$$

the Bayes factor of H_0 versus H_1 reduces to

$$\begin{aligned} \mathrm{BF}_{01}^n(\boldsymbol{X}_u^n, \boldsymbol{X}_v^n) &= \frac{m(\boldsymbol{X}_v^n)}{m(\boldsymbol{X}_v^n|\boldsymbol{X}_u^n)} \\ &= \frac{m(\boldsymbol{X}_u^n)m(\boldsymbol{X}_v^n)}{m(\boldsymbol{X}_v^n, \boldsymbol{X}_u^n)}, \end{aligned} \tag{12}$$

where $\boldsymbol{X}_u^n$ is the subvector of $\boldsymbol{X}^n$ corresponding to column u, and similarly for $\boldsymbol{X}_v^n$. Equation (12) reveals that testing for edge orientation of $u - v$ is equivalent to testing independence under the joint marginal $m(x_u, x_v)$ between data $\boldsymbol{X}_u^n$ and $\boldsymbol{X}_v^n$ observed after an intervention on X_u, in accordance with (3). Specifically, (post-intervention) independence corresponds to the edge $u \leftarrow v$; conversely, dependence corresponds to $u \rightarrow v$.

It is important to realize that, from an experimental design perspective, the BF in (12) is a function of (interventional) observations $\boldsymbol{X}^n$ yet to be collected; hence, it is a random variable whose *posterior predictive* distribution, given the past observational data $\boldsymbol{Z}$, can be derived from the corresponding distribution of $\boldsymbol{X}^n$. Details for the Gaussian case are reported in Section 4.

3.2 Sample Size Determination

Let $\mathcal{E}$ be a CPDAG with set of chain components $\mathcal{T}$. As in Section 3.1, in the following we restrict our attention to a given chain component $\tau \in \mathcal{T}$ and let $\mathcal{E}_\tau \equiv \mathcal{G}$ be the corresponding (decomposable undirected) subgraph.

3.2.1 *Single-edge orientation*. Consider an undirected edge $u - v$ in $\mathcal{G}$, whose orientation has to be determined. We argued in Section 3.1 that this can be done by testing H_0 *versus* H_1 defined in (11) leading to the BF in (12), which we write as BF_{01}^n for short.

Based on the analysis presented in Section 2.3, we can define the conditional probabilities of Decisive and Correct Evidence (DCE) as

$$\begin{aligned} p_0^{DC}(k_0, n) &= \Pr\{\mathrm{BF}_{01}^n \geq k_0 | H_0\}, \\ p_1^{DC}(k_1, n) &= \Pr\{\mathrm{BF}_{01}^n \leq 1/k_1 | H_1\}. \end{aligned} \tag{13}$$

Finally, the overall probability of DCE for the orientation of $u - v$ is

$$p_{uv}^{DC}(k_0, k_1, n) = \sum_{j \in \{0,1\}} p(H_j) p_j^{DC}(k_j, n), \tag{14}$$

and the optimal sample size to reach DCE at level $\zeta \in (0, 1)$ is

$$n_{uv}^* = \min\{n \in \mathbb{N} : p_{uv}^{DC}(k_0, k_1, n) \geq \zeta\}. \tag{15}$$

3.2.2 *Multiple-edge orientation and sequences of manipulated variables*. Consider now the decomposable UG $\mathcal{G}$ corresponding to one chain component of CPDAG $\mathcal{E}$. As recalled in Section 2.1, interventions on variables in $\mathcal{E}$ can be used to identify the orientation of undirected edges within each chain component, as they *break* the equivalence class represented by $\mathcal{E}$ into a collection of smaller (interventional) equivalence classes. The method described in Section 3.2.1 can be implemented on any sequence of variables, possibly suggested by practical considerations related to the specific experimental context. Alternatively, minimal (sufficient) sequences, presented in Section 2.2 and leading to full DAG identification, represent a natural choice, although they are not unique in general, as the following example shows.

EXAMPLE. Consider a graph $\mathcal{G} : u - v$, representing a chain component. Both $S_1 = \{u\}$ and $S_2 = \{v\}$ are sufficient sets of manipulated variables, because they allow to distinguish between $u \rightarrow v$ and $u \leftarrow v$. Since there are no other sufficient sets of smaller size, both S_1 and S_2 are also minimal according to Definition 2.1.

For a given chain-component graph $\mathcal{G}$, consider a minimal sequence $S = (u_1, \ldots, u_K)$ as in Definition 2.1. Notice that each node $u \in S$ is typically linked to a number of nodes v in the chain component, namely the *neighbors* of u in $\mathcal{G}$, $\mathrm{ne}_\mathcal{G}(u)$. Consider a node $v \in \mathrm{ne}_\mathcal{G}(u)$; from (14), we need to assign $P(H_0)$ and $p(H_1) = 1 - P(H_0)$. Recall that H_0 corresponds to $u \leftarrow v$, while the direction is reversed under H_1. A way to proceed is to consider all DAGs that are Markov equivalent to $\mathcal{G}$ and are obtained by orienting its edges; we denote this set as $[\mathcal{G}]$. Since observational data cannot distinguish among them, it is natural to regard them as equally likely. Accordingly, we set

$$p(H_0) \propto \sum_{\mathcal{D} \in [\mathcal{G}]} \mathbb{1}_{u \leftarrow v}\{\mathcal{D}\}, \tag{16}$$

where $\mathbb{1}_{u \leftarrow v}\{\mathcal{D}\} = 1$ iff $\mathcal{D}$ has $u \leftarrow v$, so that $p(H_0)$ is proportional to the number of DAGs in $[\mathcal{G}]$ containing $u \leftarrow v$.

From (15), we can determine the optimal sample size n_{uv}^* for each $v \in \mathrm{ne}_\mathcal{G}(u)$. Furthermore, the optimal sample size associated with an intervention on u becomes

$$n_u^* = \max\{n_{uv}^*, v \in \mathrm{ne}_\mathcal{G}(u)\}.$$

For a given sequence of manipulated variables, our strategy for SSD is summarized in Algorithm 1.

Recall from Definition 2.1 that a minimal sequence of manipulated variables need not be unique, and define $\{S_1, \ldots, S_L\}$ to be the collection of all such sequences. By applying Algorithm 1 to a given sequence S_l, we obtain the corresponding vector of optimal sample sizes

Algorithm 1: Sequence of optimal sample sizes

Input: A sequence of manipulated variables $S = (u_1, \ldots, u_K)$; a threshold for probability of DCE $\zeta \in (0, 1)$

Output: A collection of optimal sample sizes $\boldsymbol{n}^*$

```
for u ∈ S do
    Construct the set of neighbors of u in G, ne_G(u);
    for v ∈ ne_G(u) do
        Find n*_uv = min{n ∈ ℕ : p^DC_uv(k0, k1, n) ≥ ζ}
    end
    Compute n*_u = max{n*_uv, v ∈ ne_G(u)}
end
Return n* = (n*_{u1}, ..., n*_{uK})
```

$\boldsymbol{n}^{*(l)} = (n_{u_1}^{*(l)}, \ldots, n_{u_{K_l}}^{*(l)})$ for which we can compute the total sample size

$$N^{*(l)} = \sum_{k=1}^{K_l} n_{u_k}^{*(l)}.$$

Hence, the Best Minimal Sequence of manipulated variables (BMS) is naturally defined as the minimal sequence S^* having the smallest total sample size $N^* = \min\{N^{*(l)}, l = 1, \ldots, L\}$.

Finally, notice that in principle the smallest total sample size could be achieved for a sufficient sequence, which is not minimal; however, this would be revealed by our method provided it were applied to *all* sufficient sequences.

4. BAYES FACTOR AND PREDICTIVE DISTRIBUTIONS FOR GAUSSIAN DAGS

As in Section 3.2, consider the subgraph $\mathcal{G}$ corresponding to a chain component τ of the CPDAG $\mathcal{E}$. We assume for observations $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,T})^\top$, $i = 1, \ldots, n$, that

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \boldsymbol{\Omega} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_T(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}), \quad \boldsymbol{\Omega} \in \mathcal{P}_{\mathcal{G}}, \tag{17}$$

where $\mathcal{P}_{\mathcal{G}}$ is the space of symmetric and positive definite matrices Markov w.r.t. the decomposable graph $\mathcal{G}$. We show in Appendix 6 that, based on an objective prior approach, the Bayes factor in (12) takes the value

$$\mathrm{BF}_{01}^n(\boldsymbol{X}_u^n, \boldsymbol{X}_v^n) = g(n)\big[1 - (r_{uv}^n)^2\big]^{\frac{n-1}{2}}, \tag{18}$$

where

$$g(n) = \frac{n}{\sqrt{\pi}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})}$$

and $r_{u,v}^n$ is the sample correlation coefficient between $\boldsymbol{X}_u^n$ and $\boldsymbol{X}_v^n$, that is,

$$(r_{uv}^n)^2 = \frac{[(\boldsymbol{X}_u^n)^\top (\boldsymbol{X}_v^n)]^2}{(\boldsymbol{X}_u^n)^\top (\boldsymbol{X}_u^n) \cdot (\boldsymbol{X}_v^n)^\top (\boldsymbol{X}_v^n)}, \tag{19}$$

which can be written as

$$(r_{uv}^n)^2 = \frac{[\sum_{h=1}^n x_{h,u} x_{h,v}]^2}{\sum_{h=1}^n x_{h,u}^2 \sum_{h=1}^n x_{h,v}^2},$$

where $x_{h,u}$ is the h-component of vector $\boldsymbol{X}_u^n$.

