The Annals of Applied Statistics
2019, Vol. 13, No. 4, 2289–2311
https://doi.org/10.1214/19-AOAS1275

OBJECTIVE BAYES MODEL SELECTION OF GAUSSIAN INTERVENTIONAL ESSENTIAL GRAPHS FOR THE IDENTIFICATION OF SIGNALING PATHWAYS[1]

BY FEDERICO CASTELLETTI AND GUIDO CONSONNI

Università Cattolica del Sacro Cuore

A signalling pathway is a sequence of chemical reactions initiated by a stimulus which in turn affects a receptor, and then through some intermediate steps cascades down to the final cell response. Based on the technique of flow cytometry, samples of cell-by-cell measurements are collected under each experimental condition, resulting in a collection of interventional data (assuming no latent variables are involved). Usually several external interventions are applied at different points of the pathway, the ultimate aim being the structural recovery of the underlying signalling network which we model as a causal Directed Acyclic Graph (DAG) using intervention calculus. The advantage of using interventional data, rather than purely observational one, is that identifiability of the true data generating DAG is enhanced. More technically a Markov equivalence class of DAGs, whose members are statistically indistinguishable based on observational data alone, can be further decomposed, using additional interventional data, into smaller distinct Interventional Markov equivalence classes. We present a Bayesian methodology for structural learning of Interventional Markov equivalence classes based on observational and interventional samples of multivariate Gaussian observations. Our approach is objective, meaning that it is based on default parameter priors requiring no personal elicitation; some flexibility is however allowed through a tuning parameter which regulates sparsity in the prior on model space. Based on an analytical expression for the marginal likelihood of a given Interventional Essential Graph, and a suitable MCMC scheme, our analysis produces an approximate posterior distribution on the space of Interventional Markov equivalence classes, which can be used to provide uncertainty quantification for features of substantive scientific interest, such as the posterior probability of inclusion of selected edges, or paths.

1. Introduction. In biology, a signalling *pathway* (or cascade) is a sequence of chemical reactions initiated by a stimulus acting on a receptor, that is subsequently passed to the cell interior, and next to effector molecules, ultimately resulting in a cell response. Different reagents, capable of inhibiting or activating signalling nodes, are applied at several points of the pathway leading to distinct *interventional* settings, which define a collection of pathways corresponding to distinct triggers. Rather than considering them in isolation, one should embed such

Received April 2018; revised February 2019.
[1]Supported in part by UCSC (Research grant track D1).
Key words and phrases. Graphical model, causal inference, Markov equivalence class.

pathways into a *network*, in order to explicate potential interactions as well as other complexities. In this context the protein signalling study of Sachs et al. (2005) represents an important contribution. The resulting data set contains measurements of the abundance of 11 phosphoproteins and phospholipids recorded under different experimental settings in primary human immune system cells using flow cytometry.

More generally however, experimental data may contain both *interventional* and *observational* data, the latter arising through measurements on variables without external interventions. Data having this mixed configuration define the scope of this paper. Specifically we consider continuous multivariate observations whose joint distribution belongs to a *graphical model*, namely a family of probability distributions satisfying conditional independencies encoded by a graph (Lauritzen (1996)). In particular we rely on *Directed Acyclic Graphs* (DAGs), namely graphs having only directed edges between pairs of vertices; see, for instance, Friedman (2004), Shojaie and Michailidis (2009), and assume that the joint distribution of each observation is multivariate Gaussian. Each vertex of the DAG represents a variable in the system, and the goal is recovering the structure of the DAG which supposedly generated the observations (*structural learning*). When all the variables are discrete, DAGs are often referred to as *Bayesian networks*.

An important issue we need to address at this stage is *identifiability*. If only observational data are available, then one cannot in general distinguish between *Markov equivalent* DAGs, that is distinct DAGs encoding the same set of conditional independencies; see, for instance, Verma and Pearl (1991). Such DAGs can however be collected into Markov equivalence classes, each one represented by an *Essential Graph* (EG), also called *Completed Partially Directed Acyclic Graph* (CPDAG); see Andersson, Madigan and Perlman (1997b) and Chickering (2002). Structural learning then entails the identification of the underlying EG.

DAG models can be used purely to structure the joint distribution of a set of random variables leading to inference in terms of association. *Causal* statements however require assumptions beyond the nature of the joint distribution (Pearl (1995)). One possibility is to assume that the data were generated by a DAG, and that there are no unmeasured confounders. This is a strong assumption which is required however because graphical Markov models are not closed under marginalization (Richardson and Spirtes (2002)). If not satisfied, it may lead to erroneous consequences.

The notion of a *causal* DAG can be made precise by means of the "do" calculus (Pearl (2000)) and the allied notion of *interventional* distribution. Notice that this is predicated on a given DAG and represents an assumption about the behavior of the data generating mechanism. Summaries of the interventional distribution, such as the direct causal effect of a variable onto another can then be used. At this stage a couple of complications arise. On the one hand a DAG is generally not identifiable using observations alone even under the assumption of faithfulness (Spirtes, Glymour and Scheines (2000)). This leads to an equivalence class of DAGs. For each

of these a distinct interventional distribution holds: as a consequence we obtain a collection of causal effects for the same intervention which we can try to summarize in some meaningful way. Additionally and importantly, the structure of the true underlying graph is typically unknown, and consequently one first needs to estimate an equivalence class and then carry out a causal analysis followed by some summarization (e.g., establishing a lower bound on the causal effect of a variable onto a response); this program is admirably carried out in Maathuis, Kalisch and Bühlmann (2009).

A further possibility, which is the focus of this paper, is to assume that, besides observational data, we also collect *interventional data*. We can then couple the interventional distribution with the standard *observational* distribution for DAGs to obtain the overall *joint* distribution for observational and interventional data (Hauser and Bühlmann (2015)).

As for the purely observational case, DAGs can still be partitioned into *interventional* Markov equivalence classes (Hauser and Bühlmann (2012)), each class being represented by an *Interventional Essential Graph* (I-EG). Since the size of an equivalence class represents a measure of complexity of *causal learning* (He and Geng (2008)), it is important to emphasize that interventional Markov equivalence classes are smaller than the corresponding observational classes. Accordingly, identifiability of the true data generating model will be enhanced through interventions, and this represents an important motivation for the methodology developed in this paper.

A comprehensive and rigorous treatment of joint Gaussian modeling of observational and interventional data for structural learning of I-EGs is presented in Hauser and Bühlmann (2012) where the notion of interventional Markov equivalence class and I-EG is presented together with several characterizations. Additionally they derive maximum likelihood estimators and provide algorithmic operations to efficiently traverse the search space of I-EGs leading to their GIES algorithm. The companion paper Hauser and Bühlmann (2015) proves in particular consistency of the BIC model selection procedure for the identification of the true underlying I-EG. We refer to these two papers for a variety of technical results on inference, model selection consistency as well as computation.

In this paper we present an objective Bayes methodology for structural learning of Gaussian I-EGs based on observational and interventional data. In this way we are able to exhibit a measure of uncertainty for several quantities of interest through their corresponding posterior distribution. In particular we can compute an approximate posterior probability of specific graphical structures of interest, as well as the posterior inclusion probability of any edge, thus providing a more informative answer to substantive scientific queries. We build on results obtained in Castelletti et al. (2018) for structural learning of Gaussian EGs based on observational data only, extending their methodology to manage observational and interventional data jointly. We derive a closed-form expression for the marginal

likelihood of an interventional essential graph, and propose an MCMC strategy to explore the space of I-EGs.

The rest of the paper is organized as follows. In Section 2 we give a basic overview of graphical models, and discuss interventions on DAGs together with the notion of interventional Markov equivalence as formalized in Hauser and Bühlmann (2012). In Section 3 we present our methodology for the computation of the marginal likelihood of an I-EG. In Section 4 we summarize the MCMC strategy that we adopt to perform structural learning of interventional Markov equivalence classes of DAGs. Then, we apply the proposed methodology to some simulation settings (Section 5) and to the analysis of the protein-signaling data (Section 6). Finally Section 7 offers a brief discussion together with possible future developments. To ease the flow of ideas, a few technical results are presented in the Supplementary Material (Castelletti and Consonni (2019)).

2. Background.

2.1. *Graphical models.* A graph $\mathcal{G}$ is a pair (V, E) where $V = \{1, \ldots, q\}$ is a set of vertices (or nodes) and $E \subseteq V \times V$ a set of edges. Let $u, v \in V$, $u \neq v$. If $(u, v) \in E$ and $(v, u) \notin E$ we say that $\mathcal{G}$ contains the directed edge $u \to v$. If instead $(u, v) \in E$ and $(v, u) \in E$ we say that $\mathcal{G}$ contains the undirected edge $u - v$. Two vertices u, v are adjacent if they are connected by an edge (directed or undirected). Moreover, if $u - v$ is in $\mathcal{G}$ we say that u is a *neighbor* of v in $\mathcal{G}$. The neighbor set of v is denoted by $\text{ne}_{\mathcal{G}}(v)$; the common neighbor set of u and v is then $\text{ne}_{\mathcal{G}}(u, v) = \text{ne}_{\mathcal{G}}(u) \cap \text{ne}_{\mathcal{G}}(v)$. For any pair of distinct nodes $u, v \in V$, we say that u is a *parent* of v if $u \to v$. Conversely, we say that v is a *son* of u. The set of all parents of u in $\mathcal{G}$ is denoted by $\text{pa}_{\mathcal{G}}(u)$.

