This is the peer reviewd version of the followng article:

Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection / Consonni, Guido; LA ROCCA, Luca; Peluso, Stefano. - In: SCANDINAVIAN JOURNAL OF STATISTICS. - ISSN 0303-6898. - 44:3(2017), pp. 741-764. [10.1111/sjos.12273]

*Terms of use:*

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

12/10/2024 07:32

(Article begins on next page)

# Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection[§]

Consonni, Guido
*Department of Statistical Sciences*
*Università Cattolica del Sacro Cuore*

La Rocca, Luca [¶]
*Department of Physics, Informatics and Mathematics*
*Università di Modena e Reggio Emilia*

Peluso, Stefano
*Department of Statistical Sciences*
*Università Cattolica del Sacro Cuore*

December 23, 2016

[§]*Running headline*: Covariate-Adjusted Sparse DAG Selection
[¶]Corresponding author (`luca.larocca@unimore.it`)

**Abstract**

We present an objective Bayes method for covariance selection in Gaussian multivariate regression models having a sparse regression and covariance structure, the latter being Markov with respect to a Directed Acyclic Graph (DAG). Our procedure can be easily complemented with a variable selection step, so that variable and graphical model selection can be performed jointly. In this way, we offer a solution to a problem of growing importance especially in the area of genetical genomics (eQTL analysis). The input of our method is a single default prior, essentially involving no subjective elicitation, while its output is a closed form marginal likelihood for every covariate-adjusted DAG model, which is constant over each class of Markov equivalent DAGs; our procedure thus naturally encompasses covariate-adjusted decomposable graphical models. In realistic experimental studies our method is highly competitive, especially when the number of responses is large relative to the sample size.

*Keywords*: Bayesian model selection; covariance selection; decomposable graphical model; directed acyclic graphical model; fractional Bayes factor; Gaussian graphical model; Gaussian multivariate regression; marginal likelihood; model sparsity; variable selection.

# 1 Introduction

Graphical models are a well-established tool in multivariate statistics. They allow to simplify high-dimensional distributions, both in terms of computations and in terms of interpretation, on the basis of a graph representing independencies between variables. We assume the reader is familiar with the basic theory of undirected and acyclic directed graphical models, as presented for instance in Cowell *et al.* (1999), or Lauritzen (1996); see also Whittaker (1990). A brief summary of our graph terminology is available online as *supporting information* for this article.

Our interest lies in a collection of $q$ random variables whose joint distribution, having density with respect to a product measure, embodies a conditional independence structure which can be represented by a Directed Acyclic Graph (DAG). This means that each variable is conditionally independent of its non-descendants given its parents; see Cowell *et al.* (1999, sect. 5.3). Such a distribution is said to be Markov with respect to the DAG. A DAG model is a (parametric) family of multivariate distributions which are Markov with respect to a DAG. We will consider in particular Gaussian DAG models. Then, the DAG structure will be reflected in the covariance matrix $\mathbf{\Sigma}$: if the DAG is complete, $\mathbf{\Sigma}$ will be unconstrained; for an incomplete DAG, $\mathbf{\Sigma}$ will present constrained entries. Notice that an unconstrained covariance matrix still has to be s.p.d. (symmetric positive definite).

Typically, the DAG structure is unknown, and we want to infer it from $n$ joint observations of the $q$ variables. From a Bayesian viewpoint one starts with a prior distribution on the collection of all DAGs (prior on model space), as well as with a prior distribution on the parameter space of each given DAG (parameter prior). Given these prior inputs, Bayesian inference produces a posterior probability on the space of all DAGs, which summarizes all the uncertainty in the light of the available data. Several papers have addressed this problem for the case in which the $n$ observations are i.i.d. (independent and identically distributed) conditionally on the parameters of the model; see for instance Dawid & Lauritzen (1993); Spiegelhalter *et al.* (1993); Heckerman *et al.* (1995); Madigan *et al.* (1996). Of special interest for this paper is the work by Geiger & Heckerman (2002); see also Consonni & La Rocca (2012) and Kuipers *et al.* (2014) for a correction. Geiger & Heckerman (2002) listed a set of assumptions on the collection of parameter priors (across DAGs) which permit their construction starting from a single parameter prior under a *complete* DAG. This represents a dramatic simplification because: i) the specification of only one distribution is required, while all the remaining priors are derived from this one;