To perform Bayesian SSD, we need to compute the posterior predictive distribution of $\mathrm{BF}_{01}^n(\boldsymbol{X}_u^n, \boldsymbol{X}_v^n)$ under the two model hypothesis H_0, H_1. To ease notation, we simply write BF_{01}^n instead of $\mathrm{BF}_{01}^n(\boldsymbol{X}_u^n, \boldsymbol{X}_v^n)$ for the remainder of this section. From (18), it is enough to obtain the posterior predictive distribution of $(r_{u,v}^n)^2$, and then derive the corresponding distribution for the BF.

4.1 Posterior Predictive Under H_0

Recall that under DAG $\mathcal{D}_0$ and an intervention on variable X_u we have $X_u \perp\!\!\!\perp X_v$, so that the post-intervention model distribution of (X_u, X_v) can be written as

$$\begin{aligned} &p(x_u, x_v | \boldsymbol{\Sigma}_{\{u,v\},\{u,v\}}, \mathrm{do}(X_u = \widetilde{X}_u), \mathcal{D}_0) \\ &\quad = \widetilde{f}_u(x_u) f(x_v | \boldsymbol{\Sigma}_{v,v}), \end{aligned}$$

where $f(x_v|\boldsymbol{\Sigma}_{v,v})$ is $\mathcal{N}(0, \boldsymbol{\Sigma}_{v,v})$. Using Lemma 5.1.1 and Corollary 5.1.2 of Muirhead ([45], p. 147), we obtain

$$\begin{aligned} &(r_{u,v}^n)^2 | \boldsymbol{\Sigma}_{v,v}, \mathrm{do}(X_u = \widetilde{X}_u), \mathcal{D}_0) \\ &\quad \sim \mathrm{Beta}\Big(\frac{1}{2}, \frac{n-1}{2}\Big), \end{aligned} \tag{20}$$

so that $(r_{u,v}^n)^2$ is an ancillary statistic. As a consequence, (20) coincides with the posterior predictive distribution, which we can simply write as $p((r_{u,v}^n)^2 | \mathrm{do}(X_u = \widetilde{X}_u), \mathcal{D}_0)$. Hence, the posterior predictive of BF_{01}^n under H_0 is analytically available and can be easily sampled from because

$$\begin{aligned} &\mathrm{BF}_{01}^n = g(n)\big[1 - (r_{u,v}^n)^2\big]^{\frac{n-1}{2}} \\ &\quad \text{with } \big(1 - (r_{u,v}^n)^2\big) \sim \mathrm{Beta}\Big(\frac{n-1}{2}, \frac{1}{2}\Big). \end{aligned} \tag{21}$$

4.2 Posterior Predictive Under H_1

Under DAG $\mathcal{D}_1$ and an intervention on variable X_u, the post-intervention model distribution of (X_u, X_v) is

$$\begin{aligned} &p(x_u, x_v | \boldsymbol{\Sigma}_{\{u,v\},\{u,v\}}, \mathrm{do}(X_u = \widetilde{X}_u), \mathcal{D}_1) \\ &\quad = \widetilde{f}_u(x_u) f(x_v | x_u, \boldsymbol{L}_{u,v}, \boldsymbol{D}_{v,v}), \end{aligned}$$

where

$$\boldsymbol{L}_{u,v} = -(\boldsymbol{\Sigma}_{u,u})^{-1} \boldsymbol{\Sigma}_{u,v}, \quad \boldsymbol{D}_{v,v} = \boldsymbol{\Sigma}_{v|u} \tag{22}$$

correspond to the regression coefficient associated to x_u, respectively the variance, in the conditional distribution $f(x_v|x_u, \boldsymbol{L}_{u,v}, \boldsymbol{D}_{v,v})$.

Letting $\boldsymbol{Z}$ be the (N, T) matrix of available observational data, we choose as *design* prior for $(\boldsymbol{L}_{u,v}, \boldsymbol{D}_{v,v})$ the posterior $p(\boldsymbol{L}_{u,v}, \boldsymbol{D}_{v,v} | \boldsymbol{Z}, \mathcal{D}_1)$, which can be derived from the posterior of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, as detailed in Appendix 6.

Algorithm 2: Approximate posterior predictive of BF_{01}^n under H_1

Input: Observational (N, T) data matrix $\boldsymbol{Z}$; interventional density $\widetilde{f}_u(\cdot)$; prior hyperparameter a_Ω; required sample size n; number of Monte Carlo draws B

Output: A sample of size B from the posterior predictive distribution of BF_{01}^n under H_1

Compute $\boldsymbol{S} = \boldsymbol{Z}^\top \boldsymbol{Z}$;
for $b = 1, \ldots, B$ **do**
  Draw $\boldsymbol{Q}_{u,v}^{(b)} \sim \mathcal{W}_2(a_\Omega + N - (T-2), \boldsymbol{S}_{\{u,v\},\{u,v\}})$;
  Compute $\boldsymbol{L}_{u,v}^{(b)}, \boldsymbol{D}_{v,v}^{(b)}$;
  for $h = 1, \ldots, n$ **do**
    sample $x_u^{(h)(b)} \sim \widetilde{f}_u(\cdot)$;
    sample $x_v^{(h)(b)} \sim \mathcal{N}(\cdot \mid -\boldsymbol{L}_{u,v}^{(b)} x_u^{(h)(b)}, \boldsymbol{D}_{v,v}^{(b)})$;
  end
  obtain $\boldsymbol{X}_u^{n(b)} = (x_u^{(1)(b)}, \ldots, x_u^{(n)(b)})^\top$
  and $\boldsymbol{X}_v^{n(b)} = (x_v^{(1)(b)}, \ldots, x_v^{(n)(b)})^\top$;
  Compute $(r_{uv}^n)^{2(b)}$ using $\boldsymbol{X}_u^{n(b)}, \boldsymbol{X}_v^{n(b)}$ as in (19) ;
  Compute $\mathrm{BF}_{01}^{n(b)} = g(n)[1 - (r_{uv}^n)^{2(b)}]^{\frac{n-1}{2}}$ as in (18)
end
Return $\{\mathrm{BF}_{01}^{n(1)}, \ldots, \mathrm{BF}_{01}^{n(B)}\}$

Returning to the posterior distribution of the unconstrained matrix $\boldsymbol{\Omega}$, consider the model and objective prior

$$z_1, \ldots, z_N | \boldsymbol{\Omega} \overset{\text{iid}}{\sim} \mathcal{N}_T(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}),$$
$$\boldsymbol{\Omega} \sim p(\boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_\Omega - T - 1}{2}}.$$

We obtain

$$\boldsymbol{\Omega} | \boldsymbol{Z} \sim \mathcal{W}_T(a_\Omega + N, \boldsymbol{S}), \tag{23}$$

where $\boldsymbol{S} = \boldsymbol{Z}^\top \boldsymbol{Z}$, and this acts as the generating design prior for the parameters in (22). Now

$$(\boldsymbol{\Sigma}_{\{u,v\},\{u,v\}})^{-1} = \boldsymbol{\Omega}_{\{u,v\},\{u,v\}|\tau\setminus\{u,v\}} := \boldsymbol{Q}_{u,v}$$

so that

$$\boldsymbol{Q}_{\{u,v\}} \sim \mathcal{W}_2(a_\Omega + N - (T-2), \boldsymbol{S}_{\{u,v\},\{u,v\}}),$$

using distributional properties of the Wishart distribution [54], Theorem 5.1.4. Hence, the posterior predictive of BF_{01}^n under H_1 can be approximated by Monte Carlo simulation following Algorithm 2.

