Altomare, Davide; Consonni, Guido; La Rocca, Luca

Working Paper

Objective Bayesian Search of Gaussian DAG Models with Non-local Priors

Quaderni di Dipartimento, No. 140

Provided in Cooperation with:
University of Pavia, Department of Economics and Quantitative Methods (EPMQ)

Suggested Citation: Altomare, Davide; Consonni, Guido; La Rocca, Luca (2011) : Objective Bayesian Search of Gaussian DAG Models with Non-local Priors, Quaderni di Dipartimento, No. 140, Università degli Studi di Pavia, Dipartimento di Economia Politica e Metodi Quantitativi (EPMQ), Pavia

This Version is available at:
https://hdl.handle.net/10419/95281

Quaderni di Dipartimento

Objective Bayesian Search of Gaussian DAG Models with Non-local Priors

Davide Altomare
(Università di Pavia)

Guido Consonni
(Università di Pavia)

Luca La Rocca
(Università di Modena e Reggio Emilia)

140 (03-11)

Dipartimento di economia politica
e metodi quantitativi
Università degli studi di Pavia
Via San Felice, 5
I-27100 Pavia

Marzo 2011

Abstract

Directed Acyclic Graphical (DAG) models are increasingly employed in the study of physical and biological systems, where directed edges between vertices are used to model direct influences between variables. Identifying the graph from data is a challenging endeavor, which can be more reasonably tackled if the variables are assumed to satisfy a given ordering; in this case, we simply have to estimate the presence or absence of each possible edge, whose direction is established by the ordering of the variables. We propose an objective Bayesian methodology for model search over the space of Gaussian DAG models, which only requires default non-local priors as inputs. Priors of this kind are especially suited to learn sparse graphs, because they allow a faster learning rate, relative to ordinary local priors, when the true unknown sampling distribution belongs to a simple model. We implement an efficient stochastic search algorithm, which deals effectively with data sets having sample size smaller than the number of variables. We apply our method to a variety of simulated and real data sets.

Keywords Fractional Bayes factor; High-dimensional sparse graph; Moment prior; Non-local prior; Objective Bayes; Pathway based prior; Regulatory network; Stochastic search; Structural learning.

1 Introduction

Graphical models represent a powerful statistical tool in multivariate analysis and probabilistic expert systems; see, for instance, the monographs by Whittaker (1990); Cowell et al. (1999); Edwards (2000). Several classes of graphs can be used to define graphical models. Among them, Undirected graphs (UGs), Directed Acyclic Graphs (DAGs) and Chain Graphs (CGs) are well-known. In this paper, we concentrate on models defined by means of DAGs.

High-dimensional DAG models are becoming increasingly popular in the study of biological systems, including cell signalling pathways and gene regulatory networks; see Markowetz and Spang (2007) for a review. In particular, as discussed by Shojaie and Michailidis (2010), *sparse* high-dimensional DAGs appear especially suited for these applications. This is both because they provide satisfactory explanations of biological processes and because they can be learned from data (*structural learning*) even when the sample size is smaller than the number of variables, which is often the case in this application context.

In this paper, we deal with structural learning for (Gaussian) DAG models from an *objective* Bayesian perspective. This entails assigning a prior distribution on the space of DAGs, together with a parameter prior within each DAG. We discuss both issues, but focus primarily on parameter priors. In particular, we suggest using *non-local* (Johnson and Rossell, 2010) parameter priors, which appear to be better suited for learning simple models than ordinary *local* priors. We also implement a suitable search algorithm over the space of DAG models, and compare our results to currently available frequentist methods: the PC-algorithm (Kalisch and Buhlmann, 2007), the Lasso (Meinshausen and Bulhmann, 2006), the Adaptive Lasso (Shojaie and Michailidis, 2010), and SIN (Drton and Perlman, 2004, 2007).

We assume that there exists *a priori* a total ordering of the variables involved (temporal, logical, or other) so that we do not have to infer edge

directions. This is clearly a limitation in scope, but it greatly simplifies search over model space, while still allowing to analyze a variety of interesting examples. Geiger and Heckerman (2002) consider the general case from a *subjective* Bayesian viewpoint (with local priors).

The rest of the paper is organized as follows. In section 2 we introduce the needed background material on Gaussian DAG models, non-local priors, fractional Bayes factors (upon which our objective approach relies) and model priors. Section 3 presents our *path-based stochastic search* algorithm, which is evaluated in section 4 on simulated and real data sets. Then, in section 5, we deal with structural learning for sparse DAGs (both on simulated and real data). Finally, section 6 contains a brief discussion. Some techincal material has been moved to an Appendix.

2 Background

We assume the reader is familiar with basic graph definitions and terminology as presented, e.g., by Cowell et al. (1999).

2.1 Gaussian directed acyclic graphical models

Let $\mathcal{D} = (V, E)$ be a DAG, where $V = \{1, \ldots, q\}$ is the set of its vertices and $E \subseteq V \times V$ is the set of its directed edges. We assume a total ordering of the vertices, and that the vertices of $\mathcal{D}$ be well-numbered, so that, if a directed path goes from vertex i to vertex j in $\mathcal{D}$, then $i < j$. Each vertex corresponds to a variable, and for $W \subseteq V$ we denote by u_W the set of all variables u_j with $j \in W$. The Gaussian graphical model corresponding to $\mathcal{D}$ is the family of all q-variate normal distributions such that, if there is no edge $i \to j$ in $\mathcal{D}$, then u_j is conditionally independent of u_i given the set of variables $u_{\{1,\ldots,j\}\setminus\{i,j\}}$. For the sake of simplicity, we denote both the DAG and the corresponding Gaussian graphical model with the same symbol ($\mathcal{D}$ say). Although this can be a source of confusion, in general, because two DAGs can give rise to the same model, in our case this is perfectly safe, because the variable ordering is given once and for all.

In the model $\mathcal{D}$ the joint density of $(u_1, \ldots, u_q)$ can be written as

$$f(u_1, \ldots, u_q | \beta, \gamma) = \prod_{j=1}^{q} f(u_j | u_{\mathrm{pa}(j)}; \beta_j, \gamma_j), \tag{1}$$

where $\mathrm{pa}(j)$ denotes the parents of j in $\mathcal{D}$, i.e., the vertices of $\mathcal{D}$ preceding j and joined to j by an edge. Since each conditional distribution in (1) is a univariate normal, the vector parameter β_j represents the regression coefficients in the conditional expectation of u_j given $u_{\mathrm{pa}(j)}$, namely $(1, u'_{\mathrm{pa}(j)})\beta_j$, while γ_j is the corresponding conditional precision (inverse of variance). By convention, the first element of β_j is the intercept β_{j0}, while the remaining elements are written as β_{jk} with $k \in \mathrm{pa}(j)$. If $\mathbb{E}(u_j|\beta, \gamma) = 0$, then $\beta_{j0} = 0$, $j = 1, \ldots, q$, and the intercept can be dropped, so that β_j has dimension $|\mathrm{pa}(j)|$, the cardinality of the set $\mathrm{pa}(j)$.

2.2 Non-local priors

For data y, consider two models $\mathcal{M}_k$, $k = 0, 1$, with sampling density $f(y|\theta_k)$ and prior $p(\theta_k)$. We assume that $\mathcal{M}_0$ is *nested* in $\mathcal{M}_1$, so that each distribution in $\mathcal{M}_0$ coincides with some $f(y|\theta_1)$ in $\mathcal{M}_1$. We also assume that model comparison takes place through the Bayes Factor (BF) and write $BF_{10}(y) = m_1(y)/m_0(y)$ for the BF of $\mathcal{M}_1$ against $\mathcal{M}_0$ (or simply in favor of $\mathcal{M}_1$) where $m_k(y)$ is the marginal likelihood of $\mathcal{M}_k$, i.e., $m_k(y) = \int f(y|\theta_k)p(\theta_k)d\theta_k$. Usually $p(\theta_1)$, $\theta_1 \in \Theta_1$, is a *local* prior, i.e., assuming continuity, it is strictly positive over the subspace $\Theta_0 \subset \Theta_1$ characterizing the smaller model $\mathcal{M}_0$.