A graph is called *directed* (*undirected*) if it contains only directed (undirected) edges. A sequence of distinct vertices $\{v_0, v_1, \ldots, v_k\}$ in $\mathcal{G}$ is a *path* from v_0 to v_k if $\mathcal{G}$ contains $v_{j-1} - v_j$ or $v_{j-1} \to v_j$ for all $j = 1, \ldots, k$. A path is directed (undirected) if all edges are directed (undirected). Moreover, we say that a path is *partially directed* if it contains at least one directed edge. A path such that $v_0 = v_k$ is called a *cycle*. Let $A \subseteq V$. We denote with $\mathcal{G}_A = (A, E_A)$ the *subgraph* of $\mathcal{G} = (V, E)$ induced by A, whose edge set is $E_A = E \cap (A \times A)$.

An undirected (sub)graph is complete if its vertices are all adjacent. A particular class of undirected graphs is represented by *decomposable* graphs, also called *chordal* or *triangulated*. An undirected graph is decomposable if every path of length $l \geq 4$ contains a chord, that is two nonconsecutive adjacent vertices; see Lauritzen (1996).

A graph with only directed edges is called a *Directed Acyclic Graph* (DAG for short, denoted by $\mathcal{D}$) if it does not contain cycles. A graph with no directed cycles that may contain both directed and undirected edges is called a *Chain Graph* (CG) or simply *Partially Directed Acyclic Graph* (PDAG). For a chain graph $\mathcal{G}$ we call

chain component $\tau \subseteq V$ a set of nodes that are joined by an undirected path. The set of chain components of a CG is denoted by $\mathcal{T}$.

A subgraph of the form $u \to z \leftarrow v$, where there are no edges between u and v, is called a *v-structure* (or *immorality*). The *skeleton* of a graph $\mathcal{G}$ is the undirected graph on the same set of vertices obtained by removing the orientation of all its edges.

Let now $\mathcal{D}$ be a DAG on the set of vertices V. We denote with $[\mathcal{D}]$ its Markov equivalence class, that is, the set of all DAGs with vertex set V encoding the same conditional independencies of $\mathcal{D}$. We know from Verma and Pearl (1991), Theorem 2, that $\mathcal{D}' \in [\mathcal{D}]$ if and only if $\mathcal{D}$ and $\mathcal{D}'$ have the same skeleton and v-structures. Moreover, the Markov equivalence class $[\mathcal{D}]$ can be uniquely represented by the EG $\mathcal{D}^\star = \bigcup\{\mathcal{D}' | \mathcal{D}' \in [\mathcal{D}]\}$, where the union is to be interpreted over the edge sets, so that $u \to v$ in $\mathcal{D}$ and $v \to u$ in $\mathcal{D}'$ gives $u - v$ in $\mathcal{D}^\star$. An important result of Andersson, Madigan and Perlman (1997a) is the characterization of an EG as a CG with decomposable chain components.

THEOREM 2.1 (Andersson, Madigan and Perlman (1997a), Theorem 4.1). *A graph $\mathcal{G} = (V, E)$ is the EG $\mathcal{D}^*$ for some DAG $\mathcal{D}$ with vertex set V if and only if $\mathcal{G}$ satisfies the following four conditions*:

(i) *$\mathcal{G}$ is a CG*;
(ii) *for each chain component $\tau \in \mathcal{T}$ the subgraph $\mathcal{G}_\tau$ is a decomposable UG*;
(iii) *$\mathcal{G}$ has no flags* (*no induced subgraphs of the form $u \to v - z$*);
(iv) *each directed edge $u \to v$ contained in $\mathcal{G}$ is strongly protected* (*as illustrated in Definition* 3.3 *of Andersson, Madigan and Perlman* (1997a)).

2.2. *Interventions on DAGs*. In this section we introduce interventions on DAGs and summarize the main results about interventional Markov equivalence developed by Hauser and Bühlmann (2012).

Let $Y_1, \ldots, Y_q$ be a set of random variables from which we collect n multivariate observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, $\mathcal{D} = (V, E)$ a DAG. As we associate each variable Y_j to a vertex in $\mathcal{D}$, we constrain the distribution of each $\boldsymbol{y}_i$ by the edges in $\mathcal{D}$. We then write

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

where $\mathrm{pa}_{\mathcal{D}}(j)$ is the set of parents of node j in $\mathcal{D}$. Let now $I \subseteq V$ and $Y_I = \{Y_j, j \in I\}$ the corresponding subset of random variables. Following Pearl (2000), an intervention on I is defined as the action of setting or forcing Y_I to the value of a random variable U_I with density $\tilde{f}(\cdot)$ such that U_I is independent of Y_j, for each $j \in \mathrm{pa}_{\mathcal{D}}(I)$. We call I an *intervention target*. For a given DAG $\mathcal{D}$, an intervention on I *destroys* the original dependence between the intervened variable Y_I and its parents in $\mathcal{D}$ and leads to the definition of *intervention graph*.

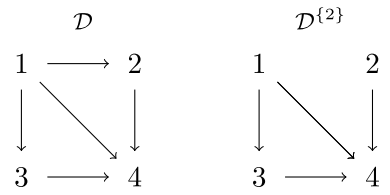

FIG. 1. *A DAG $\mathcal{D}$ and the intervention DAG $\mathcal{D}^{\{2\}}$ for the target $I = \{2\}$.*

DEFINITION 2.1 (Hauser and Bühlmann (2012), Defn. 5). Let $\mathcal{D} = (V, E)$ be a DAG, $I \subseteq V$ an intervention target. We call *intervention graph* of $\mathcal{D}$ the DAG $\mathcal{D}^I = (V, E^I)$, with $E^I := \{(u, v) : (u, v) \in E, v \notin I\}$.

In the following we assume for simplicity *single node* interventions, that is $|I| = 1$, which is also the case of the Sachs data. For example, in Figure 1 we have a DAG $\mathcal{D}$ and, given the target $I = \{2\}$, the corresponding intervention DAG $\mathcal{D}^{\{2\}}$, obtained by removing all edges $u \to 2$. The *post-intervention joint distribution* of $(Y_1, \ldots, Y_q)|Y_I \leftarrow U_I$ is then obtained using the truncated factorization

$$f_{\mathcal{D}^I}(\boldsymbol{y}|Y_I \leftarrow U_I) = \prod_{j \neq I} f(y_j|\boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)}) \cdot \tilde{f}(y_I). \tag{2}$$

With $I = \emptyset$ (no interventions) and using the convention $f_{\mathcal{D}^\emptyset}(\boldsymbol{y}|Y_\emptyset \leftarrow U_\emptyset) = f_{\mathcal{D}}(\boldsymbol{y})$, equation (2) reduces to equation (1), which holds in the observational setting.

A crucial aspect of (2) is that the manipulated variable and its parents are now *marginally independent*. This feature can be exploited if a post-intervention sample is available and provided faithfulness is assumed—that is, any conditional independence relation in the joint distribution is entailed by the factorization (1)—in order to orient undirected edges of an EG. One approach is presented in He and Geng (2008). They first estimate an EG based on observational data, and then carry out an independence test between the intervened variable and its parents using interventional data, whose outcome in turn determines the orientation of edges connecting node I, as well as other edges whose reverse orientation creates v-structures or cycles. For instance, suppose an undirected edge $I - j$ occurs in the estimated essential graph. Then $I - j$ can be oriented as $I \leftarrow j$ if the interventional sample suggests $Y_I \perp\!\!\!\perp Y_j$ (because the arrow would drop in the ensuing intervention DAG; see Definition 2.1), otherwise it is oriented as $I \to j$ (because the arrow would remain in the ensuing intervention DAG). Given the newly determined EG one can sequentially apply the same principle using further interventions until eventually all edges are oriented, so that the EG becomes a DAG. This method, called *active learning*, proceeds in stages and is especially suited from an experimental design perspective, where at each stage one should determine what is the best course of action in terms of the variables to intervene upon.

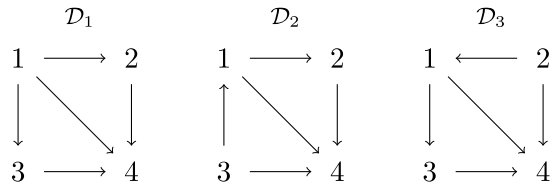

FIG. 2. *A Markov equivalence class with three Markov equivalent DAGs $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$. Assuming $\mathcal{I} = \{\emptyset, \{2\}\}$, $\mathcal{D}_3$ is no longer $\mathcal{I}$-Markov equivalent to $\mathcal{D}_1$ and $\mathcal{D}_2$.*

An alternative approach followed by Hauser and Bühlmann (2015)—as well as ourselves—*jointly* models all the data (observational and interventional) simultaneously. The goal however is the same, that is making edges distinguishable so that they can be oriented given interventional data. However in this setting the goal is achieved through a characterization of the model space which incorporates upfront the information originating from the collection of intervention targets, and is encapsulated in the notion of interventional Markov equivalence class; see below for details.