5. ILLUSTRATION AND REAL DATA ANALYSIS

In this section, we first illustrate the proposed method on a simple example with chain components having two nodes and then apply it to a high-dimensional data set on riboflavin production by *Bacillus subtilis*.

5.1 Two-Node Chain Component

Consider a chain component $u - v$. The objective is to determine the optimal sample size for an intervention on u. An observational data set $\boldsymbol{Z}$ is first generated as follows. Assuming the true DAG generating model is $u \to v$, we consider the system of linear equations

$$\begin{cases} X_u = \varepsilon_u, & \varepsilon_u \sim \mathcal{N}(0,1), \\ X_v = 0.5 X_v + \varepsilon_u, & \varepsilon_v \sim \mathcal{N}(0,1), \end{cases}$$

with $\varepsilon_u \perp\!\!\!\perp \varepsilon_v$, and then generate $N = 50$ i.i.d. observations collected in the data matrix $\boldsymbol{Z}$.

We first focus on the predictive distribution of BF_{01}^n, the Bayes Factor defined in (12). Results are summarized in Figure 2, which reports the (approximate) predictive distribution of $\log_{10}\mathrm{BF}_{01}^n$ under each of the two hypotheses, for values of $n \in \{10, 50\}$. To ease legibility, values on the horizontal axis are expressed as BF_{01}^n, and thresholds corresponding to values in $\{1/10, 1/3, 1, 3, 10\}$ are reported as vertical dotted lines. From this output, we can compute the probabilities that the BF favors the true hypothesis for each of the degree evidence categories in Table 1. Specifically, we focus on "moderate evidence" for H_0 and H_1, which corresponds to $3 < \mathrm{BF}_{01}^n < 10$ and $1/10 < \mathrm{BF}_{01}^n < 1/3$, respectively. We also consider "strong-to-extreme evidence" for H_0 and H_1, corresponding to $\mathrm{BF}_{01}^n > 10$ and $\mathrm{BF}_{01}^n < 1/10$, respectively. Results, for different sample sizes $n \in \{10, 50, 100\}$ are summarized in Table 2. For $n = 10$, the two probabilities are both zero under H_0, which is coherent with Figure 2 where the BF distribution does not exceed the threshold 3. By increasing the sample size to $n = 50$ and $n = 100$, the probability of moderate evidence increases up to about 84%, while the probability of strong-to-extreme evidence is only around 2% for $n = 100$. Conversely, when the true hypothesis is H_1, we have strong evidence with a probability higher than 80% even for a moderate sample size, $n = 50$. The latter probability grows up to 96% when the sample size increases to $n = 100$. We thus see an imbalance between the learning rate between H_0 and H_1, a phenomenon which is not new but still worth of consideration; see, for instance, Johnson and Rossell [36].

Consider now the probabilities of DCE as in equations (13) and (14). We compute $p_0^{DC}(k_0, n)$ and $p_1^{DC}(k_1, n)$ for $k_0 = k_1 = k$ by varying $k \in \{3, 6, 10\}$, and for a grid of sample sizes $n \in \{1, 2, \ldots, 1000\}$. The behavior of the two probabilities as a function of n is summarized in the first two plots of Figure 3 where each curve refers to one

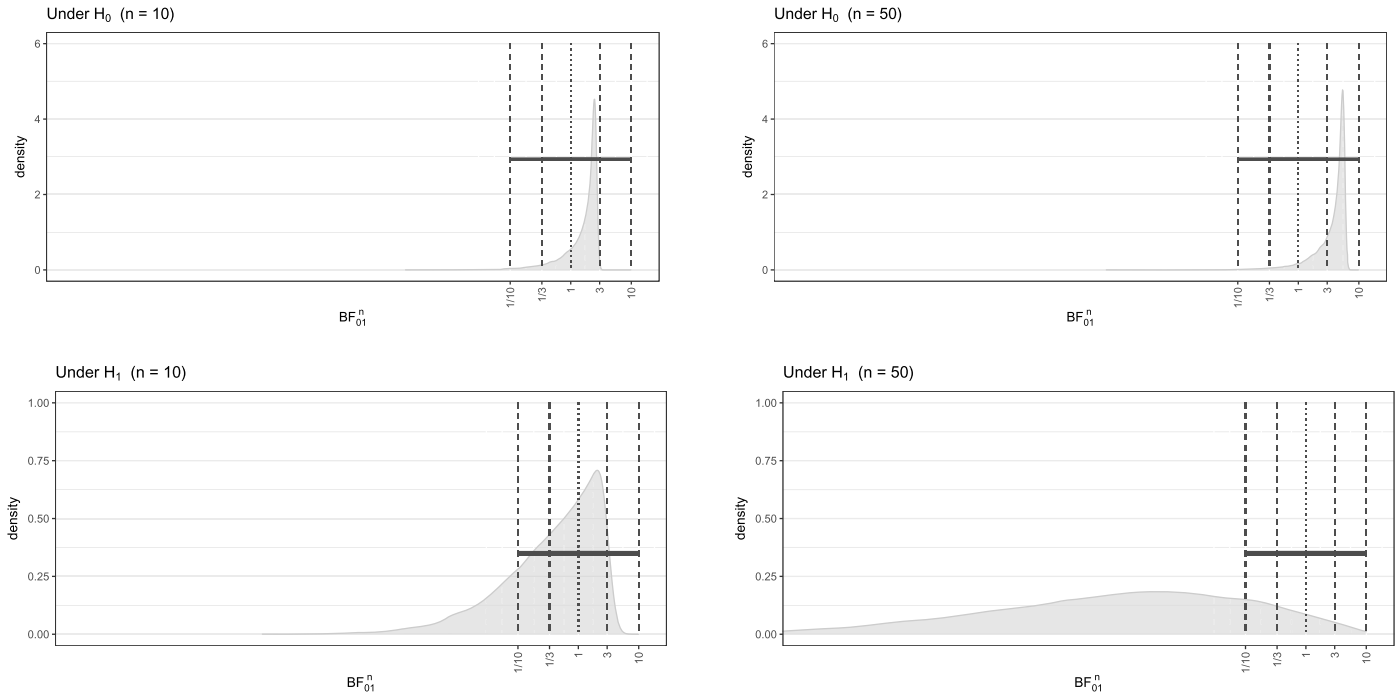

FIG. 2. *Two-node chain component*: *Predictive distribution of* $\log_{10}\mathrm{BF}_{01}^n$ *under* H_0 *and* H_1 *for sample sizes* $n = 10$ *and* $n = 50$. *Vertical dotted lines represent BF thresholds* $1/10, 1/3, 1, 3, 10$.

value of k (from dark to light grey for increasing levels of the threshold). Consider, for instance, $p_0^{DC}(k_0, n)$: this probability exceeds 80% when $k = 3$ for a sample size $n = 100$, consistently with the results of Table 2. When instead $k = 6$, the same sample size only guarantees that $p_0^{DC}(k_0, n)$ is approximately equal to 65%; moreover, to reach a level of 80% the sample size must increase to about $n = 400$. Notice that $p_0^{DC}(k_0 = 10, n)$ is zero for n up to 150; this explains the elbow in the bottom panel of Figure 3.

A similar behavior is observed for $p_1^{DC}(k_1, n)$, where however the distance between the three curves is much smaller, especially for moderate-to-large values of n. This is coherent with the results in Figure 2, which suggests that the area to the left of 1/10 of the BF is already appreciable for small values of n such as 10. In addition, the area to the right of $k = 3$ or $k = 10$ are somewhat similar (and small), which explains the reason why the curves for the probabilities of DCE are close. Finally, the bottom panel of Figure 3 reports the overall probability of DCE in equation (14), which averages $p_0^{DC}(k_0, n)$ and $p_1^{DC}(k_1, n)$ with $P(H_0) = P(H_1) = 0.5$ following equation (16).

TABLE 2

Two-node chain component. Predictive probabilities of BF_{01}^n *resulting in "moderate" or "strong-to-extreme" evidence in favor of the true hypothesis* H_0 *and* H_1, *for sample sizes* $n \in \{10, 50, 100\}$

True	n	Moderate	Strong-to-extreme
	10	0.00%	0.00%
H_0	50	73.64%	0.00%
	100	84.39%	2.10%
	10	18.50%	23.21%
H_1	50	7.70%	82.90%
	100	0.16%	96.21%

We now move to SSD and obtain the optimal sample size n_{uv}^* for an intervention on u based on (15). The latter quantity is computed for each value of the BF threshold $k \in \{3, 6, 10\}$ and for distinct thresholds for the probability of DCE $\zeta \in [0.5, \ldots, 0.95]$. Results are summarized in Figure 4, which reports the behavior of n_{uv}^* as a function of ζ for the three increasing levels of k (from dark to light gray). Clearly, the optimal sample size required for DCE increases with the threshold ζ. The behavior of the three curves as k varies is similar; however, it becomes much steeper beyond $\zeta = 0.85$ for $k = 10$ (the cutoff which separates moderate from strong evidence). As an example, if we fix $\zeta = 0.8$, we obtain an optimal sample size $n_{uv}^* \simeq 50$ for $k = 3$, and this value triples when $k = 6$ and reaches $n_{uv}^* \simeq 300$ for $k = 10$. The latter sample size would instead guarantee a probability of DCE higher than 95% when $k = 3$.