Suppose the data $y^{(n)} = (y_1, \ldots, y_n)$ arise under i.i.d. sampling from some (unknown) distribution Q. We say that the smaller model holds if Q belongs to $\mathcal{M}_0$, while we say that the larger model holds if Q belongs to $\mathcal{M}_1$ but not to $\mathcal{M}_0$. Since $p(\theta_1)$ is a local prior, we have the following learning behaviour of the BF: if $\mathcal{M}_0$ holds, then $BF_{10}(y^{(n)}) = O_p(n^{-(d_1-d_0)/2})$, as $n \to \infty$, where d_k is the dimension of $\mathcal{M}_k$, $k = 0, 1$, and $d_1 > d_0$; if $\mathcal{M}_1$ holds, then $BF_{01}(y^{(n)}) = e^{-Kn+O_p(\sqrt{n})}$, as $n \to \infty$, for some $K > 0$ (Kullback-Leibler divergence of $\mathcal{M}_0$ from Q); see Dawid (1999).

Johnson and Rossell (2010) defined *non-local* priors under $\mathcal{M}_1$ in order to reduce the above described imbalance in the asymptotic learning rate of the BF. We focus here on a specific family of non-local priors. Let $g(\theta_1)$ be a continuous function vanishing on Θ_0. For a given local prior $p(\theta_1)$, define a new non-local prior as $p^M(\theta_1) \propto g(\theta_1)p(\theta_1)$, which we name a *generalized moment prior*. For instance, if θ_1 is a *scalar* parameter in $\mathbb{R}$ and $\Theta_0 = \{\theta_0\}$, with θ_0 a *fixed* value, we may take $g(\theta_1) = (\theta_1 - \theta_0)^{2h}$, where h is a positive integer ($h = 0$ returns the starting local prior); this is precisely the *moment prior* introduced by Johnson and Rossell (2010) for testing a sharp hypothesis on a scalar parameter. In this case $BF_{10}(y^{(n)}) = O_p(n^{-h-1/2})$ when $\mathcal{M}_0$

holds, while $BF_{01}(y^{(n)}) = e^{-Kn+O_p(\sqrt{n})}$ when $\mathcal{M}_1$ holds. Thus, if $h = 1$, the rate changes from sub-linear to super-linear. While the above argument is asymptotic, its effect is already apparent for moderate sample sizes; see Consonni et al. (2010).

2.3 Fractional Bayes factors

Objective priors are often improper and thus defined up to a multiplicative constant. Hence, they cannot be used to compute BFs, even when the marginal likelihoods are positive and finite for all data realizations. A few solutions to this difficulty have been proposed: fractional Bayes factors (O'Hagan, 1995), intrinsic Bayes factors (Berger and Pericchi, 1996), intrinsic priors (Moreno, 1997), and expected posterior priors (Perez and Berger, 2002). Pericchi (2005) provides a comprehensive review. In this paper we focus on the Fractional Bayes Factor (FBF), which we find especially attractive in our context because its expression is available in closed-form.

Consider a model $\mathcal{M}_k$ with sampling density $f(\cdot|\theta_k)$ and prior $p(\theta_k)$. Let $0 < g < 1$ be a quantity depending on the sample size n, and define for data y

$$w_k(y; g) = \frac{\int f(y|\theta_k)p(\theta_k)d\theta_k}{\int f^g(y|\theta_k)p(\theta_k)d\theta_k},$$

where $f^g(y|\theta_k)$ is the sampling density raised to the g-th power, and the integrals are assumed to be finite and nonzero. We refer to $w_k(y; g)$ as the *fractional marginal likelihood* of the k-th model. Consider now two models, $\mathcal{M}_0$ and $\mathcal{M}_1$. The FBF (in favor of $\mathcal{M}_1$) is defined as $FBF_{10}(y; g) = w_1(y; g)/w_0(y; g)$. It is not difficult to see that the fractional marginal likelihood $w_k(y; g)$ can be computed as $\int f^{(1-g)}(y|\theta_k)p^F(\theta_k|y)$, where $p^F(\theta_k|b, y)$ is an implied data-dependent fractional prior proportional to $p(\theta_k)f^g(y|\theta_k)$, that is, actually a posterior based on a fraction g of the likelihood; usually g will be small, so that the dependence of the prior on the data will be weak.

Consistency of the FBF is achieved as long as $g \to 0$ for $n \to \infty$. O'Hagan (1995, sect. 6) suggests $g = n_0/n$ as a default choice of g, where n_0 is the minimal (integer) training sample size for which the fractional marginal likelihood is well defined, together with two other choices for cases when robustness is a major concern. Moreno (1997) has an argument in favour of the defalut choice, and we stick to it in this paper.

2.4 Moment fractional Bayes factors

The advantages of FBFs can be usefully combined with those of moment priors to obtain an objective Bayesian testing methodology with enhanced learning behavior (Consonni and La Rocca, 2011). We now present this approach.

Because of the recursive structure of the likelihood (1), it is natural to assume that $p(\beta, \gamma)$ satisfies the assumption of global parameter independence: $p(\beta, \gamma) = \prod_j p(\beta_j, \gamma_j)$; see Geiger and Heckerman (2002). A natural default prior is then $p^D(\beta_j, \gamma_j) \propto \gamma_j^{-1}$. Now consider two Gaussian DAG models $\mathcal{D}_0$ and $\mathcal{D}_1$ with the same vertex set, and vertex ordering, and with $\mathcal{D}_0$ nested in $\mathcal{D}_1$. For each vertex j, let L_j be the subset of parents pointing to j in $\mathcal{D}_1$ but not in $\mathcal{D}_0$, and define the corresponding default moment prior of order h as $p_1^M(\beta_j, \gamma_j) \propto \gamma_j^{-1} \prod_{l \in L_j} \beta_{jl}^{2h}$, where h is a positive integer; notice that $h = 0$ returns the initial default prior. The overall moment prior will be obtained by multiplying together the priors $p_1^M(\beta_j, \gamma_j)$:

$$p_1^M(\beta, \gamma) \propto \prod_{j=1}^{q} \left\{ \gamma_j^{-1} \prod_{l \in L_j} \beta_{jl}^{2h} \right\}. \tag{2}$$

In order to compute the FBF based on the moment prior (2), which we call a Moment FBF (MFBF) of order h, we need an expression for the fractional marginal likelihood pertaining to vertex j both under $\mathcal{D}_0$ and under $\mathcal{D}_1$. The former is standard, because it is based on the default prior, while the latter is provided in Theorem 1 of Consonni and La Rocca (2011) and is reported for completeness in the Appendix. Notice that letting $h = 0$ we obtain an Ordinary FBF (OFBF).

2.5 Priors over graph families

A seemingly objective prior over the space of all graphs with given vertex set is represented by the uniform prior. However, it is well known that this prior is strongly biased in favor of "medium-size" graphs; see, e.g., Giudici and Green (1999, sect. 1.3). A further issue is of some concern. Graphical model search can be viewed as a multiple testing problem, because it amounts to testing repeatedly for the presence of each potential edge. In this sense, graphical model search is akin to variable selection, for which Scott and Berger (2010) suggest a prior with an automatic adjustment for multiple

testing as the number of possible predictors grows. In our graphical modelling context this translates to

$$p(\mathcal{D}) = \frac{1}{m+1}\binom{m}{k}^{-1} \tag{3}$$

for a DAG $\mathcal{D}$ having k edges out of the m possible ones. This prior yields very strong control over the number of false edges admitted into the model, which typically remains bounded even as the number of spurious tests grows without bound; see also Scott and Carvalho (2008).

