Bayesian Analysis (2018) **13**, Number 4, pp. 1235–1260

Learning Markov Equivalence Classes of Directed Acyclic Graphs: An Objective Bayes Approach

Federico Castelletti*, Guido Consonni†, Marco L. Della Vedova‡, and Stefano Peluso§

Abstract. A Markov equivalence class contains all the Directed Acyclic Graphs (DAGs) encoding the same conditional independencies, and is represented by a Completed Partially Directed Acyclic Graph (CPDAG), also named Essential Graph (EG). We approach the problem of model selection among noncausal sparse Gaussian DAGs by directly scoring EGs, using an objective Bayes method. Specifically, we construct objective priors for model selection based on the Fractional Bayes Factor, leading to a closed form expression for the marginal likelihood of an EG. Next we propose a Markov Chain Monte Carlo (MCMC) strategy to explore the space of EGs using sparsity constraints, and illustrate the performance of our method on simulation studies, as well as on a real dataset. Our method provides a coherent quantification of inferential uncertainty, requires minimal prior specification, and shows to be competitive in learning the structure of the data-generating EG when compared to alternative state-of-the-art algorithms.

Keywords: Bayesian model selection, CPDAG, essential graph, fractional Bayes factor, graphical model.

1 Introduction

Graphical models based on Directed Acyclic Graphs (DAGs) are widely used to represent dependency relationships among potentially many variables; see Lauritzen (1996), Cowell et al. (1999), Koller and Friedman (2009). Applications of DAG models in various scientific areas abound, especially in genomics; see for instance Friedman (2004), Sachs et al. (2005), Shojaie and Michailidis (2009), Nagarajan and Scutari (2013). DAGs are also used for causal inference; see Pearl (2000) for a scholarly treatment and Pearl (2003) for a more expository overview. For the benefit of the reader we summarize in Section 2.1 the basics of graph theory used in this paper.

Within a noncausal framework, a DAG encodes conditional independencies of the variables in the graph determined using the notion of d-separation (Pearl, 2000). Under faithfulness, these independencies are exactly those entailed by the joint distribution of the variables. However, it is well known that different DAGs can encode the same set of conditional independencies, and thus one cannot distinguish between DAGs using observational data; see Chickering (2002). All DAGs encoding the same conditional independencies form a Markov equivalence class, which can be represented by a completed

*Università Cattolica del Sacro Cuore, Milan, Italy, federico.castelletti@unicatt.it
†Università Cattolica del Sacro Cuore, Milan, Italy, guido.consonni@unicatt.it
‡Università Cattolica del Sacro Cuore, Milan, Italy, marco.dellavedova@unicatt.it
§Università Cattolica del Sacro Cuore, Milan, Italy, stefano.peluso@unicatt.it

 https://doi.org/10.1214/18-BA1101

partially directed acyclic graph (CPDAG) (Chickering, 2002), also called essential graph (EG) by Andersson et al. (1997a). An EG is a particular chain graph (CG) whose chain components are decomposable undirected graphs (UG) linked by arrowheads; see Lauritzen (1996) for all relevant graph-theoretic definitions. Typically the structure of a DAG governing the joint distribution of the observations is unknown; accordingly our goal, from a noncausal perspective, is to learn the underlying EG.

Although there are fewer EGs than DAGs, their number still increases super-exponentially with the number of vertices (Gillispie and Perlman, 2002). For this reason, structural learning in the space of EGs has been confined to small graphs. An early paper which explores the space of EGs by means of Markov chains is Madigan et al. (1996), followed by Castelo and Perlman (2004), and more recently by Sonntag et al. (2015). He et al. (2013) propose a reversible irreducible Markov chain for *sparse* EGs, having fewer edges than a small multiple of the number of vertices.

From a statistical perspective, learning an EG is a problem in model selection which we tackle through the Bayes factor (Kass and Raftery, 1995), and adopting an objective Bayes (OB) approach, requiring minimal input from the user; see Berger and Pericchi (1996) and Pericchi (2005) for an overview. Additionally, we rely on a method for the construction of parameter priors for the selection of DAGs, originally laid out in Geiger and Heckerman (2002), subsequently revisited and implemented from an OB perspective in Consonni and La Rocca (2012) for Gaussian DAG models, and extended to the multivariate regression setting in Consonni et al. (2017).

The contribution of this paper is twofold: i) we enhance the OB methodology for sparse DAG selection by learning directly the Markov equivalence class generating the observations, that is the corresponding EG; ii) we develop an MCMC strategy to explore the space of EGs.

The rest of this paper is organized as follows. Section 2 contains some background material on EGs, on objective priors for model selection (with special emphasis on the fractional Bayes factor), and on recent results on marginal likelihoods for Gaussian multivariate regression DAG models with objective priors. The latter are used in Section 3 to compute the marginal likelihood of an EG, whilst Section 4 contains a detailed description of the MCMC method adopted to explore the space of EGs. Section 5 applies the methodology to a few simulation settings and to the analysis of the protein-signaling data presented in Sachs et al. (2005). Finally, Section 6 presents some points for discussion.

2 Background

2.1 Essential graphs

We consider a multivariate setting comprising q variables, wherein interest centers on the dependencies among such variables. The data are represented by q-dimensional i.i.d. observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ from a parametric family of sampling distributions, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})^\top$, $i = 1, \ldots, n$. We use graphs to encode dependencies. To this end, we

provide below some basic graph terminology which is useful to understand our method; further notions and details may be found for instance in the book of Lauritzen (1996).

A graph $\mathcal{G}$ is a pair (V, E), where V is a finite set of *vertices*, and $E \subseteq V \times V$ is a set of *edges*. If $(u, v) \in E$, but $(v, u) \notin E$, we say that $\mathcal{G}$ contains the directed edge $u \to v$. If $(u, v) \in E$, and $(v, u) \in E$, we say that $\mathcal{G}$ contains the undirected edge $u - v$. We assume that $\mathcal{G}$ contains no loop $u - u$. An undirected graph (UG) contains only edges of type $u - v$, while a directed graph contains only directed edges. A *path* in $\mathcal{G}$ is a sequence of vertices $v_0, v_1, \ldots, v_m$ such that, for all $i = 1, \ldots, m$, either $\mathcal{G}$ contains $v_{i-1} - v_i$ or $\mathcal{G}$ contains $v_{i-1} \to v_i$. A path is undirected when it only consists of undirected edges, otherwise it is semidirected. A *cycle* is a path such that $v_0 = v_m$. If a directed graph has no cycles, then it is a directed acyclic graph (DAG), and we denote it by $\mathcal{D}$. A chain graph (CG) is a graph that may contain both directed and undirected edges, but is *adicyclic*, that is, it has no semidirected cycles. A *consistent extension* of a CG $\mathcal{G}$ is a DAG on the same underlying set of edges, with the same orientations on the directed edges of $\mathcal{G}$ and the same set of v-structures (Dor and Tarsi, 1992). Two distinct vertices of a CG $\mathcal{G}$ belong to the same *chain component* when they are joined by an undirected path. Let $\mathcal{T}$ denote the set of chain components of $\mathcal{G}$. Because of adicyclicity, we can regard $\mathcal{T}$ as the "vertex" set of a DAG containing $\tau \to \tau'$, $\tau \in \mathcal{T}$, $\tau' \in \mathcal{T}$ if and only if $\mathcal{G}$ contains $u \to v$ for some $u \in \tau$ and $v \in \tau'$. We denote with $\mathcal{G}_A = (A, E_A)$, $A \subseteq V$, the *subgraph* of $\mathcal{G} = (V, E)$ induced by A, whose edge set is $E_A = E \bigcap A \times A$. The subgraph of $\mathcal{G}$ induced by any given chain component $\tau \in \mathcal{T}$, $\mathcal{G}_\tau$, is an UG.

Each of the q variables is associated to a vertex of a graph $\mathcal{G}$ whose structure will constrain the distribution of each observation $\boldsymbol{y}_i$. More specifically, we assume that the distribution of $\boldsymbol{y}_i$ satisfies conditional independencies which are all encoded in the graph (Markov property determined by $\mathcal{G}$); see Lauritzen (1996, sect. 3.2). The resulting sampling family is called a graphical model, which for simplicity we may still label as $\mathcal{G}$.

Let $\mathcal{D}$ be a DAG with vertex set V, and let $[\mathcal{D}]$ denote its Markov equivalence class, that is, the set of all DAGs with vertex set V that determine the same graphical model. It is known that $\mathcal{D}' \in [\mathcal{D}]$ if and only if $\mathcal{D}$ and $\mathcal{D}'$ have the same skeleton (that is are equal as UG) and immoralities (induced subgraphs of the form $u \to v \leftarrow z$); see Verma and Pearl (1991). Additionally, the class $[\mathcal{D}]$ is uniquely determined by the EG $\mathcal{D}^\star = \cup\{\mathcal{D}' \,|\, \mathcal{D}' \in [\mathcal{D}]\}$, 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 et al. (1997a) is the characterization of those graphs that may occur as an EG.

Theorem 1. (Andersson et al., 1997a, Thm. 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 in Definition 3.3 of Andersson et al., 1997a)*

From the theory presented in Andersson et al. (2001) and Drton and Eichler (2006), the joint density of $\boldsymbol{y}_i$ relative to the CG $\mathcal{G}$ factorizes as

$$f_{\mathcal{G}}(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_{\mathcal{G}}) = \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_\tau}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}), \tag{1}$$

where $\boldsymbol{y}_{i,\tau} = (y_{ij},\, j \in \tau)^\top$ denotes the subvector of $\boldsymbol{y}_i$ whose components are indexed by the vertices in $\tau \subseteq V$, and similarly for $\boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}$, with $\mathrm{pa}_{\mathcal{G}}(\tau)$ the *parents* of τ in $\mathcal{G}$, i.e., the set of all $u \in V$ such that $u \to v$ is contained in $\mathcal{G}$ for some $v \in \tau$. In expression (1) $\boldsymbol{\theta}_{\mathcal{G}}$ is the global parameter indexing the graphical model, whereas $\boldsymbol{\theta}_{\mathcal{G}_\tau}$ is a local parameter indexing the conditional sampling distribution of $\boldsymbol{y}_{i,\tau}$ given $\boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}$. If we let $\boldsymbol{\theta}_{\mathcal{G}} \in \boldsymbol{\Theta}_{\mathcal{G}}$ and $\boldsymbol{\theta}_{\mathcal{G}_\tau} \in \boldsymbol{\Theta}_{\mathcal{G}_\tau}$, we find $\boldsymbol{\Theta}_{\mathcal{G}} = \times_{\tau \in \mathcal{T}} \boldsymbol{\Theta}_{\mathcal{G}_\tau}$, i.e., the components $\boldsymbol{\theta}_{\mathcal{G}_\tau}$s of $\boldsymbol{\theta}_{\mathcal{G}}$ are variation independent (Drton and Eichler, 2006).