5.2 Riboflavin Data

In this section, we apply our method for SSD to a data set on riboflavin (vitamin B2) production by *Bacillus subtilis*. The data set is publicly available within the `R` package [66] `hdi` and includes $q = 4089$ variables,

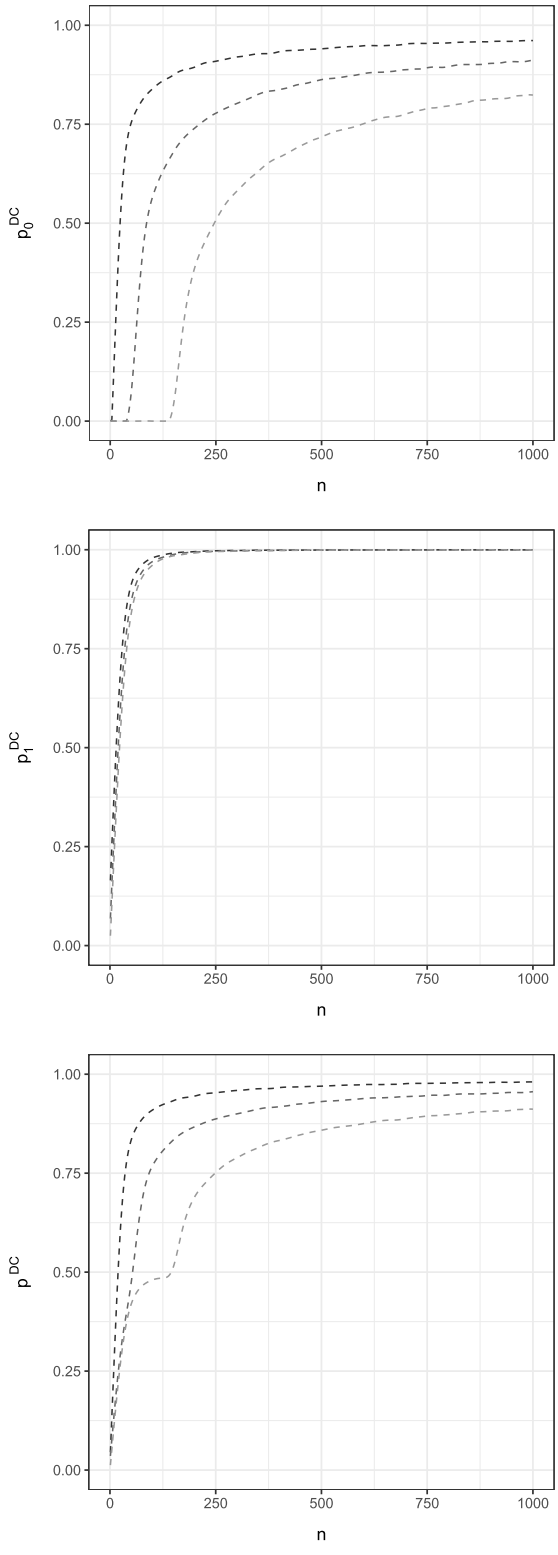

FIG. 3. *Two-node chain component*: *Probability of Decisive and Correct* (*DC*) *evidence in favour of* H_0, H_1 *and overall probability* (*from top to bottom plots*) *as a function of the sample size* n, *for different values of* $k \in \{3, 6, 10\}$ (*from dark to light grey*) *and* $k_0 = k_1 = k$.

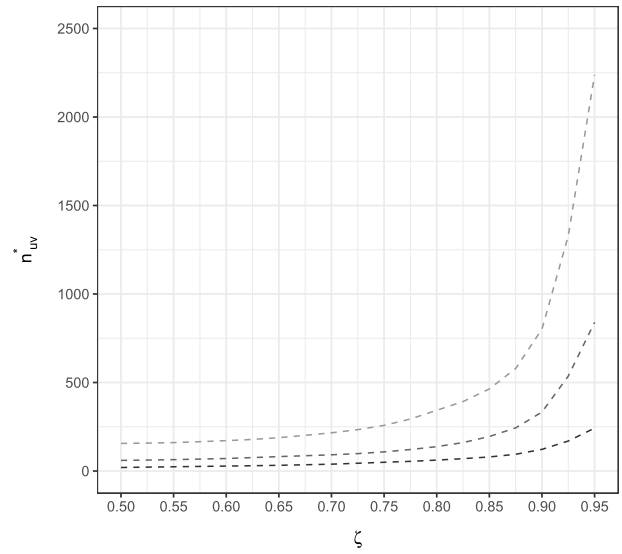

FIG. 4. *Two-node chain component*: *Optimal sample size* n^*_{uv} *as a function of threshold* $\zeta \in [0.50, 0.95]$ *for different values of* $k \in \{3, 6, 10\}$ (*from dark to light grey*) *and* $k_0 = k_1 = k$.

namely the logarithm of riboflavin production rate and the log-expression level of 4088 genes that cover essentially the whole genome of *Bacillus Subtilis*. The sample size is $N = 71$. This observational data set was analyzed by Maathuis, Kalisch and Bühlmann [43] to infer causal effects on riboflavin production rate due to single gene manipulations. To this end, the authors first estimate a CPDAG using the PC algorithm [63]. Then, using do-calculus theory, they provide an estimate of the causal effect on the riboflavin rate following an hypothetical intervention on each of the 4088 nodes. Each causal effect is not unique in general because it depends on the specific set of parents of the node (also called the *adjustment set*), and ultimately on the underlying DAG structure; see also Maathuis, Kalisch and Bühlmann [43], Algorithm 1. Since typically many DAGs are compatible with the input CPDAG, their procedure reports a collection of possible causal effects, and finally the corresponding average.

Maathuis, Kalisch and Bühlmann [43] provide a discussion supporting the hypotheses behind their causal DAG approach, in particular with respect to the Gaussian assumption and the lack of unmeasured confounders. Nevertheless, we should be cautious and refrain from overinterpreting the results of our method because other important assumptions, starting from the very causal interpretation of the DAG, encoded in the definition of the interventional distribution are hard or impossible to check in practice; see also Pearl [51] for a critical appraisal of the nature of causal concepts which cannot be derived from statistical association alone.

Our method determines the optimal sample size for a sequence of interventions on selected variables. These can be practically implemented using CRISPR-Cas9 gene editing technology, which has been extensively adapted

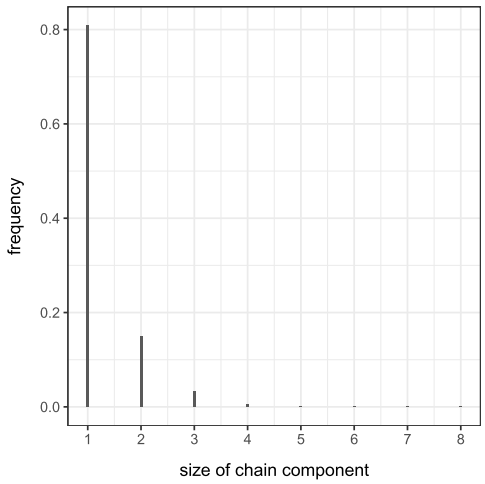

FIG. 5. *Riboflavin data*: *Distribution of the size of chain components in the estimated CPDAG.*

for genome engineering of multiple bacteria. Specifically, CRISPR-Cpf1 was used for gene manipulation in *Bacillus Subtilis*, including the flexible deletion of a single gene and multiple genes or a gene cluster, and gene knock-in; see Hao et al. [28] and Zhang, Duan and Wu [73].

