Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Bayesian learning of multiple directed networks from observational data

Federico Castelletti[1] | Luca La Rocca[2] | Stefano Peluso[1] | Francesco C. Stingo*[3] | Guido Consonni[1]

[1]Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy
[2]Department of Physics, Informatics and Mathematics, Università degli Studi di Modena e Reggio Emilia, Modena, Italy
[3]Department of Statistics, Computer Science, Applications "G. Parenti", Università degli Studi di Firenze, Florence, Italy

Correspondence
*Francesco C. Stingo, Department of Statistics, Computer Science, Applications "G. Parenti", Viale Morgagni 65, 50134 Florence, Italy. Email: francescoclaudio.stingo@unifi.it

Summary

Graphical modeling represents an established methodology for identifying complex dependencies in biological networks, as exemplified in the study of co-expression, gene regulatory, and protein interaction networks. The available observations often exhibit an intrinsic heterogeneity, which impacts on the network structure through the modification of specific pathways for distinct groups, such as disease subtypes. We propose to infer the resulting multiple graphs jointly in order to benefit from potential similarities across groups; on the other hand our modeling framework is able to accommodate group idiosyncrasies. We consider Directed Acyclic Graphs (DAGs) as network structures, and develop a Bayesian method for structural learning of multiple DAGs. We explicitly account for Markov equivalence of DAGs, and propose a suitable prior on the collection of graph spaces that induces selective borrowing strength across groups. The resulting inference allows in particular to compute the posterior probability of edge inclusion, a useful summary for representing flow directions within the network. Finally, we detail a simulation study addressing the comparative performance of our method, and present an analysis of two protein networks together with a substantive interpretation of our findings.

KEYWORDS:
essential graph, Markov equivalence, Markov random field, objective Bayes, protein network

1 | INTRODUCTION

Understanding the complex functions of the genes, proteins and other aspects of the genome is at the foundation of genomic medicine [1;2]. Additional knowledge on these mechanisms can aid the development of novel treatment strategies for the underlying disease. A major role in this effort is played by flexible and efficient quantitative models for the analysis of dependence structures of omics variables. Graphical models have been widely applied in genomics and proteomics to infer various types of networks, including co-expression, gene regulatory, and protein interaction networks [3;4;5].

While the standard setting of Gaussian graphical modeling assumes that each observation is drawn from the same population, it is often the case that data originate from several distinct groups. For instance, gene expression measurements can relate to cancer tissue samples as well as normal tissue samples. Neither the choice of assuming that the two graphs are the same, nor that they are distinct and hence should be analyzed separately, is satisfactory, as the former fails to account for differences which may be of interest, while the latter does not allow to take into account similarities. For example, in genomics it is fundamental to understand whether disease subtypes can be characterized by alterations in key signaling pathways. To address this issue, a few

proposals for the simultaneous analysis of multiple graphs have been presented. With reference to undirected graphs, Danaher et al[6] develop the joint graphical lasso for estimating the inverse covariances across multiple groups. An interesting feature of their approach is that it does not require sparsity of the individual covariance matrices; this is also the perspective of Zhao et al[7], who propose direct estimation of the difference between two networks. The joint lasso is used by Pircalabelu et al[8] to analyze brain activity measurements at voxel level based on magnetic resonance imaging.

Turning to Bayesian approaches, Peterson et al[9] propose a method for structural learning of multiple networks, whose building block is represented by a Markov random field prior on the space of graph models that encourages common structures. Specifically, their prior favors the inclusion of an edge in the graph for a particular group if the same edge is included in the graphs of related sample groups. Additionally, their method is able to learn which groups have a shared structure through parameters that measure network relatedness. In this way information is shared among sample groups only when appropriate. Tan et al[10] consider metabolic association networks. Their prior on graph structures is an extension of the multiplicative (or Chung-Lu random graph) model to multiple Gaussian graphical models, linking the probability of edge inclusion through logistic regression. Jalali et al[11] develop a scalable approach to jointly estimate multiple related Gaussian graphical models that exhibit complex edge connectivity patterns across models for different subsets of edges. To achieve this goal, they introduce a novel subset-specific-prior that for each edge aims to select the subset of models it is common to. Williams et al[12] discuss multiple graphical models with the aim of detecting differences or demonstrate replicability. To this end, they introduce two methods for comparing networks; one is based on the posterior predictive distribution, with Kullback-Leibler divergence as the discrepancy measure; the second approach instead relies on the more traditional Bayes factor.

As exemplified above, most papers assume that similarities and differences between networks are driven by individual *edges*. This approach however might not be appropriate in some cases, as discussed in Mohan et al[13]. They suggest instead to to take a *node-based* approach to assess shared network structures, and identify two instances: highly-connected hub-nodes in all network groups; and perturbed nodes exhibiting different connectivity structures across groups.

The great majority of works on multiple graphs deal with the undirected case, as in the references above. However, Directed Acyclic Graphs (DAGs) are often preferred in genetic analyses where directed pathways are of particular interest; additionally they represent the natural graphical framework to perform causal reasoning[14]. Motivated by reverse phase protein array data from a study on acute myeloid leukemia, Yajima et al[15] discuss a modeling approach based on Gaussian DAGs to contrast refractory versus relapsed patients targeting specific biological pathways. Mitra et al[16] also deal with two group structures, and construct a prior which addresses group heterogeneity while allowing for the possibility of borrowing strength. Oates et al[17] extend recent developments in exact estimation of DAGs using integer linear programming to joint estimation over multiple DAGs, without requiring that the vertices in each DAG share a common ordering.

In this paper we also deal with DAGs. We start however from the fact that Markov-equivalent DAGs, namely those which share the same conditional independencies, cannot be distinguished through observational data alone[18]. This leads us to consider a representative for each class, named the essential graph. The main contribution of this manuscript is threefold: i) we develop a methodology for Bayesian structural learning over multiple essential graphs, one for each of several groups; ii) we propose a novel prior for multiple essential graphs based on the graphs' skeletons; iii) we propose a method that identifies the direction of the associations in the (multiple) biological networks, when possible. Note that the last feature is of practical relevance in follow-up experiments, since, for example, the proposed method can suggest which one of a pair of connected genes should be knocked down. We follow an objective Bayes approach with regard to parameter priors, whereas a Markov random field prior is placed on the space of multiple essential graphs.

The rest of this paper is organized as follows. Section 2 provides the main concepts and terminology for graphical models and model selection. Section 3 contains our model formulation in terms of likelihood, parameter priors and prior on the graph space. Section 4 deals with the development of a computational algorithm for posterior sampling and with posterior summaries, while Section 5 presents a detailed simulation study and compares our results with those produced by a few current competing methods. Section 6 is devoted to the analysis of two real datasets, the protein signaling data and the leukemia data. Finally, Section 7 contains a brief summary and discussion. All codes are written in R[19] and are available upon request to the Authors.

2 | BACKGROUND

In this section we provide the background material we rely on for further developments. More information on graphs and graphical models and on objective Bayesian model selection is provided by Lauritzen[20] and Berger and Pericchi[21], respectively.

2.1 | Simple graphs

We denote a graph by $\mathcal{G} = (V, E)$, where $V = \{1, \dots, q\}$ is a set of vertices (or nodes) and $E \subseteq V \times V$ is a set of edges (or arcs) such that $(u, u) \notin E$ for all $u \in V$ (no loops). Note that, by ruling out loops and parallel edges, which would require E to be a multiset, we are assuming that $\mathcal{G}$ is *simple*. Figure 1 depicts three simple graphs with $q = 6$ nodes, which we use to illustrate the background material in this section. The idea is that arcs represent first-hand connections between nodes, which will give rise to second-hand connections, as described by paths and cycles (defined below).

If $(u, v) \in E$ but $(v, u) \notin E$ we say that $\mathcal{G}$ contains the directed edge (or arrow) $u \to v$. If instead $(u, v) \in E$ and $(v, u) \in E$ we say that $\mathcal{G}$ contains the undirected edge (or line) $u - v$. For instance, in Figure 1, the left graph contains the arrow $4 \to 5$, while the middle graph contains the line $4 - 5$. In both cases there is a first-hand connection between 4 and 5, but the two connections are of different type. We say that two vertices u, v are *adjacent* in $\mathcal{G}$ if there is any such first-hand connection between them, be it directed or undirected. Specifically, if $u - v$ is in $\mathcal{G}$ we say that u and v are *neighbors*, while we say that u is a *parent* of v if $u \to v$ is in $\mathcal{G}$. For instance, in the right graph of Figure 1 node 3 is a parent of node 4 and a neighbor of node 1, while in the middle graph of the same figure node 3 is a neighbor of both node 1 and node 4. We denote the parent set of v by $\mathrm{pa}_{\mathcal{G}}(v)$, so that $\mathrm{pa}_{\mathcal{G}}(6) = \{4, 5\}$ in the left graph of Figure 1, whereas $\mathrm{pa}_{\mathcal{G}}(6) = \{4\}$ in the right graph of the same figure.

A sequence of distinct vertices $v_0, v_1, \dots, v_k$ in $\mathcal{G}$ is a *path* (of length k) from v_0 to v_k if $\mathcal{G}$ contains $v_{j-1} - v_j$ or $v_{j-1} \to v_j$ for all $j = 1, \dots, k$. A *cycle* is defined in the same way as a path, but with $v_0 = v_k$. A path is undirected if all its edges are undirected; a cycle is directed if it contains at least one directed edge. There are no directed cycles in Figure 1, but its middle graph contains the cycles $1, 2, 4, 3, 1$ and $4, 5, 6, 4$, whose lengths are 4 and 3, respectively. There are undirected paths in the middle and right graphs of Figure 1: for instance, non-adjacent vertices 2 and 3 are joined by the undirected path $2, 1, 3$, which represents a second-hand connection between 2 and 3. We will use directed cycles and undirected paths to define and represent chain graphs (introduced below) but we first present a few additional concepts.

Let A be a non-empty subset of V. We denote by $\mathcal{G}_A = (A, E_A)$ the *subgraph* of $\mathcal{G}$ induced by A, whose edge set is $E_A = E \cap (A \times A)$. We say that a (sub)graph is *complete* if all its pairs of vertices are adjacent. If $\mathcal{G}_A$ is complete, we also say that A is complete (in $\mathcal{G}$). A complete subset of V that is maximal with respect to inclusion is called a *clique* of $\mathcal{G}$. In the middle graph of Figure 1, both $\{5, 6\}$ and $\{4, 5, 6\}$ are complete, whereas $\{1, 2, 3\}$ is not, because 2 and 3 are not adjacent; it can be seen that $\{4, 5, 6\}$ is a clique, while $\{5, 6\}$ is not, because it is strictly included in $\{4, 5, 6\}$. A *flag* is any subgraph of the form $u \to v - w$, while an *immorality* (or v-structure) is any subgraph of the form $u \to v \leftarrow w$; note that in both cases u and w are not adjacent. There are no flags in Figure 1, but the left and right graphs of the figure both contain the immorality $2 \to 4 \leftarrow 3$.

A graph is called *directed* (*undirected*) if it contains only directed (undirected) edges. The undirected graph obtained from $\mathcal{G}$ by removing the orientation of all its edges (replacing all its arrows with lines) is called the *skeleton* of $\mathcal{G}$. In Figure 1, the left graph is a directed graph and the middle graph is its skeleton (as well as the skeleton of the right graph). The skeleton of a graph preserves its first-hand connections, but drops their directions. A special class of undirected graphs is represented by *decomposable* graphs, also called *chordal* or *triangulated* graphs [20, Ch. 2]: an undirected graph is decomposable if every cycle of length $k \geq 4$ has a *chord*, that is two non-consecutive adjacent vertices. The middle graph of Figure 1 is not decomposable, because its cycle $1, 2, 4, 3, 1$ has length four but no chords; in the right graph of the same figure the subgraphs induced by $\{1, 2, 3\}$ and $\{5, 6\}$ are trivially decomposable (having no cycles). A special class of directed graphs is formed by *acyclic* directed graphs: a directed graph with no cycles, such as the left graph of Figure 1, is called a *Directed Acyclic Graph* (DAG) and it is typically denoted by $\mathcal{D}$ (in place of $\mathcal{G}$).

Both undirected graphs and DAGs are special cases of *chain graphs*, defined as graphs with no directed cycles (possibly containing both directed and undirected edges). All graphs in Figure 1 are chain graphs. For a chain graph $\mathcal{G}$ we call *chain component* a maximal (with respect to inclusion) set of nodes $\tau \subseteq V$ such that all its pairs of (distinct) nodes are joined by an undirected path. The set of all chain components of a chain graph is denoted by $\mathcal{T}$; it forms a partition of the vertex set. As a matter of fact, a chain graph can be seen as a DAG with vertex set $\mathcal{T}$ [22]. In particular, an undirected graph is a chain graph with a single chain component, while a DAG is a chain graph with singleton chain components.