In the sequel we collect all n observations into a single $n \times q$ matrix $\boldsymbol{Y}$ by stacking the n row-vectors $\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top$:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1^\top \\ \vdots \\ \boldsymbol{y}_n^\top \end{pmatrix}. \tag{2}$$

A similar definition holds for the $n \times |\tau|$ matrix $\boldsymbol{Y}_\tau$, and for the $n \times |\mathrm{pa}_{\mathcal{G}}(\tau)|$ matrix $\boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}$, where $|\tau|$ is the cardinality of τ, and similarly for $|\mathrm{pa}_{\mathcal{G}}(\tau)|$.

Recall that the observations, conditionally on $\boldsymbol{\theta}_{\mathcal{G}}$, are i.i.d.; whence

$$\begin{aligned} f_{\mathcal{G}}(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_{\mathcal{G}}) &= \prod_{i=1}^{n} \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_\tau}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) \\ &= \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}). \end{aligned} \tag{3}$$

Since the $\boldsymbol{\theta}_\tau$s are variation independent, we can further assume that the prior on $\boldsymbol{\theta}_{\mathcal{G}}$ factorizes as

$$p(\boldsymbol{\theta}_{\mathcal{G}}) = \prod_{\tau \in \mathcal{T}} p(\boldsymbol{\theta}_{\mathcal{G}_\tau}); \tag{4}$$

see also Castelo and Perlman (2004). Condition (4) extends the assumption of global (parameter) independence, which is typical for DAG models (Cowell et al., 1999, p. 193), to CG models. In this way we obtain

$$\begin{aligned} m_{\mathcal{G}}(\boldsymbol{Y}) &= \int_{\boldsymbol{\Theta}_{\mathcal{G}}} f_{\mathcal{G}}(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_{\mathcal{G}}) p(\boldsymbol{\theta}_{\mathcal{G}}) d\boldsymbol{\theta}_{\mathcal{G}} \\ &= \prod_{\tau \in \mathcal{T}} \int_{\boldsymbol{\Theta}_{\mathcal{G}_\tau}} f_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) p(\boldsymbol{\theta}_{\mathcal{G}_\tau}) d\boldsymbol{\theta}_{\mathcal{G}_\tau} \\ &= \prod_{\tau \in \mathcal{T}} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}), \end{aligned} \tag{5}$$

so that the marginal distribution for the data matrix admits the same CG factorization that holds under the sampling distribution (3).

2.2 Fractional marginal likelihoods

We assume that the reader is familiar with the basic concepts of model selection from the Bayesian perspective, as described for instance in O'Hagan and Forster (2004, ch. 7). Here we provide some background on *objective Bayes* model selection.

Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be a collection of Bayesian models for the data matrix $\boldsymbol{Y}$. Each model $\mathcal{M}_k$, $k = 1, \ldots, K$, consists of a family of sampling densities $f_{\mathcal{M}_k}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k)$, indexed by a model specific parameter $\boldsymbol{\theta}_k$, and of a prior density $p(\boldsymbol{\theta}_k)$ on $\boldsymbol{\theta}_k$, which we assume to be *proper*. We focus on the computation of $m_{\mathcal{M}_k}(\boldsymbol{Y}) = \int f_{\mathcal{M}_k}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k$, the marginal density of $\boldsymbol{Y}$ under $\mathcal{M}_k$, also known as the marginal likelihood of $\mathcal{M}_k$. We set $p(\boldsymbol{\theta}_k) = p^D(\boldsymbol{\theta}_k)$, where the latter is some default objective parameter prior, such as the Jeffreys' prior or the more general reference prior (Berger et al., 2009). However, objective priors are often improper and they cannot be naively used to compute marginal likelihoods, even when the result is finite and non-zero, because of the presence of arbitrary constants which do not cancel out in their ratios; see Pericchi (2005) for a review of several proposals to address this issue. In this paper, we adopt the fractional Bayes factor originally introduced by O'Hagan (1995).

Let $b = b(n)$, $0 < b < 1$, be a fraction of the number of observations n. Define the *fractional marginal likelihood* of model $\mathcal{M}_k$ as

$$m_{\mathcal{M}_k}(\boldsymbol{Y}; b) = \frac{\int f_{\mathcal{M}_k}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p^D(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k}{\int f^b_{\mathcal{M}_k}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p^D(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k}, \tag{6}$$

where $f^b_{\mathcal{M}_k}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) = (f_{\mathcal{M}_k}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k))^b$ is the sampling density under model $\mathcal{M}_k$ raised to the b-th power, and the two integrals are assumed to be finite and non-zero. Equation (6) can be rewritten as

$$m_{\mathcal{M}_k}(\boldsymbol{Y}; b) = \int f^{1-b}_{\mathcal{M}_k}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p^F(\boldsymbol{\theta}_k \mid b, \boldsymbol{Y}) d\boldsymbol{\theta}_k,$$

where $p^F(\boldsymbol{\theta}_k \mid b, \boldsymbol{Y}) \propto f^b_{\mathcal{M}_k}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p^D(\boldsymbol{\theta}_k)$ is the implied *fractional prior* (actually a "posterior" based on the *fractional likelihood* and the default prior). Notice that the fractional likelihood utilizes all the data, and not part of the data; it is actually a discounted full likelihood.

Usually b is chosen to be small, so that the dependence of the prior on the data will be weak. Model selection consistency is achieved provided $b \to 0$ for $n \to \infty$. (O'Hagan, 1995, sect. 4). A default choice is $b = n_0/n$, where n_0 is the minimal (integer) training sample size which makes the induced fractional prior proper. Other choices are possible, but Moreno (1997) argues in favor of the default choice, and we follow suit. In the sequel we simply write $m_{\mathcal{M}_k}(\boldsymbol{Y})$ when the choice of b is understood.

2.3 Gaussian multivariate regression DAG models

We say that the random matrix $\boldsymbol{Y}$ follows the *matrix normal distribution* with mean matrix $\boldsymbol{M}$, row covariance matrix $\boldsymbol{\Phi}$, and column covariance matrix $\boldsymbol{\Sigma}$, written

$$\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma} \sim \mathcal{N}_{n,q}(\boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}), \tag{7}$$

when $\text{vec}(\boldsymbol{Y})$ follows the multivariate normal distribution with mean vector $\text{vec}(\boldsymbol{M})$ and covariance matrix $\boldsymbol{\Sigma} \otimes \boldsymbol{\Phi}$; see Gupta and Nagar (2000, p. 55), and Dawid (1981) for more details. For any two matrices $\mathbf{A}$ and $\mathbf{B}$, $\text{vec}(\mathbf{A})$ denotes the column vector obtained by stacking the columns of $\mathbf{A}$, while $\mathbf{A} \otimes \mathbf{B}$ is the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$.

Let $\boldsymbol{\Omega}$ be a $q \times q$ *unconstrained* s.p.d. random matrix. We will write $\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{R})$ to mean that $\boldsymbol{\Omega}$ follows a *Wishart distribution* with density

$$p(\boldsymbol{\Omega}) = \frac{1}{2^{\frac{aq}{2}} \Gamma_q(\frac{a}{2})} |\boldsymbol{R}|^{\frac{a}{2}} |\boldsymbol{\Omega}|^{\frac{a-q-1}{2}} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{\Omega}\boldsymbol{R})\right\}, \tag{8}$$

when $\boldsymbol{\Omega}$ s.p.d., and $p(\boldsymbol{\Omega}) = 0$, otherwise. In (8) $\boldsymbol{R}$ is a $q \times q$ s.p.d. matrix, a is a scalar strictly greater than $q-1$, and $\Gamma_q(\frac{a}{2}) = \pi^{\frac{q(q-1)}{4}} \prod_{j=1}^{q} \Gamma(\frac{a}{2} + \frac{1-j}{2})$ is the q-dimensional gamma function evaluated at $a/2$ (generalizing $\Gamma(a/2) = \int_0^\infty z^{\frac{a}{2}-1} e^{-z} dz$).

Consonni et al. (2017) describe an objective Bayes method for model selection within the class of Gaussian multivariate regression DAG models

$$\boldsymbol{Y} \,|\, \boldsymbol{B}, \boldsymbol{\Omega}_{\mathcal{D}} \sim \mathcal{N}_{n,q}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_n, \boldsymbol{\Omega}_{\mathcal{D}}^{-1}), \tag{9}$$

where $\boldsymbol{Y}$ is the $n \times q$ matrix of observations on the responses, $\boldsymbol{X}$ the $n \times (p+1)$ design matrix of the observations on the p exogenous variables (plus an additional column vector with all entries equal to 1 to account for the intercept term), $\boldsymbol{B}$ the $(p+1) \times q$ matrix of unconstrained regression parameters, and $\boldsymbol{\Omega}_{\mathcal{D}}$ the $q \times q$ precision matrix (inverse of the covariance matrix) assumed to be Markov with respect to a DAG $\mathcal{D}$. This means that, if there is no directed edge from vertex u to vertex v, then $\rho_{uv \cdot \{1,\ldots,v\}} = 0$, $(1 \leq u < v \leq q)$ in any well-numbering of the vertices, where $\rho_{uv \cdot K}$ is the sampling partial correlation between y_{iu} and y_{iv} given $(y_{ik} \,|\, k \in K \setminus \{u, v\})$ for $\{u, v\} \subseteq K \subseteq V$; see Drton and Perlman (2008, formula (2.7)). Alternative ways of expressing the Markov property of a Gaussian DAG model are available, e.g., in terms of the Choleski decomposition of $\boldsymbol{\Omega}_{\mathcal{D}}$.

To start with, assume that the DAG $\mathcal{D}$ in (9) is *complete*, i.e., all pairs of edges are present, so that there are no conditional independencies among the q responses. In this case, the precision matrix, which we denote simply with $\boldsymbol{\Omega}$, is symmetric and positive definite (s.p.d.) but otherwise unconstrained. The methodology presented in Geiger and Heckerman (2002) for the construction of parameter priors under DAG models was employed in Consonni and La Rocca (2012) for Gaussian DAGs, and extended by Consonni et al. (2017) to the Gaussian multivariate regression setup, leading to an objective Bayes methodology for graphical model selection based on the fractional Bayes factor. The key-point is that the fractional prior on $(\boldsymbol{B}, \boldsymbol{\Omega})$ is conjugate to the likelihood under the *complete* DAG model. A very important point to be noticed is that, when the likelihood is written recursively as the product of the conditional density of each node given its predecessors, the local parameters indexing each conditional density become stochastically independent under the fractional prior, allowing the application of the methodology of Geiger and Heckerman (2002) for obtaining the prior under *any* DAG model, and eventually its marginal likelihood.