To apply our method for sample size determination (Section 3.2), we start from an input CPDAG $\mathcal{E}$ estimated using the PC algorithm with the tuning parameter α set at level 0.01. We then fix the threshold for the probability of DCE $\zeta = 0.8$, while $k_0 = k_1 = 6$. Recall now that our objective is SSD for an intervention needed to orient those edges, which are undirected in the input CPDAG. Since undirected edges can only occur between (two ore more) nodes belonging to the same chain component, we focus on those chain components whose size is larger than one. Figure 5 summarizes the distribution of the size of chain components in the input CPDAG $\mathcal{E}$. Most of these (about 80%) have size equal to one, in which case there are no edges whose orientation needs to be determined. We then implement our method on each of the remaining chain components separately. As an example, Figure 6 reports two chain component subgraphs $\mathcal{G}_1, \mathcal{G}_2$. with the corresponding Best Minimal Sequences of manipulated variables (BMS) represented as grey nodes. For $\mathcal{G}_1$, there are actually two minimal sequences according to Definition 2.1, namely $S_1 = (2, 3)$ and $S_2 = (3, 4)$, as we report

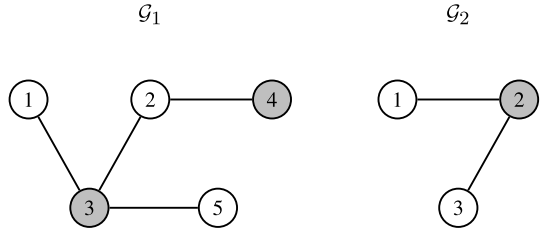

FIG. 6. *Riboflavin data*: *Two chain-component subgraphs,* $\mathcal{G}_1, \mathcal{G}_2$, *with grey dots representing BMS of manipulated variables.*

TABLE 3
Riboflavin data: *Minimal sequences of manipulated nodes and corresponding sample sizes for each of the two chain-component subgraphs,* $\mathcal{G}_1, \mathcal{G}_2$ *in Figure 6*

	Minimal sequence of nodes	Optimal sample size
$\mathcal{G}_1$	$S_1 = (2, 3)$	$\boldsymbol{n}^{*(1)} = (28, 88)$
	$S_2 = (3, 4)$	$\boldsymbol{n}^{*(2)} = (4, 86)$
$\mathcal{G}_2$	$S_1 = 2$	$\boldsymbol{n}^{*(1)} = 130$

in Table 3 with the corresponding optimal sample sizes computed as in Algorithm 1. Clearly, S_2 is the BMS because the *total* sample size under S_2 is smaller than under S_1. On the other hand, there is only one minimal sequence for $\mathcal{G}_2$, whose corresponding optimal sample size is 130 (see again Table 3).

As an overall summary, we also report in Table 4, for each size of the chain components of $\mathcal{E}$ and size of the BMS, the corresponding average total sample size (Average N^*) and average sample size *per* intervention (Average n^*), with the average computed across sequences of manipulated variables. It appears that while in general the number of interventions needed for edge orientation (Size of sequence S^*) increases with the size of the chain component, the total sample size is not typically higher for larger (chain components and) sizes of S^*; compare, for instance, $s = 4$ and $s = 5$. In addition, the average sample size *per* intervention, for each given value of s for which different sizes of S^* are observed, is smaller for those sequences of manipulated variables having larger sizes. Accordingly, while more interventions are needed to orient edges in chain components of larger dimension, the (optimal) number of interventional data (sample size) required by each intervention is in general smaller. This suggests the existence of a trade-off between the number of manipulated variables and the optimal sample size *per* intervention.

6. DISCUSSION

Observational data cannot distinguish in general among different DAG-models representing the same conditional independence assertions. This is a serious drawback for causal inference, which is predicated on a given DAG representing the data generating process, as required by do-calculus theory. Active learning methods perform manipulations on selected variables, leading to the collection of interventional data. This greatly improves structure learning and sharpen causal inference. In this context, a question, which has so far been neglected, is the determination of the sample size (SSD) of the interventional data required to achieve desirable inferential properties, which is the subject of this paper.

TABLE 4

Riboflavin data: *average total sample size* (*Average* N^*) *and average sample size per manipulated variable* (*Average* n^*) *cross-classified by size of chain component of the input CPDAG and size of the best minimal sequence of manipulated variables* S^*

Size of chain component	$s=2$	$s=3$		$s=4$		$s=5$	$s=6$		$s=7$	$s=8$
Size of sequence S^*	1	1	2	1	2	2	2	3	3	4
Average N^*	48.3	131.4	100.0	241.3	248.4	81.0	67.0	37.0	65.5	117.0
Average n^*	48.3	131.4	50.0	241.3	124.2	40.5	33.5	12.3	21.8	29.3

Our theoretical framework for SSD is based on causal sufficiency (no latent confounders) and absence of selection bias or feedbacks. A paper that tries to alleviate the assumption of no latent variables when the goal is structure learning is Frot, Nandy and Maathuis [26]. We reiterate that these assumptions cannot be typically checked when only observational data are available; however, interventions may be used to assess the causal DAG hypothesis, which underlies the notion of post-intervention distribution.

We follow the default approach of active learning methods, and start from an equivalence class of DAGs, equivalently its representative CPDAG, which typically has been estimated from an observational data set. As a consequence, our method does not accommodate for uncertainty of the initial graph structure. In principle, uncertainty on CPDAG can be accounted through the posterior distribution over the space of Markov equivalence classes; see, for example, Castelletti et al. [8]. This could be of some help to evaluate the strength of the evidence in favor of the chosen CPDAG, for instance, the highest posterior probability model or any other CPDAG estimate, such as the Median Probability Model (MPM); see again Castelletti et al. [8]. One could also use the posterior distribution to select a few most probable CPDAGs. In general, however, each of them may lead to a distinct sequence of manipulated variables and optimal sample sizes. Operationally, one would still need to implement interventions on a *single* sequence of variables, and a method should be devised to combine such distinct outcomes. Despite these limitations, our approach is still useful at least from an exploratory analysis perspective.

Our method for SSD is based on a sequence of manipulated variables arising from a *batch* intervention experiment. A *sequential* approach would instead proceed by choosing the intervention nodes one at a time and collecting new data after each intervention [65, 32]. In this way, the optimal sample size associated with a target node could be computed, at each stage, using all the samples collected up to that step, thus increasing the amount of information used for prediction at the design level. Another advantage of the sequential method is to alleviate the danger inherent in the choice of the starting Markov equivalence class, namely that the true DAG could be outside the class. Hauser and Bühlmann [30] investigate this aspect and illustrate by simulation that methods that do take into account observational as well as interventional data show a better performance in recovering the true data-generating DAG. We observe that our method could be tailored to a sequential setup by incorporating in the predictive distribution of the BF not only the initial observational data but also the newly interventional observations collected at each step. For inferential methods on DAG models dealing jointly with observational and interventional data, see Hauser and Bühlmann [31]; a Bayesian perspective is provided in Castelletti and Consonni [5] and Castelletti and Peluso [10].

In our procedure, experimental data are generated under hard interventions and in the absence of latent variables. Under hard (or perfect) interventions, dependencies between targeted variables and their direct causes are removed. This assumption may not hold in some settings where dependencies can only be altered without being fully deleted. An instance is genomic medicine, where gene manipulation through repression or activation of selected genes is performed to better understand the complex functioning of the pathway. Intervention experiments for gene regulation are meant to be perfect but in practice may not be uniformly successful across a cell population, in which case the dependence between manipulated genes and their direct causes in the network is only weakened but maintained. Identifiability of causal DAGs from soft interventions is investigated from a theoretical perspective by Yang, Katcoff and Uhler [72] who propose a consistent algorithm for DAG structure learning.

APPENDIX: PRIORS FOR DAG MODEL COMPARISON AND BAYES FACTOR COMPUTATION

Our elicitation scheme for parameter priors under a general Gaussian DAG model is based on the procedure introduced by Geiger and Heckerman [27] (G&H). This is used both to construct the analysis prior and the design prior. The former is needed to obtain the Bayes factor, whose predictive distribution is generated under the latter.

A.1 General Assumptions

The method of G&H is based on a set of assumptions, which drastically simplifies the elicitation of priors; additionally, it ensures *compatibility* of priors across DAG models, in the sense that DAGs belonging to the same equivalence class score the same marginal likelihood. This feature is important when DAG model comparison is based on observational data, because the latter cannot distinguish in general among Markov equivalent DAGs. This however is no longer the case when interventional data are also employed, as we do in Section 3.1.

The method assumes some regularity conditions on the likelihood, namely *complete model equivalence, regularity and likelihood modularity*, which are satisfied by any Gaussian model. In addition, two assumptions on the prior distributions are introduced. The first assumption (*prior modularity*) states that, for any two distinct DAG models with the *same* set of parents for vertex j, the prior for the node-parameter $\boldsymbol{\theta}_j$ must be the same under both models. Moreover, the second assumption (*global parameter independence*) states that for every DAG model $\mathcal{D}$, the parameters $\{\boldsymbol{\theta}_j;\ j = 1, \ldots, q\}$ should be a priori independent, that is,

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

Based on these assumptions, Theorem 1 of Geiger and Heckerman [27] shows that the parameter priors of *all* DAG models are completely determined by a *unique* prior on the parameter of *any* of the (equivalent) complete DAGs.