Let $\mathcal{G}$ be a decomposable graph with cliques $C_1, \dots, C_m$ and define recursively the corresponding *histories*, *residuals* and *separators*, as $H_j = H_{j-1} \cup C_j$, $R_j = C_j \setminus H_{j-1}$ and $S_j = C_j \cap H_{j-1}$, for $j = 2, \dots, m$, from the base cases $H_1 = R_1 = C_1$ and $S_1 = \emptyset$. Since $\mathcal{G}$ is decomposable, it is always possible [20, Prop. 2.17] to order its cliques in such a way that $C_1, \dots, C_m$ is a *perfect sequence* [20, p. 14]. Given that each S_j is necessarily complete (being a clique subset) this amounts to saying that $C_1, \dots, C_m$ satisfies the *running intersection property*: for all $j > 1$, there is $i < j$ such that $S_j \subseteq C_i$. Therefore, if we number the vertices in V, based on a perfect sequence of cliques, in such a way that $i < j$ implies lower numbers in R_i than in R_j, and then we direct the edges in E from lower numbered vertices to higher numbered vertices, we obtain a directed graph $\mathcal{G}^{<}$ with $\mathcal{G}$ as skeleton

and no immoralities (a *perfect directed version* of $\mathcal{G}$). For instance, in the right graph in Figure 1, for the subgraph induced by $\{1,2,3\}$, we have $m = 2$, $C_1 = \{1,2\}$ and $C_2 = \{1,3\}$, therefore $H_1 = R_1 = \{1,2\}$ and $S_1 = \emptyset$, while $H_2 = \{1,2,3\}$, $R_2 = \{3\}$ and $S_2 = \{1\}$. In the same graph, for the subgraph induced by $\{5,6\}$, we have $m = 1$ and $C_1 = \{5,6\}$, therefore $H_1 = R_1 = \{5,6\}$ and $S_1 = \emptyset$. The left graph in Figure 1 can thus be seen (chain component by chain component) as a perfect directed version of the right one.

2.2 | Directed graphical models

Let $Y_1, \ldots, Y_q$ be q random variables, such as gene or protein expressions, whose joint probability density function is $f(\boldsymbol{y}) = f(y_1, \ldots, y_q)$, where $\boldsymbol{y} = (y_1, \ldots, y_q)^\top$ is a column vector. Consider a DAG $\mathcal{D} = (V, E)$ with $V = \{1, \ldots, q\}$. To each vertex j in V we associate a variable Y_j. We say that $f(\boldsymbol{y})$ factorizes according to $\mathcal{D}$, or is Markov with respect to $\mathcal{D}$, if

$$f(\boldsymbol{y}) = \prod_{j \in V} f(y_j \mid \boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)}), \tag{2.1}$$

where $f(y_j \mid \boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)})$ is the conditional density function of Y_j given the subvector of $\boldsymbol{y}$ corresponding to the vertices in $\mathrm{pa}_{\mathcal{D}}(j)$, which we denote by $\boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)}$. Equation (2.1) constrains $f(\boldsymbol{y})$ so that the structure of $\mathcal{D}$ determines conditional independencies among the random variables. Specifically, assuming faithfulness[23], all independencies can be read off from $\mathcal{D}$ using the notion of *d-separation*[14, Ch. 1] or the moral graph approach[20, Ch. 3]. Moreover, if the joint distribution of $Y_1, \ldots, Y_q$ is multivariate Gaussian, such conditional independencies correspond to constraints on the covariance matrix of $Y_1, \ldots, Y_q$.

It is well known that distinct DAGs can encode the same conditional independencies. In such case they are called *Markov equivalent*. Verma and Pearl[24] showed that two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent if and only if they have the same skeleton and the same v-structures. If the data are only observational (i.e. passively observed) as in this paper, Markov equivalent DAGs cannot be distinguished. We can then partition the DAG space into Markov equivalence classes, each represented by a special chain graph called *Essential Graph* (EG)[18] or *Completed Partially Directed Acyclic Graph* (CPDAG)[25]. The representative EG of a Markov equivalence class is obtained as the union (with respect to the edge sets) of the DAGs contained in the class. Such union implies that if an edge occurs with different orientations inside the class, e.g. $u \to v$ and $u \leftarrow v$, then the corresponding EG will contain the undirected edge $u - v$. According to an important result[18, Theorem 4.1], EGs are characterized as chain graphs with decomposable chain components, no flags, and strongly protected arrows[18, Definition 3.3]. A different formulation of the same result is given by Roverato[26, Theorem 13] in the context of a unified characterization of representative chain graphs. If $\mathcal{G}$ is an EG with set of chain components $\mathcal{T}$, the probability density function of $Y_1, \ldots, Y_q$ constrained by any DAG in the class represented by $\mathcal{G}$ can be written as

$$f_{\mathcal{G}}(\boldsymbol{y}) = \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}}(\boldsymbol{y}_\tau \mid \boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}), \tag{2.2}$$

where $\boldsymbol{y}_\tau$ denotes the subvector of $\boldsymbol{y} = (y_1, \ldots, y_q)$ indexed by τ[27] and $\mathrm{pa}_{\mathcal{G}}(\tau) = \bigcup_{v \in \tau} \mathrm{pa}_{\mathcal{G}}(v)$ is the parent set of τ. It should be noted that all vertices in any given chain component of an EG have exactly the same parents, because the EG has no flags.

2.3 | Model comparison through marginal likelihoods

Let $\mathcal{M}_1, \ldots, \mathcal{M}_r$ be r competing statistical models for the $n \times q$ data matrix $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$. Under model $\mathcal{M}_h$, conditionally on a model specific parameter $\boldsymbol{\theta}_h \in \boldsymbol{\Theta}_h$, the vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are assumed to be independent and identically distributed observations from a q-dimensional distribution with probability density function $f_{\mathcal{M}_h}(\boldsymbol{y}|\boldsymbol{\theta}_h)$. Hence, the likelihood under model $\mathcal{M}_h$ is given by $f_{\mathcal{M}_h}(\boldsymbol{Y}|\boldsymbol{\theta}_h) = \prod_{i=1}^n f_{\mathcal{M}_h}(\boldsymbol{y}_i|\boldsymbol{\theta}_h)$. The goal is to compare the r models based on the support they receive from the data, and eventually to select one. If $\mathcal{M}_h$ prescribes that $f_{\mathcal{M}_h}(\cdot|\boldsymbol{\theta}_h)$ is Markov with repect to a graph $\mathcal{G}_h$, $\mathcal{M}_h$ is called a graphical model. In this case model comparison amounts to learning a graphical structure from the data.

We follow a Bayesian approach and introduce a parameter prior for all model specific parameters. Let $p_{\mathcal{M}_h}(\boldsymbol{\theta}_h)$ be the prior probability density of $\boldsymbol{\theta}_h$ under $\mathcal{M}_h$. The marginal likelihood of $\mathcal{M}_h$ is then defined as $m_{\mathcal{M}_h}(\boldsymbol{Y}) = \int f_{\mathcal{M}_h}(\boldsymbol{Y}|\boldsymbol{\theta}_h) p_{\mathcal{M}_h}(\boldsymbol{\theta}_h) d\boldsymbol{\theta}_h$ and provides a measure of support for $\mathcal{M}_h$ (based on $\boldsymbol{Y}$). Indeed, if prior model probabilities are introduced, posterior model probabilities can be obtained from the marginal likelihoods of all models. Specifically, if $\Pr(\mathcal{M}_h)$ is the prior probability of $\mathcal{M}_h$, the posterior probability of $\mathcal{M}_h$ can be obtained as $\Pr(\mathcal{M}_h|\boldsymbol{Y}) = m_{\mathcal{M}_h}(\boldsymbol{Y})\Pr(\mathcal{M}_h) / \sum_{k=1}^r m_{\mathcal{M}_k}(\boldsymbol{Y})\Pr(\mathcal{M}_k)$. Then, a single model supported by the data can be selected, if needed, based on a summary of the posterior distribution on the model space. Alternatively, strategies based on model averaging can be easily implemented.

In lack of substantive prior information, which is typically the case when different large graphical structures are to be compared, an objective approach to the specification of $\Pr(\mathcal{M}_h)$ and, particularly, $p_{\mathcal{M}_h}(\boldsymbol{\theta}_h)$ is recommended; for example, in many applications it is typically difficult to specify an informative prior on the strength of the connections in the network. The choice of the latter is especially critical, because default estimation priors are typically improper and their undefined normalizing constants will make marginal likelihoods, hence posterior model probabilities, meaningless. In this paper, we specify parameter priors following Consonni et al[28], who solve this problem using priors based on the *fractional Bayes factor*[29] and compute marginal likelihoods for directed graphical models following the procedure presented by Geiger and Heckerman[30]. The choice of $\Pr(\mathcal{M}_h)$ is also important, and will be discussed in Section 3 in the context of EG-models.

3 | MODEL FORMULATION

In this section we introduce our full Bayesian model for the data of interest, describing its likelihood, parameter prior and prior on model space.

3.1 | Likelihood

Consider observations from K groups, such as the disease subtype or disease stage. For each $k = 1, \dots, K$, let $\boldsymbol{Y}_{[k]}$ be the $n_k \times q$ data matrix for group k having rows $\boldsymbol{y}_{[k]1}^\top, \dots, \boldsymbol{y}_{[k]n_k}^\top$, where n_k is the sample size of group k. Note that we observe the same set of random variables for all groups. Within group k, given $\boldsymbol{\mu}_{[k]}$ (a vector) and $\boldsymbol{\Sigma}_{[k]}$ (a symmetric positive definite matrix), we assume that $\boldsymbol{y}_{[k]1}, \dots, \boldsymbol{y}_{[k]n_k}$ are independent and identically distributed random vectors following a multivariate (q-dimensional) Gaussian distribution with mean $\boldsymbol{\mu}_{[k]}$ and covariance matrix $\boldsymbol{\Sigma}_{[k]}$. The matrix $\boldsymbol{\Sigma}_{[k]}$ will be constrained by a group specific EG $\mathcal{G}_k = (V, E_k)$, representing the conditional dependence relationships existing in the group between the q observables, where $V = \{1, \dots, q\}$ as in Section 2. We also assume independence across groups, conditionally on the model parameters, which results in the sampling probability density

$$f(\boldsymbol{Y}_{[1:K]} | \boldsymbol{\mu}_{[1:K]}, \boldsymbol{\Sigma}_{[1:K]}, \mathcal{G}_{[1:K]}) = \prod_{k=1}^{K} \prod_{i=1}^{n_k} f_{\mathcal{G}_k}(\boldsymbol{y}_{[k]i} | \boldsymbol{\mu}_{[k]}, \boldsymbol{\Sigma}_{[k]}) \tag{3.1}$$

for the full data matrix $\boldsymbol{Y}_{[1:K]}$ obtained by stacking the group data matrices $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_K$ one upon the other, where $\mathcal{G}_{[1:K]}$ is the collection of EG's, while $\boldsymbol{\mu}_{[1:K]}$ and $\boldsymbol{\Sigma}_{[1:K]}$ collect all mean and covariance parameters and

$$f_{\mathcal{G}_k}(\boldsymbol{y}_{[k]i} | \boldsymbol{\mu}_{[k]}, \boldsymbol{\Sigma}_{[k]}) = (2\pi)^{-q/2} (\det(\boldsymbol{\Sigma}_{[k]}))^{-1/2} \exp\{-(\boldsymbol{y}_{[k]i} - \boldsymbol{\mu}_{[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\boldsymbol{y}_{[k]i} - \boldsymbol{\mu}_{[k]})/2\}. \tag{3.2}$$

Since $\boldsymbol{\Sigma}_{[k]}$ is constrained by $\mathcal{G}_k$, factorization (2.2) holds. Specifically, we have

$$f_{\mathcal{G}_k}(\boldsymbol{y}_{[k]i} | \boldsymbol{\mu}_{[k]}, \boldsymbol{\Sigma}_{[k]}) = \prod_{\tau \in \mathcal{T}_k} f_{\mathcal{G}_k}(\boldsymbol{y}_{[k]i\tau} | \boldsymbol{y}_{[k]i\mathrm{pa}_k(\tau)}, \boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]}), \tag{3.3}$$

where $\boldsymbol{B}_{[k,\tau]}$ is an unconstrained $(|\mathrm{pa}_k(\tau)| + 1) \times |\tau|$ matrix, $\boldsymbol{\Omega}_{[k,\tau]}$ is a positive definite $|\tau| \times |\tau|$ matrix constrained by $\mathcal{G}_{k\tau}$, and $f_{\mathcal{G}_k}(\boldsymbol{y}_{[k]i\tau} | \boldsymbol{y}_{[k]i\mathrm{pa}_k(\tau)}, \boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]})$ is the $|\tau|$-dimensional Gaussian density with mean $\boldsymbol{B}_{[k,\tau]}^\top (1, \boldsymbol{y}_{[k]i\mathrm{pa}_k(\tau)})$ and covariance matrix $\boldsymbol{\Omega}_{[k,\tau]}^{-1}$; note the shorthand $\mathrm{pa}_k(\tau)$ for $\mathrm{pa}_{\mathcal{G}_k}(\tau)$. Plugging equation (3.3) in equation (3.1), we finally obtain

$$f(\boldsymbol{Y}_{[1:K]} | \boldsymbol{\mu}_{[1:K]}, \boldsymbol{\Sigma}_{[1:K]}, \mathcal{G}_{[1:K]}) = \prod_{k=1}^{K} \prod_{\tau \in \mathcal{T}_k} f_{\mathcal{G}_k}(\boldsymbol{Y}_{[k]\tau} | \boldsymbol{Y}_{[k]\mathrm{pa}_k(\tau)}, \boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]}), \tag{3.4}$$

where $\boldsymbol{Y}_{[k]\tau} = (\boldsymbol{y}_{[k]1\tau}, \dots, \boldsymbol{y}_{[k]n_k\tau})^\top$ and $\boldsymbol{Y}_{[k]\mathrm{pa}_k(\tau)} = (\boldsymbol{y}_{[k]1\mathrm{pa}_k(\tau)}, \dots, \boldsymbol{y}_{[k]n_k\mathrm{pa}_k(\tau)})^\top$ are component specific data matrices, and $f_{\mathcal{G}_k}(\boldsymbol{Y}_{[k]\tau} | \boldsymbol{Y}_{[k]\mathrm{pa}_k(\tau)}, \boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]}) = \prod_{i=1}^{n_k} f_{\mathcal{G}_k}(\boldsymbol{y}_{[k]i\tau} | \boldsymbol{y}_{[k]i\mathrm{pa}_k(\tau)}, \boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]})$.