3 Graphical model determination

We are now ready to address the issue of objective Bayesian model determination over the space of all DAGs consistent with a fixed ordering of the variables. Our method for assigning priors is based on a *pairwise* comparison of two nested models, which in turn produces a BF determining the posterior probability of the two models. Now suppose we entertain a finite *collection* of DAGs $\{\mathcal{D}_k\}$. The posterior proabability over the family $\{\mathcal{D}_k\}$ can still be deduced from a collection of pairwise BFs as follows. We single out a *reference* DAG $\mathcal{D}_0$ and obtain the collection of BFs of $\mathcal{D}_k$ against $\mathcal{D}_0$, namely $\{BF_{k0}(y)\}$. Because of the way we construct our priors, a natural requirement is that $\mathcal{D}_0$ be *nested* into every other model $\mathcal{D}_k$; this is called *encompassing from below*. If *all* possible DAGs belong to the collection, this means choosing $\mathcal{D}_0$ as the *complete independence* DAG (DAG with no edges). We derive the posterior probability of model $\mathcal{D}_k$ as

$$p(\mathcal{D}_k|y) = \frac{BF_{k0}(y)p(\mathcal{D}_k)}{\sum_j BF_{j0}(y)p(\mathcal{D}_j)}, \tag{4}$$

where index j in the denominator runs over all possible model-indexes. Notice that formula (4) is valid because the parameter prior under $\mathcal{D}_0$ is the same in *all* pairwise comparisons. In the following we describe our MFBF search algorithm, where we let BF be an MFBF of order $h \geq 1$. Since the description also works for $h = 0$, we also have an OFBF search algorithm.

3.1 Stochastic search

Even under the assumption of a fixed ordering of the variables, the number of DAGs grows exponentially in the number of variables, so that enumerating all

of them is not feasible already for moderately sized vertex sets. Therefore, one must resort to some form of search algorithm to identify the most valuable models. With regard to this point, it is by now recognized that MCMC methods are not efficient, because of the sheer cardinality of model space and of its nature, which typically gives rise to irregularly multi-modal posteriors; see Friedman and Koller (2003). Nevertheless, other forms of stochastic search can help in this context. In particular, since we have the score (unnormalized posterior probability) of each model available in closed-form, we can look directly for a list of models with high score and re-normalize their scores to obtain an assessment of their posterior probabilities. More on these issues can be found in Scott and Carvalho (2008).

In order to describe our algorithm, we need the notion of *inclusion probability* of a potential edge $e \in \{e_1, \ldots, e_m\}$, defined as the posterior probability of e being present in the unknown structure; this can be deduced from (4) by summing the posterior probabilities of all DAGs including e:

$$p(e|y) \equiv q_e = \sum_{j:\, e \in \mathcal{D}_j} p(\mathcal{D}_j|y). \tag{5}$$

The Median Probability (MP)-DAG is the graph containing those edges whose inclusion probability is at least 0.5. The concept of MP model was introduced by Barbieri and Berger (2004) in the context of regression models, where it was shown to perform better, in terms of prediction, than the Highest Posterior Probability (HPP)-model.

In practice, expression (4) is not immediately helpful, because its denominator cannot be evaluated due to the excessive number of terms. However, if a collection of high scoring DAGs is available, we can estimate the posterior probability of $\mathcal{D}_k$ by summing over the models in the collection to compute the denominator of (4); we denote this estimate by $\hat{p}(\mathcal{D}_k|y)$, and the corresponding estimate of (5) by $\hat{p}(e|y)$.

Our search algorithm is similar to that proposed by Scott and Carvalho (2008), which in turn draws on the general ideas presented in Berger and Molina (2005) for the case of variable selection. It includes resampling moves, local moves and global moves. The rationale is very simple and can be summarized by saying that edge moves which have improved some models are more likely (than a randomly chosen move) to improve other models as well. A step-by-step description follows.

1. Start with a base DAG $\mathcal{D}_B$ and obtain deterministically $m \equiv q(q-1)/2$ distinct new DAGs each one differing from $\mathcal{D}_B$ by exactly one

edge. Based on this initial collection of DAGs, compute the (estimated) graph posterior probabilities $\hat{p}(\mathcal{D}_j|y)$, $j = 1, \ldots, m$, and edge inclusion probabilities $\hat{q}_e$, $e \in \{e_1, \ldots, e_m\}$. Initialize a counter $t = m$.

2. At iteration t return to one of the previously visited graphs, say $\mathcal{D}_R$, according to the posterior probabilities $\hat{p}(\mathcal{D}_j|y)$, $j = 1, \ldots, t$. This is called a *resampling move*.

3. Identify the possible *local moves* from $\mathcal{D}_R$ and choose one of them: change (i.e., add if not present and delete if present) edge e with probability

$$\propto \left(\frac{\hat{q}_e + C}{1 - \hat{q}_e + C} \right)^{2[1-D_R(e)]-1}, \quad e = z_1, \ldots, z_w, \tag{6}$$

where $D_R(e) = 1$ if edge e belongs to $\mathcal{D}_R$ while $D_R(e) = 0$ otherwise, and $z_1, \ldots, z_w$ are the edges that can be changed in such a way that the movement chosen is directed towards a new model. The constant $C > 0$ is introduced to keep the probability of all local moves bounded away from 0 and 1, even when all or none of the models visited include a certain edge. Update t to $t + 1$ and the posterior graph and edge inclusion probabilities.

4. Usually return directly to step 2 (while $t \leq T$) but periodically make a *global move*: define $\mathcal{D}_R$ to be the current MP-DAG and return to step 3.

3.2 Path-based search

At each iteration t of our algorithm, the posterior probabilities of all visited models must be updated using formula (4). This requires the MFBF of each visited model relative to the reference model $\mathcal{D}_0$; see equation (12). Computing the MFBF of $\mathcal{D}_t$ relative to $\mathcal{D}_0$ is slow when $\mathcal{D}_t$ is "far apart" from $\mathcal{D}_0$. This has to do with the definition of function H appearing in (10); see formula (see 11). On the other hand, the computation of MFBF is extremely fast when the comparison is restricted to *adjacent* (i.e., differing exactly by one edge) DAGs; see (13). Bearing in mind the great importance of speed in computations, and the fact that our search strategy operates locally on adjacent pairs of DAGs, it is natural to propose an alternative

computation of MFBF through the following "chain rule":

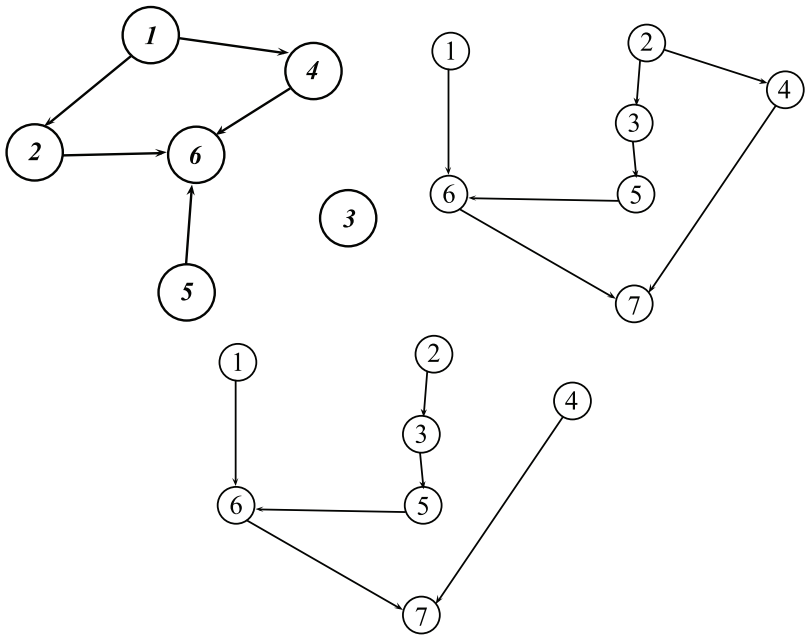
$$MFBF_{t0}(y;g) = \prod_{s=1}^{t} MFBF_{s,s-1}(y;g). \tag{7}$$

Each factor $MFBF_{s,s-1}(y;g)$ in (7) involves adjacent DAGs and is evaluated using a fractional moment prior under the larger model and an ordinary fractional prior under the smaller one. Formula (7) is used for comparing non-adjacent DAGs by suitably defining a path of adjacent DAGs connecting the two DAGs.