Specifically, in the zero-mean Gaussian framework, all priors across DAG models can be shown to be driven by a *single* Wishart distribution on an *unconstrained* precision matrix. Most importantly, a direct consequence of the method is that each marginal data distribution in equation (12) corresponds to the marginal data distribution computed under any complete DAG model; see the next section for more details.

A.2 Marginal Data Distributions and Bayes Factor

Consider a multivariate Gaussian model of the form

$$\begin{aligned} \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\Omega} &\overset{\text{iid}}{\sim} \mathcal{N}_T(\mathbf{0}, \boldsymbol{\Omega}^{-1}), \\ \boldsymbol{\Omega} &\sim \mathcal{W}_T(a, \boldsymbol{U}), \end{aligned} \tag{24}$$

where $\mathcal{W}_T(a, \boldsymbol{U})$ denotes a Wishart distribution having expectation $a\boldsymbol{U}^{-1}$ and $a > T - 1$. Let also $\boldsymbol{S} = \sum_{h=1}^{n} \boldsymbol{x}_h \boldsymbol{x}_h^\top$. The marginal data distribution restricted to variables in $A \subseteq \{1, \ldots, T\}$ is given by

$$m(\boldsymbol{X}_A) = \frac{\prod_{j=1}^{|A|} \Gamma(\frac{a-|\bar{A}|+n+1-j}{2})}{\prod_{j=1}^{|A|} \Gamma(\frac{a-|\bar{A}|+1-j}{2})} \times \frac{|\boldsymbol{U}_{A,A}|^{\frac{a-|\bar{A}|}{2}}}{|\boldsymbol{U}_{A,A} + \boldsymbol{S}_{A,A}|^{\frac{a-|\bar{A}|+n}{2}}}, \tag{25}$$

where $\boldsymbol{U}_{A,A}$ denotes the submatrix of $\boldsymbol{U}$ with rows and columns indexed by A and $\bar{A} = \{1, \ldots, T\} \setminus A$; see, for instance, Consonni and La Rocca [15], equation (12). Moreover, for simplicity in this section, we omit superscript n from data matrices. Under the Gaussian setting of Section 4, the BF in equation (12) can be evaluated using the marginal likelihood (25) for $A = u$, $A = v$ and $A = \{u, v\}$. We thus obtain

$$m(\boldsymbol{X}_u) = \frac{\Gamma(\frac{a-(T-1)+n}{2})}{\Gamma(\frac{a-(T-1)}{2})} \frac{|\boldsymbol{U}_{u,u}|^{\frac{a-(T-1)}{2}}}{|\boldsymbol{U}_{u,u} + \boldsymbol{S}_{u,u}|^{\frac{a-(T-1)+n}{2}}} \tag{26}$$

and similarly for $A = v$, while for $A = \{u, v\}$,

$$\begin{aligned} m(\boldsymbol{X}_{u,v}) = {} & \frac{\Gamma(\frac{a-(T-2)+n}{2})}{\Gamma(\frac{a-(T-2)}{2})} \frac{\Gamma(\frac{a-(T-2)+n-1}{2})}{\Gamma(\frac{a-(T-2)-1}{2})} \\ & \times \frac{|\boldsymbol{U}_{\{u,v\},\{u,v\}}|^{\frac{a-(T-1)}{2}}}{|\boldsymbol{U}_{\{u,v\},\{u,v\}} + \boldsymbol{S}_{\{u,v\},\{u,v\}}|^{\frac{a-(T-1)+n}{2}}}. \end{aligned} \tag{27}$$

Therefore, the BF in (12) reduces to

$$\begin{aligned} \mathrm{BF}_{01}^{n} = {} & \frac{\Gamma(\frac{a-(T-1)+n}{2})}{\Gamma(\frac{a-(T-1)}{2})} \cdot \frac{\Gamma(\frac{a-(T-2)}{2})}{\Gamma(\frac{a-(T-2)+n}{2})} \\ & \times \frac{[\boldsymbol{U}_{u,u}\boldsymbol{U}_{v,v}]^{\frac{a-(T-1)}{2}}}{|\boldsymbol{U}_{\{u,v\},\{u,v\}}|^{\frac{a-(T-2)}{2}}} \\ & \times \frac{|\boldsymbol{U}_{\{u,v\},\{u,v\}} + \boldsymbol{S}_{\{u,v\},\{u,v\}}|^{\frac{a-(T-2)+n}{2}}}{[(\boldsymbol{U}_{u,u} + \boldsymbol{S}_{u,u})(\boldsymbol{U}_{v,v} + \boldsymbol{S}_{v,v})]^{\frac{a-(T-1)+n}{2}}}. \end{aligned} \tag{28}$$

So far, results were obtained under a subjective prior on $\boldsymbol{\Omega}$. We now consider an *objective* framework based on the notion of *Fractional Bayes Factor* (FBF) [47]. Specifically, we start from the default objective prior

$$p^D(\boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_\Omega - T - 1}{2}}. \tag{29}$$

The (data dependent) fractional prior on $\boldsymbol{\Omega}$ is defined as

$$p^F(\boldsymbol{\Omega}) \propto \{p(\boldsymbol{X}|\boldsymbol{\Omega})\}^b p^D(\boldsymbol{\Omega}),$$

where $b \in (0, 1)$ is typically chosen as the smallest value s.t. the fractional prior is proper. After some calculations, we obtain

$$\boldsymbol{\Omega} \sim \mathcal{W}_T(a_\Omega + n_0, n_0\bar{\boldsymbol{S}}),$$

where $\bar{\boldsymbol{S}} = \frac{1}{n}\boldsymbol{S}$ and $n_0 = bn$, a distribution which is a proper provided $a_\Omega + n_0 > T - 1$; see Consonni and La Rocca [15] for full details. Also, the posterior distribution of $\boldsymbol{\Omega}$ is

$$\begin{aligned} p(\boldsymbol{\Omega}|\boldsymbol{X}) &\propto \{p(\boldsymbol{X}|\boldsymbol{\Omega})\}^{1-b} p^F(\boldsymbol{\Omega}) \\ &= p(\boldsymbol{X}|\boldsymbol{\Omega}) p^D(\boldsymbol{\Omega}). \end{aligned}$$

The FBF is obtained by specializing (28) with

$$a \mapsto a_\Omega + n_0, \quad n \mapsto n - n_0,$$
$$\boldsymbol{U} \mapsto \frac{n_0}{n}\boldsymbol{S}, \quad \boldsymbol{S} \mapsto \frac{n-n_0}{n}\boldsymbol{S},$$

which after some calculations leads to

$$\mathrm{BF}_{01}^n = \frac{\Gamma(\frac{a_\Omega-(T-1)+n}{2})}{\Gamma(\frac{a_\Omega+n_0-(T-1)}{2})} \cdot \frac{\Gamma(\frac{a_\Omega+n_0-(T-2)}{2})}{\Gamma(\frac{a_\Omega-(T-2)+n}{2})} \times \left(\frac{n_0}{n}\right)^{-1} \left[\frac{|\boldsymbol{S}_{\{u,v\},\{u,v\}}|}{\boldsymbol{S}_{u,u}\boldsymbol{S}_{v,v}}\right]^{\frac{n-n_0}{2}}.$$

Now notice that

$$|\boldsymbol{S}_{\{u,v\},\{u,v\}}| = \sum_{h=1}^n x_{h,u}^2 \sum_{h=1}^n x_{h,v}^2 - \left(\sum_{h=1}^n x_{h,u}x_{h,v}\right)^2$$

and

$$\boldsymbol{S}_{u,u} = \sum_{h=1}^n x_{h,u}^2, \quad \boldsymbol{S}_{v,v} = \sum_{h=1}^n x_{h,v}^2.$$

Therefore, we can write

$$\frac{|\boldsymbol{S}_{\{u,v\},\{u,v\}}|}{\boldsymbol{S}_{u,u}\boldsymbol{S}_{v,v}} = 1 - \frac{(\sum_{h=1}^n x_{h,u}x_{h,v})^2}{\sum_{h=1}^n x_{h,u}^2 \sum_{h=1}^n x_{h,v}^2} = 1 - \big(r_{u,v}^n\big)^2 \tag{30}$$

where $r_{u,v}^n$ denotes the sample correlation coefficient between X_u and X_v. In the sequel, we choose $a_\Omega = T - 1$ so that the prior is proper even with a training sample size n_0 equal to one, and we obtain

$$\mathrm{BF}_{01}^n = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})} n\big[1 - \big(r_{uv}^n\big)^2\big]^{\frac{n-1}{2}}. \tag{31}$$

A.3 Posterior Distribution of DAG Model Parameters

The design prior for $(\boldsymbol{L}_{u,v}, \boldsymbol{D}_{v,v})$ that we adopt in Section 4.2 corresponds to the posterior $p(\boldsymbol{L}_{u,v}, \boldsymbol{D}_{v,v}|\boldsymbol{Z}, \mathcal{D}_1)$. The latter can be recovered from the posterior on $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, the (unconstrained) precision matrix of a complete DAG, following the procedure of G&H, which we detail below.