Consider now a set of intervention targets $\mathcal{I} = \{I_k, k = 1, \ldots, K\}$, also called a *family of targets*. In particular, we say that $\mathcal{I}$ is *conservative* if for each $j \in V$ there is at least one $I_k \in \mathcal{I}$ such that $j \neq I_k$; see Hauser and Bühlmann (2012) for a detailed discussion about the importance of such property. Each target is then associated to a density $\tilde{f}_k(\cdot)$ assigned to Y_{I_k}. We also assume U_{I_h} independent of U_{I_k}, for each $h \neq k$. We know that two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are (observationally) Markov equivalent if $f_{\mathcal{D}_1}(\cdot)$ and $f_{\mathcal{D}_2}(\cdot)$ encode the same conditional independencies (Section 2.1). Markov equivalence with respect to a family of intervention targets $\mathcal{I}$ states that $\mathcal{D}_1$ and $\mathcal{D}_2$ are *interventionally Markov equivalent* if $f_{\mathcal{D}_1^I}(\cdot)$ and $f_{\mathcal{D}_2^I}(\cdot)$ encode the same conditional independencies for each $I \in \mathcal{I}$. Theorem 2.2 provides a graphical criterion to establish if $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent under the conservative family of targets $\mathcal{I}$.

THEOREM 2.2 (Hauser and Bühlmann (2012), Theorem 10). *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two DAGs and $\mathcal{I}$ a conservative family of targets. Then, $\mathcal{D}_1$ and $\mathcal{D}_2$ are $\mathcal{I}$-Markov equivalent ($\mathcal{D}_1 \sim_{\mathcal{I}} \mathcal{D}_2$) if for each $I \in \mathcal{I}$, $\mathcal{D}_1^I$ and $\mathcal{D}_2^I$ have the same skeleton and v-structures.*

Let now $[\mathcal{D}]_{\mathcal{I}}$ be the $\mathcal{I}$-Markov equivalence class of $\mathcal{D}$, that is the set of all DAGs that are $\mathcal{I}$-Markov equivalent to $\mathcal{D}$. An important consequence of Theorem 2.2 is that interventions based on a conservative family of targets define a finer partition of DAGs into equivalence classes; see also Hauser and Bühlmann (2012) for details. For instance, in Figure 2 we have a Markov equivalence class with three Markov equivalent DAGs, $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$. However, under the family of targets $\mathcal{I} = \{\emptyset, \{2\}\}$, $\mathcal{D}_3$ is not $\mathcal{I}$-Markov equivalent to $\mathcal{D}_1$ and $\mathcal{D}_2$.

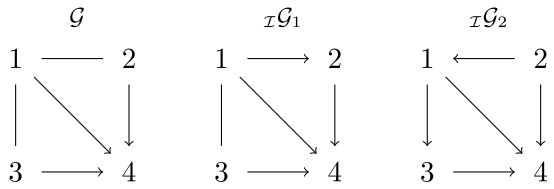

FIG. 3. *An EG $\mathcal{G}$ and two $\mathcal{I}$-EGs $_{\mathcal{I}}\mathcal{G}_1$ and $_{\mathcal{I}}\mathcal{G}_2$ for $\mathcal{I} = \{\emptyset, \{2\}\}$. $\mathcal{G}$ is the representative of the Markov equivalence class $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$ in Figure 2 which is partitioned into two $\mathcal{I}$-Markov equivalence classes: $\{\mathcal{D}_1, \mathcal{D}_2\}$ represented by $_{\mathcal{I}}\mathcal{G}_1$ and $\{\mathcal{D}_3\}$ represented by $_{\mathcal{I}}\mathcal{G}_2 \equiv \mathcal{D}_3$.*

2.2.1. *Interventional essential graphs*. As for the observational case (Section 2.1), each interventional Markov equivalence class can be uniquely represented by a special CG called *Interventional Essential Graph* ($\mathcal{I}$-EG).

DEFINITION 2.2 (Hauser and Bühlmann (2012), Defn. 11). Let $\mathcal{D}$ be a DAG and $\mathcal{I}$ a conservative family of targets. The $\mathcal{I}$-essential graph of $\mathcal{D}$ is defined as $_{\mathcal{I}}\mathcal{G}(\mathcal{D}) = \bigcup_{\mathcal{D}^* \in [\mathcal{D}]_{\mathcal{I}}} \mathcal{D}^*$.

Figure 3 shows the EG $\mathcal{G}$ representing the Markov equivalence class $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$ of Figure 2 and the $\mathcal{I}$-EGs $_{\mathcal{I}}\mathcal{G}_1, _{\mathcal{I}}\mathcal{G}_2$ for the two $\mathcal{I}$-Markov equivalence classes $\{\mathcal{D}_1, \mathcal{D}_2\}$ and $\{\mathcal{D}_3\}$. Clearly, $_{\mathcal{I}}\mathcal{G}(\mathcal{D}) = _{\mathcal{I}}\mathcal{G}(\mathcal{D}^*)$ for each $\mathcal{D}^* \in [\mathcal{D}]_{\mathcal{I}}$. In the sequel we often use $\mathcal{G}$ to identify an $\mathcal{I}$-EG without making explicit its originating DAG and family of targets $\mathcal{I}$. The following theorem gives the characterization of an $\mathcal{I}$-EG.

THEOREM 2.3 (Hauser and Bühlmann (2012), Theorem 18). *Let $\mathcal{D}$ be a DAG on the set of vertices V and $\mathcal{I}$ a conservative family of targets. A graph $\mathcal{G}$ is the $\mathcal{I}$-essential graph of $\mathcal{D}$ if and only if*:

(i) *$\mathcal{G}$ is a chain graph*;
(ii) *for each chain component $\tau \in \mathcal{T}$, $\mathcal{G}_\tau$ is chordal*;
(iii) *$\mathcal{G}$ has no induced subgraphs of the form $u \to v - z$ (flags)*;
(iv) *every arrow $u \to v$ is strongly $\mathcal{I}$-protected (as illustrated in Definition 14 of Hauser and Bühlmann (2012))*;
(v) *$\mathcal{G}$ has no line $u - v$ for which there exists some $I \in \mathcal{I}$ such that $|I \cap \{u, v\}| = 1$.*

Recall that a chordal graph (also called *decomposable*) can be uniquely represented through its set of separators and cliques, which are denoted by $\mathcal{C}$ and $\mathcal{S}$, respectively; see, for instance, (Lauritzen (1996), p. 18). While Conditions (i), (ii), (iii) are the same as in Theorem 2.1, Condition (iv) is a natural extension of the corresponding one. Condition (v) is instead specific to the interventional setting and of particular interest for this work. It says that, given a family of targets $\mathcal{I}$,

each $\mathcal{I}$-EG $\mathcal{G}$ is such that it has no chain components containing nodes on which at least one intervention was performed together with nodes on which no interventions were made.

3. Gaussian interventional essential graphs. In this section we focus on Gaussian interventional essential graphs. The objective is the computation of the marginal likelihood of an $\mathcal{I}$-EG given a collection of (observational and) interventional data. Results are based on the marginal likelihood of Gaussian models obtained by adopting the notion of Fractional Bayes Factor. Please refer to the Supplementary Material for a concise background on such results.

3.1. *Likelihood and prior factorization.* In the interventional setting of Section 2.2 a dataset consists of a collection of multivariate observations each one associated to an intervention target. More specifically, we assume to have, for each target $I_k \in \mathcal{I}$ $n^{(k)}$ i.i.d. q-variate observations from the sampling distribution of $(Y_1, \dots, Y_q)|Y_{I_k} \leftarrow U_{I_k}$ collected in the $(n^{(k)}, q)$ matrix $\boldsymbol{Y}^{I_k} \equiv \boldsymbol{Y}^k$. By *row-binding* the matrices $\boldsymbol{Y}^k$, $I_k \in \mathcal{I}$, we then obtain the (n, q) data matrix $\boldsymbol{Y}$, where $n = \sum_k n^{(k)}$. Recall from Theorem 2.3 that an $\mathcal{I}$-EG $\mathcal{G}$ is a CG with set of chain components $\mathcal{T}$. Therefore, according to Andersson, Madigan and Perlman (2001), we can write the factorization

$$f_{\mathcal{G}}(\boldsymbol{Y}|\boldsymbol{\Theta}_{\mathcal{G}}) = \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau|\boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Theta}_{\mathcal{G}_\tau}), \tag{3}$$

where $\boldsymbol{Y}_\tau$ denotes selected columns of the data matrix $\boldsymbol{Y}$ corresponding to the subset $\tau \subseteq V$. $\boldsymbol{\Theta}_{\mathcal{G}}$ is instead a global parameter indexing the graphical model $\mathcal{G}$ and $\boldsymbol{\Theta}_{\mathcal{G}_\tau}$ a local parameter for chain component τ.