Of special interest for this paper is the case in which the precision matrix in (9) is Markov with respect to a decomposable UG $\mathcal{G}$. Since the class of decomposable UG models is strictly smaller than the class of DAG models (Andersson et al., 1997b) the above fractional Bayes factor methodology allows to compute also the marginal likelihood of $\mathcal{G}$,

$$m_{\mathcal{G}}(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\prod_{C\in\mathcal{C}} m(\boldsymbol{Y}_C|\boldsymbol{X})}{\prod_{S\in\mathcal{S}} m(\boldsymbol{Y}_S|\boldsymbol{X})}, \tag{10}$$

see Consonni et al. (2017, formula (28)), where conditioning on $\boldsymbol{X}$ is tacitly assumed. In (10) $\mathcal{C}$ is the set of cliques, and $\mathcal{S}$ the set of separators, of the decomposable graph $\mathcal{G}$ (Lauritzen, 1996), while $\boldsymbol{Y}_C$ is the submatrix of responses belonging to $C \in \mathcal{C}$, with a similar interpretation for $\boldsymbol{Y}_S$. It is crucial to remark that, in the right-hand-side of (10), $m(\boldsymbol{Y}_C|\boldsymbol{X})$ is computed under any complete graph (hence the lack of a subscript) and similarly for $m(\boldsymbol{Y}_S|\boldsymbol{X})$.

Expression (10) will be used in Subsection 3.2 to obtain the marginal likelihood of an EG.

3 Gaussian essential graphs

3.1 Likelihood factorization

We consider observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ which, conditionally on their mean vector $\boldsymbol{\mu}$ and their precision matrix $\boldsymbol{\Omega}_{\mathcal{D}}$ are i.i.d. $\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}_{\mathcal{D}}^{-1})$, with $\boldsymbol{\Omega}_{\mathcal{D}}$ Markov with respect to a DAG $\mathcal{D}$.

Now consider the EG $\mathcal{G}$ for the equivalence class of $\mathcal{D}$, and the factorization in the first display of (3). It is easy to verify that

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

or equivalently, letting $\boldsymbol{\alpha}_\tau = \boldsymbol{\mu}_\tau - \boldsymbol{\Gamma}_\tau \boldsymbol{\mu}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}$,

$$f_{\mathcal{G}_\tau}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) = \mathcal{N}_{|\tau|}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{\alpha}_\tau + \boldsymbol{\Gamma}_\tau \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}), \tag{12}$$

where $\boldsymbol{\mu}_\tau = \mathbb{E}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{\mu}, \boldsymbol{\Omega}_{\mathcal{D}})$, $\boldsymbol{\Gamma}_\tau$ is the matrix of regression parameters and $\boldsymbol{\Omega}_{\mathcal{G}_\tau}$ is the conditional precision matrix, i.e. $\boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1} = \mathbb{V}\mathrm{ar}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Omega}_{\mathcal{G}_\tau})$.

Collecting terms we can write

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

where

$$\boldsymbol{x}_{i,\tau} = \begin{bmatrix} 1 \\ \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)} \end{bmatrix}; \quad \boldsymbol{B}_\tau = \begin{bmatrix} \boldsymbol{\alpha}_\tau^\top \\ \boldsymbol{\Gamma}_\tau^\top \end{bmatrix}. \tag{14}$$

Notice that the matrix $\boldsymbol{B}_\tau$ consists of unconstrained components; this happens because the EG $\mathcal{G}$ has no flags (Theorem 1).

Letting

$$\boldsymbol{X}_\tau = \begin{pmatrix} \boldsymbol{x}_{1,\tau}^\top \\ \vdots \\ \boldsymbol{x}_{n,\tau}^\top \end{pmatrix}, \tag{15}$$

we can write

$$\boldsymbol{Y}_\tau \,|\, \boldsymbol{X}_\tau, \boldsymbol{B}_\tau, \boldsymbol{\Omega}_{\mathcal{G}_\tau} \sim \mathcal{N}_{n,|\tau|}(\boldsymbol{X}_\tau \boldsymbol{B}_\tau, \boldsymbol{I}_n, \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}), \tag{16}$$

so that

$$f_{\mathcal{G}}(\boldsymbol{Y} \,|\, \boldsymbol{\mu}, \boldsymbol{\Omega}_{\mathcal{G}}) = \prod_{\tau \in \mathcal{T}} \mathcal{N}_{n,|\tau|}(\boldsymbol{Y}_\tau \,|\, \boldsymbol{X}_\tau \boldsymbol{B}_\tau, \boldsymbol{I}_n, \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}), \tag{17}$$

where $\boldsymbol{X}$ is the collection (column binding) of the $\boldsymbol{X}_\tau$s.

3.2 Marginal likelihood

Formula (17) shows that the *formal* structure of each term indexed by τ is that of a multivariate Gaussian regression model whose precision matrix is Markov w.r.t. a decomposable UG; this is the setting described in Section 2.3. We now detail the calculations leading to the marginal likelihood $m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \,|\, \boldsymbol{X}_\tau)$.

Because of global parameter independence (4), we only need to specify priors separately under each chain component τ. Let $\boldsymbol{\Omega}_\tau$ denote the precision matrix of the variables in τ under a complete graph. A default prior on $(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau)$, with $\boldsymbol{\Omega}_\tau$ s.p.d., is

$$p^D(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau) \propto |\boldsymbol{\Omega}_\tau|^{\frac{a_D - |\tau| - 1}{2}}, \tag{18}$$

which is flexible enough to accommodate different default choices. In particular, setting $a_D = |\tau| - 1$, gives $p^D(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau) \propto |\boldsymbol{\Omega}_\tau|^{-1}$, which is the prior discussed in Geisser and Cornfield (1963).

Using the prior (18), and setting the fraction b equal to n_0/n, $n_0 < n$, the fractional prior for model (16) is given by

$$\begin{aligned} p^F(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau) \quad &\propto \quad |\boldsymbol{\Omega}_\tau|^{\frac{a_D + n_0 - |\tau| - 1}{2}} \\ &\cdot \quad e^{-\frac{n_0}{2} \operatorname{tr}\left(\boldsymbol{\Omega}_\tau \left\{ (\boldsymbol{B}_\tau - \hat{\boldsymbol{B}}_\tau)^\top \tilde{\boldsymbol{C}}_\tau (\boldsymbol{B}_\tau - \hat{\boldsymbol{B}}_\tau) + \tilde{\boldsymbol{R}}_\tau \right\}\right)}, \end{aligned} \tag{19}$$

where $\hat{\boldsymbol{B}}_\tau = (\boldsymbol{X}_\tau^\top \boldsymbol{X}_\tau)^{-1} \boldsymbol{X}_\tau^\top \boldsymbol{Y}_\tau$, $\hat{\boldsymbol{E}}_\tau = (\boldsymbol{Y}_\tau - \boldsymbol{X}_\tau \hat{\boldsymbol{B}}_\tau)$, $\tilde{\boldsymbol{C}}_\tau = n^{-1} \boldsymbol{X}_\tau^\top \boldsymbol{X}_\tau$, and $\tilde{\boldsymbol{R}}_\tau = n^{-1} \hat{\boldsymbol{E}}_\tau^\top \hat{\boldsymbol{E}}_\tau$. Formula (19) can be expressed as

$$\boldsymbol{B}_\tau \,|\, \boldsymbol{\Omega}_\tau \sim \mathcal{N}_{|\mathrm{pa}_{\mathcal{G}}(\tau)|+1,|\tau|}\big(\hat{\boldsymbol{B}}_\tau, (n_0 \tilde{\boldsymbol{C}}_\tau)^{-1}, \boldsymbol{\Omega}_\tau^{-1}\big), \tag{20a}$$

$$\boldsymbol{\Omega}_\tau \sim \mathcal{W}_{|\tau|}\big(a_D + n_0 - |\mathrm{pa}_{\mathcal{G}}(\tau)| - 1, n_0 \tilde{\boldsymbol{R}}_\tau\big). \tag{20b}$$

The prior characterized by the density in (19), or by the hierarchical structure in (20), belongs to the family of *matrix normal Wishart* distributions, and is conjugate to the sampling model (16). It is proper under two conditions: i) $a_D + n_0 - |\mathrm{pa}_{\mathcal{G}}(\tau)| > |\tau|$; ii) $n > |\tau| + |\mathrm{pa}_{\mathcal{G}}(\tau)|$; see Consonni et al. (2017). Condition ii) is a sparsity condition on the graph structure. Condition i) becomes $n_0 > |\mathrm{pa}_{\mathcal{G}}(\tau)| + 1$, upon setting $a_D = |\tau| - 1$.

As described below in (22), to compute the marginal density of $\boldsymbol{Y}_\tau$ given $\boldsymbol{X}_\tau$ we need an expression for the density of selected columns of $\boldsymbol{Y}_\tau$.

In general, let $J \subseteq \tau$ and denote with $\boldsymbol{Y}_{J,\tau}$ the submatrix of $\boldsymbol{Y}_\tau$ containing the columns corresponding to the variables in J. Using formula (22) of Consonni et al. (2017), we get

$$\begin{aligned} m_\tau(\boldsymbol{Y}_{J,\tau} \,|\, \boldsymbol{X}_\tau) \;&=\; \pi^{-\frac{(n-n_0)|J|}{2}} \frac{\Gamma_{|J|}\left(\frac{a_D+n-|\mathrm{pa}_{\mathcal{G}}(\tau)|-1-|\bar{J}|}{2}\right)}{\Gamma_{|J|}\left(\frac{a_D+n_0-|\mathrm{pa}_{\mathcal{G}}(\tau)|-1-|\bar{J}|}{2}\right)} \\ &\cdot\; \left(\frac{n_0}{n}\right)^{\frac{|J|(a_D+n_0-|\bar{J}|)}{2}} |\hat{\boldsymbol{E}}_{J,\tau}^\top \hat{\boldsymbol{E}}_{J,\tau}|^{-\frac{n-n_0}{2}}, \end{aligned} \tag{21}$$

where $\bar{J} = \tau \setminus J$, so that $|\bar{J}| = |\tau| - |J|$, and $\hat{\boldsymbol{E}}_{J,\tau} = (\boldsymbol{Y}_{J,\tau} - \boldsymbol{X}_\tau \hat{\boldsymbol{B}}_{J,\tau})$, with $\hat{\boldsymbol{B}}_{J,\tau} = (\boldsymbol{X}_\tau^\top \boldsymbol{X}_\tau)^{-1} \boldsymbol{X}_\tau^\top \boldsymbol{Y}_{J,\tau}$. Recall that $\mathcal{G}_\tau$ is decomposable; let $\mathcal{C}_\tau$ be the set of (maximal) cliques of $\mathcal{G}_\tau$, and let $\mathcal{S}_\tau$ be the corresponding set of separators. Then using (10)

$$m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \,|\, \boldsymbol{X}_\tau) = \frac{\prod_{C \in \mathcal{C}_\tau} m_\tau(\boldsymbol{Y}_{C,\tau} \,|\, \boldsymbol{X}_\tau)}{\prod_{S \in \mathcal{S}_\tau} m_\tau(\boldsymbol{Y}_{S,\tau} \,|\, \boldsymbol{X}_\tau)}. \tag{22}$$

We compute $m_\tau(\boldsymbol{Y}_{C,\tau} \,|\, \boldsymbol{X}_\tau)$ and $m_\tau(\boldsymbol{Y}_{S,\tau} \,|\, \boldsymbol{X}_\tau)$ in (22) by setting $J = C$ and $J = S$, respectively, in (21). Finally, using (5), we can recover the overall marginal distribution of $\boldsymbol{Y}$ under $\mathcal{G}$ by multiplying the terms given in (22)

$$m_{\mathcal{G}}(\boldsymbol{Y}) = \prod_{\tau \in \mathcal{T}} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \,|\, \boldsymbol{X}_\tau), \tag{23}$$

which from a model choice perspective represents the marginal likelihood of $\mathcal{G}$.