3.2 | Parameter priors

We assume that model parameters are *a priori* independent across groups and chain components, conditionally on the group specific EGs, so that we can write the marginal density of the data as

$$f(\boldsymbol{Y}_{[1:K]} | \mathcal{G}_{[1:K]}) = \prod_{k=1}^{K} \prod_{\tau \in \mathcal{T}_k} m_{\mathcal{G}_k}(\boldsymbol{Y}_{[k]\tau} | \boldsymbol{Y}_{[k]\mathrm{pa}_k(\tau)}), \tag{3.5}$$

where $m_{\mathcal{G}_k}(\boldsymbol{Y}_{[k]\tau}|\boldsymbol{Y}_{[k]\mathrm{pa}_k(\tau)})$ is the conditional marginal likelihood of $\mathcal{G}_{k\tau}$, representing the contribution of the chain component τ to the marginal likelihood of $\mathcal{G}_k$:

$$m_{\mathcal{G}_k}(\boldsymbol{Y}_{[k]\tau}|\boldsymbol{Y}_{[k]\mathrm{pa}_k(\tau)}) = \int f_{\mathcal{G}_k}(\boldsymbol{Y}_{[k]\tau}|\boldsymbol{Y}_{[k]\mathrm{pa}_k(\tau)}, \boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]}) p_{\mathcal{G}_k}(\boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]}) d(\boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]}). \tag{3.6}$$

We specify the parameter prior $p_{\mathcal{G}_k}(\boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]})$, for all $k = 1, \dots, K$ and $\tau \in \mathcal{T}_k$, as recommended by Consonni et al[28]. In this way, given that $\mathcal{G}_{k\tau}$ is decomposable, we obtain

$$m_{\mathcal{G}_k}(\boldsymbol{Y}_{[k]\tau}|\boldsymbol{Y}_{[k]\mathrm{pa}_k(\tau)}) = \frac{\prod_{C\in\mathcal{C}_{k,\tau}} m(\boldsymbol{Y}_{[k]C})}{\prod_{S\in\mathcal{S}_{k,\tau}} m(\boldsymbol{Y}_{[k]S})}, \tag{3.7}$$

where $\mathcal{C}_{k,\tau}$ is a perfect sequence of cliques for $\mathcal{G}_{k\tau}$, $\mathcal{S}_{k,\tau}$ is the corresponding sequence of separators, and, following Castelletti et al[31], we can compute

$$m(\boldsymbol{Y}_{[k]\alpha}) = \frac{\Gamma_{|\alpha|}\left(\frac{|\alpha|+n_k-|\mathrm{pa}_k(\alpha)|-2}{2}\right)}{\pi^{\frac{|\alpha|(n_k-|\mathrm{pa}_k(\alpha)|-2)}{2}}\Gamma_{|\alpha|}\left(\frac{|\alpha|}{2}\right)}\left(\frac{|\mathrm{pa}_k(\alpha)|+2}{n_k}\right)^{\frac{|\alpha|(|\alpha|+|\mathrm{pa}_k(\alpha)|+1)}{2}}\det\left(\widehat{\boldsymbol{E}}_{[k]\alpha}^\top\widehat{\boldsymbol{E}}_{[k]\alpha}\right)^{-\frac{(n_k-|\mathrm{pa}_k(\alpha)|-2)}{2}}, \tag{3.8}$$

for all nonempty $\alpha \subseteq \tau$, from the partial residual matrix $\widehat{\boldsymbol{E}}_{[k]\alpha} = \boldsymbol{Y}_{[k]\alpha} - \boldsymbol{X}_{[k,\tau]}\widehat{\boldsymbol{B}}_{[k,\tau]\alpha}$ determined by the design matrix $\boldsymbol{X}_{[k,\tau]} = (\boldsymbol{1}_{n_k}, \boldsymbol{Y}_{[k]\mathrm{pa}_k(\tau)})$ and the corresponding ordinary least squares estimator $\widehat{\boldsymbol{B}}_{[k,\tau]\alpha} = (\boldsymbol{X}_{[k,\tau]}^\top\boldsymbol{X}_{[k,\tau]})^{-1}\boldsymbol{X}_{[k,\tau]}^\top\boldsymbol{Y}_{[k]\alpha}$ for the partial response $\boldsymbol{Y}_{[k]\alpha}$. Note that $\Gamma_{|\alpha|}$ in (3.8) denotes the multivariate gamma function, which is defined by $\Gamma_q(x/2) = \pi^{q(q-1)/4}\prod_{j=1}^q\Gamma((x+1-j)/2)$, where Γ is the ordinary (univariate) Euler's gamma function.

3.3 | Linking essential graphs with a Markov prior

We here present a Bayesian hierarchical model that links the skeletons of the group specific EGs. We expect the different EGs to have similar skeletons, but we will be able to learn from the data whether this is the case or rather skeletons should be considered independently. We find it useful to represent the skeleton $\widetilde{\mathcal{G}}_k$ through the upper triangular part of its adjacency matrix, denoted by $\widetilde{\boldsymbol{G}}_{[k]}$ and consisting of $q(q-1)/2$ elements. Specifically, to encourage similar skeleton structures, we assign a Markov Random Field (MRF) prior to the elements of $\widetilde{\boldsymbol{G}}_{[1]}, \dots, \widetilde{\boldsymbol{G}}_{[K]}$. More precisely, we let the binary vectors of edge inclusion indicators $\boldsymbol{s}_{ij} = (\widetilde{g}_{[1]ij}, \dots, \widetilde{g}_{[K]ij})$, for $1 \le i < j \le q$, have prior probability mass function

$$p(\boldsymbol{s}_{ij}|\nu_{ij}, \boldsymbol{\Theta}) = C(\nu_{ij}, \boldsymbol{\Theta})^{-1}\exp(\nu_{ij}\boldsymbol{1}_K^\top\boldsymbol{s}_{ij} + \boldsymbol{s}_{ij}^\top\boldsymbol{\Theta}\boldsymbol{s}_{ij}), \tag{3.9}$$

where ν_{ij} is a sparsity parameter specific to the set of edges indicated by $\boldsymbol{s}_{ij}$, $\boldsymbol{\Theta}$ is a $K \times K$ symmetric matrix denoting pairwise associations, $\boldsymbol{1}_K$ is the unit vector of dimension K, and

$$C(\nu_{ij}, \boldsymbol{\Theta}) = \sum_{\boldsymbol{s}_{ij}\in\{0,1\}^K}\exp(\nu_{ij}\boldsymbol{1}_K^\top\boldsymbol{s}_{ij} + \boldsymbol{s}_{ij}^\top\boldsymbol{\Theta}\boldsymbol{s}_{ij}) \tag{3.10}$$

is the normalizing constant, which can be analytically calculated if the number of groups K is reasonably small. Each off-diagonal element θ_{km} of $\boldsymbol{\Theta}$ allows us to create dependency between sample groups k and m: $\theta_{km} = 0$ implies that groups m and k are conditionally independent, given the other groups, whereas non-zero values in $\boldsymbol{\Theta}$ define a measure of relative skeleton similarity across groups. Then, conditionally on $\boldsymbol{\nu}$ (upper triangular matrix with elements ν_{ij}) and $\boldsymbol{\Theta}$, we assume the vectors $\boldsymbol{s}_{ij}$ independent over $1 \le i < j \le q$, thus obtaining

$$p(\widetilde{\boldsymbol{G}}_{[1]}, \dots, \widetilde{\boldsymbol{G}}_{[K]}|\boldsymbol{\nu}, \boldsymbol{\Theta}) = \prod_{i<j} p(\mathbf{s}_{ij}|\nu_{ij}, \boldsymbol{\Theta}) \tag{3.11}$$

for the joint prior on the skeletons. Following Castelletti et al[31] for single (not multiple) Bayesian inference on EGs, among others in the literature, we impose a prior on $\mathcal{G}_k$, $k = 1, \dots, K$, solely dependent on specific features of the graph. We then choose $p(\mathcal{G}_1, \dots, \mathcal{G}_K|\boldsymbol{\nu}, \boldsymbol{\Theta}) \propto p(\widetilde{\boldsymbol{G}}_{[1]}, \dots, \widetilde{\boldsymbol{G}}_{[K]}|\boldsymbol{\nu}, \boldsymbol{\Theta})$, assigning equal prior probability to all EGs with a given skeleton. Alternative priors, specifically targeted to EGs, to our knowledge are not available in the literature, and beyond the scope of the present paper.

Under the above described prior, the conditional probability of inclusion of edge (i, j) in $\widetilde{\mathcal{G}}_k$, given the inclusion of that edge in all remaining skeletons, can be expressed as

$$p(s_{ijk}|(s_{ijm})_{m\neq k}, \nu_{ij}, \boldsymbol{\Theta}) = \frac{\exp(s_{ijk}\{\nu_{ij} + 2\sum_{m\neq k}\theta_{km}s_{ijm}\})}{1 + \exp(\nu_{ij} + 2\sum_{m\neq k}\theta_{km}s_{ijm})}, \tag{3.12}$$

which shows that the parameter θ_{km} indicates pairwise similarity of the two skeletons $\widetilde{\mathcal{G}}_k$ and $\widetilde{\mathcal{G}}_m$. Then, following Peterson et al[9], we select our prior as a spike and slab prior on the off-diagonal entries θ_{km}. The non-zero component is set to have a positive support, because we want to encourage only skeleton similarity between two related groups. In detail, we specify the following prior:

$$p(\theta_{km}|\gamma_{km}) = (1-\gamma_{km})\delta_0(\theta_{km}) + \gamma_{km}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta_{km}^{\alpha-1}e^{-\beta\theta_{km}}, \tag{3.13}$$

where γ_{km} is a latent indicator representing the event that skeleton k is related to skeleton m, $\delta_0(\theta_{km})$ is the probability mass function of a random variable almost surely equal to zero (the spike), and the probability density function of the second component in the mixture (the slab) corresponds to a gamma distribution with fixed hyperparameters α and β. Assuming prior independence across $1 \le k < m \le q$, and identifying $\boldsymbol{\Theta}$ with its upper triangular part, we obtain

$$p(\boldsymbol{\Theta}|\boldsymbol{\gamma}) = \prod_{k<m} p(\theta_{km}|\gamma_{km}), \tag{3.14}$$

where $\boldsymbol{\gamma}$ denotes the upper triangular matrix with elements γ_{km}. We complete the specification of our prior on $(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ by defining an independent Bernoulli prior on $\boldsymbol{\gamma}$ (with hyperparameter $w \in [0,1]$):

$$p(\boldsymbol{\gamma}) = \prod_{k<m} w^{\gamma_{km}}(1-w)^{(1-\gamma_{km})}. \tag{3.15}$$

The proposed prior borrows strength between groups when appropriate without enforcing similarity if groups have different skeleton structures.

Finally, we specify a prior for the edge inclusion probabilities ν_{ij}, $1 \le i < j \le q$, to encourage sparsity of the skeletons $\widetilde{\mathcal{G}}_1, \dots, \widetilde{\mathcal{G}}_K$. We take $\boldsymbol{\nu}$ independent of $\boldsymbol{\Theta}$ and also assume independence across edges:

$$p(\boldsymbol{\nu}) = \prod_{i<j} p(\nu_{ij}), \tag{3.16}$$

where the choice of $p(\nu_{ij})$ is based on the following considerations. Small values of ν_{ij} correspond to small prior probability of inclusion for edge (i,j) in each skeleton $\widetilde{\mathcal{G}}_k$, and consequently a prior favoring smaller values of $\boldsymbol{\nu}$ will lead to a preference for model sparsity, which can be attractive in high-dimensional applications. In contrast, larger values of ν_{ij} make edge (i,j) more likely to be selected. If, for all $m \neq k$, either $\theta_{km} = 0$ or s_{ijm} is not selected, the probability of inclusion of edge (i,j) in $\widetilde{\mathcal{G}}_k$ can be written as

$$p(s_{ijk}|\nu_{ij}) = \frac{e^{\nu_{ij}}}{1+e^{\nu_{ij}}} = q_{ij}; \tag{3.17}$$

see (3.12). Following Peterson et al[9], we impose a prior $q_{ij} \sim \text{Beta}(a,b)$ and set the hyperparameters a and b to reflect our prior assumption of sparsity. The implied prior on ν_{ij} is then

$$p(\nu_{ij}) = \frac{1}{B(a,b)}\frac{e^{a\nu_{ij}}}{(1+e^{\nu_{ij}})^{a+b}}. \tag{3.18}$$

When a reference skeleton is available, for example from a curated database that provides the connections within a signaling pathways for a normal cell, this prior can be used to incorporate prior knowledge. For example, higher prior probability can be given to edges belonging to the reference skeleton.

4 | MODEL FITTING

4.1 | MCMC algorithms

In Section 3.2 we derived the marginal likelihood of our model in closed form. The resulting target distribution is the posterior of the parameters $(\mathcal{G}_1, ..., \mathcal{G}_k, \Theta, \nu)$. In this section we describe how to implement a Metropolis-Hastings algorithm that explores the parameter space of interest, with particular emphasis on the complex space of EGs.