The use of (7) was suggested by Berger and Molina (2005) in the context of variable selection for linear models and using a variant of the intrinsic Bayes factor introduced by Berger and Pericchi (1996). Berger and Molina (2005) recommended (7) not only for computational reasons, but also on conceptual grounds: since dependence of the BF on parameter priors is huge when comparing models of significantly varying dimension, this dependence is potentially mitigated when comparing adjacent models. We concur with Berger and Molina, although we are aware that $t+1$ priors play a role in (7).

The "chain-rule" (7) can be criticized because it is not "coherent", in the sense that the fractional marginal likelihood based on a moment prior for DAG $\mathcal{D}_{t-1}$, say, may depend on the specific pairwise comparison under consideration. To see why, suppose that $\mathcal{D}_t$ and $\mathcal{D}_{t-2}$ are distinct but both nested within $\mathcal{D}_{t-1}$. Then, the fractional moment prior under $\mathcal{D}_{t-1}$ will be different in the two comparisons, and similarly for the fractional marginal likelihood of $\mathcal{D}_{t-1}$. As a consequence, the product of $MFBF_{t,t-1}$ and $MFBF_{t-1,t-2}$ will not provide $MFBF_{t,t-2}$, as it ought to, because the two marginal likelihoods for $\mathcal{D}_{t-1}$ do not cancel out. However, since $\mathcal{D}_{t-1}$ is compared with adjacent models, we can expect that the difference between the two fractional marginal likelihoods be modest, so that the "chain rule" will provide a reasonable approximation to the actual MFBF. Moreover, this seems to be a minor problem, when compared to the more general problem of stochastic search algorithms, which determine their inferences solely on the basis of visits to a (very) modest number of models. Finally, it must be emphasized that the "chain rule" would be perfectly coherent if one were to use the OFBF, whose definition does not involve a pairwise comparison.

Figure 1: True DAG for the six-variable simulation example (left panel); MP-DAGs found by UFBF (center panel) and MFBF (right panel) for the publishing productivity example.



4 Search method evaluation

In this section we evaluate our method on a couple of small data sets (first on simulated data and then on real data).

4.1 Simulated data from a six-variable DAG

We considered the DAG in Figure 1 (left panel) whose Gaussian graphical model corresponds to the structural equations

$$\begin{gathered} u_1 = \varepsilon_1,\ u_2 = \beta_{21}u_1 + \varepsilon_2,\ u_3 = \varepsilon_3,\ u_4 = \beta_{41}u_1 + \varepsilon_4, \\ u_5 = \varepsilon_5,\ u_6 = \beta_{62}u_2 + \beta_{64}u_4 + \beta_{65}u_5 + \varepsilon_6, \end{gathered} \tag{8}$$

where $\varepsilon_j \sim \mathcal{N}(\beta_{j0}, \sigma_j^2)$ independently over $j = 1, \ldots, 6$. Then, we let $\beta_{21} = \beta_{41} = \beta_{65} = \beta_{64} = \beta_{62} = 1$, $\beta_{j0} = 0$ and $\sigma_j^2 = 1$, for all $j = 1, \ldots, 6$,

and simulated 100 six-dimensional data sets from (8), each with sample size $n = 50$. Notice that, for the six variables in the given order, there are $m = 6(6-1)/2 = 15$ potential edges, and thus $2^{15} = 32768$ possible DAGs.

We ran a first order ($h = 1$) MFBF search on each simulated data set. Concerning the choice of g, recall the result in Theorem A.1 in the Appendix requiring $gn > p + 2h|L|$ for each vertex, where p is the number of parents and $|L|$ is the cardinality of the set of "extra" parents in the larger model. We actually took $ng = n_0$, with $n_0 = p + 2h|L| + 1$. For adjacent models $|L| = 1$, and substituting $h = 1$ gives $n_0 = p + 3$. Notice that the value of g is specific to each vertex. Each search was stopped when a total of $k = 100$ *distinct* models had been visited. We started the algorithm with a random base DAG. We also experienced with a fixed base DAG (complete DAG or complete independence DAG) finding similar results, with possibly smaller variability due to using the same starting point across simulations. For comparison purposes, we also ran an OFBF search, setting $g = n_0/n$ with $n_0 = p + 1$.

For all 15 potential edges, both for the OFBF and the MFBF search, we computed the average edge inclusion probabilities over the 100 data sets. Additionally, for each estimated edge inclusion probability, we computed the 95% coverage interval whose end points are the 2.5% and 97.5% quantiles relative to the 100 simulations. Table 1 reports our results. It also contains, for comparison purposes, the average (over the 100 simulated data sets) of the *p-values* given by the frequentist procedure SIN (Drton and Perlman, 2004, 2007).

It is apparent that both the OFBF and the MFBF search are able to capture the absence/presence of an edge extremely well. Notwithstanding this remark, the MFBF search scores much better when the edges are absent, because the average posterior inclusion probabilities are almost always below 2%, while never exceeding 3%. On the other hand, the corresponding average values for the OFBF search fall in the range 7–12%. The comparison between OFBF and MFBF is clearly in favor of the latter when comparing the distributions of edge inclusion probabilities, as expressed by the 95% coverage intervals for the edges absent in the true DAG: it is not uncommon for OFBF to have upper endpoints in the range 30%–40%, and in three cases they even exceed the 50% value, while for MFBF these values are always below 25%.

We also found that the MFBF search picks up more decisevely a single model, while the OFBF search dilutes the posterior mass on the model space.

Table 1: Six-variable simulation results. Estimated edge inclusion probabilities, searching 100 distinct models (starting from a random base DAG) with OFBF and first order MFBF search, compared to p-values given by SIN. Averages (A) and 95% coverage intervals (I) are reported, based on 100 independent simulated data sets. Starred edges are those present in the true DAG.

Edge	OFBF A	OFBF I	MFBF A	MFBF I	SIN A	SIN I
$(5,6)\star$	0.9992	(0.9893, 1.0000)	0.9965	(0.9380, 1.0000)	0.0002	(0.0000, 0.0031)
$(4,6)\star$	1.0000	(0.9997, 1.0000)	1.0000	(0.9993, 1.0000)	0.0001	(0.0000, 0.0006)
$(3,6)$	0.1167	(0.0299, 0.5900)	0.0328	(0.0031, 0.2372)	0.8335	(0.1978, 1.0000)
$(2,6)\star$	0.9998	(0.9980, 1.0000)	0.9992	(0.9955, 1.0000)	0.0004	(0.0000, 0.0017)
$(1,6)$	0.0911	(0.0300, 0.5462)	0.0205	(0.0031, 0.2088)	0.8710	(0.1511, 1.0000)
$(4,5)$	0.0942	(0.0261, 0.4384)	0.0231	(0.0028, 0.1606)	0.8571	(0.3032, 1.0000)
$(3,5)$	0.0765	(0.0309, 0.3003)	0.0177	(0.0028, 0.0714)	0.8958	(0.3264, 1.0000)
$(2,5)$	0.0878	(0.0285, 0.3945)	0.0259	(0.0028, 0.2121)	0.8623	(0.2861, 1.0000)
$(1,5)$	0.0816	(0.0343, 0.3344)	0.0208	(0.0029, 0.1513)	0.8307	(0.1393, 1.0000)
$(3,4)$	0.0703	(0.0253, 0.3449)	0.0122	(0.0029, 0.0806)	0.9052	(0.4210, 1.0000)
$(2,4)$	0.1040	(0.0343, 0.5283)	0.0272	(0.0029, 0.1811)	0.8787	(0.2082, 1.0000)
$(1,4)\star$	0.9876	(0.8129, 1.0000)	0.9828	(0.7647, 1.0000)	0.0145	(0.0000, 0.1693)
$(2,3)$	0.0756	(0.0307, 0.2366)	0.0183	(0.0028, 0.0597)	0.8980	(0.2007, 1.0000)
$(1,3)$	0.0781	(0.0296, 0.3756)	0.0201	(0.0029, 0.1165)	0.8845	(0.1099, 1.0000)
$(1,2)\star$	0.9996	(0.9914, 1.0000)	0.9984	(0.9661, 1.0000)	0.0009	(0.0000, 0.0015)

More in detail, we found an average posterior probability of the HPP-DAG (coinciding with the true DAG) equal to 0.8167 for MFBF and to 0.3545 for OFBF, while the corresponding 95% coverage intervals were $(0.5357, 0.9568)$ and $(0.1667, 0.5047)$, respectively.