4 An MCMC algorithm on equivalence classes of DAGs

In this section we describe in detail an MCMC algorithm to investigate the posterior distribution on the space of essential graphs, which we name EG-space. Because the number of Essential Graphs grows super-exponentially in the number of nodes (Gillispie and Perlman, 2002), and a full enumeration of the EGs is not feasible (Madigan et al., 1996), the posterior probability of each EG is only available up to a normalizing constant. We therefore resort to an MCMC algorithm to approximate the posterior distribution across EGs. Specifically, we first sample an EG from a candidate distribution, which is accepted with a probability given by a Metropolis-Hastings ratio defined to guarantee the convergence of the algorithm to the correct posterior distribution. The starting point of our sampler is He et al. (2013), who propose a reversible irreducible Markov chain on Markov equivalence classes of DAGs. We use their Markov chain as a proposal distribution in a Metropolis-Hastings algorithm whose target is the posterior distribution on the EG-space.

Let $\mathcal{S}_q$ be the set of all EGs having q nodes and $\mathcal{S}$ any subset of $\mathcal{S}_q$. In a sparse setting, $\mathcal{S}$ can be the set of all EGs on q nodes having fewer edges than a specified threshold

M. Let $\mathcal{G} \in \mathcal{S}$ denote an EG belonging to the space $\mathcal{S}$. He et al. (2013) introduce a suitable set of operators that determine the transition from $\mathcal{G}$ to $\mathcal{G}' \in \mathcal{S}$ through a local modification of $\mathcal{G}$. We say that $\mathcal{G}'$ is a *direct successor* of $\mathcal{G}$ if $\mathcal{G}'$ can be reached from $\mathcal{G}$ in a single transition. Given an EG $\mathcal{G}$ they consider six types of operators: inserting an undirected edge (denoted by 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) and converting a v-structure in two adjacent undirected edges (RemoveV). Each operator is then characterized by two features: its type, and the edges it modifies (notice that MakeV and RemoveV may modify two edges). The *modified graph* of an operator on $\mathcal{G}$ is the same as $\mathcal{G}$ except for the modified edges. The modified graph of an operator on $\mathcal{G}$ need not be an EG in general (see He et al. 2013, Supplement). Nevertheless such an operator can still be *valid*, in the sense that it might *result* in a transition to an EG. This is substantially different from other authors, e.g. Madigan et al. (1996), who only allow moves leading directly to an EG, thus significantly reducing the number of possible transitions from each state (an EG) of the chain. In summary: the modified graph is not required to be an EG, but only a chain graph that admits a consistent extension, whilst the direct successor is the EG corresponding to the Markov Equivalence class of such consistent extension, and represents the final output of a Markov chain move. For detailed examples we refer the reader to He et al. (2013, Supplement).

The collection of operators in He et al. (2013) is *perfect*; this means that such operators induce a Markov chain on a set of EGs guaranteed to have the following desirable properties: a) for any $\mathcal{G}'$ direct successor of $\mathcal{G}$, the modified graph of an operator is a CG with a consistent extension and all modified edges in the modified graph occur in $\mathcal{G}'$ (*validity*), b) there is a unique operator that transforms $\mathcal{G}$ in $\mathcal{G}'$ (*distinguishability*), c) starting from $\mathcal{G}$, there is a positive probability of reaching any other EG in $\mathcal{S}$ via a sequence of operators (*irreducibility*), d) if $\mathcal{G}'$ is a direct successor of $\mathcal{G}$, then $\mathcal{G}$ is also a direct successor of $\mathcal{G}'$ (*reversibility*). Let $\mathcal{O}_{\mathcal{G}}$ be the set of perfect operators on $\mathcal{G}$ and $\mathcal{O} = \cup_{\mathcal{G} \in S} \mathcal{O}_{\mathcal{G}}$. Each operator $o_{\mathcal{G}} \in \mathcal{O}_{\mathcal{G}}$ determines the transition of $\mathcal{G}$ into another EG $\mathcal{G}'$ (one of its direct successors). The Markov chain $\{\mathcal{G}_t\}$ defined on $\mathcal{S}$ is such that the probability of transition from $\mathcal{G}$ to $\mathcal{G}'$ is

$$p_{\mathcal{G},\mathcal{G}'} = 1/|\mathcal{O}_{\mathcal{G}}|, \tag{24}$$

if $\mathcal{G}'$ is a direct successor of $\mathcal{G} \in \mathcal{S}$ and 0 otherwise. It follows from (24) that all direct successors of $\mathcal{G}$ have the same probability of being reached for any given $\mathcal{G}$. Furthermore, we remark that the adoption of different types of operators, as those proposed by Madigan et al. (1996), might result in *extra-connectivity* among the states of the model space (Chickering, 2002), that is in a higher number of direct successors, but it could cause a loss of some of the above-mentioned properties of the Markov chain.

Let $m_{\mathcal{G}}(\boldsymbol{Y})$ be the marginal likelihood for model $\mathcal{G}$; see (23). Additionally, let $p(\mathcal{G})$ be a prior on $\mathcal{G}$, and $q(\cdot \,|\, \mathcal{G})$ a proposal distribution on $\mathcal{S}$ when the chain is in state $\mathcal{G}$. In order to have an appropriate posterior sample of EGs, the transition from $\mathcal{G}$ to $\mathcal{G}'$, is accepted with probability

$$\alpha = \min \left\{ 1; \frac{m_{\mathcal{G}'}(\boldsymbol{Y})}{m_{\mathcal{G}}(\boldsymbol{Y})} \cdot \frac{p(\mathcal{G}')}{p(\mathcal{G})} \cdot \frac{q(\mathcal{G} \,|\, \mathcal{G}')}{q(\mathcal{G}' \,|\, \mathcal{G})} \right\}. \tag{25}$$

In the sequel a proposal is taken to be a step in the Markov chain of He et al. (2013), so that $q(\mathcal{G}' \,|\, \mathcal{G}) = 1/|\mathcal{O}_{\mathcal{G}}|$ in accord with (24)

Briefly, an MCMC algorithm on the space $\mathcal{S}$ can be constructed as follows. Starting from an arbitrary $\mathcal{G}_0$, for $t = 1, \ldots, T$: (1) set $\mathcal{G} = \mathcal{G}_{t-1}$; (2) generate $\mathcal{G}'$ from the proposal $q(\mathcal{G}' \,|\, \mathcal{G})$; (3) compute the probability of acceptance α in (25); (4) update $\mathcal{G}_t = \mathcal{G}'$ with probability α, $\mathcal{G}_t = \mathcal{G}_{t-1}$ with probability $1 - \alpha$.

We assign a prior to $\mathcal{G}$ through a prior on the adjacency matrix of $\mathcal{G}^u$, where $\mathcal{G}^u$ is the skeleton of $\mathcal{G}$ (same edges as in $\mathcal{G}$, but with no orientation):

$$\begin{aligned} \mathcal{G}^u_{(j)} \,|\, \pi &\overset{i.i.d}{\sim} \text{Ber}(\pi), \quad j = 1, \ldots, q(q-1)/2, \\ \pi &\sim \text{Beta}(a, b), \end{aligned} \tag{26}$$

where $\mathcal{G}^u_{(j)}$ is the j-th element of the vectorized lower triangular part of the adjacency matrix of $\mathcal{G}^u$, and $q(q-1)/2$ is the maximum number of edges in an EG on q nodes. A similar prior on the space of decomposable UG was also implemented in Bhadra and Mallick (2013) in the context of covariate-adjusted graphical model selection. Notice that the prior on $\mathcal{G}$ only depends on the skeleton of the graph: two EGs with the same number of edges (directed or undirected) will be assigned the same prior probability. Alternative priors, specifically targeted to EGs, are not available in the literature to our knowledge, and would be beyond the scope of the present paper. The ratio between the prior assigned to the proposed EG $\mathcal{G}'$ and the current $\mathcal{G}$ is then

$$\frac{p(\mathcal{G}')}{p(\mathcal{G})} = \frac{\Gamma(|\mathcal{G}'| + a)}{\Gamma(|\mathcal{G}| + a)} \cdot \frac{\Gamma\left(\frac{q(q-1)}{2} - |\mathcal{G}'| + b\right)}{\Gamma\left(\frac{q(q-1)}{2} - |\mathcal{G}| + b\right)}, \tag{27}$$

where $|\mathcal{G}|$ denotes the number of edges in $\mathcal{G}$. A common choice is $a = b = 1$ so that $\pi \sim \text{Unif}(0, 1)$. However, to *favor* sparsity, we can set $a < b$, so that $\mathbb{E}(\pi) < 0.5$. Our choice is $a = 1$, $b = (2q-2)/3-1$, whence $\mathbb{E}(\pi) = 3/(2q-2)$, which resembles the sparse simulation setting as defined in Peters and Bühlmann (2014). We note however that all the results given below are generally insensitive to the choice of a and b, because the ratio of marginal likelihoods $m_{\mathcal{G}'}(\boldsymbol{Y})/m_{\mathcal{G}}(\boldsymbol{Y})$ is by far the leading factor in the acceptance probability of the proposed EG given in (25). Since the marginal likelihood of the EG appears to be the driving force in the MCMC algorithm, one can reasonably expect that results will also be insensitive to prior specifications on graph space alternative to (26).