Let $\mathcal{S}_q$ be the set of all EGs with vertex set $V = \{1, \dots, q\}$. If some knowledge of *sparsity* is available, we can limit $\mathcal{S}_q$ to those EGs having at most M edges[32]; we denote the resulting model space as $\mathcal{S}_q^M$. To construct a Markov chain on EGs we first need to define the transitions among them. We start from the set of operators introduced by Chickering[25] and He et al[32]. Such operators can modify *locally* an EG, each involving a pair (or a triple) of nodes only. We consider seven types of operators: inserting an undirected edge (InsertU), deleting an undirected edge (DeleteU), inserting a directed edge (InsertD), deleting a

directed edge (DeleteD), converting two adjacent undirected edges in a *v*-structure (MakeV), converting a *v*-structure in two adjacent undirected edges (RemoveV) and reversing a directed edge (ReverseD). Each operator is then determined by two parts: the type and the modified edges. Chickering[25] and He et al[32] introduce a set of conditions that must be satisfied by these seven operators to guarantee that the resulting Markov chain has good theoretical properties. This leads to the definition of *perfect* operator. For each EG $\mathcal{G}$ we can then construct the corresponding set of perfect operators that determine the transition from $\mathcal{G} \in \mathcal{S}_q^M$ to $\mathcal{G}^* \in \mathcal{S}_q^M$ (one of its *direct successors*). Let $\mathcal{O}_{\mathcal{G}_k}$ be the set of perfect operators on $\mathcal{G}_k$. The probability of transition from $\mathcal{G}_k$ to $\mathcal{G}_k^*$, for each $\mathcal{G}_k^*$ direct successor of $\mathcal{G}_k$, is then

$$q(\mathcal{G}_k^* \mid \mathcal{G}_k) = 1/|\mathcal{O}_{\mathcal{G}_k}|, \tag{4.1}$$

where $|\mathcal{O}_{\mathcal{G}_k}|$ is computed following the accelerated version of the algorithm in He et al.[32].

In order to sample from the joint posterior distribution of $(\mathcal{G}_1, \dots, \mathcal{G}_K)$, $\boldsymbol{\Theta}$ and $\boldsymbol{\nu}$, we adopt the algorithm presented by Peterson et al[9] and based on the proposal of Gottardo & Raftery[33]. Accordingly, we sample the graph similarity and selection parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\gamma}$ from their joint full conditional distribution. We then update $\boldsymbol{\nu}$, and finally $\mathcal{G}_1, \dots, \mathcal{G}_K$, from their marginal full conditionals. Specifically, in the graph similarity step, if in the current state $\gamma_{km} = 1$, we propose $\gamma_{km}^* = 0$ and $\theta_{km}^* = 0$; conversely, if in the current state $\gamma_{km} = 0$, we propose $\gamma_{km}^* = 1$ and sample θ_{km}^* from $q(\theta_{km}^*) = \text{Gamma}(\theta_{km}^* \mid \alpha^*, \beta^*)$, for some proposal hyperparameters α^* and β^*. For the update of $\boldsymbol{\nu}$ we instead propose q^* from $\text{Beta}(a^*, b^*)$ and then set $\nu_{ij} = \text{logit}(q^*)$ for each $1 \le i < j \le q.$, for some proposal hyperparameters a^* and b^*. More details are given by Peterson et al[9]. Finally, in the graph selection step, we sample an EG $\mathcal{G}_k^*$ from the proposal (4.1) for each $k = 1, \dots, K$.

Relative to the acceptance rate of $\mathcal{G}_k^*$, we distinguish three cases:

(i) if we move from $\mathcal{G}_k$ to $\mathcal{G}_k^*$ by adding an edge between i and j, through an operator $o_{\mathcal{G}_k}$ of type InsertU or InsertD, then

$$\frac{p(\mathcal{G}_k^* \mid \nu, \boldsymbol{\Theta})}{p(\mathcal{G}_k \mid \nu, \boldsymbol{\Theta})} = \frac{p(\mathcal{G}_k^* \mid \widetilde{\mathcal{G}}_k^*)}{p(\mathcal{G}_k \mid \widetilde{\mathcal{G}}_k)} \exp\left\{ \nu_{ij} + 2 \sum_{m \neq k} \theta_{km} s_{ijm} \right\};$$

(ii) if we move from $\mathcal{G}_k$ to $\mathcal{G}_k^*$ by removing an edge between i and j, that is the operator $o_{\mathcal{G}_k}$ is of type DeleteU or Delete D, then

$$\frac{p(\mathcal{G}_k^* \mid \nu, \boldsymbol{\Theta})}{p(\mathcal{G}_k \mid \nu, \boldsymbol{\Theta})} = \frac{p(\mathcal{G}_k^* \mid \widetilde{\mathcal{G}}_k^*)}{p(\mathcal{G}_k \mid \widetilde{\mathcal{G}}_k)} \exp\left\{ -\nu_{ij} - 2 \sum_{m \neq k} \theta_{km} s_{ijm} \right\};$$

(iii) if we move from $\mathcal{G}_k$ to $\mathcal{G}_k^*$ without modifying the skeleton of $\mathcal{G}_k$, which happens when $o_{\mathcal{G}_k}$ is of type ReverseD, MakeV or RemoveV, then

$$\frac{p(\mathcal{G}_k^* \mid \nu, \boldsymbol{\Theta})}{p(\mathcal{G}_k \mid \nu, \boldsymbol{\Theta})} = 1.$$

Since all EGs with a given skeleton have equal prior probability, the term $p(\mathcal{G}_k^* \mid \widetilde{\mathcal{G}}_k^*)/p(\mathcal{G}_k \mid \widetilde{\mathcal{G}}_k)$ is the ratio of the number of EGs with skeleton $\widetilde{\mathcal{G}}_k$ to those with skeleton $\widetilde{\mathcal{G}}_k^*$. To compute exactly the number of EGs with a given skeleton, we refer to the freely available algorithm of Radhakrishnan et al[34], which counts the EGs per skeleton by looking at the number of possible allocations of *v*-structures. However, as the number of nodes increases, say for $q > 12$, the computational time required by the algorithm becomes prohibitive. For bigger problems, we suggest the following heuristic approximation.

Let $N_{\widetilde{\mathcal{G}}}$ and $N_{\widetilde{\mathcal{G}}^*}$ be the number of EGs with skeletons $\widetilde{\mathcal{G}}$ and $\widetilde{\mathcal{G}}^*$, respectively. Let also $n_{\mathcal{D}|\mathcal{G}}$ be the number of DAGs in the equivalence class of the EG $\mathcal{G}$. Accordingly, the number $n_{\mathcal{D}|\widetilde{\mathcal{G}}}$ of DAGs compatible with the skeleton $\widetilde{\mathcal{G}}$ can be obtained as the sum $n_{\mathcal{D}|\widetilde{\mathcal{G}}} = \sum_{\mathcal{G} \in S_{\widetilde{\mathcal{G}}}} n_{\mathcal{D}|\mathcal{G}}$, where $S_{\widetilde{\mathcal{G}}}$ is the set of all EGs with skeleton $\widetilde{\mathcal{G}}$. Moreover, we can define the average number of DAGs per EG in $S_{\widetilde{\mathcal{G}}}$ as $n_{\widetilde{\mathcal{G}}} = \sum_{\mathcal{G} \in S_{\widetilde{\mathcal{G}}}} n_{\mathcal{D}|\mathcal{G}} / N_{\widetilde{\mathcal{G}}}$. Therefore, the number of EGs with skeleton $\widetilde{\mathcal{G}}$ can be written as $N_{\widetilde{\mathcal{G}}} = n_{\mathcal{D}|\widetilde{\mathcal{G}}}/n_{\widetilde{\mathcal{G}}}$ (similarly for $\widetilde{\mathcal{G}}^*$) and the ratio $N_{\widetilde{\mathcal{G}}}/N_{\widetilde{\mathcal{G}}^*}$ becomes

$$N_{\widetilde{\mathcal{G}}}/N_{\widetilde{\mathcal{G}}^*} = \frac{n_{\mathcal{D}|\widetilde{\mathcal{G}}}/n_{\widetilde{\mathcal{G}}}}{n_{\mathcal{D}|\widetilde{\mathcal{G}}^*}/n_{\widetilde{\mathcal{G}}^*}}.$$

Since in our MCMC move $\mathcal{G}^*$ and $\mathcal{G}$ differ at most by one edge (and so the corresponding skeletons), we can assume that their average number of DAGs per EG are reasonably close, that is $n_{\widetilde{\mathcal{G}}^*} \approx n_{\widetilde{\mathcal{G}}}$. Accordingly, the previous ratio simplifies to $r = N_{\widetilde{\mathcal{G}}}/N_{\widetilde{\mathcal{G}}^*} \approx n_{\mathcal{D}|\widetilde{\mathcal{G}}}/n_{\mathcal{D}|\widetilde{\mathcal{G}}^*}$. As before, we can then distinguish the two cases (i) and (ii). In particular, if we move from $\mathcal{G}$ to $\mathcal{G}^*$ by inserting an edge, we can observe that $n_{\mathcal{D}|\widetilde{\mathcal{G}}}$ is at most twice $n_{\mathcal{D}|\widetilde{\mathcal{G}}^*}$, because the edge insertion can be made with two orientations

at maximum; it follows that $r \in [1, 2]$ in case (i). Conversely, if we move as in case (ii) from $\mathcal{G}$ to $\mathcal{G}^*$ by deleting an edge, we have $r \in [1/2, 1]$. By considering the geometric mean of the two ranges we obtain the approximation

$$\frac{p(\mathcal{G}_k^* \mid \widetilde{\mathcal{G}}_k^*)}{p(\mathcal{G}_k \mid \widetilde{\mathcal{G}}_k)} \approx \sqrt{2} \quad \text{in (i)}, \qquad \frac{p(\mathcal{G}_k^* \mid \widetilde{\mathcal{G}}_k^*)}{p(\mathcal{G}_k \mid \widetilde{\mathcal{G}}_k)} \approx \frac{1}{\sqrt{2}} \quad \text{in (ii)}. \tag{4.2}$$

To assess the appropriateness of our approximation, we run $T = 3000$ iterations of the Markov chain proposal on the EG space or number of nodes $q = 10$ and compute the exact ratio r. Results are reported in Figure 2, where we highlight different applied operators in different shades of gray: dark gray (light gray) dots correspond to operators of type Insert (Delete), while middle gray to all the remaining operators, which do not modify the skeleton; horizontal lines correspond to the approximated values $\sqrt{2}$ and $1/\sqrt{2}$. It appears that the suggested approximation works reasonably well.

Finally, for $k = 1, \dots, K$, the Metropolis-Hastings ratio for the acceptance of a newly proposed $\mathcal{G}_k^*$, conditionally on $\boldsymbol{\nu}$ and $\boldsymbol{\Theta}$, and given the graph $\mathcal{G}_k$ at the current MCMC iteration, is

$$r_{\text{MH}} = \frac{f(\boldsymbol{Y}_{[k]} \mid \mathcal{G}_k^*)}{f(\boldsymbol{Y}_{[k]} \mid \mathcal{G}_k)} \cdot \frac{p(\mathcal{G}_k^* \mid \boldsymbol{\nu}, \boldsymbol{\Theta})}{p(\mathcal{G}_k \mid \boldsymbol{\nu}, \boldsymbol{\Theta})} \cdot \frac{q(\mathcal{G}_k \mid \mathcal{G}_k^*)}{q(\mathcal{G}_k^* \mid \mathcal{G}_k)}. \tag{4.3}$$

4.2 | Posterior summaries

The main output of our methodology is the collection of multiple graphs visited by the MCMC at each iteration. This can be used to approximate posterior model probabilities or to compute measures of uncertainty, such as the marginal posterior probability of inclusion of specific edges. The marginal posterior probability of inclusion of $u \to v$ in group k is defined as

$$p_{k,u\to v}(\boldsymbol{Y}) = \sum_{\mathcal{G}_k \mid (u,v) \in E_k} p(\mathcal{G}_k \mid \boldsymbol{Y}), \tag{4.4}$$

where E_k is the edge set of $\mathcal{G}_k$. This quantity can be approximated from our MCMC output as

$$p_{k,u\to v}(\boldsymbol{Y}) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{u\to v}\{\mathcal{G}_k^{(t)}\}, \tag{4.5}$$

where $\mathbb{1}_{u\to v}\{\mathcal{G}_k^{(t)}\}$ is the indicator function taking value 1 if $\mathcal{G}_k^{(t)}$ contains $u \to v$ and 0 otherwise. Note that the undirected edge $u - v$ is equivalent to the union of $u \to v$ and $u \leftarrow v$.

Starting from the above probabilities, we can also provide an estimate of the true EGs, for comparison purposes. In this regard, we adopt the *projected median probability graph model*[31], constructed as a consistent extension to the EG space of the median probability (graph) model. The latter is obtained by including all edges whose posterior probability exceeds 0.5, as in the median probability model introduced by Barbieri & Berger[35]. The median probability graph model is not in general an EG (nor a DAG), but a partially directed graph. If we require our point estimate to be an EG, one possibility is to first construct a consistent extension of the median probability model, which is now a DAG, and then consider the EG representing its Markov equivalence class. The final output is called the projected median probability (graph) model. Since all consistent extensions belong to the same Markov equivalence class[36], the projected median probability model is unambiguously defined. However, the median probability model may not have any consistent extension, e.g. because by orienting some edges we necessarily introduce additional v-structures or cycles. Nevertheless, in our applications it always existed. As for uniqueness, since all consistent extensions belong to the same Markov equivalence class[36], the projected median probability model does not introduce any degree of arbitrariness in the resulting EG.

5 | SIMULATIONS

In the current section we perform simulation studies under diverse settings, to test the validity of the proposed approach. In more details, we construct different scenarios by varying the group sample size $n_k \in \{50, 100\}$ and the distance $s \in \{0, 4, \infty\}$ of each group-specific graph from a common unique DAG. We use the Structural Hamming Distance (SHD), defined as the number of edge insertions, deletions or flips needed to transform a graph into another. Clearly, if $s = 0$ all DAGs are equal, while by convention $s = \infty$ corresponds to four independently generated graphs. Under each of these scenarios the number of nodes and groups are fixed to $q = 20$ and $K = 4$ respectively. For each replication we first generate a DAG $\mathcal{D}_0$ with probability

of edge inclusion $p_{edge} \approx 0.08$, as in the sparse setting presented by Castelletti et al[31], and then we construct four DAGs having distance s from $\mathcal{D}_0$. In the intermediate case $s = 4$, we perform, separately for each group, four local moves from $\mathcal{D}_0$ to create the group-specific DAG. Under each DAG $\mathcal{D}$ data are generated according to

$$y_{ij} = \mu_j + \sum_{l \in \text{pa}_{\mathcal{D}}(j)} \beta_{lj} y_{il} + \varepsilon_{ij}, \tag{5.1}$$

for $i = 1, \dots, n$, $j = 1, \dots, q$, where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$ independently. For each j we fix $\mu_j = 0$ and $\sigma_j^2 = 1$, while regression coefficients β_{lj} are uniformly chosen in the interval $[-1, -0.1] \cup [0.1, 1]$[37]. Accordingly, we generate $N = 12$ multiple datasets each consisting of K distinct $n_k \times q$ data matrices $\boldsymbol{Y}_{[1]}, \dots, \boldsymbol{Y}_{[K]}$.