ii) the latter distribution is placed on an unconstrained parameter space describing the model with no conditional independencies. In the Gaussian case ii) means that one can use a standard Inverse Wishart on the covariance matrix, equivalently a Wishart on the corresponding precision matrix (defined as the inverse of the covariance matrix) so that the marginal likelihood can be expressed in closed form.

Different DAGs may define the same set of conditional independencies, in which case they are called Markov equivalent. Accordingly, the set of all DAGs for the $q$ variables can be partitioned into Markov equivalence classes (corresponding to distinct statistical models). If DAGs are meant to specify exclusively conditional independencies, as opposed to causal relationships (Lauritzen, 2001; Dawid, 2003), then all DAGs within the same equivalence class should be regarded as indistinguishable using observational data. The method by Geiger & Heckerman (2002) ensures that DAGs belonging to the same equivalence class obtain the same marginal likelihood. As a consequence, their method can also be used to infer decomposable graph structures, by simply replacing each structure with an equivalent DAG (no matter which).

Despite its many advantages, the inferential procedure proposed by Geiger & Heckerman (2002) still requires the specification of a potentially high-dimensional parameter prior (especially in large $q$ settings). This naturally suggests an objective Bayes approach, which is virtually free from prior elicitation. We carried out this program in Consonni & La Rocca (2012) for Gaussian DAG models, using the method of the fractional Bayes factor (O'Hagan, 1995). Our findings were consistent with, and extended, those presented in Carvalho & Scott (2009) for Gaussian decomposable graphical models, which relied on the use of the hyper-inverse Wishart distribution (Letac & Massam, 2007).

More recently, research has shifted towards *covariate-adjusted* estimation of covariance matrices. Motivation for this research stems from the analysis of genetical genomics data (eQTL analysis) where the aim is to study conditional dependence structures of gene expressions after the confounding genetic effects are taken into account. Indeed, an important finding from many genetical genomics experiments is that the gene expression level of many genes is inheritable and can be partially explained by genetic variation; see e.g. Brem & Kruglyak (2005). Since some genetic variants have effects on the expression of multiple genes, they act as confounders when trying to learn the association between the genes. Accordingly, ignoring the effects of genetic variants on the gene expression levels can lead to both false positive and false negative associations in the gene network graph. The effect of genetic variants on gene expression therefore needs to be adjusted in estimating the high-dimensional

precision matrix. Work in this direction was carried out by Rothman *et al.* (2010); Yin & Li (2011); Sohn & Kim (2012); Zhang & Kim (2014); Chen *et al.* (2016). Sohn & Kim (2012) also considered a financial application, while Wytock & Kolter (2013) dealt with large-scale energy forecasting in the same framework.

The problem is usually formulated as one of joint estimation of multiple regression coefficients and a precision matrix, with the latter assumed to be Markov with respect to some graph. Since these models are used in high-dimensional settings, both the regression and the covariance structure are assumed to be sparse. All of the above work assumes that the error term is multivariate normal; this assumption is relaxed in the paper by Cai *et al.* (2013). The literature in the area, as exemplified by the above papers, is carried out within a penalized likelihood maximization approach (under a suitable norm). Bayesian contributions are still very limited; a notable exception is Bhadra & Mallick (2013) who perform variable and covariance selection jointly, using decomposable graphs and weakly informative hierarchical priors.