5 Simulations and real data analysis

In the present section we compare our methodology, which henceforth we name *Objective Bayes Essential graph Search* (OBES) for easier reference, with a few benchmark methods, namely: i) Greedy Equivalence Search (GES); ii) PC algorithm (PC); iii) Greedy DAG Search (GDS). The GES algorithm is a search-and-score method which provides an estimate of the true EG using the greedy equivalence search algorithm of Chickering (2002). Through additions and deletions of single edges, GES maximizes a score function in the space of the EGs, with a modification introduced by Hauser and

Bühlmann (2012) to improve estimation performance. The PC algorithm is a constraint-based method (Spirtes et al., 2000) which outputs an estimate of the true EG using a sequence of conditional independence tests. The skeleton is estimated using an order-independent modified version (Colombo and Maathuis, 2014) of the original PC algorithm, and the edges are directed following the orientation rules in Algorithm 2 of Kalisch and Bühlmann (2007). Finally, the GDS algorithm (Chickering, 2002; Hauser and Bühlmann, 2012) estimates the true EG by greedily optimizing a score function in the space of DAGs. The GDS is viewed as a suboptimal alternative, because greedy search takes place in the space of DAGs instead of EGs. As a consequence it is more prone to be stuck in local optima of the score function, and is expected to yield worse results than GES. We include GDS in our comparative study to highlight the usefulness of operating directly on the space of EGs. We remark that, while each of the above algorithms results in a single graph, OBES provides a richer output, that is an approximation of the posterior distribution on the space of EGs. This in turn allows to quantify probabilistically our uncertainty not only on the structure of the true EG, but also on interesting related features, such as the number of directed or undirected edges, the number of v-structures, as well as multiple-edge inclusion probabilities. For comparison purposes we analyse a variety of simulation scenarios, and a real data set.

5.1 Simulation studies

A simulation framework is characterized by the pair (q, n), where $q \in \{5, 10, 20\}$ is the number of nodes and $n \in \{50, 100, 200\}$ the sample size, giving rise to nine scenarios. A total of 50 datasets, corresponding to 50 true EGs, are generated in each scenario. Following Peters and Bühlmann (2014), each dataset is obtained as follows: we randomly generate a topologically ordered DAG with probability of edge inclusion $p_{edge} = 3/(2q-2)$. The DAG thus obtained implies the following set of equations

$$Y_{i,j} = \mu_j + \sum_{k \in pa(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, q, \tag{28}$$

where $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ independently. For each j we fix $\mu_j = 0$ and generate each σ_j^2 from a $Unif(0, 2)$, while the regression coefficients $\beta_{k,j}$ are uniformly chosen in the interval $[-1, -0.1] \cup [0.1, 1]$. Datasets of size n are then generated accordingly. In order to compare the EG estimates we need to obtain, for each randomly generated DAG, the corresponding EG (the *true EG*). This is done through the function `dag2essgraph` in the `R` package `pcalg`; see Hauser and Bühlmann (2012) for details.

With regard to OBES, a few pilot runs are used as a diagnostic tool to evaluate MCMC convergence and mixing relative to some graph feature. OBES relies on the method by He et al. (2013), which allows to constrain the EG-space to a subspace with no more than a given number of edges, so that sparsity of the EG can be introduced to improve structural learning. Specifically, we require that the number of edges is not higher than 1.5 the number of nodes. This threshold is well above the number of edges expected in the true EG in each simulation scenario described above (4, 8 and 15 edges respectively for 5, 10 and 20 nodes), and it is in line with the average sparsity constraint

in the simulation scenarios of He et al. (2013). For comparison purposes with the other methods, an estimate of the true EG is also provided under OBES. To this end, define the inclusion probability of the directed edge $u \to v$ as

$$p_{u\to v}(\boldsymbol{Y}) = \sum_{\mathcal{G}\in\mathcal{S}_{u\to v}} p(\mathcal{G} \mid \boldsymbol{Y}),$$

where $\mathcal{S}_{u\to v}$ is the class of EGs containing the directed edge $u \to v$ (recall that an undirected edge $u - v$ is equivalent to $u \to v$ and $v \to u$). The *median probability (graph) model* is defined as the graph containing only those directed edges $u \to v$ such that $p_{u\to v}(\boldsymbol{Y}) \geq 0.5$. This definition is in accord with that of median probability model introduced in a Gaussian regression setting by Barbieri and Berger (2004) where it was shown to be predictively optimal. Peterson et al. (2015) also use the median probability model for learning graph structures. In general, the median probability model is not guaranteed to be an EG, but it is a partially directed acyclic graph (PDAG). Accordingly, our final estimate is the *projected* median probability model built by first generating a consistent extension of the median probability model and then deriving the corresponding EG. According to Theorem 1 in Verma and Pearl (1991), all consistent extensions of a PDAG, if they exist, belong to a unique Markov equivalence class, and therefore any consistent extension of the median probability model will produce the same EG. OBES only requires to specify the value of the hyperparameters a and b in the beta prior for the edge inclusion probability. However we found that results are robust to the choice of a and b even in scenarios with small sample sizes. Based on sparsity considerations, we use $a = 1$ and $b = (2q - 2)/3 - 1$.

With regard to the alternative benchmark methods under consideration, we note that the output of the PC algorithm depends on the choice of the significance level employed in a sequence of conditional tests. We present results for significance levels 1%, 5% and 10%. The GES approach is computed for three different optimization criteria: the Bayesian Information Criterion (Schwarz, 1978) and the Extended Bayesian Information Criterion with tuning coefficient $\gamma \in \{0.5, 1\}$ recommended in Foygel and Drton (2010); see also Chen and Chen (2008).

Each method under comparison is evaluated using several performance indicators measuring its effectiveness in recovering the true underlying EG. This is achieved by comparing some features of the true EG with the corresponding ones in the estimated EG produced by the method. We start with the Structural Hamming Distance (SHD), which represents the number of edge insertions, deletions or flips needed to transform the estimated EG into the true EG. Clearly lower values of SHD correspond to better performances. The SHD for each method is reported in Figure 1 as a boxplot of the SHD values over the 50 replicates. For OBES we report not only the final estimate *OB Proj* (projected median probability model), but also the intermediate median probability model (*OB Med*): this is done to highlight the impact of the consistent extension of the PDAG to the CPDAG. For $q = 5$, the space of EGs is relatively small, and all methodologies perform similarly, regardless of the sample size, with our method always better or equal to the best alternative. For higher number of nodes q, it becomes clear that GDS is not competitive and suffers from exploring the space of DAGs instead of the

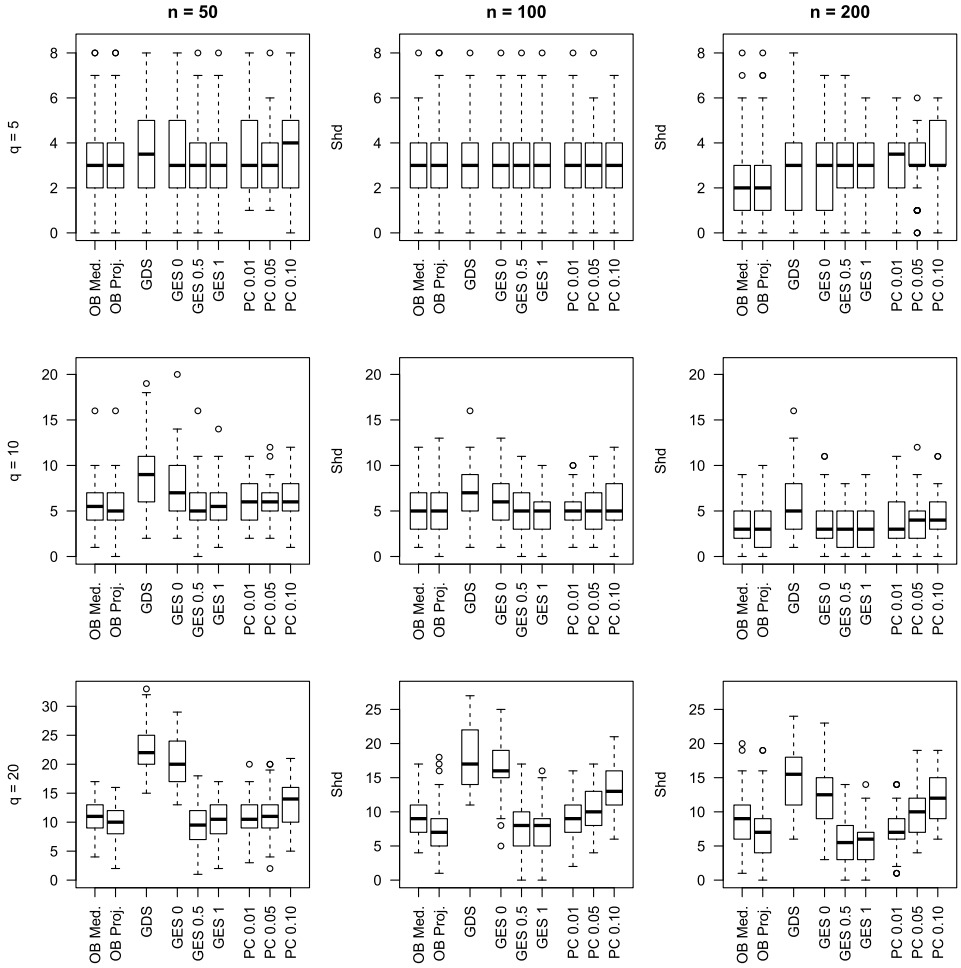

Figure 1: Simulation studies. Structural Hamming distances between the estimated EGs and the true EG, over 50 datasets, for number of nodes $q \in \{5, 10, 20\}$ and sample size $n \in \{50, 100, 200\}$. The performances are measured for our intermediate output, the *median probability model* (OB Med), our final estimate, the *projected median probability model* (OB Proj), the GDS algorithm (GDS), the GES algorithm with tuning parameter equal to 0, 0.5 and 1 (respectively GES 0, GES 0.5 and GES 1), and the PC algorithm at significance levels 1%, 5% and 10% (respectively PC 0.01, PC 0.05 and PC 0.1).

space of EGs. All methods improve their performance as the sample size increases. OBES remains highly competitive in all scenarios, only slightly underperfoming relative to the best GES methods in one scenario. The very slight difference between OB Med and OB Proj shows that the impact of the consistent extension is minimal. For each scenario and method we also evaluate the performance in learning the graphical structure of