To speed up MCMC mixing, we constrain the EG space by fixing the sparsity parameter $M = 40$ (Section 4.1), that is we constrain the model space to those graphs having at most 40 edges. Such a threshold is not restrictive since well above the expected number of edges in the true graphs (about 15). As hyperparameters for the slab portion of $p(\theta_{km} \mid \gamma_{km})$ we choose $\alpha = 2$ and $\beta = 5$, while we set $w = 0.9$ in the Bernoulli prior on γ_{km}. In the Beta prior on q_{ij} we instead choose $a = 0.5, b = (2q - 2)/3 - 1$. For $q = 20$ this results in a prior edge inclusion probability of about 0.04 which is smaller than the expected level of sparsity ($p_{edge} = 0.08$) as recommended by Peterson et al[9]. We finally set the proposal parameters to $a^* = 2, b^* = 4$ and $\alpha^* = 1, \beta^* = 0.5$.

We compare our method with two benchmarks: the first is the *Objective Bayes Essential graph Search* method (OBES)[31], which is equivalent to our multiple EG search method without accounting for possible shared structures between graphs (equivalently with each element in $\boldsymbol{\Theta}$ fixed equal to 0); the second benchmark is the *Greedy Equivalence Search* method (GES)[25;38]. GES is computed for three different optimization criteria: the Bayesian Information Criterion[39] and the Extended Bayesian Information Criterion with tuning coefficient $\gamma \in \{0.5, 1\}$[40].

Under each scenario and for each method we evaluate the performance in learning the graphical structure of the true EG in terms of misspecification rate, specificity, sensitivity, precision and Matthews correlation coefficient, defined as

$$\text{MISR} = \frac{FN+FP}{q(q-1)}, \quad \text{SPE} = \frac{TN}{TN+FP}, \quad \text{SEN} = \frac{TP}{TP+FN},$$
$$\text{PRE} = \frac{TP}{TP+FP}, \quad \text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TP, TN, FP, FN are the numbers of true positives, true negatives, false positives and false negatives (respectively). Results for $n \in \{50, 100\}$ are reported in Tables 1 and 2 respectively.

In both tables, in the scenario $s = \infty$, we find that OBES performs in line with GES, with the choice of the tuning parameter strongly affecting the performance of the latter; similar results were obtained by Castelletti et al[31]. Furthermore, the proposed Multiple OBES (MOBES) does not suffer relative to the other benchmarks, a result we consider satisfactory since, contrary to the others, our method is not specifically designed for independent graphs. Once we move from $s = \infty$ to $s = 4$ and to $s = 0$, more and more dependence among the graphs is introduced: while all the other methods worsen all their measured performance indicators, MOBES reveals its superiority, with respect to the benchmarks and to MOBES itself in scenarios with less dependence among the graphs.

Furthermore, we measure the Structural Hamming Distance (SHD) between the four graphs estimated under each method and the corresponding true EGs, that is the EGs representing the Markov equivalence classes of the four true DAGs. Results for $n \in \{50, 100\}$ and $s \in \{0, 4, \infty\}$ are reported in the boxplots of Figure 3. As expected, GES with BIC criterion is the worst performer, while the GES methods with EBIC are comparable to OBES in the case of independent graphs. MOBES is clearly the closest to the true graphs in the scenarios with shared structures, having the edge over OBES and GES with BIC also in the independent settings.

We also consider two additional biologically-driven simulated scenarios. We first fix $K = 4$ true EGs equal to the EG estimates obtained from the Leukemia data in Section 6.2 below. For each true EG we then consider a DAG within its equivalence class and generate n_k i.i.d. observations, $k = 1, \dots, K$, from the (DAG-constrained) set of linear equations in (5.1), everything repeated for $N = 12$ multiple datasets. In the first additional scenario, we fix all sample sizes as being equal, by setting $n_k = 50$ for all k, while in the second scenario we differentiate sample sizes among groups with $n_1 = 25, n_2 = 50, n_3 = 100$ and $n_4 = 50$. We compare the results with the OBES method (where the K EGs are estimated independently) in terms of SHD between true and estimated EGs, and summarize the results in the first row of Figure 4. Finally, on the same scenarios, we also implement the method of Peterson et al[9], which returns multiple estimates of the graph skeletons. We compare it with OBES and MOBES and evaluate the performance of the various methods in terms of SHD between the skeletons of estimates and true EGs (Figure 4,

second row). All comparisons in these biologically motivated scenarios, with both homogeneous and heterogeneous sample sizes, confirm results that favor our methodology.

Note that, despite the complexity of the model, there are only five hyperparameters to be set, all related to the MRF prior. Three hyperparameters w, α, β concern network similarity, and the other two (a and b) network sparsity. As described in Li and Zhang (2010)[41], α and β should be set to avoid phase transition, i.e. to avoid that larger values of parameters θ_{km} lead to a extremely sharp increase in the expected number of edges included in all networks. In our context, phase transition would result in the selection of the same identical graph for all groups. As noted by Peterson et al[9], $\alpha = 2$ and $\beta = 5$ can be considered a default choice that avoids phase transition. We then perform sensitivity analyses with respect to w and to a, b, by applying our method to a single fixed (multiple) dataset, as generated in our new simulation setting. We first vary $w \in \{0.3, 0.5, 0.7, 0.9\}$, while keeping the other hyperparameters fixed as in the original setting, and evaluate the effect on the average Posterior Probability of Inclusion (PPI) for the elements θ_{km}, $k \neq m$. We then vary the mean of the prior Beta(a, b) in (3.18) in the range $[0.05, 0.30]$, with variance fixed as in the original setting, to assess the impact of the prior probability of edge inclusion on the average PPI. Results are reported in Figure 5 and at the end of Section 5 in the paper. The average edge PPIs showed a steady increase from just below 0.100 to 10.125 in whole range considered for the prior mean, and from 0.42 to 0.47 in the set of studied values of w. The direction of the effect in both cases is expected, and the overall difference in levels is not strong.

Finally, we investigate the computational time of our method as a function of the number of variables q, the number of groups K and the sample size n, as measured on a PC Intel(R) Core(TM) i7-8550U 1,80 GHz. We report the computational time (averaged over 12 multiple datasets) *per* iteration for $q = 20$ as a function of $K \in \{2, 3, 4, 5\}$ (Figure 6, right panel) and for $K = 4$ as a function of $q \in \{5, 10, 20, 40\}$ (Figure 6, left panel), with $n = 50$. The behavior of all curves suggests a polynomial dependence of the computational time from both q and K, while we do not show for brevity that processing times are insensitive to the group sample sizes n_k.

6 | DATA ANALYSIS

6.1 | Perturbed protein signaling networks

We first investigate the datasets of Sachs et al[42] on multiple phosphorylated proteins and phospholipid components in individual primary human immune system cells. Observations are obtained from flow cytometry which also allows to measure protein modification states. Specifically, measurements of $q = 11$ phosphorylated proteins and phospholipids are collected after a series of stimulatory and inhibitory interventions obtained from the administration of different reagents. This results in a collection of distinct datasets. Some of these can be related to interventions on observed variables and were analyzed by Castelletti and Consonni[43] to infer a unique graph called *interventional* essential graph which reflects modifications in the edge structure due to interventions on nodes. The same dataset was instead analyzed by Peterson et al[9] from a multiple undirected graphs perspective. In our study we include $K = 5$ datasets that are not linked to interventions on specific variables, but rather to general perturbations of the system. The sample size of each dataset ranges between 700 and 1000 observations. We set the prior and proposal parameters as in the simulation setting of Section 5 and run $T = 25000$ iterations of the MCMC scheme presented in Section 4.1. The first 5000 iterations are discarded as a burn-in period.

We report the five EGs estimated by MOBES in Figure 7, where similarities in the skeletons are highlighted with dotted edges. An edge is included in our graph estimate if its marginal posterior probability of inclusion is estimated to be higher than 0.5. Such probabilities are reported in the heatmaps of Figure 8. The posterior probabilities of inclusion for the elements θ_{km} in $\boldsymbol{\Theta}$, given by $\text{PPI}_{km}(\boldsymbol{Y}) = \Pr(\theta_{km} \neq 0 | \boldsymbol{Y})$, for $k \neq m = 1, \ldots, K$, and the number of shared edges between estimated graphs across groups are the following:

$$\text{PPI}(\boldsymbol{Y}) = \begin{pmatrix} * & 0.508 & 0.516 & 0.522 & 0.492 \\ & * & 0.483 & 0.509 & 0.466 \\ & & * & 0.435 & 0.448 \\ & & & * & 0.459 \\ & & & & * \end{pmatrix}, \qquad \text{Shared edge count} = \begin{pmatrix} 10 & 10 & 9 & 10 & 9 \\ & 10 & 9 & 10 & 9 \\ & & 11 & 9 & 8 \\ & & & 10 & 9 \\ & & & & 10 \end{pmatrix};$$

note that the shared edge counts on the diagonal correspond to the number of edges in the estimated graphs.

We also compare MOBES with alternative approaches for structural learning of EGs which do not account for similarities between graphs, namely the OBES method and the GES algorithm as presented in the simulation setting of Section 5. Results

for each group are shown in Table 3, where we report the SHD between estimated graphs. Differences between OBES and MOBES are relatively small, so that there seems to be no substantial gain from the adoption of a "multiple graphs" approach if the sample size of each group is sufficiently large as in this case. Conversely, the GES algorithm is more sensitive to the choice of the tuning parameter; in particular the differences are more noticeable between GES 0 ($\gamma = 0$) and MOBES.

In Figure 9 a proper mixing of the sampling algorithm is shown by the MCMC chains of a few features of the visited EGs, for the first dataset: number of undirected and directed edges, of υ-structures and of chain components. From the five EGs, it is clear that there is a common structure shared by the graphs. Convergence diagnostics results based on Geweke statistics[44], not reported for brevity, show appropriate convergence for all groups, for various features of the graphical structures visited by the MCMC: number of edges, directed edges, undirected edges, v-structures and chain components.

In comparison with the results of Peterson et al[9], our analysis reveals some similarities, such as the identification of the same chain components, and some unique findings, notably the directed arrows that link PIP3, and its component, as well as Erk, and its component, to PKC in the network of group 3 (top-right graph in Figure 7). These findings suggest that the perturbation corresponding to group 3 may have triggered an alternative regulatory cascade.

6.2 | Leukemia Protein Networks

In this subsection we analyze data on protein levels for 213 Acute Myeloid Leukemia (AML) patients presented in the supplementary material of Kornblau et al[45]. Subtypes of statistical units are based on cytogenetics and cellular morphology criteria, among which it is reasonable to expect interactions, justifying an estimation method that accounts for heterogeneity[9]. Accordingly, we infer an EG for each of the following subtypes: M0 (17 subjects), M1 (34 subjects), M2 (68 subjects), and M4 (59 subjects). We emphasize that sharing information among EGs is particularly appropriate in this setting, because of the small to moderate sample sizes. We exclude from the analysis further subtypes whose sample sizes are even smaller[9].

The estimated EGs for the four subtypes are shown in Figure 10. The inter-dependency is reflected in the matrix PPI($\boldsymbol{Y}$) containing the posterior probabilities of inclusion of common edges, and the number of shared edges between estimated graphs across groups

$$\text{PPI}(\boldsymbol{Y}) = \begin{pmatrix} * & 0.552 & 0.564 & 0.522 \\ & * & 0.543 & 0.527 \\ & & * & 0.498 \\ & & & * \end{pmatrix}, \qquad \text{Shared edge count} = \begin{pmatrix} 9 & 6 & 7 & 6 \\ & 9 & 6 & 5 \\ & & 15 & 6 \\ & & & 12 \end{pmatrix}.$$

The inferential advantage of MOBES is apparent in comparison with OBES and other alternative methods that do not account for commonalities among graphs, and therefore are bounded to rely on small to moderate number of observations for each subtype. As a matter of fact, the same graphs estimated with OBES are too sparse, while GES estimates are strongly affected in terms of sparsity by the tuning parameter $\gamma \in \{0, 0.5, 1\}$. As an example, with reference to the group M0, GES 0 ($\gamma = 0$) returns 31 edges, while GES 1 ($\gamma = 1$) only one edge. Such information is included in the main diagonals of Table 4. In the same table, the off-diagonal elements report the SHDs between graphs estimated using the five methods under comparison. Again, appropriate MCMC convergence diagnostics is confirmed, for all groups and features, by Geweke statistics[44].

Some of the findings of our analysis are of particular interest. For example, GSK3 was found to regulate, or be regulated by, AKT in the M2 network. The correlation of GSK3.p with a number of proteins including AKT was recently established by Ruvolo et al[46]; the same authors reached the conclusion that AKT/GSK3 is a critical axis in AML, which may be a therapeutic target in AML patients with intermediate cytogenetics, i.e. M2 patients. In agreement with Peterson et al[9], we identified associations between the BAD and PTEN proteins (specifically PTEN-BAD.p155 and PTEN.p-BAD.p136) to be present for all four groups; beyond the findings reported also by Peterson et al[9], the proposed method detected a direct effect of BAD.p136 on PTEN.p, in M1 and M4 patients.