In this paper we deal with covariate-adjusted selection of Gaussian DAG models within an objective Bayes framework. Specifically, we reconsider the foundations of the approach by Geiger & Heckerman (2002), originally presented for the case of i.i.d. sampling, and show that it can be meaningfully extended to the multivariate regression setting. We provide closed-form expressions for the marginal likelihood of any DAG, then we propose an objective Bayes procedure, based on the fractional Bayes factor, which works for DAGs with small parent sets. Our results extend to the regression setup those of Consonni & La Rocca (2012) and Carvalho & Scott (2009); they also complement those of Bhadra & Mallick (2013), because they are derived within an objective framework, and cope with a broader family of graphs, while requiring a theoretically simpler setup.

The paper is organized as follows. Section 2 reviews the matrix distributions used in the paper, and section 3 presents the Gaussian multivariate regression setup. Section 4 illustrates our objective framework, while section 5 contains our proposal for covariance selection. In section 6 our method is compared through simulations to available alternative approaches. Finally, section 7 briefly discusses our work.

# 2 Matrix distributions

Consider $n$ independent observations on $q$ continuous dependent variables, arranged in an $n \times q$ response matrix:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1^\top \\ \vdots \\ \boldsymbol{y}_n^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{Y}_1 & \dots & \boldsymbol{Y}_q \end{pmatrix}, \tag{1}$$

where $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iq})$ is the $i$-th observation, and $\boldsymbol{Y}_j = (y_{1j}, \dots, y_{nj})$ represents the observations on the $j$-th variable. Let $\boldsymbol{X}$ be a design matrix with $n$ rows and $p+1$ columns ($p$ predictors plus intercept) which we assume known without error; denote by $\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_n^\top$ its rows. We model the observations as $\boldsymbol{y}_i \mid \boldsymbol{B}, \boldsymbol{\Sigma} \sim \mathcal{N}_q(\boldsymbol{B}^\top \boldsymbol{x}_i, \boldsymbol{\Sigma})$, independently over $i = 1, \dots, n$, where $\boldsymbol{B}$ is an unconstrained $(p+1) \times q$ matrix, $\boldsymbol{\Sigma}$ is an unconstrained (s.p.d.) $q \times q$ matrix, and $\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the $q$-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The $j$-th column of $\boldsymbol{B}$, namely $\boldsymbol{B}_j$, is the vector of regression coefficients for the $j$-th variable, and $\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Sigma}) = \boldsymbol{X}\boldsymbol{B}$. The distribution of $\boldsymbol{Y}$, given $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, is a special case of the matrix normal distribution; the general case, reviewed in section 2.1, will give a conjugate prior for $\boldsymbol{B}$ (given $\boldsymbol{\Sigma}$). A conjugate prior for $\boldsymbol{\Sigma}^{-1}$ will be given by the Wishart distribution, which is reviewed in section 2.2.

## 2.1 Matrix normal

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 $\boldsymbol{\Sigma}$, when $\mathrm{vec}(\boldsymbol{Y})$ follows the multivariate normal distribution with mean vector $\mathrm{vec}(\boldsymbol{M})$ and covariance matrix $\boldsymbol{\Sigma} \otimes \boldsymbol{\Phi}$; recall that $\mathrm{vec}(\boldsymbol{Y})$ is the vector obtained from $\boldsymbol{Y}$ by stacking its columns on top of one another, while $\otimes$ denotes the Kronecker product. If $\boldsymbol{Y}$ is an $n \times q$ matrix, $\boldsymbol{M}$ will be an $n \times q$ matrix, $\boldsymbol{\Phi}$ an s.p.d. $n \times n$ matrix, $\boldsymbol{\Sigma}$ an s.p.d. $q \times q$ matrix, and we will write

$$\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma} \sim \mathcal{N}_{n,q}(\boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}); \tag{2}$$

see Gupta & Nagar (2000, p. 55), and Dawid (1981), for more information. We obtain the special case where $\boldsymbol{Y}$ is the response matrix described above by letting $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{B}$, and $\boldsymbol{\Phi} = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix; this will be taken up in section 3.