		OB Proj	GDS	GES 0	GES 0.5	GES 1	PC 0.01	PC 0.05	PC 0.1
$n = 50$	MISR	**8.24**	13.07	11.04	8.44	8.64	9.20	8.78	8.98
	SPE	97.81	93.24	94.70	97.82	**98.69**	98.56	98.01	97.34
	SEN	51.63	49.43	**54.06**	49.90	40.99	38.81	46.77	50.36
	PRE	75.10	48.69	56.78	74.56	**78.79**	76.72	73.57	69.89
	MCC	**60.58**	49.35	54.98	59.11	55.17	53.29	57.20	57.93
$n = 100$	MISR	7.98	10.80	9.00	7.49	**7.44**	7.49	7.62	7.93
	SPE	96.99	93.93	95.94	97.95	98.61	**98.71**	98.10	97.37
	SEN	60.07	**61.14**	60.67	57.01	51.77	51.31	54.98	58.54
	PRE	73.00	56.60	64.29	78.21	83.04	**83.15**	77.10	72.82
	MCC	64.39	57.74	61.36	**64.74**	62.99	63.28	63.72	64.01
$n = 200$	MISR	**4.67**	8.69	5.42	4.67	4.84	5.62	5.16	5.80
	SPE	98.30	95.08	97.81	98.96	**99.12**	99.02	98.75	98.08
	SEN	**75.96**	70.10	75.21	71.56	69.13	64.38	69.80	69.92
	PRE	85.52	66.36	81.54	**89.66**	90.78	89.50	88.30	82.41
	MCC	**78.99**	66.60	76.85	78.18	77.18	73.82	76.64	74.08

Table 1: Simulation studies. Misspecification rate (MISR), specificity (SPE), sensitivity (SEN), precision (PRE) and Matthews correlation coefficient (MCC) for all methods under comparison, for number of nodes $q = 10$ and sample size $n \in \{50, 100, 200\}$.

the 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). The results in the simulation settings with number of nodes $q = 10$ and $n \in \{50, 100, 200\}$ are summarized in Table 1: with the exception of MISR, better performances correspond to higher indicators. The table decomposes the raw performances of Figure 1 in finer measures: for all indicators and scenarios, OBES is better than GES 0.5, GES 1, PC 0.01 and PC 0.05 most of the time, and it is almost uniformly better than the alternatives GDS, GES 0 and PC 0.1. Results not shown here for the sake of brevity also confirm that OBES is broadly insensitive to the choice of tuning parameters. On the other hand, from Table 1 it is evident that the tuning parameter of GES highly affects its performance, with simulation results giving no clear indication of superiority between GES 0.5 and GES 1. Similar considerations apply to the significance level of PC, with no clear-cut ranking between PC 0.01 and PC 0.05, and with PC 0.1 always outperformed. The tables for the remaining scenarios with $q = 5$ and $q = 20$ are similar, and for this reason they are not reported. Also, similar results hold in terms of learning the skeleton of the graph, that is when directionality of edges in the estimated and true graph is ignored.

We investigate the computational time of the proposed methodology as a function of the number of nodes q and of the sample size n: in the left panel of Figure 2 we report the

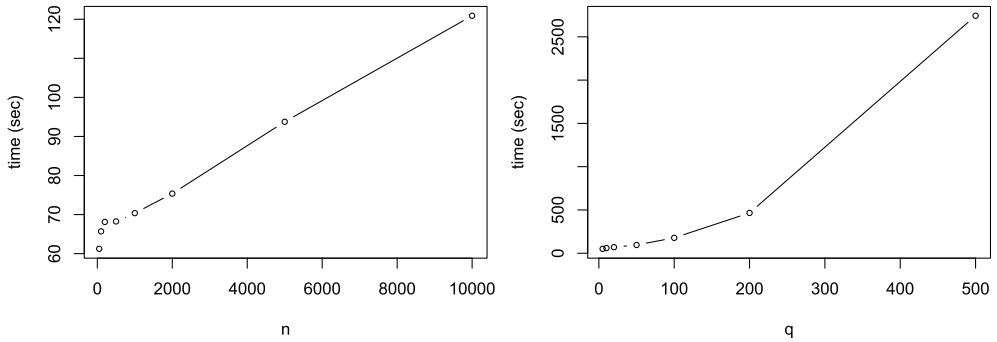

Figure 2: Simulation studies. Computational time (in seconds) of 1000 MCMC iterations of OBES, as a function of the sample size n for fixed number of nodes $q = 20$ (left panel) and as a function of q for $n = 1000$ (right panel), averaged over 50 simulated datasets.

time in seconds required by our algorithm to perform 1000 MCMC iterations for $q = 20$ and n between 50 and 10000, whilst in the right panel we show the computational time for $n = 1000$ and q between 5 and 500 (all the results we quote are averages over 50 simulated datasets). The codes are written in `R` and `Python`, and were run on a 2 x Intel(R) Xeon(R) CPU E5-2687W v3 3.10GHz machine. The computational burden of the alternative methodologies is generally lower; on the other hand OBES is based on an MCMC algorithm that not only provides a point estimate of the EG, but also an approximation of the whole posterior distribution over the EG-space. The fastest methods are the two GES approaches under the Extended Bayesian Information Criterion, whose computing times are on average three seconds in the worst scenario. Exceptions are the GES under the Bayesian Information Criterion and the PC algorithm with significance level 1%. Both algorithms are more time consuming in the scenario with $q = 500$: in particular the PC algorithm shows an average computing time of approximately 70 minutes (4200 seconds).

5.2 Protein-signaling dataset

The data, provided as a supplement to Sachs et al. (2005), include the levels of eleven phosphorylated proteins and phospholipids quantified using flow cytometry under nine different experimental conditions, each with sample size in the range 700–1000. In the original work of Sachs et al. (2005), the objective was to infer a single DAG, whilst Friedman et al. (2008) used the same dataset to learn a single undirected graph. More recently, Peterson et al. (2015) analyzed this protein dataset to infer an undirected graph for each of the nine conditions, allowing for the possibility of shared structural features among graphs. Our focus instead is on learning the structure of the generating EG, independently in each of the nine experimental conditions. In Table 2 we report the SHDs between OBES and the estimates provided by the alternative methodologies in each of the nine scenarios. To save space, we restrict the output of the analysis to the first experimental condition, and to the benchmarks GDS, GES with $\gamma = 0.5$ and PC

with significance level 0.01 (those with a better performance in the simulation studies). Similar results hold for the other experimental conditions and for the variants of GES and PC. The last line of Table 2 reports the number of edges in the EG estimate under OBES ($|\hat{\mathcal{G}}_{OBES}|$). When this information is taken into account, it appears that the number of modifications required to convert the estimate under each benchmark to the estimate under OBES is generally small.

Dataset	1	2	3	4	5	6	7	8	9
GDS	3	0	2	1	0	4	0	0	0
GES 0	0	2	5	3	4	4	0	0	0
GES 0.5	0	3	0	1	0	0	2	0	0
GES 1	1	4	0	0	3	1	5	0	0
PC 0.01	0	3	0	0	5	0	7	0	0
PC 0.05	0	3	5	0	6	4	4	0	0
PC 0.1	7	0	8	0	6	3	3	0	0
$\lvert\hat{\mathcal{G}}_{OBES}\rvert$	8	11	10	7	10	9	11	11	10

Table 2: Protein-signaling data. Structural Hamming distances between the estimated EG under OBES, and the one estimated by the alternative methods (rows of the table) under different experimental conditions (columns of the table). Number of edges in the EG estimate under OBES (last line).

As clarified in Section 5.1, our final EG estimate is the projected median probability model. The median probability model specifies a threshold for edge inclusion of 50%. By varying this threshold one obtains distinct *projected quantile probability models*. Each such model represents an EG estimate with specific graph features. Figure 3 reports for a grid of thresholds, the distribution of four selected graph features. We also indicate the value of each feature under GES 0.5, PC 0.01 and GDS. It appears that OBES always exhibits a prevailing value coinciding with that under the two best alternatives. On the other hand, GDS deviates to some extent from this pattern. The output of Figure 3 is confirmed in the graphs of Figures 4 and 5: EG estimates under OBES, GES 0.5 and PC 0.01 coincide, whilst GDS does not detect a v-structure created by $10 \rightarrow 9 \leftarrow 11$.

As already recalled, an advantage of OBES is that it accounts for model uncertainty in a principled way through the posterior distribution on the space of EGs. In particular we can evaluate the uncertainty of edge inclusion by computing the marginal posterior probability of inclusion for each directed edge $u \rightarrow v$. These probabilities are reported in the heat map of Figure 4, which exhibits sparsity with a clear indication for the presence of some edges. Specifically we can recognize the six undirected edges together with the v-structure $10 \rightarrow 9 \leftarrow 11$ Finally, we report in Figure 6 the MCMC trace plots of the number of undirected and directed edges, chain components and v-structures, for the EGs visited at each MCMC iteration.

6 Discussion

Observational data cannot distinguish among Directed Acyclic Graphs (DAGs) encoding the same set of conditional independencies, that is among Markov equivalent DAGs.

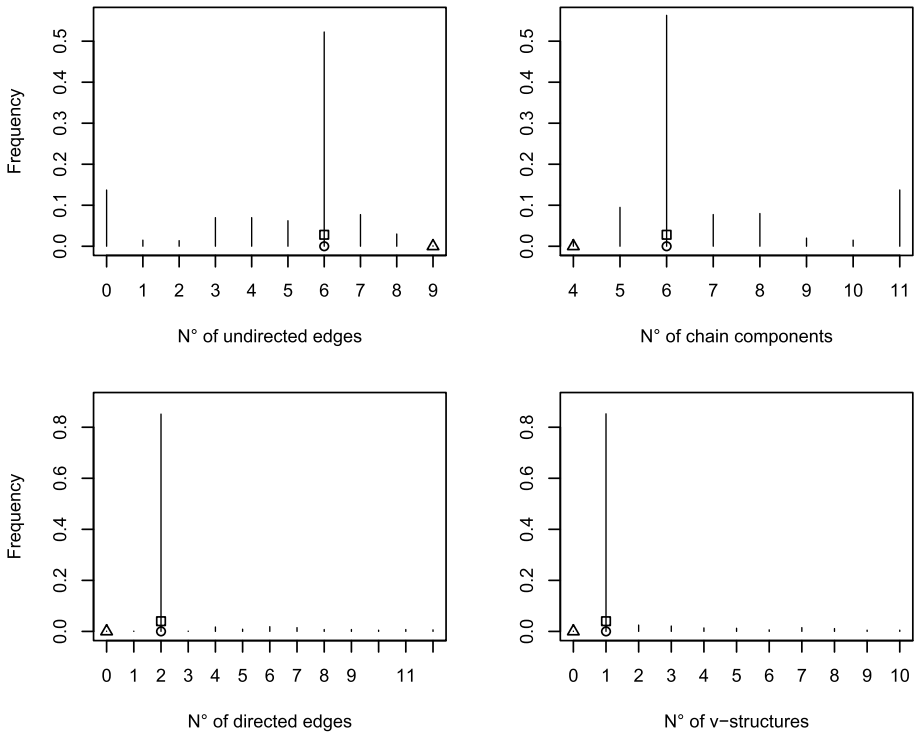

Figure 3: Protein-signaling data. First experimental condition. Frequency distribution of four graph features for a collection of distinct projected quantile models obtained by varying the edge inclusion probability threshold in [0.1,0.9]. The value of each feature for GES 0.5 (◯), PC 0.01 (□) and GDS (△) are superimposed.