We conclude this section by comparing the collections of models identified by our search algorithms using an average divergence measure:

$$\text{DVG} = \frac{1}{mk} \sum_{h=1}^{k} \sum_{i=1}^{m} |q_i - \hat{q}_{ih}|, \tag{9}$$

where q_i is 1 if edge i belongs to the true graph and 0 otherwise, and $\hat{q}_{ih}$ is the estimated inclusion probability of edge i using data set h. We also compare to SIN, by replacing $\hat{q}_{ih}$ in (9) with 1 minus the p-value for edge i based on data set h. While the rationale behind (9) is clear, both for our search algorithms and for SIN, a naive comparison of of our algorithms to SIN in terms of DVG should be regarded as purely indicative, because edge inclusion probabilities play a different role than p-values. We found an average divergence of 0.0403, with standard deviation 0.0432, for the MFBF search, an average divergence of 0.1483, with standard deviation 0.0501 for the OFBF search, and an average divergence of 0.2164, with standard deviation 0.1658 for SIN. It is apparent that MFBF outperforms OFBF. The divergence associated to SIN is quite high, but is presumably of dubious value because of the high level of the corresponding standard deviation.

4.2 Seven-variable data on publishing productivity

We considered a real data set discussed in Spirtes et al. (2000, Example 5.8.1) and analyzed by Drton and Perlman (2008) using the frequentist procedure SIN. These data are part of a larger study aimed at investigating the interrelationship among variables potentially related to publishing productivity of academics. The sample comprises $n = 162$ subjects and seven variables, which we write in the order considered by Drton and Perlman (2008): 1. subject's sex (Sex); 2. score of the subject's ability (Ability); 3. measure of the quality of the graduate program attended (GPQ); 4. preliminary measure of productivity (PreProd); 5. quality of the first job (QFJ); 6. publication rate (Pubs); 7. citation rate (Cites).

We ran both first order ($h = 1$) MFBF searches and OFBF searches. More exactly, we ran 100 times both algorithms, each time starting with a

random base DAG, in order to better evaluate the stability of our results. We stopped all searches when $k = 1000$ distinct models had been visited. Table 2 reports our findings (for the 21 potential edges) in a format similar to that of subsection 4.1. Estimated quantities were computed from the 1000 visited DAGs within a single run of the the algorithm, while averages and 95% coverage intervals refer to the 100 replications of the algorithm.

The MFBF search shows somewhat shorter intervals relative to the OFBF search, signalling that its estimates of edge inclusion probabilities are more stable over the different 100 paths. Furthermore, a feature that was already observed in subsection 4.1 is also apparent here: edges with low inclusion probability are more clearly emphasized by the MFBF search; conversely, edges with a high estimated inclusion probability score a higher value using the OFBF search. Edge $2 \to 4$ is an interesting case, because it gets a 90% inclusion probability with OFBF and only 38% using MFBF. This is the reason why the MFBF and OFBF find different MP-DAGs; see Figure 1 (right panel for MFBF and center panel for OFBF). The MFBF search identifies a more parsimonious model, which states an additional conditional independence relationship: variable 4 (PreProd) is independent of variable 2 (Ability) given variables 1 (Sex) and 3 (GPQ). This result is consistent with our expectation that MFBFs tend to favor simpler models.

The last column of Table 2 shows the p-values given by SIN; there is no column for intervals in this case, because it makes no sense to replicate a deterministic algorithm. Interestingly, if we select edges whose p-value is below 5%, the graph found by SIN coincides with the MP-DAG identified by OFBF.

Finally, we report the estimated posterior probability of the HPP-DAG both for the OFBF and the MFBF search (in the latter case this coincides with the MP-DAG). We found an average value of 0.2596 for MFBF, with 95% coverage interval $(0.2592, 0.2606)$, and an average value of 0.0431 for OFBF, with 95% coverage interval $(0.0423, 0.0442)$. Similarly to what happened in section 4, the MFBF search piles much more probability on the HPP-DAG than the OFBF search does.

5 Application to sparse high-dimensional DAGs

We now evaluate our method in the contex of sparse high-dimensional DAGs, both on simulated and real data sets.

Table 2: Publishing productivity results. Estimated edge inclusion probabilities, searching 1000 distinct models (starting from a random base DAG) with OFBF and first order MFBF search, compared to the p-values given by SIN. Averages (A) and 95% coverage intervals (I) are reported, based on 100 replications.

Edge	OFBF M	OFBF I	MFBF M	MFBF I	SIN
$(6,7)$	1.0000	(1.0000, 1.0000)	1.0000	(1.0000, 1.0000)	0.0000
$(5,7)$	0.4257	(0.4079, 0.4386)	0.0561	(0.0558, 0.0564)	0.4613
$(4,7)$	0.9658	(0.9629, 0.9685)	0.7641	(0.7632, 0.7670)	0.0142
$(3,7)$	0.2742	(0.2558, 0.2842)	0.0464	(0.0462, 0.0466)	0.9308
$(2,7)$	0.2481	(0.2333, 0.2590)	0.0625	(0.0585, 0.0634)	0.9308
$(1,7)$	0.0778	(0.0720, 0.0839)	0.0052	(0.0051, 0.0052)	0.9239
$(5,6)$	1.0000	(1.0000, 1.0000)	0.9999	(0.9999, 0.9999)	0.0000
$(4,6)$	0.2924	(0.2729, 0.3094)	0.0341	(0.0339, 0.0342)	0.6627
$(3,6)$	0.0351	(0.0325, 0.0372)	0.0008	(0.0008, 0.0008)	0.9308
$(2,6)$	0.4352	(0.4232, 0.4463)	0.0566	(0.0565, 0.0568)	0.4007
$(1,6)$	1.0000	(1.0000, 1.0000)	1.0000	(1.0000, 1.0000)	0.0000
$(4,5)$	0.0291	(0.0260, 0.0315)	0.0007	(0.0007, 0.0007)	0.9308
$(3,5)$	0.9736	(0.9719, 0.9754)	0.7294	(0.7287, 0.7302)	0.0512
$(2,5)$	0.0261	(0.0239, 0.0281)	0.0035	(0.0034, 0.0036)	1.0000
$(1,5)$	0.0605	(0.0562, 0.0646)	0.0023	(0.0022, 0.0023)	0.9207
$(3,4)$	0.0443	(0.0413, 0.0479)	0.0015	(0.0015, 0.0015)	0.9239
$(2,4)$	0.9030	(0.8989, 0.9069)	0.3846	(0.3831, 0.3858)	0.0176
$(1,4)$	0.0313	(0.0289, 0.0335)	0.0006	(0.0006, 0.0006)	0.9308
$(2,3)$	1.0000	(1.0000, 1.0000)	1.0000	(1.0000, 1.0000)	0.0000
$(1,3)$	0.0408	(0.0384, 0.0411)	0.0012	(0.0012, 0.0012)	0.9308
$(1,2)$	0.0563	(0.0524, 0.0603)	0.0021	(0.0021, 0.0021)	0.9207

5.1 Simulated data

We generated three random DAGs of size $q = 50, 100, 200$, using the random DAG generator in the R-package `pcalg`; see Kalisch and Buhlmann (2007). We obtained, in the three cases, $|E| = 110, 97, 108$ actual edges, respectively, out of $m = q(q-1)/2 = 1225, 4950, 19900$ possible edges. Clearly, sparseness increases with q: the ratio of actual to possible edges is 9% in the first DAG, 2% in the second DAG and 0.5% in the third DAG. Then, for each of the three DAGs, we generated $n = 100$ observations according to the linear structural equation model

$$u_i = \sum_{j \in \mathrm{pa}_i} \rho_{ij} u_j + \varepsilon_i \quad i = 1, \ldots, q,$$

with $\rho_{ij} = \rho = 0.8$. Furthermore, we replicated each simulated data set 10 times in order to assess variability.