Let $\phi_{ii'}$ denote the generic element of $\boldsymbol{\Phi}$, and $\sigma_{jj'}$ the generic element of $\boldsymbol{\Sigma}$. Clearly, we have $\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}) = \boldsymbol{M}$. Moreover, we have $\mathbb{C}\mathrm{ov}(y_{ij}, y_{i'j'} \mid \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}) = \phi_{ii'}\sigma_{jj'}$,

so that $\mathbb{V}\mathrm{ar}(\boldsymbol{y}_i \,|\, \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}) = \phi_{ii}\boldsymbol{\Sigma}$, $i = 1, \ldots, n$, whereas $\mathbb{V}\mathrm{ar}(\boldsymbol{Y}_j \,|\, \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}) = \sigma_{jj}\boldsymbol{\Phi}$, $j = 1, \ldots, q$, with $\mathbb{V}\mathrm{ar}(\boldsymbol{u})$ denoting the covariance matrix of the random vector $\boldsymbol{u}$. More generally, we find $\mathbb{C}\mathrm{ov}(\boldsymbol{y}_i, \boldsymbol{y}_{i'} \,|\, \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}) = \phi_{ii'}\boldsymbol{\Sigma}$ and $\mathbb{C}\mathrm{ov}(\boldsymbol{Y}_j, \boldsymbol{Y}_{j'} \,|\, \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}) = \sigma_{jj'}\boldsymbol{\Phi}$, if we denote by $\mathbb{C}\mathrm{ov}(\boldsymbol{u}, \boldsymbol{v})$ the cross-covariance matrix of $\boldsymbol{u}$ and $\boldsymbol{v}$, whose elements are the covariances between all pairs consisting of one element in $\boldsymbol{u}$ and the other in $\boldsymbol{v}$. Notice that $\mathbb{C}\mathrm{ov}(\boldsymbol{u}, \boldsymbol{u}) = \mathbb{V}\mathrm{ar}(\boldsymbol{u})$.

Reparameterizing from $\boldsymbol{\Sigma}$ s.p.d. to $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ s.p.d., and from $\boldsymbol{\Phi}$ s.p.d. to $\boldsymbol{K} = \boldsymbol{\Phi}^{-1}$ s.p.d., which we will find useful for Bayesian analysis, the density of the matrix normal distribution $\mathcal{N}_{n,q}(\boldsymbol{M}, \boldsymbol{K}^{-1}, \boldsymbol{\Omega}^{-1})$ can be written as

$$f(\boldsymbol{Y} \,|\, \boldsymbol{M}, \boldsymbol{K}, \boldsymbol{\Omega}) = \frac{|\boldsymbol{K}|^{\frac{q}{2}}|\boldsymbol{\Omega}|^{\frac{n}{2}}}{(2\pi)^{\frac{nq}{2}}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Omega}(\boldsymbol{Y} - \boldsymbol{M})^\top \boldsymbol{K}(\boldsymbol{Y} - \boldsymbol{M})\right)\right\}, \tag{3}$$

where $|\boldsymbol{\Psi}|$ denotes the determinant of the matrix $\boldsymbol{\Psi}$, and $\mathrm{tr}(\boldsymbol{\Psi})$ its trace. Formula (3) follows from the density of $\mathrm{vec}(\boldsymbol{Y}) \,|\, \mathrm{vec}(\boldsymbol{M}), \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{K}^{-1}$, keeping into account that $\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Psi}\boldsymbol{K}\boldsymbol{\Psi}^\top) = \mathrm{tr}(\boldsymbol{\Psi}^\top\boldsymbol{\Omega}\boldsymbol{\Psi}\boldsymbol{K})$ is the value at $(\boldsymbol{\Psi}, \boldsymbol{\Psi})$ of the bilinear form associated to $\boldsymbol{\Omega} \otimes \boldsymbol{K} = (\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{K}^{-1})^{-1}$, which is the precision matrix of $\mathrm{vec}(\boldsymbol{Y})$, and that $|\boldsymbol{\Omega} \otimes \boldsymbol{K}| = |\boldsymbol{\Omega}|^n|\boldsymbol{K}|^q$; see Lauritzen (1996, appendix B). We call $\boldsymbol{K}$ the row precision matrix of $\boldsymbol{Y}$, and $\boldsymbol{\Omega}$ its column precision matrix. Clearly, whenever $\boldsymbol{Y} \,|\, \boldsymbol{M}, \boldsymbol{K}, \boldsymbol{\Omega} \sim \mathcal{N}_{n,q}(\boldsymbol{M}, \boldsymbol{K}^{-1}, \boldsymbol{\Omega}^{-1})$, we have $\boldsymbol{Y}^\top \,|\, \boldsymbol{M}, \boldsymbol{K}, \boldsymbol{\Omega} \sim \mathcal{N}_{q,n}(\boldsymbol{M}^\top, \boldsymbol{\Omega}^{-1}, \boldsymbol{K}^{-1})$, which means $\mathrm{vec}(\boldsymbol{Y}^\top) \,|\, \boldsymbol{M}, \boldsymbol{K}, \boldsymbol{\Omega} \sim \mathcal{N}_{qn}(\mathrm{vec}(\boldsymbol{M}^\top), \boldsymbol{K}^{-1} \otimes \boldsymbol{\Omega}^{-1})$.