For a given family of targets $\mathcal{I}$, let $\mathcal{S}_q$ be the set of all $\mathcal{I}$-EGs on q nodes. As a consequence of Condition (v) in Theorem 2.3 we have that each $\mathcal{I}$-EG $\mathcal{G} \in \mathcal{S}_q$ contains a chain component $\tau = I_k$ for each $I_k \in \mathcal{I}$. Hence, we can write

$$f_{\mathcal{G}_\tau}\big(\boldsymbol{Y}^k_\tau|\boldsymbol{Y}^k_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Theta}_{\mathcal{G}_\tau}\big) = \begin{cases} f_{\mathcal{G}_\tau}\big(\boldsymbol{Y}^k_\tau|\boldsymbol{Y}^k_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}\big) & \text{if } I_k \neq \tau, \\ \tilde{f}_{\mathcal{G}_\tau}\big(\boldsymbol{y}^k_\tau|\boldsymbol{\psi}^k_\tau\big) & \text{if } I_k = \tau, \end{cases} \tag{4}$$

where $\boldsymbol{Y}^k_\tau$ denotes columns indexed by τ of $\boldsymbol{Y}^k$. The first case of equation (4) is the usual factorization for $\mathcal{G}_\tau$ which holds for all those observations $\boldsymbol{Y}^k_\tau$ such that $I_k \neq \tau$, that is when no interventions are performed on chain component $I_k = \tau$. The second case corresponds instead to the (intervention) density $\tilde{f}_{\mathcal{G}_\tau}(\cdot)$, where the intervention on I_k destroys the original dependence between node $I_k = \tau$ and its parents. Moreover, $\boldsymbol{\psi}^k_\tau$ is a parameter modelling the effect of the intervention on chain component τ, while $\boldsymbol{\theta}_{\mathcal{G}_\tau}$ is the chain component parameter of the conditional distribution of those observations not arising from an intervention on τ. Implicitly

we assume that $\boldsymbol{\theta}_{\mathcal{G}_\tau}$ does not depend on I_k. Assuming now $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ independent we can write each term in (3) as

$$
\begin{aligned}
&f_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau | \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Theta}_{\mathcal{G}_\tau}) \\
&\quad = \prod_{k:I_k \neq \tau} f_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau^k | \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}^k, \boldsymbol{\theta}_{\mathcal{G}_\tau}) \cdot \prod_{k:I_k=\tau} \tilde{f}_{\mathcal{G}_\tau}(\boldsymbol{y}_\tau^k | \boldsymbol{\psi}_\tau^k) \\
&\quad = f_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau^* | \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}^*, \boldsymbol{\theta}_{\mathcal{G}_\tau}) \cdot \prod_{k:I_k=\tau} \tilde{f}_{\mathcal{G}_\tau}(\boldsymbol{y}_\tau^k | \boldsymbol{\psi}_\tau^k),
\end{aligned} \tag{5}
$$

being $\boldsymbol{Y}_\tau^*$ a $(n_\tau^*, |\tau|)$ matrix collecting all the observations $\boldsymbol{Y}_\tau^k$ such that $I_k \neq \tau$. We assume that the prior on $\boldsymbol{\Theta}_{\mathcal{G}}$ factorizes as

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

while for each chain component parameter $\boldsymbol{\Theta}_{\mathcal{G}_\tau}$ we assume $\boldsymbol{\theta}_{\mathcal{G}_\tau}$ a priori independent of $\boldsymbol{\psi}_\tau^k$, for all k. Hence, we obtain

$$
p(\boldsymbol{\Theta}_{\mathcal{G}_\tau}) = p(\boldsymbol{\theta}_{\mathcal{G}_\tau}) \cdot \prod_{k:I_k=\tau} p(\boldsymbol{\psi}_\tau^k), \tag{7}
$$

which extends the assumption of global parameter independence (Cowell et al. (1999)) to an interventional setting.

3.2. *Marginal likelihood.* Under the assumptions of Section 3.1 the marginal likelihood of the $\mathcal{I}$-EG $\mathcal{G}$ given the data $\boldsymbol{Y}$ can be computed as

$$
\begin{aligned}
m_{\mathcal{G}}(\boldsymbol{Y}) &= \prod_{\tau \in \mathcal{T}} \int f_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau | \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Theta}_{\mathcal{G}_\tau}) p(\boldsymbol{\Theta}_{\mathcal{G}_\tau}) \, d\boldsymbol{\Theta}_{\mathcal{G}_\tau} \\
&= \prod_{\tau \in \mathcal{T}} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau | \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}).
\end{aligned} \tag{8}
$$

From Theorem 2.3 recall that each $\mathcal{G}_\tau$ is a decomposable (chordal) graph (possibly made up by a single node). Let $\mathcal{C}_\tau$ be its set of (maximal) cliques and $\mathcal{S}_\tau$ the corresponding set of separators; see also Lauritzen (1996). Then

$$
m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau | \boldsymbol{X}_\tau) = \frac{\prod_{C \in \mathcal{C}_\tau} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_{C,\tau} | \boldsymbol{X}_\tau)}{\prod_{S \in \mathcal{S}_\tau} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_{S,\tau} | \boldsymbol{X}_\tau)}, \tag{9}
$$

where $\boldsymbol{X}_\tau = \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}$ and

$$
m_{\mathcal{G}_\tau}(\boldsymbol{Y}_{J,\tau} | \boldsymbol{X}_\tau) = m_{\mathcal{G}_\tau}(\boldsymbol{Y}_{J,\tau}^* | \boldsymbol{X}_\tau^*) \cdot \prod_{k:I_k=\tau} m_{\mathcal{G}_\tau}(\boldsymbol{y}_\tau^k), \tag{10}
$$

where $J \subseteq \tau$ refers to a generic clique or separator of $\mathcal{G}_\tau$.

Let $\mathcal{I}$ be a family of targets, $\mathcal{G}$ an $\mathcal{I}$-EG. For each observation (row) $\boldsymbol{y}^*_{i,\tau}$ in $\boldsymbol{Y}^*_\tau$ we assume

$$f_{\mathcal{G}_\tau}(\boldsymbol{y}^*_{i,\tau}|\boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) = \mathcal{N}_{|\tau|}(\boldsymbol{y}^*_{i,\tau}|\boldsymbol{\alpha}_\tau + \boldsymbol{\Gamma}_\tau \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Omega}^{-1}_{\mathcal{G}_\tau}), \tag{11}$$

independently with respect to i, where $\boldsymbol{\alpha}_\tau + \boldsymbol{\Gamma}_\tau \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}$ denotes the conditional mean, $\boldsymbol{\alpha}_\tau = \boldsymbol{\mu}_\tau - \boldsymbol{\Gamma}_\tau \boldsymbol{\mu}_{\mathrm{pa}_{\mathcal{G}}(\tau)}$, $\boldsymbol{\Gamma}_\tau$ is the matrix of regression parameters and $\boldsymbol{\Omega}_{\mathcal{G}_\tau}$ the conditional precision matrix; see also Castelletti et al. (2018). Now, letting

$$\boldsymbol{x}^*_{i,\tau} = \begin{bmatrix} 1 \\ \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)} \end{bmatrix}, \qquad \boldsymbol{B}_\tau = \begin{bmatrix} \boldsymbol{\alpha}_\tau^\top \\ \boldsymbol{\Gamma}_\tau^\top \end{bmatrix}, \tag{12}$$

we can write

$$f_{\mathcal{G}_\tau}(\boldsymbol{y}^*_{i,\tau}|\boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) = \mathcal{N}_{|\tau|}(\boldsymbol{y}^*_{i,\tau}|\boldsymbol{B}_\tau^\top \boldsymbol{x}^*_{i,\tau}, \boldsymbol{\Omega}^{-1}_{\mathcal{G}_\tau}). \tag{13}$$

Please observe that the matrix $\boldsymbol{B}_\tau$ consists of unconstrained component since the $\mathcal{I}$-EG $\mathcal{G}$ has no flags (Condition (iii) of Theorem 2.3). In matrix normal notation we can write

$$f_{\mathcal{G}_\tau}(\boldsymbol{Y}^*_\tau|\boldsymbol{Y}^*_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_\tau) = \mathcal{N}_{n^*_\tau,|\tau|}(\boldsymbol{Y}^*_\tau|\boldsymbol{X}^*_\tau \boldsymbol{B}_\tau, \boldsymbol{I}_{n^*_\tau}, \boldsymbol{\Omega}^{-1}_{\mathcal{G}_\tau}), \tag{14}$$

being

$$\boldsymbol{X}^*_\tau = \begin{pmatrix} \boldsymbol{x}^{*\top}_{1,\tau} \\ \vdots \\ \boldsymbol{x}^{*\top}_{n^*,\tau} \end{pmatrix}.$$