7 | SUMMARY AND DISCUSSION

Statistical methods for the reconstruction of gene and protein networks under multiple conditions are a viable tool for studying the biological mechanisms underlying genomic driven diseases. To this end we have introduced a Bayesian model for structural learning of a collection of Essential Graphs (EGs), each identifying an equivalence class of DAGs. In the context of multiple networks, this represents the first attempt at learning the directionality of an arrow, whenever this is doable. The centerpiece

of our approach is a prior distribution on the structure of multiple EGs that directly models the skeletons of these graphs and encourages skeleton similarity, when supported by the data. Our modeling approach includes parameter priors that are free from hyperparameters and produce a closed-form expression for the marginal likelihood of an EG based on the fractional Bayes factor, which greatly enhances computational efficiency.

Using simulation studies, we demonstrate the superior performance of our approach in comparison with state-of-the-art methods, and the gain in network reconstruction accuracy yielded by an approach that jointly infers multiple related networks. We apply our model to two datasets. First, an analysis of protein networks from primary human immune system cells revealed that some types of perturbations may lead to alternative regulatory mechanisms that warrant further investigation; further findings were consistent with the literature. Second, our analysis of the protein networks of AML patients, grouped by subtype, revealed novel (potential) regulatory mechanisms, and the direction in which these mechanisms operate. In this application, groups have varying sample sizes with larger groups resulting in denser networks; between-group analyses of graph structure differences should be performed with particular care.

In many applications, including the ones discussed in this paper, the number of groups K is assumed to be small. If K grows larger, the normalizing constant of equation (3.10) cannot be analytically computed, and we would need to rely on computational strategies for doubly-intractable distributions, such as the one proposed by Murray et al[47], among others. From a broader perspective, our model is predicated on the assumption that the observations follow a Gaussian graphical model. When this setup is not appropriate, a robust version could be developed using Dirichlet t-distributions[48] or a more elaborate non-parametric Bayesian approach based on hierarchical normalized completely random measures[49].

ACKNOWLEDGMENTS

The work of Federico Castelletti, Guido Consonni and Stefano Peluso was partially supported by grants from UCSC (projects D1) and by the EU COSTNET project (CA15109).

Conflict of interest

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The protein signalling data of Sachs et al. (2005)[42] are provided as supplement to the original paper and publicly available at https://science.sciencemag.org/content/308/5721/523/tab-figures-data. The RPPA protein data of Kornblau et al (2009)[45] are provided as supplement to the original paper and publicly available at http://bioinformatics.mdanderson.org/supplements.html (under "RPPA Data in AML").

References

1. Chin L, Andersen J, Futreal A. Cancer genomics: from discovery science to personalized medicine. *Nature Medicine* 2011; 17(3): 297–303.

2. Kristensen V, Lingjaerde O, Russnes H, Vollan H, Frigessi A, Borresen-Dale A. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer* 2014; 14(5): 299–313.

3. Dobra A, Jones B, Hans C, Nevins J, West M. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* 2004; 90(1): 196–212.

4. Stingo FC, Chen YA, Vannucci M, Barrier M, Mirkes PE. A Bayesian graphical modeling approach to microRNA regulatory network inference. *Annals of Applied Statistics* 2010; 4(4): 2024–2048.

5. Telesca D, Müller P, Parmigiani G, Freedman RS. Modeling dependent gene expression. *Annals of Applied Statistics* 2012; 6(2): 542–560.

6. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2014; 76(2): 373–397.

7. Zhao SD, Cai TT, Li H. Direct estimation of differential networks. *Biometrika* 2014; 101(2): 253–268.

8. Pircalabelu E, Claeskens G, Waldorp LJ. Mixed scale joint graphical lasso. *Biostatistics* 2016; 17(4): 793–806.

9. Peterson C, Stingo FC, Vannucci M. Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association* 2015; 110(509): 159–174.

10. Tan LS, Jasra A, De Iorio M, Ebbels TM. Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *Annals of Applied Statistics* 2017; 11(4): 2222–2251.

11. Jalali P, Khare K, Michailidis G. A Bayesian approach to joint estimation of multiple graphical models. *arXiv e-prints* 2019: arXiv:1902.03651.

12. Williams DR, Rast P, Pericchi L, Mulder J. Comparing Gaussian graphical models with the posterior predictive distribution and Bayesian model selection. PsyArXiv, https://doi.org/10.31234/osf.io/yt386; 2019

13. Mohan K, London P, Fazel M, Witten D, Lee SI. Node-Based Learning of Multiple Gaussian Graphical Models. *Journal of Machine Learning Research* 2014; 15(13): 445-488.

14. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press . 2000.

15. Yajima M, Telesca D, Ji Y, Müller P. Detecting differential patterns of interaction in molecular pathways. *Biostatistics* 2014; 16(2): 240–251.

16. Mitra R, Müller P, Ji Y. Bayesian Graphical Models for Differential Pathways. *Bayesian Analysis* 2016; 11(1): 99–124.

17. Oates C, Smith J, Mukherjee S, Cussens J. Exact estimation of multiple directed acyclic graphs. *Statistics and Computing* 2016; 26(4): 797–811.

18. Andersson SA, Madigan D, Perlman MD. A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics* 1997; 25(2): 505–541.

19. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2017.

20. Lauritzen SL. *Graphical Models*. Oxford, UK: Oxford University Press . 1996.

21. Berger JO, Pericchi LR. Objective Bayesian methods for model selection: introduction and comparison. In: Lahiri P. , ed. *Model Selection*. 38 of *IMS Lecture Notes Monograph Series*. Beachwood, OH: Institute of Mathematical Statististics. 2001 (pp. 135–207).

22. Drton M. Discrete chain graph models. *Bernoulli* 2009; 15(3): 736–753.

23. Spirtes P, Glymour C, Scheines R. *Causation, Prediction and Search*. Cambridge, MA: The MIT Press. 2 ed. 2000.

24. Verma T, Pearl J. Equivalence and synthesis of causal models. In: Bonissone PP, Henrion M, Kanal LN, Lemmer JF. , eds. *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence.*Association for Uncertainty in Artificial Intelligence. Elsevier Science; 1991; New York, NY: 255–270.

25. Chickering DM. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research* 2002; 2(Feb): 445–498.

26. Roverato A. A unified approach to the characterization of equivalence classes of DAGs, chain graphs with no flags and chain graphs. *Scandinavian Journal of Statistics* 2005; 32(2): 295–312.

27. Andersson SA, Madigan D, Perlman MD. Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics* 2001; 28(1): 33–85.

28. Consonni G, La Rocca L, Peluso S. Objective Bayes covariate-adjusted sparse graphical model selection. *Scandinavian Journal of Statistics* 2017; 44(3): 741–764.

29. O'Hagan A. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; 57(1): 99–138.

30. Geiger D, Heckerman D. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Annals of Statistics* 2002; 30(5): 1412–1440.

31. Castelletti F, Consonni G, Della Vedova M, Peluso S. Learning Markov equivalence classes of directed acyclic graphs: an objective Bayes approach. *Bayesian Analysis* 2018; 13(4): 1231–1256.

32. He Y, Jia J, Yu B. Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *Annals of Statistics* 2013; 41(4): 1742–1779.

33. Gottardo R, Raftery AE. Markov Chain Monte Carlo with mixtures of mutually singular distributions. *Journal of Computational and Graphical Statistics* 2008; 17(4): 949–975.

34. Radhakrishnan A, Solus L, Uhler C. Counting Markov equivalence classes by number of immoralities. In: Elidan G, Kersting K, Ihler AT., eds. *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*.Association for Uncertainty in Artificial Intelligence. AUAI Press; 2017: 1–10.

35. Barbieri MM, Berger JO. Optimal predictive model selection. *Annals of Statistics* 2004; 32(3): 870–897.

36. Dor D, Tarsi M. A simple algorithm to construct a consistent extension of a partially oriented graph. Tech. Rep. R-185, Cognitive Systems Laboratory, UCLA; Los Angeles, CA: 1992.

37. Peters J, Bühlmann P. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* 2014; 101(1): 219–228.

38. Hauser A, Bühlmann P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research* 2012; 13(1): 2409–2464.

39. Schwarz GE. Estimating the dimension of a model. *Annals of Statistics* 1978; 6(2): 461-464.

40. Foygel R, Drton M. Extended Bayesian information criteria for Gaussian graphical models. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A., eds. *Proceedings of NIPS 2010*. 23 of *Advances in Neural Information Processing Systems*. Curran Associates. 2010 (pp. 604–612).

41. Li F, Zhang NR. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association* 2010; 105(491): 1202-1214.

42. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 2005; 308(5721): 523–529.

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

44. Geweke J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Tech. Rep. 148, Federal Reserve Bank of Minneapolis; Minneapolis, MN: 1991.

45. Kornblau SM, Tibes R, Qiu YH, et al. Functional proteomic profiling of AML predicts response and survival. *Blood* 2009; 113(1): 154–164.

46. Ruvolo PP, Qiu Y, Coombes KR, et al. Phosphorylation of GSK3α/β correlates with activation of AKT and is prognostic for poor overall survival in acute myeloid leukemia patients. *BBA Clinical* 2015; 4(Dec): 59-68.

47. Murray I, Ghahramani Z, MacKay DJC. MCMC for doubly-intractable distributions. In: Dechter R, Richardson T., eds. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*.Association for Uncertainty in Artificial Intelligence. AUAI Press; 2006: 359–366.

48. Finegold M, Drton M. Robust Bayesian graphical modeling using Dirichlet *t*-distributions. *Bayesian Analysis* 2014; 9(3): 521–550.

49. Cremaschi A, Argiento R, Shoemaker K, Peterson C, Vannucci M. Hierarchical normalized completely random measures for robust graphical modeling. *Bayesian Analysis* 2018; 9(3): 521–550.

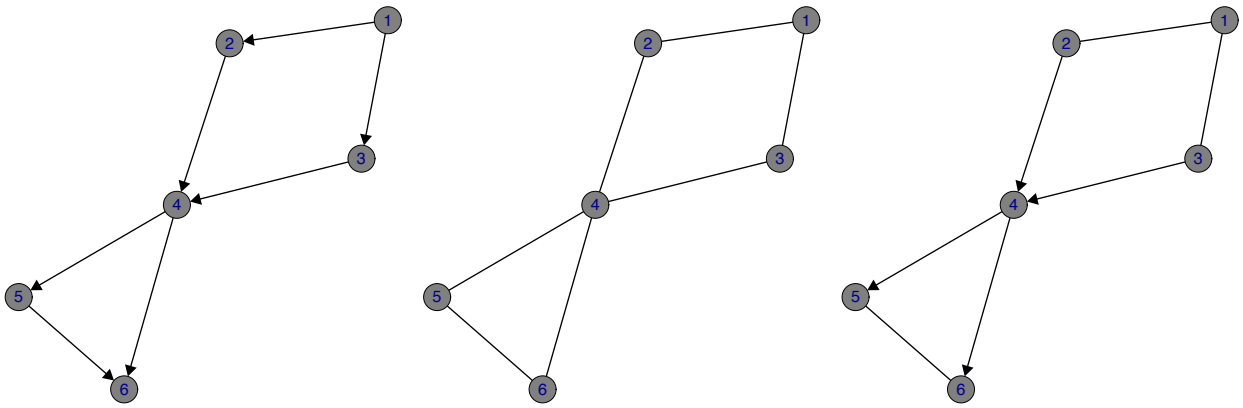

FIGURE 1 Three chain graphs with $V = \{1, 2, 3, 4, 5, 6\}$. The DAG with $E = \{(1, 2), (1, 3), (2, 4), (3, 4), (4, 5), (4, 6), (5, 6)\}$ on the left, its skeleton in the middle, and its essential graph on the right. There are six (singleton) chain components in the left graph and a single chain component V in the middle graph, while $\mathcal{T} = \{\{1, 2, 3\}, \{4\}, \{5, 6\}\}$ in the right graph.

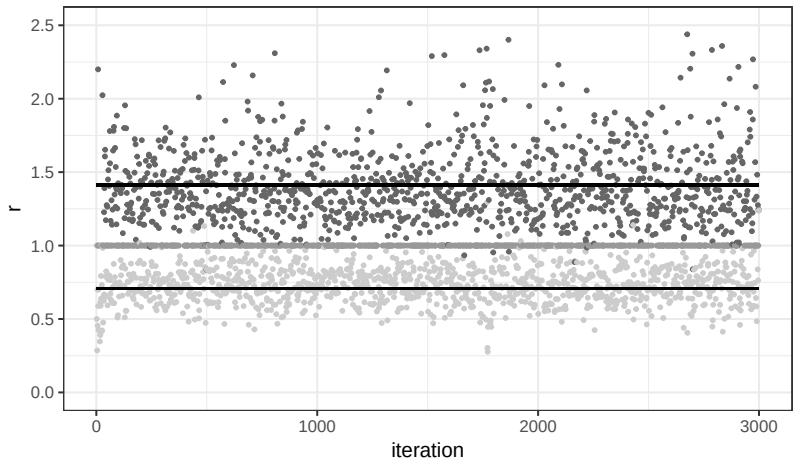

FIGURE 2 Comparison between exact and approximated ratio r. Dark grey (light grey) dots correspond to operators of type Insert (Delete), while middle grey to all the remaining operators that do not modify the skeleton. Approximated values, $\sqrt{2}$ and $1/\sqrt{2}$, are represented by horizontal lines.