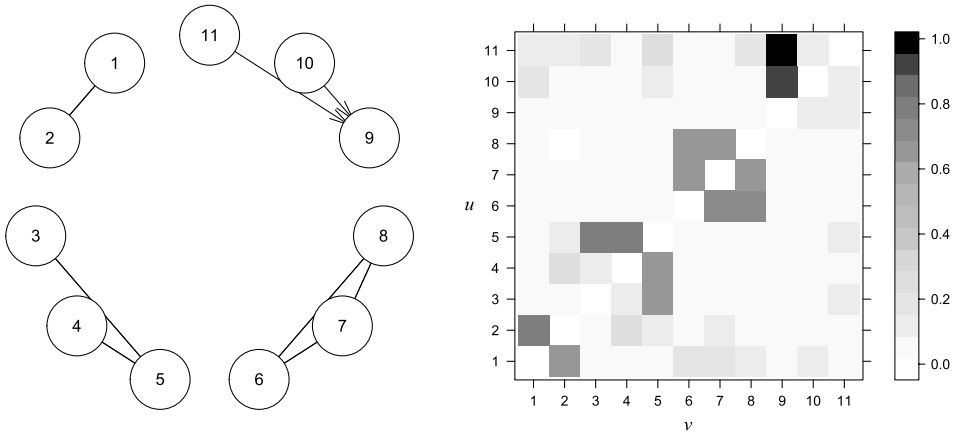

Figure 4: Protein-signaling data. First experimental condition. Estimated EG under OBES and heat map with marginal posterior probabilities of edge inclusion.

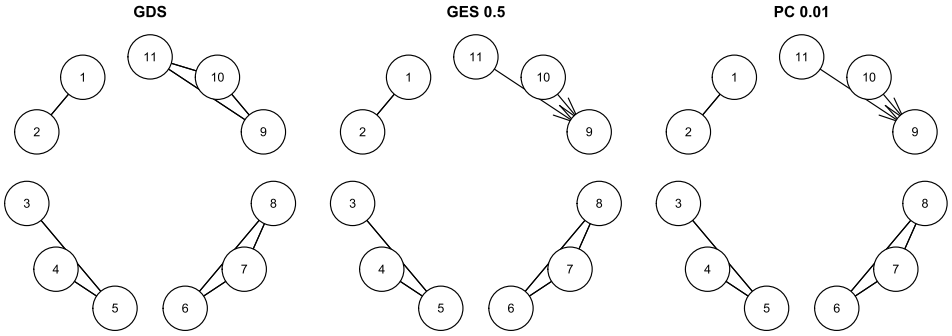

Figure 5: Protein-signaling data. First experimental condition. Estimated EGs under GDS, GES 0.5 and PC 0.01.

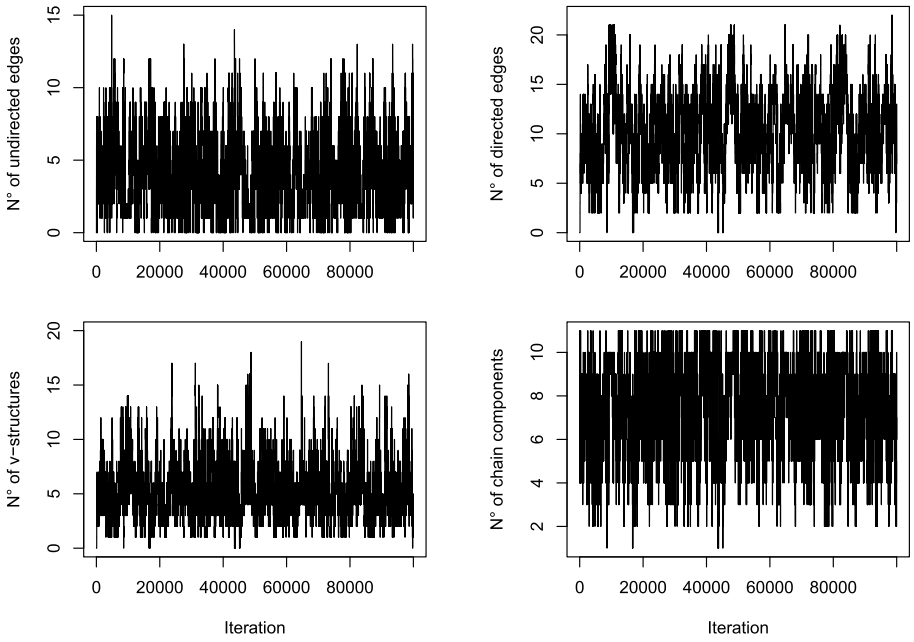

Figure 6: Protein-signaling data. First experimental condition. MCMC trace plots of visited EGs features: number of undirected and directed edges, chain components and v-structures.

Each Markov equivalence class is represented by a special chain graph, known as CPDAG or Essential Graph (EG). In this paper we have presented an objective Bayes method to learn the structure of the EG generating the data. Building on recent results for the objective Bayesian comparison of Gaussian multivariate regression graphical models (Consonni et al., 2017), we obtain a closed form expression for the marginal likelihood

of an EG. Next we construct an MCMC sampler that explores the space of EGs under sparsity constraints. We apply the proposed methodology, named OBES, to simulated and real datasets, and provide comparisons with state-of-the-art benchmark methods in the literature. We illustrate on simulated datasets that OBES is competitive with GES under the Extended BIC, and it outperforms GES under BIC, GDS and the PC algorithm in producing a point estimate of the underlying EG. On the other hand, our method yields a posterior distribution on the space of the EGs. Accordingly, it can provide not only single estimate of the EG, but also an uncertainty evaluation of other features of interest, such as the probability of inclusion of a particular edge. Finally, being objective, it is virtually free from prior specifications.

Besides the fractional Bayes factor adopted in this paper, there exist a few other *general* methods for the construction of objective priors for Bayesian model comparison, such as the Intrinsic prior (Berger and Pericchi, 1996) and the Expected Posterior Prior (EPP) of Peréz and Berger (2002), and variants thereof as in Consonni et al. (2013) or Fouskakis et al. (2015). A more principled setup is presented in Bayarri et al. (2012). An appealing feature of the priors produced by the above methods is that they are not data-dependent. They have been successfully implemented in the classic scenario of variable selection for Gaussian linear models although the expression of the model marginal likelihood may be analytically available only up to an integral; see for instance Womack et al. (2014) for the case of intrinsic priors. Outside linear regression the implementation of the above priors becomes even more burdensome; see for instance Fouskakis et al. (2017) in the setting of generalized linear models using (power) EPP.

Computational expediency is however not the only reason for using the fractional Bayes factor as an objective Bayes method in the context of graphical models. The methodology of Geiger and Heckerman (2002), which we adopt, requires that the prior on the parameter of the complete DAG induces global parameter independence; in that paper the Normal Wishart distribution is shown to be the only prior which satisfies this condition when the number of variables is at least three. Global parameter independence is crucial for the marginal likelihood to be Markov with respect to the underlying DAG, and conjugacy affords a closed-form expression. The same concept is expressed by the strong hyper-Markov property of Dawid and Lauritzen (1993). Finally Consonni et al. (2017) extended the analysis to the multivariate regression setting showing that the conjugate matrix Normal Wishart prior also induces global parameter independence, and that the fractional prior belongs to this conjugate family. This result has been instrumental in deriving the marginal likelihood of an essential graph in this paper. Of course one could argue that *any* proper matrix Normal Wishart prior would be equally valid; however its specification would be problematic because the parameter entails a covariance matrix (possibly high-dimensional). Additionally, resorting to "vagueness" assumptions, in order to alleviate the elicitation task, is hardly a solution in a model selection context even when the prior is proper; see Pericchi (2005, Sect. 1.6). Finally other desirable features that priors for model selection should satisfy, such as *compatibility* (Consonni and Veronese, 2008), could be hardly expressed in a subjective elicitation, thus reinforcing our motivation for an objective approach. Specifically, the use of a fraction of the likelihood to update the default parameter prior under each model establishes a connection between distributions across models.

DAGs can be used to model data that are not only observational, as in the current paper, but also interventional, produced by exogenous perturbations of variables, or by randomized intervention experiments; see Hauser and Bühlmann (2015). The ensuing *intervention distribution* is still amenable to a factorization similar to that holding in the observational case (Pearl, 2000), and one can define the *interventional* Markov equivalence class. The latter can be appreciably smaller than the corresponding class in the observational setting, thus improving the identifiability of the true data generating DAG. Similarly, the interventional essential graph (I-EG) will contain fewer undirected edges than its observational counterpart, because some will be oriented through interventions (He and Geng, 2008). Additionally, Hauser and Bühlmann (2012) show that I-EGs are still chain graphs with decomposable chain components. Such a characterization is important because it makes our approach feasible also for the computation of the marginal likelihood when both interventional and observational data are available. From the computational viewpoint one should extend the Markov chain of He et al. (2013) to the I-EG space, thus extending OBES to this setting, too.

The protein-signaling dataset was collected under nine distinct experimental conditions. As a consequence, nine distinct graphical structures could be considered. The basic choice is to estimate them separately, as we did in Subsection 5.2. Another possibility is to analyze them jointly in order to exploit potential shared features among graphs, in the hope of improving inference through Bayesian borrowing strength. Joint structural learning for multiple Gaussian undirected graphical models is carried out in Peterson et al. (2015) through a suitable Markov random field prior which encourages common edges, as well as a spike-and-slab prior on the parameters that measure network relatedness. While an extension of our methodology to infer multiple essential graphs with a shared structure is beyond the scope of the present work, it is conceptually feasible and is currently under investigation.