Consonni, La Rocca and Peluso (2017) derive a formula to compute the marginal likelihood of a decomposable UG allowing for the presence of covariates, by adopting an objective Bayes approach based on the notion of fractional Bayes factor; see also the Supplementary Material (Castelletti and Consonni (2019), Section 1). Their approach relies on the methodology for prior construction introduced by Geiger and Heckerman (2002). According to this method, it is sufficient to specify a prior distribution under *any* complete DAG model (that is when $\boldsymbol{\Omega}_{\mathcal{G}_\tau}$ is unconstrained) which is used to compute the corresponding marginal likelihood given the data $\boldsymbol{Y}_\tau$. The marginal likelihood of any graphical model $\mathcal{G}_\tau$ with $\boldsymbol{\Omega}_{\mathcal{G}_\tau}$ Markov w.r.t. a decomposable graph can be derived automatically. Such methodology is then extended in Castelletti et al. (2018) to the EG setting. Formally, to compute each term $m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau|\boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)})$ in (8) we need a formula for the marginal likelihood of a Gaussian multivariate regression model; see (14). Moreover, because of assumption (7), we can specify priors separately for each chain component τ. Let $\boldsymbol{\Omega}_\tau$ denote the precision matrix of the variables in τ under a complete (unconstrained) graph. Assuming the default prior on $(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau)$,

$$p(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau) \propto |\boldsymbol{\Omega}_\tau|^{\frac{a_D - |\tau| - 1}{2}}, \tag{15}$$

the fractional marginal likelihood for model (14) restricted to the subset $J \subseteq \tau$ is obtained as

$$
\begin{aligned}
m_\tau(\boldsymbol{Y}^*_{J,\tau}|\boldsymbol{X}^*_\tau) = (\pi)^{-\frac{(n^*_\tau - n^*_0)|J|}{2}} \cdot \frac{\Gamma_{|J|}(\frac{a_D - |\bar{J}| + n^*_\tau - |\mathrm{pa}_{\mathcal{G}}(\tau)| - 1}{2})}{\Gamma_{|J|}(\frac{a_D - |\bar{J}| + n^*_0 - |\mathrm{pa}_{\mathcal{G}}(\tau)| - 1}{2})} \\
\times \left(\frac{n^*_0}{n^*}\right)^{\frac{|J|(a_D - |\bar{J}| + n^*_0)}{2}} |\hat{\boldsymbol{E}}^{*\top}_{J,\tau} \hat{\boldsymbol{E}}^*_{J,\tau}|^{-\frac{n^* - n^*_0}{2}},
\end{aligned} \tag{16}
$$

where we will set, in general, $a_D = |\tau| - 1$ according to Geisser and Cornfield (1963); see also the Supplementary Material (Castelletti and Consonni (2019), Section 1.1). We focus now on the densities $\tilde{f}_{\mathcal{G}_\tau}(\cdot)$ of formula (5). Assuming for each $y^k_{i,\tau}$ in $\boldsymbol{y}^k_\tau$,

$$
\tilde{f}_{\mathcal{G}_\tau}(y^k_{i,\tau}|\boldsymbol{\psi}^k_\tau) = \mathcal{N}(y^k_{i,\tau}|\xi^k_\tau, (\phi^k_\tau)^{-1}), \tag{17}
$$

we obtain

$$
\tilde{f}_{\mathcal{G}_\tau}(\boldsymbol{y}^k_\tau|\boldsymbol{\psi}^k_\tau) = \mathcal{N}_{n^{(k)},1}(\boldsymbol{y}^k_\tau|\boldsymbol{\delta}^k_\tau, \boldsymbol{I}_{n^{(k)}}, (\phi^k_\tau)^{-1}), \tag{18}
$$

where $\boldsymbol{\psi}^k_\tau = (\boldsymbol{\delta}^k_\tau, \phi^k_\tau)$ is the chain component parameter modelling the effect of the intervention with target I_k on the chain component τ and

$$
\boldsymbol{\delta}^k_\tau = \mathbf{1}_{n^{(k)}} \xi^k_\tau,
$$

being $\mathbf{1}_{n^{(k)}}$ the unit vector of length $n^{(k)}$. Starting from the default prior for $(\boldsymbol{\delta}^k_\tau, \phi^k_\tau)$,

$$
p(\boldsymbol{\delta}^k_\tau, \phi^k_\tau) \propto |\phi^k_\tau|^{\frac{a^k_D - 2}{2}},
$$

we obtain (see Supplementary Material, Section 1.3)

$$
\begin{aligned}
m_{\mathcal{G}_\tau}(\boldsymbol{y}^k_\tau) = (\pi)^{-\frac{n^{(k)}_\tau - n^{(k)}_0}{2}} \cdot \frac{\Gamma(\frac{a^{(k)}_D + n^{(k)}_\tau - 1}{2})}{\Gamma(\frac{a^{(k)}_D + n^{(k)}_0 - 1}{2})} \\
\times \left(\frac{n^{(k)}_0}{n^{(k)}_\tau}\right)^{\frac{a^{(k)}_D + n^{(k)}_0}{2}} (s^{(k)}_\tau)^{-\frac{n^{(k)}_\tau - n^{(k)}_0}{2}},
\end{aligned} \tag{19}
$$

where $s^{(k)}_\tau = \sum_{i=1}^{n^{(k)}} (e^{(k)}_{i,\tau})^2$, $e^{(k)}_{i,\tau} = y^k_{i,\tau} - \bar{\boldsymbol{y}}^k_\tau$ and $\bar{\boldsymbol{y}}^k$ is the sample mean of $\boldsymbol{y}^k$. Therefore, the overall marginal likelihood of a given $\mathcal{I}$-EG $\mathcal{G}$ can be recovered computing each term $m_{\mathcal{G}_\tau}(\boldsymbol{Y}_{J,\tau}|\boldsymbol{X}_\tau)$ in (10) using (16) and (19). These are combined in (9) to obtain $m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau|\boldsymbol{X}_\tau)$ and finally $m_{\mathcal{G}}(\boldsymbol{Y})$ according to (8).

4. MCMC algorithms. It is well known that the number of EGs grows super-exponentially in the number of nodes and then an exhaustive enumeration of all possible EGs on q nodes is feasible only for small values of q; see, for instance, Gillispie and Perlman (2002). The same happens in the $\mathcal{I}$-EG space which is generally *larger* than the corresponding EG space. Hence, model selection of $\mathcal{I}$-EGs requires the adoption of MCMC algorithms. In Castelletti et al. (2018) an MCMC on the EG space is implemented using the Markov chain on (observational) Markov equivalence classes of DAGs proposed by He, Jia and Yu (2013). Similarly, we construct an MCMC on *interventional* Markov equivalence classes of DAGs, which is resumed in Section 2 of the Supplementary Material. The output of Algorithm 1 (Supplementary Material, Section 2) consists of a collection of $\mathcal{I}$-EGs $\{\mathcal{G}^{(0)}, \ldots, \mathcal{G}^{(T)}\}$. Typically this is used to approximate the posterior distribution across models or equivalently to *estimate* posterior model probabilities. The number of visits of each model over the total number of iterations T is generally used as such approximation. Moreover, we can approximate the marginal posterior probability of inclusion of $u \to v$ as

$$
\begin{aligned}
p_{u\to v}(\boldsymbol{Y}) &= \sum_{\mathcal{G}\in\mathcal{S}_{u\to v}} p(\mathcal{G}|\boldsymbol{Y}) \\
&\approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{u\to v}(\mathcal{G}^{(t)}),
\end{aligned}
\tag{20}
$$

where $\mathcal{S}_{u\to v}$ is the class of $\mathcal{I}$-EGs containing the directed edge $u \to v$ (recall that an undirected edge $u - v$ is equivalent to $u \to v$ and $v \to u$) and $\mathbb{I}_{u\to v}$ is the indicator function taking value 1 if and only if $\mathcal{G}^{(t)}$ contains $u \to v$. We can also use these probabilities to summarize the MCMC output with a single model estimate. In model selection of EGs (Castelletti et al. (2018)) we adopt the so-called *projected median probability graph model*. This is constructed by including all edges $u \to v$ such that $p_{u\to v}(\boldsymbol{Y}) > 0.5$ (*median probability graph model*) and then constructing *any* consistent extension of the completely partially directed graph thus obtained. The final output is then the EG representing the Markov equivalent class of such consistent extension; see Castelletti et al. (2018) for details. The use of the projected median probability graph model in the EG setting does not introduce any discretion because all consistent extension belong to the same Markov equivalence class and so the resulting EG is unique. This is not guaranteed in general in the $\mathcal{I}$-EG context. To obtain a point estimate which summarizes our MCMC output we then proceed as follows.

Starting from the median probability graph model, which may contain both directed and undirected edges, we first obtain a directed version (DAG). Specifically, we first take all the directed edges in $\mathcal{G}$. Then, for each undirected edge $u - v$ in $\mathcal{G}$ we take $u \to v$ if and only if $p_{u\to v}(\boldsymbol{Y}) > p_{v\to u}(\boldsymbol{Y})$. Finally, from the DAG $\mathcal{D}$ thus obtained, we construct the corresponding $\mathcal{I}$-EG, that is the representative of the

interventional equivalence class of $\mathcal{D}$. We call again such result *projected median probability (graph) model*.