To evaluate the performance of a particular method at reconstructing the generating DAG, different measures of structural difference between graphs can be used. Baldi et al. (2000) provide an overview of techniques for assessing the accuracy of prediction algorithms for classification; we refer the reader to this paper for further details. In our case, it is enough to consider *binary* classifiers, because each edge is either present or absent in the true graph. A first index is the Structural Hamming Distance (SHD), which counts the number of edges *not* in common between the estimated and the true graph. The main drawback of this measure is its dependence on the number of vertices; in addition, it does not appear to be suitable for sparse graphs. An alternative measure of performance for binary classifiers is *Matthew's Correlation Coefficient* (Matthews, 1975):

$$\mathrm{MCC} = \frac{(\mathrm{TP} \cdot \mathrm{TN}) - (\mathrm{FP} \cdot \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}},$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives; here TP(TN) means that an edge which is present(absent) in the true DAG is *also* present(absent) in the selected DAG, while FP(FN) means that an edge which is absent(present) in the true DAG is *instead* present(absent) in the selected DAG. The index MCC varies between -1 and $+1$. A value of $-1(+1)$ indicates total disagreement(agreement) between the true DAG and the selected DAG. The value zero corresponds to a random prediction. Finally, one can use the divergence measure DVG defined in section 4.

Table 3: Sparse DAG simulation results. False positives (FP) and false negatives (FN), Matthew's correlation coefficient (MCC), and divergence index (DVG) for the MFBF search with $h = 1, 2, 4$, the OFBF search, and SIN with significance level $\alpha = 0.01, 0.02, 0.05$. Sample size $n = 100$. All indexes are averaged over 10 simulated data sets.

q		FP	FN	MCC	DVG
50	SIN($\alpha = 0.01$)	0.000 (0.000)	42.90 (4.977)	0.774 (0.030)	0.065 (0.004)
50	SIN($\alpha = 0.02$)	0.000 (0.000)	39.80 (5.138)	0.791 (0.030)	0.065 (0.004)
50	SIN($\alpha = 0.05$)	0.000 (0.000)	33.80 (6.015)	0.826 (0.034)	0.065 (0.004)
50	OFBF	2.200 (1.400)	0.000 (0.000)	0.990 (0.007)	0.214 (0.181)
50	MFBF($h = 1$)	0.500 (0.707)	0.100 (0.316)	0.997 (0.004)	0.090 (0.126)
50	MFBF($h = 2$)	0.500 (1.269)	0.400 (1.265)	0.996 (0.012)	0.054 (0.031)
50	MFBF($h = 4$)	0.000 (0.000)	0.900 (0.994)	0.996 (0.004)	0.044 (0.001)
100	OFBF	4.300 (1.767)	0.000 (0.000)	0.978 (0.009)	0.103 (0.041)
100	MFBF($h = 1$)	1.000 (0.471)	0.100 (0.316)	0.994 (0.003)	0.014 (0.005)
100	MFBF($h = 2$)	0.100 (0.352)	0.200 (0.422)	0.998 (0.003)	0.010 (0.000)
100	MFBF($h = 4$)	0.100 (0.316)	0.400 (0.966)	0.997 (0.007)	0.009 (0.000)
200	OFBF	6.800 (2.700)	0.000 (0.000)	0.970 (0.011)	0.316 (0.353)
200	MFBF($h = 1$)	1.200 (0.471)	0.000 (0.000)	0.995 (0.006)	0.140 (0.222)
200	MFBF($h = 2$)	0.500 (0.707)	0.100 (0.316)	0.997 (0.004)	0.039 (0.063)
200	MFBF($h = 4$)	0.000 (0.000)	1.300 (1.567)	0.994 (0.007)	0.004 (0.003)

Table 3 compares the performance of OFBF and MFBF (with $h = 1, 2, 4$) on the simulated data, when the MP-DAG is selected. For $q = 50$ the performance of SIN is also shown, when the DAG containg edges significant at a given level ($\alpha = 0.01, 0.02, 0.05$) is selected. This is admittedly a rather crude implementation of SIN: a refined version would require to look at graphs representing each edge "significance" and then choose α accordingly (based on subjective judgment); see Drton and Perlman (2008). There are no SIN results for $q = 100$ and $q = 200$, because SIN cannot be applied if $q \leq n$.

In Table 3 we notice a very high number of false negatives for SIN, which is outperformed by MFBF, whatever its order, with very few false positives and false negatives. Matthew's correlation coefficient for MFBF is also close to perfect agreement. Within the MFBF approach, a good compromise seems to be achieved when $h = 1$, especially if our aim is to control FN rather than

FP. This is sensible if we do not want to miss the (few) edges which in fact are present, because they might indicate an interesting signal, while we are somewhat more tolerant of adding an edge which in fact is missing.

For larger DAGs, that is, when $q = 100$ or $q = 200$, it would appear that a value of $h = 2$ is preferable relative to $h = 1$, because it strikes a good compromise on all four indicators; in particular, it controls FN better than $h = 4$. The behavior of FP and FN, as functions of h, is consistent with our expectations: as h increases, the prior privileges the smaller model (the model with fewer edges) in each pairwise comparison; this reduces FP, because declaring an edge present is less likely, while symmetrically raising FN, because declaring and edge absent is more likely.

5.2 Data on human cell signalling pathways

Sachs et al. (2003) carried out a set of flow cytometry experiments on signalling networks of human immune system cells. The ordering of the connections between pathway components was established based on perturbations in cells using molecular interventions, and we consider it to be known *a priori*. The data set includes $q = 11$ proteins and $n = 7466$ observations. Figure 2 (left panel) shows the (assumed) known regulatory network, while Table 4 shows the performance measures described in the previous subsection for the PC-algorithm (Kalisch and Buhlmann, 2007), the Lasso (Meinshausen and Bulhmann, 2006), the Adaptive Lasso (Shojaie and Michailidis, 2010), SIN, UFBF and MFBF. It appears from Table 4 that the Adaptive Lasso, which yields the best performance among the three frequentist methods presented, has a false positive rate (rFP) of 12% and a false negative rate (rFN) of 26%.

Moving to the MFBF search, we produce results corresponding to the HPP-DAG, having probability around 90% and coinciding in this application with the MP-DAG. Small values of h, say up to 10, are not able to reproduce the performance of Adaptive Lasso. However, letting h grow to higher values, such as 15–25, we can considerably lower rFP down to 12% or even 8%, while rFN remains at an approximate level of 32%. We can thus obtain, relative to Adaptive Lasso, a reduction in FP of about one third (from 12% to 8%) while suffering an increase in FN of the order of less than a quarter (from 26% to 32%). If we now focus on the actual structure of the selected DAG and consider, for comparison purposes, the results of MFBF with $h = 15$, it appears that the HPP-graph has 20 edges differing from those in the graph selected by Adaptive Lasso: the two structures are thus appreciably different.

Table 4: Sparse real data results. False positives (FP) and false negatives (FN), false positive rate (rFP) and false negative rate (rFN), Matthew's correlation coefficient (MCC), structural Hamming distance (SHD), and divergence from true model (DVG).