Now let $J$ be a proper subset of $\{1, \ldots, q\}$, and denote by $\boldsymbol{Y}_J$ the submatrix of $\boldsymbol{Y}$ consisting of the columns indexed by $J$. It is immediate to check that $\mathrm{vec}(\boldsymbol{Y}_J)$ is multivariate normal with mean vector $\mathrm{vec}(\boldsymbol{M}_J)$ and covariance matrix $\boldsymbol{\Sigma}_{JJ} \otimes \boldsymbol{\Phi}$, where $\boldsymbol{\Sigma}_{JJ}$ is the submatrix of $\boldsymbol{\Sigma}$ consisting of the rows and columns indexed by $J$; see Lauritzen (1996, prop. (C.4)). Hence, *column marginalization* results in

$$\boldsymbol{Y}_J \,|\, \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma} \sim \mathcal{N}_{n,|J|}(\boldsymbol{M}_J, \boldsymbol{\Phi}, \boldsymbol{\Sigma}_{JJ}). \tag{4}$$

Notice that, if $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{B}$, then $\boldsymbol{M}_J = \boldsymbol{X}\boldsymbol{B}_J$.

Finally, letting $\bar{J} = \{1, \ldots, q\} \setminus J$, it is well known that $\mathrm{vec}(\boldsymbol{Y}_J) \,|\, \mathrm{vec}(\boldsymbol{Y}_{\bar{J}})$ is multivariate normal with mean vector $\mathrm{vec}(\boldsymbol{M}_J) - (\boldsymbol{\Omega}_{JJ}^{-1} \otimes \boldsymbol{K}^{-1})(\boldsymbol{\Omega}_{J\bar{J}} \otimes \boldsymbol{K})\mathrm{vec}(\boldsymbol{Y}_{\bar{J}} - \boldsymbol{M}_{\bar{J}})$, and precision matrix $\boldsymbol{\Omega}_{JJ} \otimes \boldsymbol{K}$, where $\boldsymbol{\Omega}_{JJ}^{-1} = (\boldsymbol{\Omega}_{JJ})^{-1}$; see Lauritzen (1996, prop. C.5). Since $(\boldsymbol{\Omega}_{JJ}^{-1} \otimes \boldsymbol{K}^{-1})(\boldsymbol{\Omega}_{J\bar{J}} \otimes \boldsymbol{K}) = (\boldsymbol{\Omega}_{JJ}^{-1}\boldsymbol{\Omega}_{J\bar{J}}) \otimes (\boldsymbol{K}^{-1}\boldsymbol{K}) = (\boldsymbol{\Omega}_{JJ}^{-1}\boldsymbol{\Omega}_{J\bar{J}}) \otimes \boldsymbol{I}_n$, we find