Let $\mathcal{D}$ be an arbitrary DAG and let $\prec j \succ = \mathrm{pa}(j)$, and $\prec j] = \mathrm{pa}(j) \times j$. Consider the (generalized Cholesky) reparameterization $\boldsymbol{\Omega} \mapsto (\boldsymbol{L}, \boldsymbol{D})$ where, for $j \in \{1, \ldots, q\}$,

$$\boldsymbol{D}_{jj} = \boldsymbol{\Sigma}_{jj|\mathrm{pa}_\mathcal{D}(j)}, \boldsymbol{L}_{\prec j]} = -\boldsymbol{\Sigma}_{\prec j\succ}^{-1}\boldsymbol{\Sigma}_{\prec j]}$$

For each node $j \in \{1, \ldots, q\}$, let $\{\boldsymbol{D}_{jj}, \boldsymbol{L}_{\prec j]}\}$ be the parameters associated to node j, and identify a complete DAG $\mathcal{D}^{C(j)}$ such that $\mathrm{pa}_{\mathcal{D}^{C(j)}}(j) = \mathrm{pa}_\mathcal{D}(j)$. Let $\{\boldsymbol{D}_{jj}^{C(j)}, \boldsymbol{L}_{\prec j]}^{C(j)}\}$ be the parameters of node j under the complete DAG $\mathcal{D}^{C(j)}$. We then assign to $\{\boldsymbol{D}_{jj}, \boldsymbol{L}_{\prec j]}\}$ the same prior of $\{\boldsymbol{D}_{jj}^{C(j)}, \boldsymbol{L}_{\prec j]}^{C(j)}\}$. However, because our interest is in obtaining the posterior of DAG parameters $(\boldsymbol{D}, \boldsymbol{L})$, we can compute first the posterior on the unconstrained $\boldsymbol{\Omega}$, which by conjugacy is still Wishart, and then recover the posterior on $(\boldsymbol{D}, \boldsymbol{L})$.

Consider a random sample of size N, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$, with $\boldsymbol{z}_i = (x_{i,1}, \ldots, x_{i,q})^\top$ and $\boldsymbol{z}_i|\boldsymbol{\Omega} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$, $i = 1, \ldots, N$, where $\boldsymbol{\Omega}$ is unconstrained. Let also $\boldsymbol{Z}$ be the (N, q) data matrix, obtained by row-binding the individual $\boldsymbol{z}_i^\top$'s. The posterior distribution of $\boldsymbol{\Omega}$ computed under the default prior (29) is given by

$$\boldsymbol{\Omega}|\boldsymbol{Z} \sim \mathcal{W}_T(a_\Omega + n, \boldsymbol{Z}^\top\boldsymbol{Z}).$$

To obtain draws from the posterior of $(\boldsymbol{L}_{u,v}, \boldsymbol{D}_{v,v})$, set $(\boldsymbol{\Sigma}_{\{u,v\},\{u,v\}})^{-1} := \boldsymbol{Q}_{u,v}$, and using properties of the Wishart distribution deduce

$$\boldsymbol{Q}_{\{u,v\}} \sim \mathcal{W}_2\big(a_\Omega + N - (T-2), \boldsymbol{S}_{\{u,v\},\{u,v\}}\big); \tag{32}$$

see Press [54], Theorem 5.1.4, and Consonni and La Rocca [15], Section 2.1. Consider now a draw from (32) and compute $\boldsymbol{\Sigma}_{\{u,v\},\{u,v\}}$. Finally, recover

$$\boldsymbol{L}_{u,v} = -(\boldsymbol{\Sigma}_{u,u})^{-1}\boldsymbol{\Sigma}_{u,v}, \quad \boldsymbol{D}_{v,v} = \boldsymbol{\Sigma}_{v|u},$$

where $\boldsymbol{\Sigma}_{v|u} = \boldsymbol{\Sigma}_{v,v} - (\boldsymbol{\Sigma}_{v,u})^2\boldsymbol{\Sigma}_{u,u}^{-1}$.

ACKNOWLEDGMENTS

The authors would like to thank the Editor, an Associate Editor and two anonymous reviewers for their useful comments which helped improve the clarity of the paper.

FUNDING

Work partially supported by UCSC (D1 and 2019-D.3.2 research grants).

SOFTWARE

Algorithms for sample size determination have been implemented in the R language [66] and are available at https://github.com/FedeCastelletti/ssd_causal_discovery.

FUNDING

Work partially supported by MUR-PRIN grant 2022 SMNNKY CUP J53D23003870008 and UCSC (D1 and 2019-D.3.2 research grants).

REFERENCES

[1] ADCOCK, C. J. (1997). Sample size determination: A review. *J. R. Stat. Soc., Ser. D, Stat.* **46** 261–283.

[2] ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312 https://doi.org/10.1214/aos/1031833662

[3] ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Stat.* **28** 33–85. MR1844349 https://doi.org/10.1111/1467-9469.00224

[4] BANDYOPADHYAY, P. S. and FORSTER, M. R., eds. (2011). Posterior model probabilities. In *Philosophy of Statistics*. *Handbook of the Philosophy of Science* **7**. Elsevier/North-Holland, Amsterdam. MR3295937 https://doi.org/10.1016/B978-0-444-51862-0.50001-0

[5] CASTELLETTI, F. and CONSONNI, G. (2019). Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways. *Ann. Appl. Stat.* **13** 2289–2311. MR4037431 https://doi.org/10.1214/19-aoas1275

[6] CASTELLETTI, F. and CONSONNI, G. (2020). Discovering causal structures in Bayesian Gaussian directed acyclic graph models. *J. Roy. Statist. Soc. Ser. A* **183** 1727–1745. MR4157833

[7] CASTELLETTI, F. and CONSONNI, G. (2021). Bayesian inference of causal effects from observational data in Gaussian graphical models. *Biometrics* **77** 136–149. MR4229727 https://doi.org/10.1111/biom.13281

[8] CASTELLETTI, F., CONSONNI, G., DELLA VEDOVA, M. L. and PELUSO, S. (2018). Learning Markov equivalence classes of directed acyclic graphs: An objective Bayes approach. *Bayesian Anal.* **13** 1231–1256. MR3855370 https://doi.org/10.1214/18-BA1101

[9] CASTELLETTI, F. and PELUSO, S. (2021). Equivalence class selection of categorical graphical models. *Comput. Statist. Data Anal.* **164** Paper No. 107304. MR4280200 https://doi.org/10.1016/j.csda.2021.107304

[10] CASTELLETTI, F. and PELUSO, S. (2023). Bayesian learning of network structures from interventional experimental data. *Biometrika* asad032.

[11] CASTELO, R. and PERLMAN, M. D. (2004). Learning essential graph Markov models from data. In *Advances in Bayesian Networks*. *Stud. Fuzziness Soft Comput.* **146** 255–269. Springer, Berlin. MR2090887 https://doi.org/10.1007/978-3-540-39879-0_14

[12] CHALONER, K. and VERDINELLI, I. (1995). Bayesian experimental design: A review. *Statist. Sci.* **10** 273–304. MR1390519

[13] CHICKERING, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Uncertainty in Artificial Intelligence* (*Montreal*, *PQ*, 1995) 87–98. Morgan Kaufmann, San Francisco, CA. MR1615012

[14] CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2** 445–498. MR1929415 https://doi.org/10.1162/153244302760200696

[15] CONSONNI, G. and LA ROCCA, L. (2012). Objective Bayes factors for Gaussian directed acyclic graphical models. *Scand. J. Stat.* **39** 743–756. MR3000846 https://doi.org/10.1111/j.1467-9469.2011.00785.x

[16] CONSONNI, G. and VERONESE, P. (2008). Compatibility of prior specifications across linear models. *Statist. Sci.* **23** 332–353. MR2483907 https://doi.org/10.1214/08-STS258

[17] COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems*. *Statistics for Engineering and Information Science*. Springer, New York. MR1697175

[18] DASGUPTA, A. (1996). Review of optimal Bayes designs. In *Design and Analysis of Experiments*. *Handbook of Statist.* **13** 1099–1147. North-Holland, Amsterdam. MR1492591 https://doi.org/10.1016/S0169-7161(96)13031-5

[19] DAWID, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In *Bayesian Statistics*, 4 (*Peñíscola*, 1991) 109–125. Oxford Univ. Press, New York. MR1380273

[20] DAWID, A. P. (2010). Beware of the DAG!. In *Proceedings of Workshop on Causality*: *Objectives and Assessment at NIPS* 2008 (I. Guyon, D. Janzing and B. Schölkopf, eds.). *Proceedings of Machine Learning Research* **6** 59–86. PMLR, Whistler, Canada.