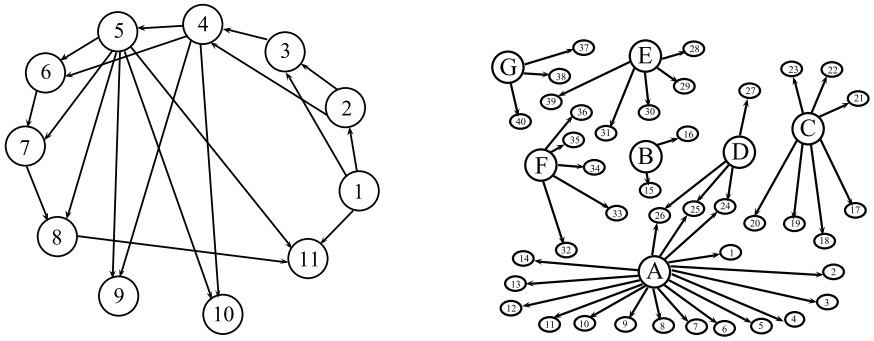
Human cell signalling pathways

	FP	FN	rFP	rFN	MCC	SHD	DVG
PC-ALG	14	8	0.137	0.421	0.397	22	
LASSO	13	6	0.128	0.316	0.493	19	
ADA LASSO	13	5	0.128	0.263	0.532	18	
UFBF	28	4	0.275	0.211	0.391	32	0.258
MFBF ($h = 1$)	23	4	0.226	0.211	0.442	27	0.218
MFBF ($h = 2$)	19	5	0.186	0.263	0.450	24	0.172
MFBF ($h = 3$)	18	6	0.177	0.316	0.423	24	0.194
MFBF ($h = 4$)	16	6	0.157	0.316	0.450	22	0.182
MFBF ($h = 5$)	16	6	0.157	0.316	0.450	22	0.180
MFBF ($h = 10$)	14	6	0.137	0.316	0.478	20	0.165
MFBF ($h = 15$)	12	6	0.118	0.316	0.509	18	0.149
MFBF ($h = 20$)	9	6	0.088	0.316	0.562	15	0.124
MFBF ($h = 25$)	8	6	0.078	0.316	0.582	14	0.112
MFBF ($h = 30$)	8	8	0.078	0.421	0.500	16	0.129
MFBF ($h = 50$)	7	9	0.069	0.474	0.479	16	0.132

Transcription regulatory network of *E. coli*

	FP	FN	rFP	rFN	MCC	SHD	DVG
PC-ALG	1	42	0.004	0.977	0.082	43	
LASSO	10	30	0.042	0.698	0.342	40	
ADA LASSO	16	27	0.068	0.628	0.345	43	
UFBF	72	20	0.304	0.465	0.176	92	0.395
FBF ($h = 1$)	38	24	0.160	0.558	0.252	62	0.255
FBF ($h = 2$)	28	27	0.118	0.628	0.252	55	0.218
FBF ($h = 3$)	24	28	0.101	0.651	0.258	52	0.197
FBF ($h = 4$)	19	28	0.080	0.651	0.297	47	0.182
FBF ($h = 5$)	16	28	0.068	0.651	0.323	44	0.172
FBF ($h = 6$)	11	30	0.046	0.698	0.330	41	0.164
FBF ($h = 7$)	4	33	0.017	0.767	0.357	37	0.160
FBF ($h = 8$)	2	35	0.008	0.814	0.345	37	0.160
FBF ($h = 9$)	1	37	0.004	0.861	0.313	38	0.173
FBF ($h = 10$)	1	41	0.004	0.953	0.150	42	0.186

Figure 2: Known regulatory network of human cell signalling pathway data (left panel) and transcription regulatory network of *E. coli* data (right panel).



5.3 Data on *E. coli* transcription regulatory network

Transcriptional regulatory networks play an important role in controlling the gene expression in cells, and incorporating the underlying regulatory network results in more efficient estimation and inference (Shojaie and Michailidis, 2009). Shojaie and Michailidis (2004) proposed the network component analysis method to infer the transcriptional regulatory network of *Escherichia coli* (*E. coli*). They provided information about the known regulatory network (Figure 2, right panel) together with gene expression data for 7 transcription factors and 40 regulated genes, so that the total number of variables is $q = 47$. By contrast, the sample size is $n = 24$. We tried to reconstruct the network structure using our MFBF search and summarize its performance in Table 4, along with that of the PC-algorithm, Lasso, Adaptive Lasso, and UFBF search.

We experimented with MFBFs of several orders ($h = 0, \ldots, 10$) and report results corresponding to the HPP-DAG. As previously remarked, when h increases rFP decreases while rFN increases. In essence, the MFBF search is able to reproduce the values of rFP and rFN achieved by any of the three frequentist methods listed at the top of Table 4. In particular, to reach approximately the values yielded by Adaptive Lasso we need $h = 5$, whereas $h = 6$ produces values close to those of Lasso, and those of the PC-algorithm can be obtained by letting $h = 10$. This does not mean that the resulting graphs are identical; indeed the number of edges differing between the DAG

selected by MFBF and the DAG chosen by the PC-algorithm, Lasso, or Adaptive Lasso lies between 5 and 12.

6 Discussion

We presented a novel objective Bayesian method for searching the space of Gaussian DAG models under the assumption of a fixed ordering of the variables. Our approach only requires default parameter priors, thus drastically simplifying the daunting task of prior elicitations in moderate to complex problems. In order to enhance the learning performance of our method, we transform these default priors into corresponding non-local (specifically moment) priors. Finally, we use an FBF approach to make our procedure operational.

Relative to alternative frequentist approaches, our method produces a posterior distribution on the space of DAG models, thus providing a better appreciation, and quantification, of the uncertainty inherent in model search; additionally, it allows inference on specific features such as edge inclusion probabilities and prediction based on model averaging. Within the Bayesian paradigm, we show that our method outperforms FBF searches based on conventional default local priors.

Our proposal is sufficiently flexible to accommodate varying degrees of model separation, through the regulation of a tuning parameter (the order exponent h of the moment prior). We found that, as sparseness of the network increases, so should h in order to favor, in each pairwise comparison, the smaller model. Indeed, by suitably modifying the value of h, we were able to reproduce typical performance summaries (such as false positive and negative rates) of some state-of-the-art frequentist methods recently proposed in the literature.

Our procedure can also be run when the sample size n is smaller than the number of variables q. The only requirement is that, in each pairwise comparison between adjacent DAGs, n be bigger than the size of the maximal parent set augmented by twice the value of h. In this sense, our method can deal with data sets characterized by a number of variables possibly much larger than the sample size, provided the underlying family of DAGs can be assumed to be sufficiently sparse.

Acknowledgements

Part of this work is based on the unpublished PhD dissertation by Davide Altomare (University of Pavia) titled *Priors for Bayesian model choice with applications to graphical models*.

References

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16,** 412–424.

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

Berger, J. and Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica* **59,** 3–15.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91,** pp. 109–122.

Consonni, G., Forster, J., and La Rocca, L. (2010). Enhanced objective Bayesian testing for the equality of two proportions. Technical Report M10-13, University of Southampton. Southampton Statistical Sciences Research Institute. Submitted.

Consonni, G. and La Rocca, L. (2011). On moment priors for Bayesian model choice with applications to directed acyclic graphs. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 9 – Proceedings of the Ninth Valencia International Meeting*. Oxford University Press. In press.

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

Dawid, A. (1999). The trouble with Bayes factors. *Research Report 202, University College London, Department of Statistical Science* .

Drton, M. and Perlman, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **91,** 591–602.

Drton, M. and Perlman, M. D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statist. Sci.* **22,** 430–449.

Drton, M. and Perlman, M. D. (2008). A SINful approach to Gaussian graphical model selection. *J. Statist. Plann. Inference* **138,** 1179–1200.

Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer, New York.

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

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,** pp. 1412–1440.

Giudici, P. and Green, P. (1999). Decomposable graphical gaussian model determination. *Biometrika* **86,** 785–801.

Johnson, V. and Rossell, D. (2010). On the use of non-local prior densities in bayiesian hypotesis tests. *Journal of the royal Statistical Society, series B* **72,** 143–170.

Kalisch, M. and Buhlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.* **8,** 613–36.