		MISR	SPE	SEN	PRE	MCC
	MOBES	**2.36** (0.82)	**99.33** (0.49)	**71.68** (11.18)	**87.61** (8.75)	**78.24** (8.38)
	OBES	4.56 (0.97)	99.33 (0.37)	36.15 (12.96)	77.02 (11.97)	51.60 (11.70)
$s = 0$	GES 0	6.80 (1.73)	95.97 (1.24)	51.59 (10.73)	46.36 (11.49)	48.89 (9.42)
	GES 0.5	3.99 (1.26)	98.95 (0.60)	51.36 (13.05)	76.26 (12.93)	61.66 (11.70)
	GES 1	4.29 (1.05)	99.32 (0.38)	40.63 (13.51)	79.05 (11.08)	55.45 (11.96)
	MOBES	**2.74** (0.93)	99.16 (0.54)	**66.79** (11.37)	82.26 (10.71)	**73.18** (9.37)
	OBES	4.30 (0.83)	99.44 (0.40)	34.64 (10.87)	78.99 (12.00)	51.01 (9.56)
$s = 4$	GES 0	7.41 (1.25)	95.63 (0.96)	46.68 (10.30)	38.27 (8.82)	42.67 (8.15)
	GES 0.5	3.76 (1.00)	99.12 (0.44)	49.84 (10.83)	76.66 (10.03)	61.01 (9.52)
	GES 1	3.89 (0.86)	**99.46** (0.42)	41.50 (11.26)	**82.66** (11.20)	57.25 (9.55)
	MOBES	3.96 (1.23)	99.10 (0.61)	45.65 (13.87)	75.46 (14.43)	57.79 (12.94)
	OBES	4.42 (0.94)	99.45 (0.39)	31.25 (13.51)	78.18 (12.84)	47.79 (12.28)
$s = \infty$	GES 0	7.52 (1.29)	95.17 (0.91)	48.26 (11.95)	37.61 (9.79)	43.06 (9.09)
	GES 0.5	**3.68** (0.96)	99.05 (0.58)	**51.00** (10.81)	77.92 (11.29)	**61.99** (8.76)
	GES 1	4.13 (1.05)	**99.47** (0.44)	36.14 (14.06)	**81.50** (13.83)	52.66 (12.89)

TABLE 1 Simulations. Misspecification rate (MISR), specificity (SPE), sensitivity (SEN), precision (PRE) and Matthews correlation coefficient (MCC) for MOBES, OBES and GES, for number of nodes $q = 20$, sample size $n = 50$ and distance of the data generating DAGs from a common unique DAG $s \in \{0, 4, \infty\}$. Average values (standard deviations) are computed over $N = 12$ multiple datasets.

		MISR	SPE	SEN	PRE	MCC
	MOBES	**1.55** (0.94)	99.16 (0.67)	**88.06** (9.28)	**87.77** (9.53)	**87.17** (7.81)
	OBES	2.93 (0.90)	99.24 (0.38)	63.92 (11.92)	84.32 (8.31)	72.41 (9.04)
$s = 0$	GES 0	5.60 (1.36)	96.72 (0.97)	59.29 (10.94)	54.55 (10.59)	56.34 (9.14)
	GES 0.5	2.80 (0.98)	99.34 (0.40)	64.69 (11.33)	86.83 (7.82)	73.90 (8.29)
	GES 1	3.01 (0.96)	**99.37** (0.41)	60.38 (12.35)	86.00 (9.70)	71.01 (9.96)
	MOBES	**1.57** (0.97)	99.17 (0.65)	**86.69** (10.52)	85.86 (10.32)	**85.53** (8.86)
	OBES	2.46 (0.99)	99.43 (0.45)	67.16 (12.38)	87.31 (8.77)	75.58 (9.23)
$s = 4$	GES 0	4.88 (1.31)	97.14 (0.82)	64.42 (11.41)	56.26 (11.32)	59.64 (10.33)
	GES 0.5	2.50 (0.97)	99.27 (0.47)	68.81 (12.28)	84.14 (10.31)	75.24 (10.36)
	GES 1	2.67 (0.79)	**99.50** (0.36)	61.68 (11.28)	**87.76** (9.23)	72.54 (8.76)
	MOBES	**2.42** (1.21)	99.25 (0.65)	**69.97** (13.82)	85.11 (12.22)	**76.35** (11.32)
	OBES	2.69 (0.77)	99.38 (0.43)	62.55 (11.59)	85.87 (8.55)	72.32 (8.64)
$s = \infty$	GES 0	5.07 (1.46)	96.95 (1.05)	61.89 (11.60)	55.37 (11.78)	58.03 (10.32)
	GES 0.5	2.52 (1.07)	99.27 (0.54)	67.79 (12.85)	85.02 (9.47)	75.08 (9.85)
	GES 1	2.71 (0.80)	**99.52** (0.48)	59.87 (11.09)	**88.59** (9.30)	71.74 (8.20)

TABLE 2 Simulations. Misspecification rate (MISR), specificity (SPE), sensitivity (SEN), precision (PRE) and Matthews correlation coefficient (MCC) for MOBES, OBES and GES, for number of nodes $q = 20$, sample size $n = 100$ and distance of the data generating DAGs from a common unique DAG $s \in \{0, 4, \infty\}$. Average values (standard deviations) are computed over $N = 12$ multiple datasets.

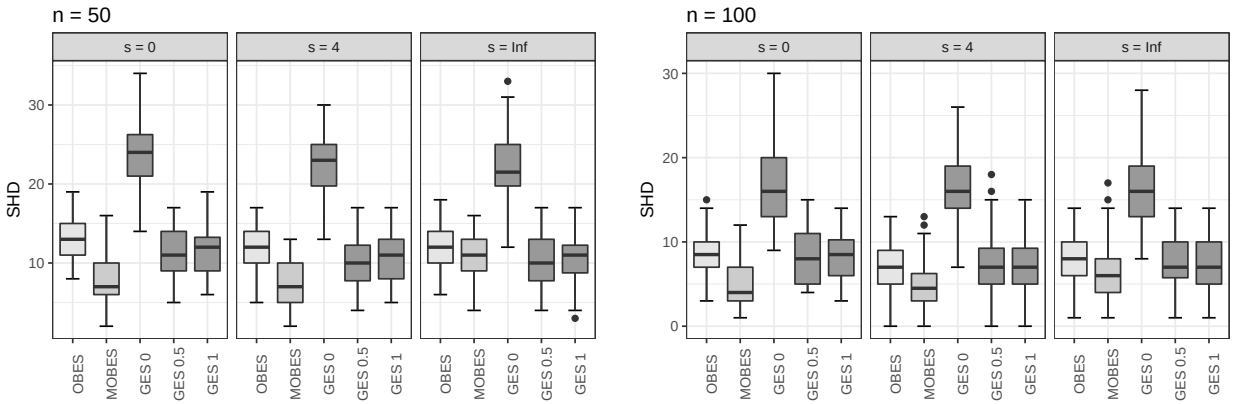

FIGURE 3 Simulations. Structural Hamming Distances between estimated EGs and true EGs, over 12 multiple datasets, for number of nodes $q = 20$, sample size $n \in \{50, 100\}$ and distance of the data generating DAGs from a common unique DAG $s \in \{0, 4, \infty\}$. The five methods under comparison are: our Multiple Objective Bayes Essential graph Search (MOBES), the Objective Bayes Essential graph Search (OBES) and the Greedy Equivalence Search (GES) computed for three different optimization criteria (BIC and EBIC with tuning parameter $\gamma \in \{0.5, 1\}$).

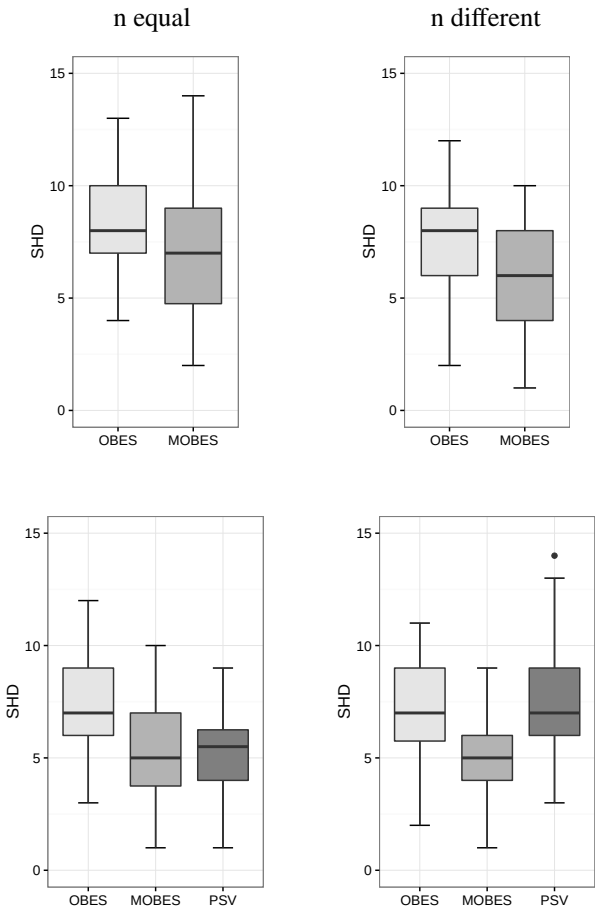

FIGURE 4 Simulations. Multiple datasets generated from the four Leukemia-data EG estimates of Section 6.2, for group sample sizes $n_1 = n_2 = n_3 = n_4 = 50$ (n equal) and $n_1 = 25, n_2 = 50, n_3 = 100, n_4 = 50$ (n different). Methods under comparison: the proposed Multiple Objective Bayes Essential graph Search (MOBES), the Objective Bayes Essential graph Search (OBES) and the method of Peterson et al[9] (PSV) for multiple UG model selection. First row. Structural Hamming Distances between estimated EGs and true EGs. Second row: Structural Hamming Distances between corresponding skeletons.

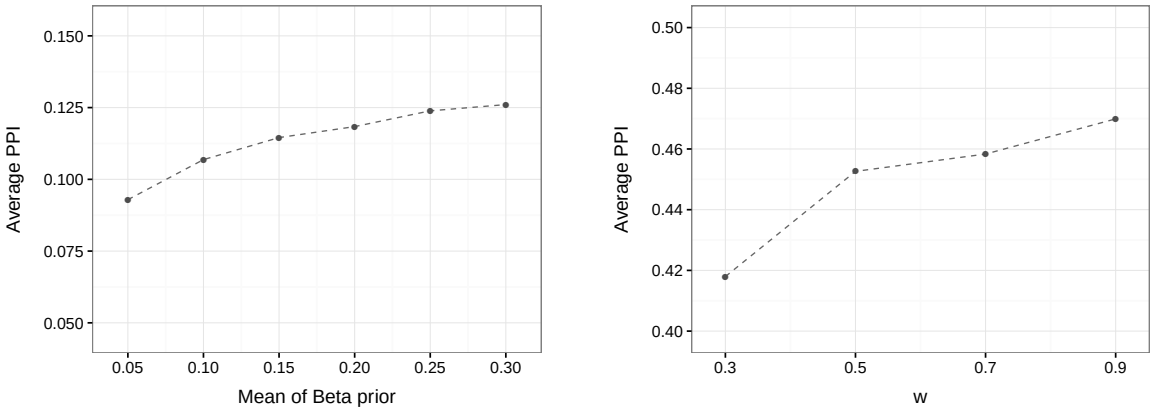

FIGURE 5 Sensitivity analysis of average Posterior Probability of Inclusion (PPI) to different values of the hyperparameters a and b (left) and w (right).

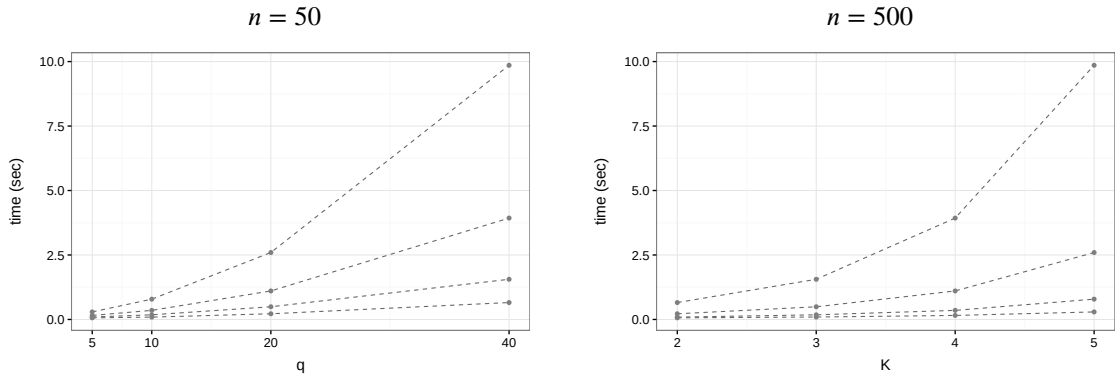

FIGURE 6 Computational time (in seconds) *per* iteration of MOBES as a function of the number of variables $q \in \{5, 10, 20, 40\}$ for fixed number of groups $K = 4$ (left panel) and as a function of the number of groups $K \in \{2, 3, 4, 5\}$ for fixed number of variables $q = 20$ (right panel); group sample size $n_k = 50$, $k = 1, \ldots, K$, averaged over 12 multiple simulated datasets.