The median probability model specifies a threshold for edge inclusion $k = 0.5$. Alternatively, it is possible to choose k by looking at the *expected false discovery rate* (FDR) (Benjamini and Hochberg (1995)). Set for simplicity $p_{u\to v}(\boldsymbol{Y}) \equiv p_{u\to v}$; Then, the expected FDR for a given threshold $k \in (0, 1)$ can be defined as

$$\text{FDR}(k) = \frac{\sum_{u=1}^{q}\sum_{u\neq v}(1 - p_{u\to v})\mathbb{I}(p_{u\to v} \geq k)}{\sum_{u=1}^{q}\sum_{u\neq v}\mathbb{I}(p_{u\to v} \geq k)}, \tag{21}$$

being $\mathbb{I}$ the indicator function. $\text{FDR}(k)$ is an increasing function of k. Hence, one can select k so that the expected FDR is below a desired level, typically 0.05. See also Peterson, Stingo and Vannucci (2015) for the adoption of the FDR in multiple Gaussian graphical model selection.

5. Simulations. In this section we apply our methodology, named *Objective Bayes Interventional Essential graph Search* (OBIES), on a few simulation scenarios. Please refer to the Supplementary Material for additional results and comparisons with the benchmark GIES method (Hauser and Bühlmann (2012)) not shown here for brevity.

5.1. *Data generation and scenarios.* We construct a collection of simulation scenarios by varying the number of nodes $q \in \{10, 20, 40\}$, the number of observational data $n^{\emptyset} \in \{100, 200, 500, 1000\}$ and the proportion of intervened nodes $p \in \{0, 0.2, 0.4, 0.8\}$. For each target of intervention I_k, we then set the number of interventional data as $n^{(k)}(q, n^{\emptyset}) = n^{\emptyset}q/100$. Under each scenario characterized by $(q, n^{\emptyset}, p)$, 40 datasets, corresponding to 40 true DAGs, are generated. Each dataset, which contains both observational and interventional data, is obtained as follows.

For a given q, we randomly generate a topologically ordered DAG $\mathcal{D}$ with probability of edge inclusion $p_{\text{edge}} = 3/(2q - 2)$ (Peters and Bühlmann (2014)). The DAG thus obtained is the responsible of a data generating process and implies the set of equations

$$Y_{i,j} = \mu_j + \sum_{k\in \text{pa}_D(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j}, \tag{22}$$

for $i = 1, \ldots, n^{\emptyset}$, $j = 1, \ldots, q$, where $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ and $Y_{i,j}$ are independent with respect to i. For each j we fix $\mu_j = 0$ and $\sigma_j^2 = 1$, while regression coefficients $\beta_{k,j}$ are uniformly chosen in the interval $[-1, -0.1] \cup [0.1, 1]$; see also Peters and Bühlmann (2014). An observational dataset of size $n^{\emptyset}$ is then generated accordingly.

Next, for a given p (proportion of intervened nodes) we randomly sample without replacement $[pq]$ nodes in $\{1, \ldots, q\}$ which represents intervention targets of

size one, $I_1, \ldots, I_{[pq]}$. Under each I_k we first obtain the corresponding intervention DAG $\mathcal{D}^{I_k}$ (see Section 2.2) which implies the set of equations

$$Y_{i,j} = \begin{cases} \mu_j + \sum_{k \in \mathrm{pa}_{\mathcal{D}}(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j} & \text{if } j \neq I_k, \\ \delta_j + \epsilon_{i,j} & \text{if } j = I_k, \end{cases} \tag{23}$$

for $i = 1, \ldots, n^{(k)}$, $j = 1, \ldots, q$, where again $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$, $\epsilon_{i,j} \sim \mathcal{N}(0, \phi_j^2)$ and $Y_{i,j}$ are independent with respect to i. For each j we fix $\delta_j = 0$ and $\phi_j^2 = 0.1$. $n^{(k)}$ interventional data are then generated accordingly.

Let $\mathcal{S}_q$ be the set of all $\mathcal{I}$-EGs on q nodes. According to our MCMC algorithm (Supplementary Material, Section 2.2), we can also fix the maximum number of edges in the $\mathcal{I}$-EG space to a *small* multiple of the number of nodes q, rq (He, Jia and Yu (2013)), so that sparsity can be introduced to improve structural learning og $\mathcal{I}$-EGs. Specifically we choose $r = 2$, that is we require that the number of edges is not higher than 2 the number of nodes. We highlight that such threshold is well above the number of edges expected in the true DAG in each simulation scenario (7.5, 15 and 30 edges, respectively, for 10, 20 and 40 nodes). For each scenario, we use few pilot runs to choose the number of iterations as well as the initial *burn-in* period. We then fix $T = \{25{,}000, 50{,}000, 100{,}000\}$ for $q = \{10, 20, 40\}$, respectively.

5.2. *Results*. In the following we report a few simulation results from the application of OBIES to the scenarios of Section 5.1. To understand how much interventions can improve the identifiability of the true DAG generating model, we measure the Structural Hamming Distance (SHD), defined as the number of edge insertions, deletions or flips needed to transform the estimated $\mathcal{I}$-EG into the true DAG. In this way, under each setting defined by q, the benchmark of our comparison is represented by the same set of (40) true DAGs. In Figure 4 we report the boxplots of the SHD values over the 40 replicates under the simulation settings defined by q, $n^{\emptyset}$ and p. For each q and $n^{\emptyset}$ we observe that as the proportion of intervened nodes p increases, the SHDs between estimated graph and true DAG become smaller. Moreover, such reduction is all the more effective as $n^{\emptyset}$ (and so n^{I}) grows. As q increases, modelling jointly observational and interventional data produced under $p = 0.2$ results in a substantial reduction of the SHDs with respect to the true DAG if compared to the scenario $p = 0$. With reference to the $q = 20$ setting, we observe that interventions on the 40% of nodes *randomly chosen* are sufficient to strongly reduce the uncertainty around the true DAG estimate, especially for large sample sizes. In Table 1 we also report summary statistics (mean and standard deviation) of the SHDs represented in Figure 4.

6. Protein-signaling data. In this section we apply OBIES to the protein-signaling data set of Sachs et al. (2005). As mentioned, data consist of a collection

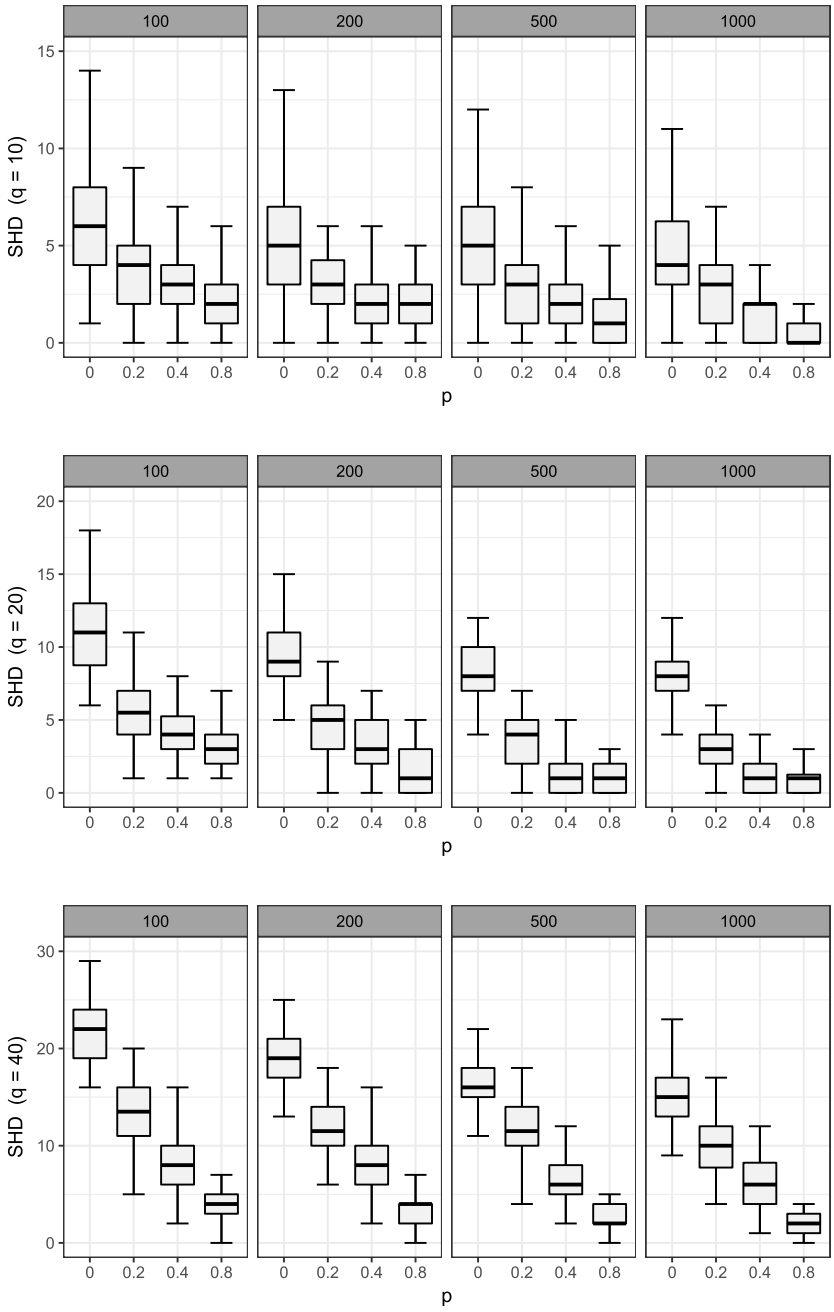

FIG. 4. *Simulations. Structural Hamming distances between the estimated $\mathcal{I}$-EG and the true DAG, over* 40 *datasets, for number of nodes* $q \in \{10, 20, 40\}$, *number of observational data* $n^{\emptyset} \in \{100, 200, 500, 1000\}$ *and proportion of intervened nodes* $p \in \{0, 0.2, 0.4, 0.8\}$.