[21] DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267 https://doi.org/10.1214/aos/1176349260

[22] DE SANTIS, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *J. Statist. Plann. Inference* **124** 121–144. MR2066230 https://doi.org/10.1016/S0378-3758(03)00198-8

[23] EBERHARDT, F. (2008). Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. *UAI'*08 161–168. AUAI Press, Arlington, VA, USA.

[24] ETZIONI, R. and KADANE, J. B. (1993). Optimal experimental design for another's analysis. *J. Amer. Statist. Assoc.* **88** 1404–1411. MR1245377

[25] FRIEDMAN, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **303** 799–805. https://doi.org/10.1126/science.1094068

[26] FROT, B., NANDY, P. and MAATHUIS, M. H. (2019). Robust causal structure learning with some hidden variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 459–487. MR3961495 https://doi.org/10.1111/rssb.12315

[27] GEIGER, D. and HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30** 1412–1440. MR1936324 https://doi.org/10.1214/aos/1035844981

[28] HAO, W., SUO, F., LIN, Q., CHEN, Q., ZHOU, L., LIU, Z., CUI, W. and ZHOU, Z. (2020). Design and construction of portable CRISPR-Cpf1-mediated genome editing in bacillus subtilis 168 oriented toward multiple utilities. *Front. Bioeng. Biotechnol.* **8**.

[29] HAUSER, A. and BÜHLMANN, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13** 2409–2464. MR2973606

[30] HAUSER, A. and BÜHLMANN, P. (2014). Two optimal strategies for active learning of causal models from interventional data. *Internat. J. Approx. Reason.* **55** 926–939. MR3178409 https://doi.org/10.1016/j.ijar.2013.11.007

[31] HAUSER, A. and BÜHLMANN, P. (2015). Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 291–318. MR3299409 https://doi.org/10.1111/rssb.12071

[32] HE, Y.-B. and GENG, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.* **9** 2523–2547. MR2460892

[33] HYTTINEN, A., EBERHARDT, F. and HOYER, P. O. (2013). Experiment selection for causal discovery. *J. Mach. Learn. Res.* **14** 3041–3071. MR3138909

[34] IMBENS, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *J. Econ. Lit.* **58** 1129–1179.

[35] JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. MR0187257

[36] JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 143–170. MR2830762 https://doi.org/10.1111/j.1467-9868.2009.00730.x

[37] KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

[38] KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572

[39] KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models*: *Principles and Techniques*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2778120

[40] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford University Press, New York. MR1419991

[41] LINDLEY, D. V. (1971). *Bayesian Statistics*, *a Review*. *Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics*, *No*. 2. SIAM, Philadelphia, PA. MR0329081

[42] LINDLEY, D. V. (1997). The choice of sample size. *J. R. Stat. Soc*., *Ser. D*, *Stat*. **46** 129–138.

[43] MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist*. **37** 3133–3164. MR2549555 https://doi.org/10.1214/09-AOS685

[44] MEGANCK, S., LERAY, P. and MANDERICK, B. (2006). Learning causal Bayesian networks from observations and experiments: A decision theoretic approach. In *Modeling Decisions for Artificial Intelligence* (V. Torra, Y. Narukawa, A. Valls and J. Domingo-Ferrer, eds.) 58–69. Springer Berlin Heidelberg, Berlin, Heidelberg.

[45] MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0652932

[46] NAGARAJAN, R., SCUTARI, M. and LÈBRE, S. (2013). *Bayesian Networks in R with Applications in Systems Biology*. *Use R!* Springer, New York. MR3059206 https://doi.org/10.1007/978-1-4614-6446-4

[47] O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57** 99–138. MR1325379

[48] O'HAGAN, A. and STEVENS, J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med. Decis. Mak*. **21** 219–230. https://doi.org/10.1177/0272989X0102100307

[49] PAN, J. and BANERJEE, S. (2021). A unifying Bayesian approach for sample size determination using design and analysis priors. ArXiv preprint. Available at arXiv:2112.03509.

[50] PEARL, J. (2000). *Causality*: *Models*, *Reasoning*, *and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

[51] PEARL, J. (2003). Statistics and causal inference: A review. *TEST* **12** 281–318. MR2044313 https://doi.org/10.1007/BF02595718

[52] PENG, S., SHEN, X. and PAN, W. (2020). Reconstruction of a directed acyclic graph with intervention. *Electron. J. Stat*. **14** 4133–4164. MR4175391 https://doi.org/10.1214/20-EJS1767

[53] PETERS, J. and BÜHLMANN, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101** 219–228. MR3180667 https://doi.org/10.1093/biomet/ast043

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

[55] RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. *Harvard Business School Publications*. Division of Research, Graduate School of Business Adminitration, Harvard Univ. MR0117844

[56] ROYALL, R. (2000). On the probability of observing misleading statistical evidence. *J. Amer. Statist. Assoc*. **95** 760–780. MR1803877 https://doi.org/10.2307/2669456

[57] ROYALL, R. M. (1997). *Statistical Evidence*: *A Likelihood Paradigm*. *Monographs on Statistics and Applied Probability* **71**. CRC Press, London. MR1629481

[58] SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. A. and NOLAN, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** 523–529. https://doi.org/10.1126/science.1105809

[59] SCHÖNBRODT, F. D. and WAGENMAKERS, E. J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychon. Bull. Rev.* **25** 128–142.

[60] SHOJAIE, A. and MICHAILIDIS, G. (2009). Analysis of gene sets based on the underlying regulatory network. *J. Comput. Biol*. **16** 407–426. MR2487566 https://doi.org/10.1089/cmb.2008.0081

[61] SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P. (2003). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, New York.

[62] SPIEGELHALTER, D. J. and FREEDMAN, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat. Med*. **5** 1–13. https://doi.org/10.1002/sim.4780050103

[63] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

[64] SQUIRES, C., MAGLIACANE, S., GREENEWALD, K., KATZ, D., KOCAOGLU, M. and SHANMUGAM, K. (2020). Active structure learning of causal DAGs via directed clique trees. In *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*. *NIPS'*20. Curran Associates Inc., Red Hook, NY, USA.

[65] STEFAN, A. M., SCHÖNBRODT, F. D., EVANS, N. J. and WAGENMAKERS, E. J. (2022). Efficiency in sequential testing: Comparing the sequential probability ratio test and the sequential Bayes factor test. *Behav. Res. Methods* **54** 1554–3528.

[66] R CORE TEAM (2021). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[67] TONG, S. and KOLLER, D. (2001). Active learning for structure in Bayesian networks. In *Proceedings of the* 17*th International Joint Conference on Artificial Intelligence - Volume* 2. *IJCAI'*01 863–869. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[68] 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. Elsevier Science Inc., New York, NY, USA.

[69] VON KÜGELGEN, J., RUBENSTEIN, P. K., SCHÖLKOPF, B. and WELLER, A. (2019). Optimal experimental design via Bayesian optimization: Active causal structure learning for Gaussian process networks. In *NeurIPS* 2019 *Workshop do the Right Thing*: *Machine Learning and Causal Inference for Improved Decision Making*.

[70] WANG, F. and GELFAND, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statist. Sci*. **17** 193–208. MR1925941 https://doi.org/10.1214/ss/1030550861

[71] WEISS, R. (1997). Bayesian sample size calculations for hypothesis testing. *J. R. Stat. Soc*., *Ser. D*, *Stat*. **46** 185–191.

[72] YANG, K., KATCOFF, A. and UHLER, C. (2018). Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proceedings of the* 35*th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 5541–5550. PMLR.

[73] ZHANG, K., DUAN, X. and WU, J. (2016). Multigene disruption in undomesticated Bacillus subtilis ATCC 6051a using the CRISPR/Cas9 system. *Sci. Rep*. **6** 27943.