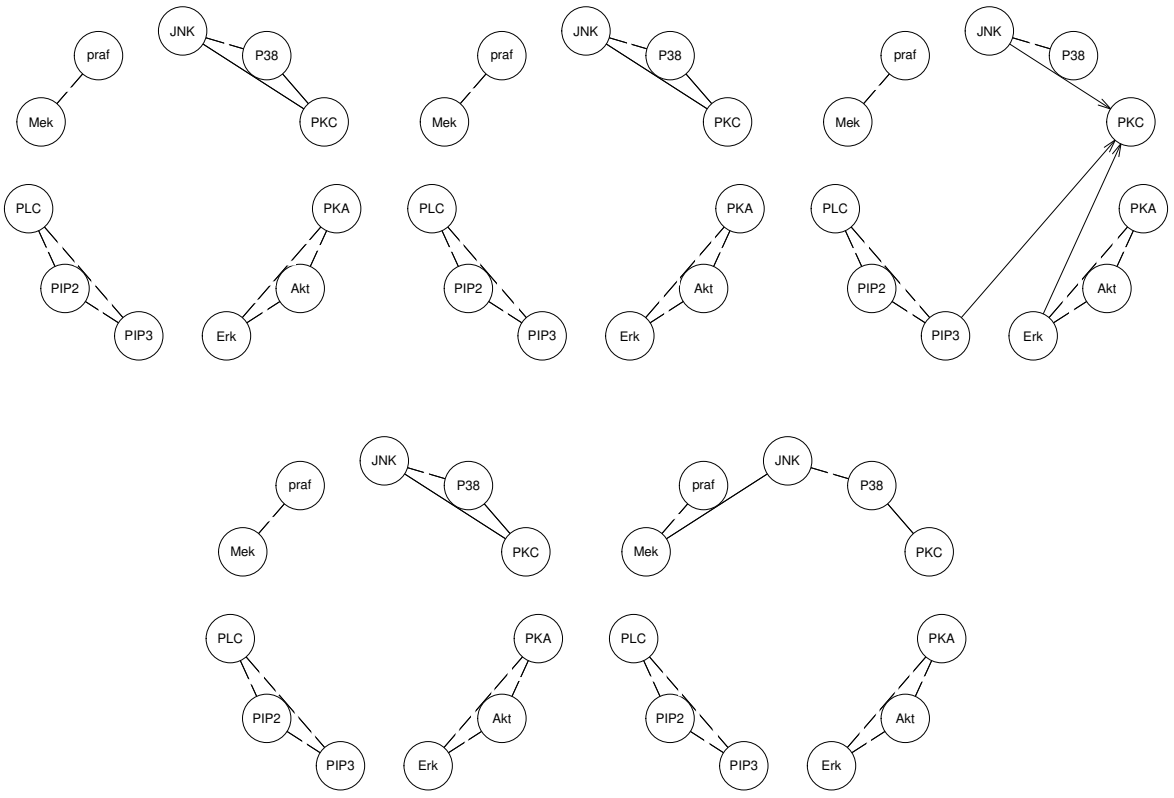

FIGURE 7 Sachs data. EGs estimated by MOBES for the five datasets included in the study.

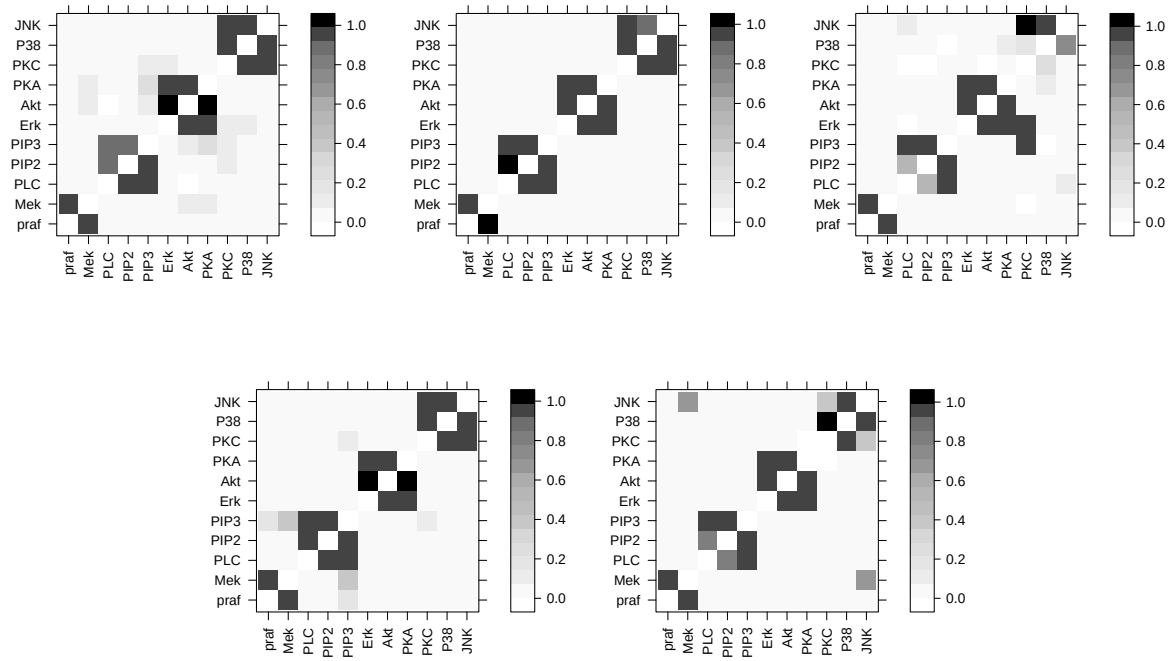

FIGURE 8 Sachs data. Heat maps of marginal posterior probabilities of edge inclusion for the five datasets included in the study.

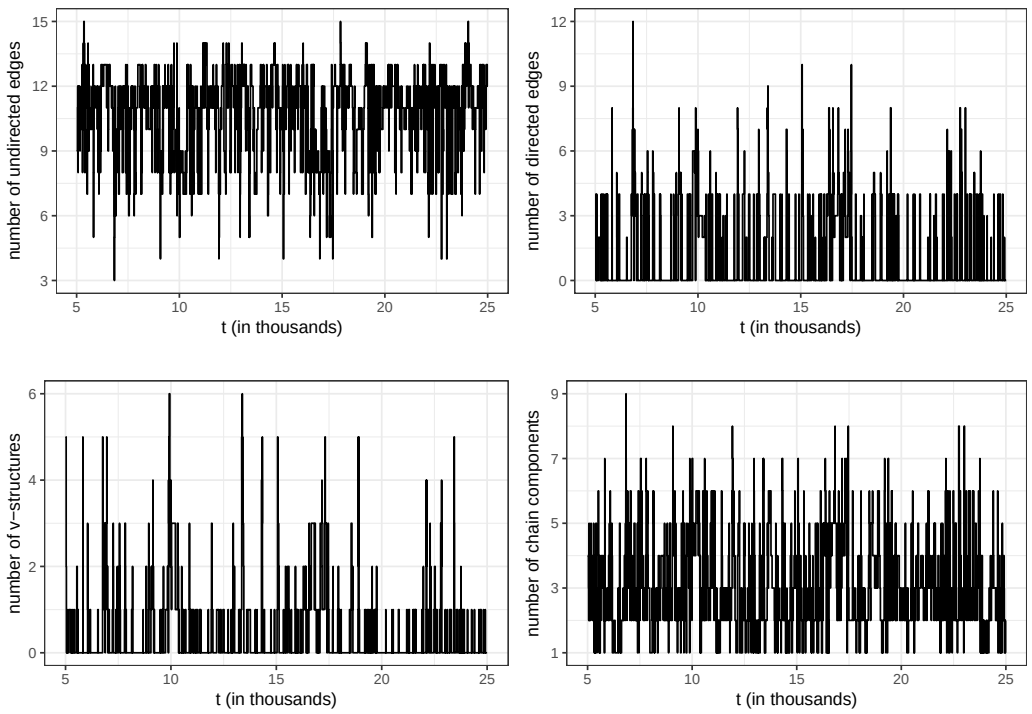

FIGURE 9 Sachs data. First dataset. MCMC trace plots of four EG features: number of undirected and directed edges, *v*-structures and chain components.

Group	Method	MOBES	OBES	GES 0	GES 0.5	GES 1
	MOBES	10	0	8	3	3
	OBES		10	8	3	3
1	GES 0			11	5	5
	GES 0.5				9	0
	GES 1					9
	MOBES	10	1	4	4	4
	OBES		9	5	3	3
2	GES 0			11	8	8
	GES 0.5				8	0
	GES 1					8
	MOBES	11	1	4	1	2
	OBES		10	3	0	1
3	GES 0			11	3	4
	GES 0.5				10	1
	GES 1					9
	MOBES	10	0	1	0	1
	OBES		10	1	0	1
4	GES 0			11	1	2
	GES 0.5				10	1
	GES 1					9
	MOBES	10	3	5	3	2
	OBES		9	4	0	1
	GES 0			11	4	5
5	GES 0.5				9	1
	GES 1					8

TABLE 3 Sachs data. Structural Hamming distances between graphs estimated with the five methods under comparison for each dataset (group); number of edges in the estimated graphs are reported on the main diagonal of each sub-table. The five methods under comparison are: our Multiple Objective Bayes Essential graph Search (MOBES), the Objective Bayes Essential graph Search (OBES) and the Greedy Equivalence Search (GES) computed for three different optimization criteria (BIC and EBIC with tuning parameter $\gamma \in \{0.5, 1\}$).

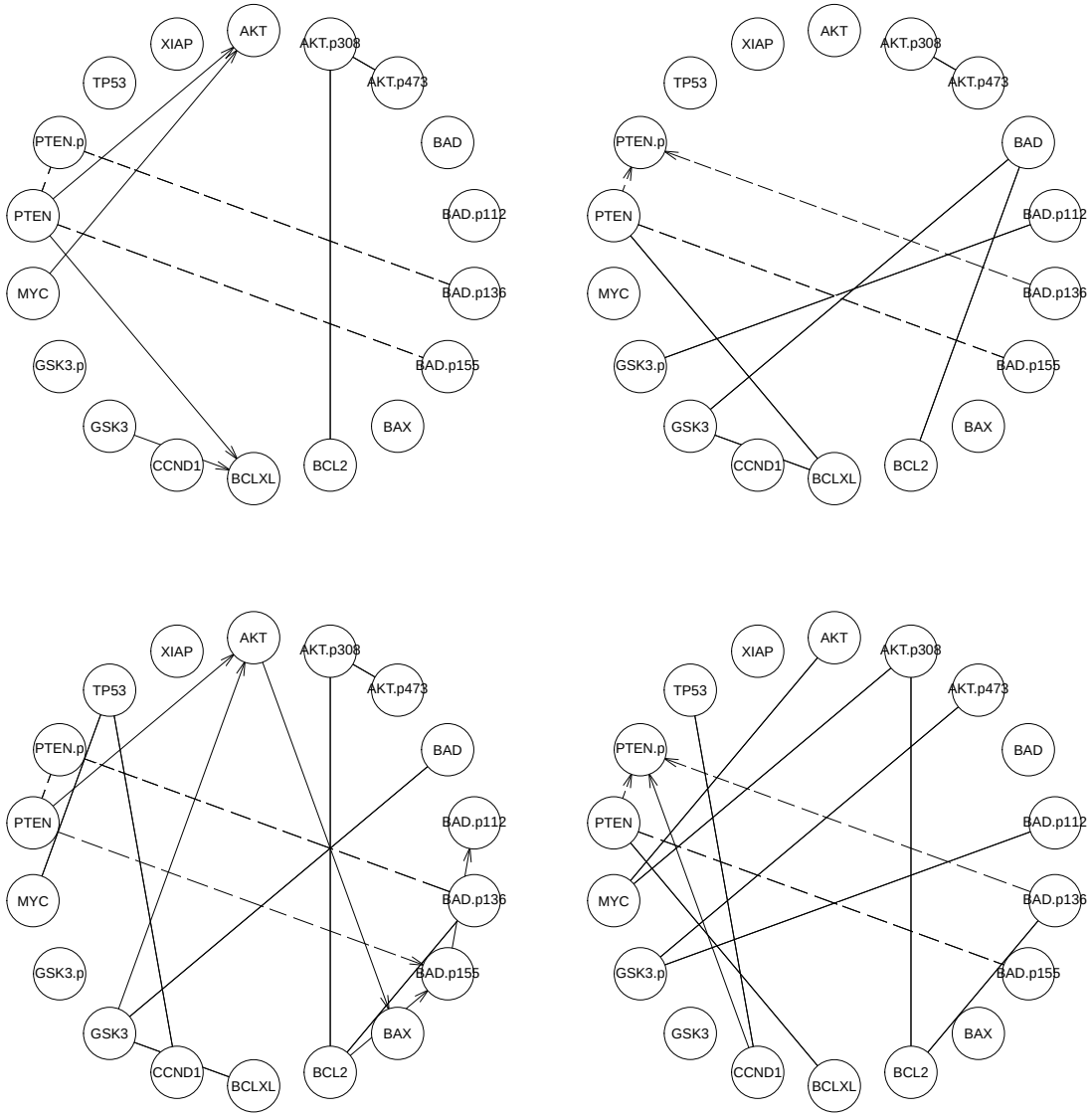

FIGURE 10 Leukemia data. Estimated EGs for subject subtypes, from top-left to bottom-right, M0, M1, M2 and M4.

Group	Method	MOBES	OBES	GES 0	GES 0.5	GES 1
	MOBES	9	9	32	11	8
	OBES		0	31	8	1
M0	GES 0			31	24	31
	GES 0.5				8	8
	GES 1					1
	MOBES	9	10	32	11	7
	OBES		2	35	12	3
M1	GES 0			34	28	32
	GES 0.5				11	9
	GES 1					2
	MOBES	15	13	22	7	13
	OBES		5	28	10	0
M2	GES 0			30	20	28
	GES 0.5				14	10
	GES 1					5
	MOBES	12	8	24	8	7
	OBES		4	26	9	1
M4	GES 0			28	18	26
	GES 0.5				12	9
	GES 1					5

TABLE 4 Leukemia data. Structural Hamming distances between graphs estimated with the five methods under comparison for each dataset (group); number of edges in the estimated graphs are reported on the main diagonal of each sub-table.