TABLE 1
Mean (standard deviation) of the structural Hamming distances between OBIES estimate and true DAG over 40 *data sets for number of nodes* $q \in \{10, 20, 40\}$, $n^{\emptyset} \in \{100, 200, 500, 1000\}$ *and* $p \in \{0, 0.2, 0.4, 0.8\}$

	$n^{\emptyset} = 100$	$n^{\emptyset} = 200$	$n^{\emptyset} = 500$	$n^{\emptyset} = 1000$
		$q = 10$		
$p = 0$	6.78 (3.69)	5.78 (3.58)	5.20 (3.32)	5.08 (3.25)
$p = 0.2$	3.70 (2.03)	3.35 (1.90)	3.17 (2.31)	2.88 (2.21)
$p = 0.4$	3.10 (1.85)	2.17 (1.74)	2.13 (1.51)	1.50 (1.13)
$p = 0.8$	2.08 (1.83)	1.88 (1.45)	1.27 (1.72)	0.45 (0.71)
		$q = 20$		
$p = 0$	11.07 (3.11)	9.57 (2.16)	8.40 (2.09)	8.00 (2.04)
$p = 0.2$	5.47 (2.72)	4.80 (2.43)	3.70 (1.91)	3.05 (1.52)
$p = 0.4$	4.25 (2.06)	3.25 (1.75)	1.32 (1.27)	1.32 (0.98)
$p = 0.8$	3.27 (1.72)	1.48 (1.34)	1.35 (1.58)	1.25 (1.12)
		$q = 40$		
$p = 0$	21.75 (3.30)	18.75 (2.93)	16.45 (2.43)	15.35 (3.37)
$p = 0.2$	13.18 (3.74)	11.97 (3.48)	12.07 (3.27)	9.98 (3.30)
$p = 0.4$	8.40 (3.49)	8.00 (3.54)	7.12 (4.77)	6.38 (3.73)
$p = 0.8$	4.10 (2.28)	3.83 (2.43)	2.67 (2.77)	2.23 (2.08)

of observations measured under different experimental conditions and then from our perspective can be considered as purely interventional. In the original work of Sachs et al. (2005) the objective was to infer a single DAG, whilst Friedman, Hastie and Tibshirani (2008) used the same dataset to learn a single undirected graph. Moreover, Luo and Zhao (2011) proposed a Bayesian hierarchical model for "causal inference among proteins from interventional data". More recently, Peterson, Stingo and Vannucci (2015) analysed the same dataset from a *multiple graphs* perspective, that is inferring an undirected graph for each experimental condition, allowing for the possibility of shared structural features among graphs.

6.1. *Data set*. The data set, provided as a supplement to Sachs et al. (2005), is based on simultaneous measurements of multiple phosphorylated proteins and phospholipid components in individual primary human immune system cells. Observations are obtained from intracellular multicolor flow cytometry, which allows for simultaneous measurements in individual cells, thus resulting in a large sample of observations. Measurements of $q = 11$ phosphorylated proteins and phospholipids are collected after a series of stimulatory cues and inhibitory interventions obtained from the administration of reagents, each one being the responsible of the perturbation of a signaling node. In addition, some interventions affect receptor enzymes instead of (measured) signalling molecules, which would require the introduction of latent (unobserved) variables. However by removing such observations the dataset becomes purely interventional, because each of the remaining

TABLE 2
Intervention targets and sample sizes for the seven datasets included in the study

k	1	2	3	4	5	6	7
I_k	Akt_1	PKC_1	PIP2	Mek	Akt_2	PKC_2	PKA
$n^{(k)}$	911	723	810	799	848	913	707

5,711 multivariate observations was produced by coercing the value of some variables in the system. We then include in our study seven datasets, each containing observations measured under the same experimental condition. In Table 2 we report, for each dataset $\boldsymbol{Y}^k$, the corresponding intervention target I_k and sample size. Please observe that different reagents can perturb the same signalling node. In particular, for each node Akt and PKC, two interventions were performed, which we denote by adding a subscript (1 or 2).

6.2. *Model searching and results*. We apply OBIES to learn the structure of an $\mathcal{I}$-EG from the dataset $\boldsymbol{Y} = \{\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^7\}$; see Table 2. The corresponding conservative family of intervention targets is

$$\mathcal{I} = \{\text{Akt}_1, \text{PKC}_1, \text{PIP2}, \text{Mek}, \text{Akt}_2, \text{PKC}_2, \text{PKA}\}.$$

We perform structural learning of interventional essential graphs focusing on the $\mathcal{I}$-EG space $\mathcal{S}_{11}^r$ with sparsity parameter $r = 2$, which corresponds to a maximum number of edges of 22. We then run $T = 10^5$ iterations of Algorithm 1 (Supplementary Material, Section 2.2) with a burn-in period of 2×10^4. As the result, we start reporting in Figure 5 the heat map with the marginal posterior probabilities of edge inclusion.

We can use such information to construct the median probability graph model, by including those edges whose probability of inclusion is greater than 0.5. Alternatively, we can compute for a grid of thresholds $k \in (0, 1)$ the expected false discovery rate $\text{FDR}(k)$ as defined in equation (21) and then choose the maximum value of k such that $\text{FDR}(k) < 0.05$. In doing so, we obtain the threshold for edge inclusion $k^* = 0.434$. The corresponding $\mathcal{I}$-EG estimate is then constructed accordingly. We observe that the resulting graph of Figure 6 *is* an $\mathcal{I}$-EG and then no projection to the $\mathcal{I}$-EG space is required; see also Section 4.

Being fully Bayesian, our OBIES method also provides a measure of uncertainty around graphical features of interest. In addition to the probabilities of edge inclusion (Figure 5), we also show in Figure 7 the trace plot of the number of edges in the $\mathcal{I}$-EGs visited by the MCMC, together with the corresponding frequency distribution. Such result can be also used as a graphical diagnostic for the convergence of our algorithm. Other diagnostic tests based on multiple chains are reported in the Supplementary Material.

Finally, we compare OBIES estimate of Figure 6 with the Greedy Interventional Equivalence Search method (Hauser and Bühlmann (2012)). GIES is computed

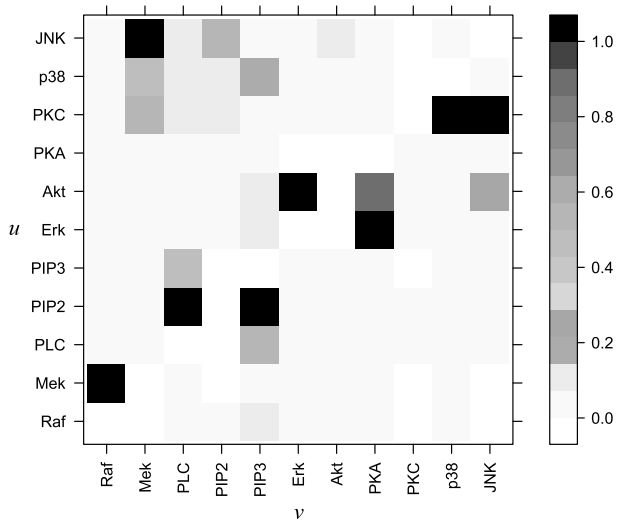

FIG. 5. *Protein-signaling data. Heat map with marginal posterior probabilities of edge inclusion* $p_{u \to v}$.

for three different optimization criteria: the Bayesian Information Criterion (GIES 0) and the Extended Bayesian Information Criterion with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5 and GIES 1, respectively); see also Foygel and Drton (2010). Results with $\mathcal{I}$-EG estimates are reported in Figure 8. It appears that the tuning parameter γ can be used to intensify sparsity of the resulting graph. If compared with the GIES 0, our OBIES estimate of Figure 6 exhibits 10 edges in common.