Markowetz, F. and Spang, R. (2007). Inferring cellular networks – a review. *BMC Bioninformatics* **8,** S5.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta* **405,** 412–451.

Meinshausen, N. and Bulhmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann.Statist.* **34,** 1436–62.

Moreno, E. (1997). Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness. In Dodge, Y., editor, L_1*-Statistical Procedures and Related Topics*, pages 257–270. Institute of Mathematical Statistics.

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

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

Perez, J. M. and Berger, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika* **89,** pp. 491–511.

Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and Bayes factors. In Dey, D. and Rao, C. R., editors, *Bayesian thinking: modeling and computation*, volume 25 of *Handbook of Statistics*, pages 115–149. Elsevier/North-Holland, Amsterdam.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2003). Casual protein-signaling networks derived from multiparameter single-cell data. *Science* **308,** 504–6.

Scott, J. and Carvalho, C. (2008). Feature-inclusion stochastic search for gaussian graphical models. *J. Comp. Graph. Stat.* **17,** 790–808.

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

Shojaie, A. and Michailidis, G. (2004). Transcriptome-based determination of multiple transcription regulator activities in escherichia coli by using network component analysis. *Proc. Nat. Acad. Sci.* **101,** 641–6.

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

Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97,** 519–538.

Spirtes, P., Glymour, C., and Scheines, R. (2000). Causation, prediction and search (2nd edition). *Cambridge, MA: The MIT Press.* pages 1–16.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

A

A.1 Formula for the moment fractional Bayes factor

In this Appendix we reproduce the expression of the MFBF for the comparison of two nested Gaussian DAG models, originally presented in Consonni and La Rocca (2011). We also present the corresponding expression for the comparison of two adjacent DAGs, which is of special interest for the algorithm developed in this paper.

Let $\mathcal{D}_0$ and $\mathcal{D}_1$ be two DAGs with $\mathcal{D}_0$ nested in $\mathcal{D}_1$. We will consider the default moment prior (2) under model $\mathcal{D}_1$, and an ordinary default prior under $\mathcal{D}_0$. Because of the recursive structure of the likelihood (1), and of the property of global independence satisfied by the priors, it is enough to concentrate on a single vertex. To simplify notation, we omit in the statement of the theorem the subscript j for the vertex; thus we use y instead of y_j for the data, while β and γ stand for β_j and γ_j.

Theorem A.1 *For a DAG model $\mathcal{D}_1$, consider a vertex and the associated conditional density $f(y \mid y_{pa}; \beta, \gamma)$, which is an n-variate normal distribution with expectation $X\beta$ and variance matrix $\gamma^{-1} I_n$, where X is an $n \times p$ matrix whose columns contain the observations on the parent variables (adding as first column the vector $\nVdash_n$ whenever appropriate). For the comparison of $\mathcal{D}_1$ with respect to a nested DAG model $\mathcal{D}_0$, assume a vertex moment prior $p_1^M(\beta, \gamma) \propto \gamma^{-1} \prod_{l \in L} \beta_l^{2h}$, where $L \subseteq pa$ is the subset of the parents pointing to the vertex in $\mathcal{D}_1$ but not in $\mathcal{D}_0$. Then, the fractional marginal likelihood based on the moment prior associated to the vertex is*

$$w_1(y \mid X, g) = \left(\pi b S^2\right)^{-\frac{n(1-g)}{2}} \cdot \cdot \frac{\sum_{i=0}^{h|L|} 4^{-i} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}) \Gamma(\frac{n-p-2i}{2}) (S^2)^i}{\sum_{i=0}^{h|L|} 4^{-i} H_i^{(h)}(\hat{\beta}, (X'X)^{-1}) \Gamma(\frac{ng-p-2i}{2}) (S^2)^i}, \tag{10}$$

where $0 < g < 1$ is the fraction satisfying $ng > p + 2h|L|$, $\hat{\beta} = (X'X)^{-1}X'y$, $S^2 = (y - \hat{y})'(y - \hat{y})$, with $\hat{y} = X\hat{\beta}$ and

$$H_i^{(h)}(\mu, \Sigma) = \sum_{j \in J_h(i)} \prod_{l=1}^{d} (2h)! \prod_{m=1}^{d} \frac{\sigma_{lm}^{j_{lm}}}{j_{lm}!} \prod_{l=1}^{d} \frac{\mu_l^{j_l^\star}}{j_l^\star!}, \tag{11}$$

having defined $j_l^\star = 2h - \sum_{m=1}^{d} j_{lm} - \sum_{m=1}^{d} j_{ml}$ and $J_h(i) = \{j : \sum_{l=1}^{d} \sum_{m=1}^{d} j_{lm} = i \;\&\; \forall l : j_l^\star \geq 0\}$.

The fractional marginal likelihood of $\mathcal{D}_1$ is thus given by $w_1(y;g) = \prod_{j=1}^{q} w_1(y_j \,|\, X_j, g)$, where each individual factor $w_1(y_j \,|\, X_j, g)$ is as in (10). It is important to realize that the quantity $w_1(y;b)$ is contingent upon the choice of the specific nested DAG model $\mathcal{D}_0$ used for the comparison: this determines the nature of the sets $L_j \subseteq \mathrm{pa}_j$ used in constructing the moment prior. The MFBF of $\mathcal{D}_1$ against $\mathcal{D}_0$ is now given by the ratio of the two fractional marginal likelihoods:

$$MFBF_{10}(y;g) = \prod_{j=1}^{q} \frac{w_1(y_j \,|\, X_{1j}, g)}{w_0(y_j \,|\, X_{0j}, g)} \tag{12}$$

where each individual $w_1(y_j \,|\, X_{1j}, b)$ is computed using formula (10), while each individual $w_0(y_j \,|\, X_{0j}, b)$ is directly available using standard calculations for the FBF in the normal linear model (O'Hagan and Forster, 2004, sect. 11.40); the latter can also be deduced from (10) upon setting $h = 0$ throughout. Of course, in order to compute the quantity $MFBF_{10}^{M}(y;g)$ one requires only those fractional marginal likelihoods referring to vertices with different parent structures under the two DAGs $\mathcal{D}_1$ and $\mathcal{D}_0$; otherwise $w_1(y_j \,|\, X_{1j}, g)/w_0(y_j \,|\, X_{0j}, g)$ is identically one.

If $\mathcal{D}_0$ and $\mathcal{D}_1$ differ exactly by edge $i \to j$, then

$$\begin{aligned} MFBF_{10}(y;g) = & \left(\frac{S^2}{S_0^2}\right)^{-\frac{n(1-g)}{2}} \frac{\Gamma\left(\frac{n-p}{2}\right)}{\Gamma\left(\frac{ng-p}{2}\right)} \cdot \\ & \cdot \frac{\sum_{z=0}^{h} 4^{-z} \frac{v_{ii}\hat{\beta}_{ji}^{(2h-2z)}}{z!\,(2h-2z)!} \Gamma(\frac{n-p-2z}{2})(S^2)^z}{\sum_{z=0}^{h} 4^{-z} \frac{v_{ii}\hat{\beta}_{ji}^{(2h-2z)}}{z!\,(2h-2z)!} \Gamma(\frac{ng-p-2z}{2})(S^2)^z}, \end{aligned} \tag{13}$$

where: p is the number of parents of j in $\mathcal{D}_1$ and $ng > p + 2h$; $\hat{\beta}_{ji}$ is the i-th component of $\hat{\beta}_j$, the least squares estimate of the coefficients in the regression model of u_j against $u_{\mathrm{pa}(j)}$, with $\mathrm{pa}(j)$ the set of parents of node j in $\mathcal{D}_1$, and S^2 is the associated residual sum of squares; $v_{ii} \equiv [X'_{\mathrm{pa}(j)} X_{\mathrm{pa}(v)}]_{ii}^{-1}$; S_0^2 is the residual sum of squares in the regression of u_j against $u_{\mathrm{pa}_0(j)}$, where $\mathrm{pa}_0(j)$ is the set of parents of node j in $\mathcal{D}_0$.