References

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997a). "A characterization of Markov equivalence classes for acyclic digraphs." *The Annals of Statistics*, 25: 505–541. MR1439312. doi: https://doi.org/10.1214/aos/1031833662. 1236, 1237

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997b). "On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs." *Scandinavian Journal of Statistics*, 24: 81–102. MR1436624. doi: https://doi.org/10.1111/1467-9469.t01-1-00050. 1241

Andersson, S. A., Madigan, D., and Perlman, M. D. (2001). "Alternative Markov properties for chain graphs." *Scandinavian Journal of Statistics*, 28: 33–85. MR1844349. doi: https://doi.org/10.1111/1467-9469.00224. 1237

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32: 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 1247

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40: 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 1254

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). "The formal definition of reference priors." *The Annals of Statistics*, 37: 905–938. MR2502655. doi: https://doi.org/10.1214/07-AOS587. 1239

Berger, J. O. and Pericchi, L. R. (1996). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91: 109–122. MR1394065. doi: https://doi.org/10.2307/2291387. 1236, 1254

Bhadra, A. and Mallick, B. K. (2013). "Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis." *Biometrics*, 69: 447–457. MR3071063. doi: https://doi.org/10.1111/biom.12021. 1245

Castelo, R. and Perlman, M. D. (2004). "Learning essential graph Markov models from data." In *Advances in Bayesian networks*, volume 146 of *Studies in Fuzziness and Soft Computing*, 255–269. Springer, Berlin. MR2090887. doi: https://doi.org/10.1007/978-3-540-39879-0_14. 1236, 1238

Chen, J. and Chen, Z. (2008). "Extended Bayesian information criteria for model selection with large model spaces." *Biometrika*, 95: 759–771. MR2443189. doi: https://doi.org/10.1093/biomet/asn034. 1247

Chickering, D. M. (2002). "Learning equivalence classes of Bayesian-network structures." *Journal of Machine Learning Research*, 2: 445–498. MR1929415. doi: https://doi.org/10.1162/153244302760200696. 1235, 1236, 1244, 1245, 1246

Colombo, D. and Maathuis, M. H. (2014). "Order-independent constraint-based causal structure learning." *Journal of Machine Learning Research*, 15: 3921–3962. MR3291411. 1246

Consonni, G., Forster, J. J., and La Rocca, L. (2013). "The Whetstone and the Alum Block: Balanced Objective Bayesian Comparison of Nested Models for Discrete Data." *Statistical Science*, 28: 398–423. MR3135539. doi: https://doi.org/10.1214/13-STS433. 1254

Consonni, G. and La Rocca, L. (2012). "Objective Bayes Factors for Gaussian Directed Acyclic Graphical Models." *Scandinavian Journal of Statistics*, 39: 743–756. MR3000846. doi: https://doi.org/10.1111/j.1467-9469.2011.00785.x. 1236, 1240

Consonni, G., La Rocca, L., and Peluso, S. (2017). "Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection." *Scandinavian Journal of Statistics*, 44: 741–764. MR3687971. doi: https://doi.org/10.1111/sjos.12273. 1236, 1240, 1241, 1242, 1243, 1253, 1254

Consonni, G. and Veronese, P. (2008). "Compatibility of Prior Specifications Across

Linear Models." *Statistical Science*, 23: 332–353. MR2483907. doi: https://doi.org/10.1214/08-STS258. 1254

Cowell, R. G., Dawid, P. A., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. New York: Springer. MR1697175. 1235, 1238

Dawid, A. P. (1981). "Some matrix-variate distribution theory: Notational considerations and a Bayesian application." *Biometrika*, 68: 265–274. MR0614963. doi: https://doi.org/10.1093/biomet/68.1.265. 1240

Dawid, A. P. and Lauritzen, S. L. (1993). "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models." *The Annals of Statistics*, 21: 1272–1317. MR1241267. doi: https://doi.org/10.1214/aos/1176349260. 1254

Dor, D. and Tarsi, M. (1992). "Simple algorithm to construct a consistent extension of a partially oriented graph." *Technical Report R-185, Cognitive Systems Laboratory, UCLA*. 1237

Drton, M. and Eichler, M. (2006). "Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property." *Scandinavian Journal of Statistics*, 33: 247–257. MR2279641. doi: https://doi.org/10.1111/j.1467-9469.2006.00482.x. 1237, 1238

Drton, M. and Perlman, M. D. (2008). "A SINful approach to Gaussian graphical model selection." *Journal of Statistical Planning and Inference*, 138: 1179–1200. MR2416875. doi: https://doi.org/10.1016/j.jspi.2007.05.035. 1240

Fouskakis, D., Ntzoufras, I., and Draper, D. (2015). "Power-Expected-Posterior Priors for Variable Selection in Gaussian Linear Models." *Bayesian Analysis*, 10: 75–107. MR3420898. doi: https://doi.org/10.1214/14-BA887. 1254

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2017). "Power-Expected-Posterior Priors for Generalized Linear Models." *Bayesian Analysis*. Advance publication. 1254

Foygel, R. and Drton, M. (2010). "Extended Bayesian Information Criteria for Gaussian Graphical Models." In *Advances in Neural Information Processing Systems 23*, 2020–2028. 1247

Friedman, J., Hastie, T., and Tibshirani, R. (2008). "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, 9: 432–441. 1250

Friedman, N. (2004). "Inferring Cellular Networks Using Probabilistic Graphical Models." *Science*, 303: 799–805. 1235

Geiger, D. and Heckerman, D. (2002). "Parameter priors for directed acyclic graphical models and the characterization of several probability distributions." *The Annals of Statistics*, 30: 1412–1440. MR1936324. doi: https://doi.org/10.1214/aos/1035844981. 1236, 1240, 1254

Geisser, S. and Cornfield, J. (1963). "Posterior distributions for multivariate normal parameters." *Journal of the Royal Statistical Society. Series B (Methodological)*, 25: 368–376. MR0171354. 1242

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. doi: https://doi.org/10.1016/S0004-3702(02)00264-3. 1236, 1243

Gupta, A. K. and Nagar, D. K. (2000). *Matrix variate distributions*. Chapman & Hall/CRC, Boca Raton, FL. MR1738933. 1240

Hauser, A. and Bühlmann, P. (2012). "Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs." *Journal of Machine Learning Research*, 13: 2409–2464. MR2973606. 1245, 1246, 1255

Hauser, A. and Bühlmann, P. (2015). "Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs." *Journal of the Royal Statistical Society. Series B (Methodology)*, 77: 291–318. MR3299409. doi: https://doi.org/10.1111/rssb.12071. 1255

He, Y. and Geng, Z. (2008). "Active learning of causal networks with intervention experiments and optimal designs." *Journal of Machine Learning Research*, 9: 2523–2547. MR2460892. 1255

He, Y., Jia, J., and Yu, B. (2013). "Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs." *The Annals of Statistics*, 41: 1742–1779. MR3127848. doi: https://doi.org/10.1214/13-AOS1125. 1236, 1243, 1244, 1245, 1246, 1247, 1255

Kalisch, M. and Bühlmann, P. (2007). "Estimating high-dimensional directed acyclic graphs with the PC-algorithm." *Journal of Machine Learning Research*, 8: 613–36. 1246

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 1236

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press. MR2778120. 1235

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press. MR1419991. 1235, 1236, 1237, 1241

Madigan, D., Andersson, S., Perlman, M., and Volinsky, C. (1996). "Bayesian Model Averaging And Model Selection For Markov Equivalence Classes Of Acyclic Digraphs." *Communications in Statistics: Theory and Methods*, 2493–2519. MR1439312. doi: https://doi.org/10.1214/aos/1031833662. 1236, 1243, 1244

Moreno, E. (1997). "Bayes Factors for Intrinsic and Fractional Priors in Nested Models. Bayesian Robustness." In Dodge, Y. (ed.), *L_1-Statistical Procedures and Related Topics*, 257–270. Institute of Mathematical Statistics. MR1833592. doi: https://doi.org/10.1214/lnms/1215454142. 1239

Nagarajan, R. and Scutari, M. (2013). *Bayesian Networks in R with Applications in Systems Biology*. New York: Springer. MR3059206. doi: https://doi.org/10.1007/978-1-4614-6446-4. 1235

O'Hagan, A. (1995). "Fractional Bayes Factors for Model Comparison." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57: 99–138. MR1325379. 1239

O'Hagan, A. and Forster, J. J. (2004). *Bayesian Inference. Kendall's Advanced Theory of Statistics*. Arnold, 2nd edition. MR3237119. 1239

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge. MR1744773. 1235, 1255

Pearl, J. (2003). "Statistics and causal inference: A review." *Test*, 12: 281–345. MR2044313. doi: https://doi.org/10.1007/BF02595718. 1235

Peréz, J. M. and Berger, J. O. (2002). "Expected-Posterior Prior Distributions for Model Selection." *Biometrika*, 89: pp. 491–511. MR1929158. doi: https://doi.org/10.1093/biomet/89.3.491. 1254

Pericchi, L. R. (2005). "Model selection and hypothesis testing based on objective probabilities and Bayes factors." In Dey, D. and Rao, C. R. (eds.), *Bayesian thinking: modeling and computation*, volume 25 of *Handbook of Statistics*, 115–149. Elsevier. MR2490524. doi: https://doi.org/10.1016/S0169-7161(05)25004-6. 1236, 1239, 1254

Peters, J. and Bühlmann, P. (2014). "Identifiability of Gaussian structural equation models with equal error variances." *Biometrika*, 101: 219–228. MR3180667. doi: https://doi.org/10.1093/biomet/ast043. 1245, 1246

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). "Bayesian inference of multiple Gaussian graphical models." *Journal of the American Statistical Association*, 110: 159–174. MR3338494. doi: https://doi.org/10.1080/01621459.2014.896806. 1247, 1250, 1255

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. 1235, 1236, 1250

Schwarz, G. E. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 6: 461–464. MR0468014. 1247

Shojaie, A. and Michailidis, G. (2009). "Analysis of gene sets based on the underlying regulatory network." *Journal of Computational Biology*, 16: 407–26. MR2487566. doi: https://doi.org/10.1089/cmb.2008.0081. 1235

Sonntag, D., Peña, J. M., and Gómez-Olmedo, M. (2015). "Approximate Counting of Graphical Models via MCMC Revisited." *International Journal of Intelligent Systems*, 30: 384–420. 1236

Spirtes, P., Glymour, C., and Scheines, R. (2000). "Causation, Prediction and Search (2nd edition)." *Cambridge, MA: The MIT Press.*, 1–16. MR1815675. 1246

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. New York, NY, USA: Elsevier Science Inc. 1237, 1247

Womack, A. J., León-Novelo, L., and Casella, G. (2014). "Inference From Intrinsic Bayes' Procedures Under Model Selection and Uncertainty." *Journal of the American Statistical Association*, 109: 1040–1053. MR3265679. doi: https://doi.org/10.1080/01621459.2014.880348. 1254

Acknowledgments

We thank Luca La Rocca, Università di Modena e Reggio Emilia, for helpful contributions on an early draft of this paper. We also thank the Associate Editor and the Reviewer for useful comments that led to an improved version of the paper. Partial financial support was provided by UCSC (Research grant track D1).