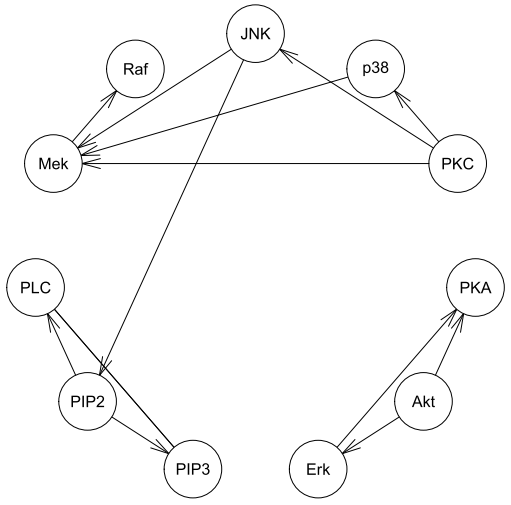

FIG. 6. *Protein-signaling data. OBIES estimate obtained from the* FDR *criterion.*

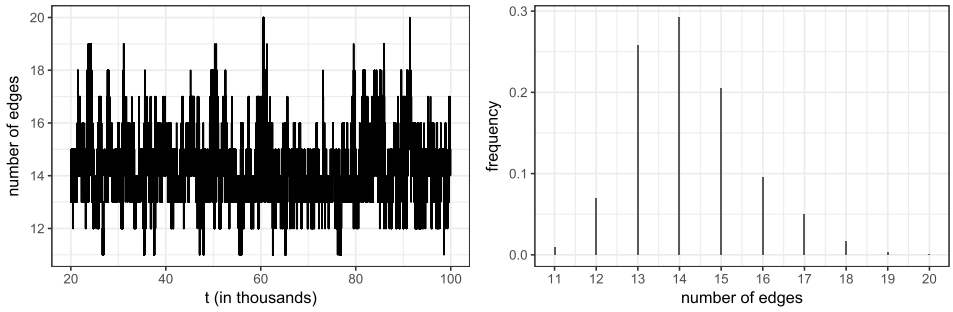

FIG. 7. *Protein-signaling data. Trace plot of the number of edges in the $\mathcal{I}$-EGs visited by the MCMC* (*left panel*) *and corresponding frequency distribution* (*right panel*).

The main difference is the presence in our OBIES estimate of two additional edges, PKC $\rightarrow$ Mek and P38 $\rightarrow$ Mek, the former being better supported (probability of inclusion 68%) while the latter (44%) is only marginal greater than the selected threshold for edge inclusion $k^* = 0.434$.

7. Discussion. In this article we presented an objective Bayes method for model determination of Gaussian DAGs in the presence of (observational and) interventional data. Specifically, we derived a closed formula for the marginal likelihood of an Interventional Essential Graph (I-EG) and then proposed an MCMC scheme for structural learning of Markov equivalence classes of DAGs. We illustrated through simulations that our Objective Bayes Interventional Essential graph Search (OBIES) method is competitive with the benchmark Greedy Interventional Equivalence Search (Hauser and Bühlmann (2012)) if the goal is model selection, that is to produce a single *model estimate*. Moreover, because of its Bayesian nature, OBIES also provides a coherent measure of uncertainty of graphical features of interest, such as the posterior inclusion probability of a specific edge.

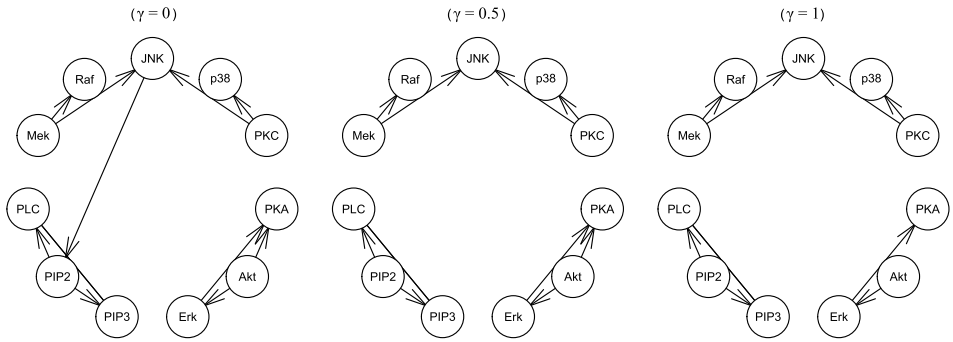

FIG. 8. *Protein-signaling data. Estimated $\mathcal{I}$-EG under GIES* 0 ($\gamma = 0$), *GIES* 0.5 ($\gamma = 0.5$) *and GIES* 1 ($\gamma = 1$).

Our method does not focus on parameter inference or prediction which however are required for the calculation of Bayesian *goodness of fit* measures based on posterior predictive model checks (Gelman, Meng and Stern (1996)). Sampling from the posterior predictive distribution of Gaussian DAG-models is possible using DAG-Wishart priors (Cao, Khare and Ghosh (2019)) coupled with techniques similar to those developed in this paper. We have carried out predictive model checks in this framework obtaining more than satisfactory results which we do not report here for lack of space. Finally, we highlight that our method for model choice, based on marginal likelihoods (equivalently Bayes factors relative to a benchmark model), takes directly care of goodness of fit and complexity (Hoijtink (2013)). In other words, the model we single out as most promising has already passed a selection procedure which accounts for its ability to fit the data.

Randomized intervention experiments can be used to improve the identifiability of the true data generating model (He and Geng (2008)). By enlarging the family of intervention targets, one can in principle reduce each (observational) Markov equivalence class to a single DAG, because all edges that are undirected in the original essential graph become directed. Ideally, one could pursue this goal by means of a small number of interventions selected according to an optimal experimental plan. Such active learning of signaling networks is pursued in Ness et al. (2017) and is currently under investigation by the Authors of this paper.

The protein-signaling dataset in Sachs et al. (2005) was collected under nine distinct experimental conditions, each corresponding to an intervention on *some* variables, either observable or latent. We analysed all the observations corresponding to an intervention on a single observable variable jointly assuming a unique graphical generating structure, namely an interventional Essential Graph. On the other hand, Peterson, Stingo and Vannucci (2015) approached the problem from a *multiple-graphs* perspective. Specifically, they allow each interventional dataset to have its own underlying graphical structure (an undirected graph). Then they analyzed the collection of datasets jointly by relating graphs across groups through a suitable Markov random field prior which encourages common edges to exploit potential shared features, as well as a spike-and-slab prior on the parameters that measure network relatedness. More recently, Tan et al. (2017) applied multiple Gaussian graphical models based on G-Wishart priors to metabolic association networks, using a logistic regression structure to link probability of edge inclusions among graphs.

SUPPLEMENTARY MATERIAL

Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways (DOI: 10.1214/19-AOAS1275SUPP; .pdf). Additional material is provided in the Supplement (Castelletti and Consonni (2019)). This includes some theoretical results on fractional marginal likelihoods for Gaussian models, a detailed treatment of the MCMC algorithm here adopted,

further simulation results and some diagnostics for the convergence of the MCMC on the protein-signaling dataset.

REFERENCES

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997a). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997b). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scand. J. Stat.* **24** 81–102. MR1436624

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Stat.* **28** 33–85. MR1844349

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

CAO, X., KHARE, K. and GHOSH, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Ann. Statist.* **47** 319–348. MR3909935

CASTELLETTI, F. and CONSONNI, G. (2019). Supplement to "Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways." DOI:10.1214/19-AOAS1275SUPP.

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

CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2** 445–498. MR1929415

CONSONNI, G., LA ROCCA, L. and PELUSO, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. *Scand. J. Stat.* **44** 741–764. MR3687971

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

FOYGEL, R. and DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Adv. Neural Inf. Process. Syst.* **23** 2020–2028.

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

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

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

GEISSER, S. and CORNFIELD, J. (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. Ser. B* **25** 368–376. MR0171354

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GILLISPIE, S. B. and PERLMAN, M. D. (2002). The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence* **141** 137–155. MR1935281

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

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

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

HE, Y., JIA, J. and YU, B. (2013). Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *Ann. Statist.* **41** 1742–1779. MR3127848

HOIJTINK, H. (2013). Objective Bayes factors for inequality constrained hypotheses. *Int. Stat. Rev.* **81** 207–229. MR3100657

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

LUO, R. and ZHAO, H. (2011). Bayesian hierarchical modeling for signaling pathway inference from single cell interventional data. *Ann. Appl. Stat.* **5** 725–745. MR2840173

MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133–3164. MR2549555

NESS, R. O., SACHS, K., MALLICK, P. and VITEK, O. (2017). A Bayesian active learning experimental design for inferring signaling networks. In *Research in Computational Molecular Biology*. *Lecture Notes in Computer Science* **10229** 134–156. Springer, Cham. MR3662679

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710. MR1380809

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

PETERS, J. and BÜHLMANN, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101** 219–228. MR3180667

PETERSON, C., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Amer. Statist. Assoc.* **110** 159–174. MR3338494

RICHARDSON, T. and SPIRTES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166

SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. and NOLAN, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** 523–529.

SHOJAIE, A. and MICHAILIDIS, G. (2009). Analysis of gene sets based on the underlying regulatory network. *J. Comput. Biol.* **16** 407–426. MR2487566

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

TAN, L. S. L., JASRA, A., DE IORIO, M. and EBBELS, T. M. D. (2017). Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *Ann. Appl. Stat.* **11** 2222–2251. MR3743295

VERMA, T. and PEARL, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, *UAI* 90 255–270. Elsevier, New York.

DEPARTMENT OF STATISTICAL SCIENCES
UNIVERSITÀ CATTOLICA DEL SACRO CUORE
LARGO GEMELLI 1, MILAN
ITALY
E-MAIL: federico.castelletti@unicatt.it
guido.consonni@unicatt.it