$$\boldsymbol{Y}_J \,|\, \boldsymbol{Y}_{\bar{J}}, \boldsymbol{M}, \boldsymbol{K}, \boldsymbol{\Omega} \sim \mathcal{N}_{n,|J|}(\boldsymbol{M}_J - (\boldsymbol{Y}_{\bar{J}} - \boldsymbol{M}_{\bar{J}})\boldsymbol{\Omega}_{\bar{J}J}\boldsymbol{\Omega}_{JJ}^{-1}, \boldsymbol{K}^{-1}, \boldsymbol{\Omega}_{JJ}^{-1}) \tag{5}$$

for *column conditioning*. In the case $\boldsymbol{K} = \boldsymbol{I}_n$, formula (5) returns $\boldsymbol{y}_{iJ} \,|\, \boldsymbol{M}, \boldsymbol{K}, \boldsymbol{\Omega} \sim \mathcal{N}_{|J|}(\boldsymbol{m}_{iJ} - \boldsymbol{\Omega}_{JJ}^{-1}\boldsymbol{\Omega}_{J\bar{J}}(\boldsymbol{y}_{i\bar{J}} - \boldsymbol{m}_{i\bar{J}}), \boldsymbol{\Omega}_{JJ}^{-1})$, independently over $i = 1, \ldots, n$, where $\boldsymbol{y}_{iJ}$ and $\boldsymbol{m}_{iJ}$ are the subvectors of $\boldsymbol{y}_i$ and $\boldsymbol{m}_i$, respectively, consisting of the elements indexed by $J$, while $\boldsymbol{m}_i^\top$ is the $i$-th row of $\boldsymbol{M}$.

## 2.2 Wishart

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

$$p(\boldsymbol{\Omega}) = \frac{1}{2^{\frac{aq}{2}}\Gamma_q(\frac{a}{2})} |\boldsymbol{R}|^{\frac{a}{2}} |\boldsymbol{\Omega}|^{\frac{a-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{R})\right\}, \tag{6}$$

$\boldsymbol{\Omega}$ s.p.d., and $p(\boldsymbol{\Omega}) = 0$, otherwise, where $\boldsymbol{R}$ is a $q \times q$ s.p.d. matrix, $a$ is a scalar strictly greater than $q-1$, and $\Gamma_q(\frac{a}{2}) = \pi^{\frac{q(q-1)}{4}} \prod_{j=1}^{q} \Gamma\left(\frac{a}{2} + \frac{1-j}{2}\right)$ is the $q$-dimensional gamma function evaluated at $a/2$ (generalizing $\Gamma(a/2) = \int_0^\infty z^{\frac{a}{2}-1} e^{-z} dz$). As for parameter interpretation, it can be shown that $\mathbb{E}[\boldsymbol{\Omega}|\boldsymbol{R}, a] = a\boldsymbol{R}^{-1}$. Our notation $\mathcal{W}_q(a, \boldsymbol{R})$ for the density (6) is essentially that of DeGroot (1970, p. 59); other authors (Press, 1982; Lauritzen, 1996) would use $\boldsymbol{R}^{-1}$ in place of $\boldsymbol{R}$.

We now recall some useful results. Let $\boldsymbol{\Omega}$ be the precision matrix of $\boldsymbol{y} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, that is, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Think of $\boldsymbol{y}$ as the generic row of the matrix $\boldsymbol{Y}$ (dropping subscript $i$). Partition $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ into the blocks corresponding to the variables indexed by $J$ and its complement $\bar{J}$, for a given proper subset $J$ of $\{1, \ldots, q\}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{JJ} & \boldsymbol{\Sigma}_{J\bar{J}} \\ \boldsymbol{\Sigma}_{\bar{J}J} & \boldsymbol{\Sigma}_{\bar{J}\bar{J}} \end{bmatrix}, \quad \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{JJ} & \boldsymbol{\Omega}_{J\bar{J}} \\ \boldsymbol{\Omega}_{\bar{J}J} & \boldsymbol{\Omega}_{\bar{J}\bar{J}} \end{bmatrix}. \tag{7}$$

The block $\boldsymbol{\Sigma}_{JJ}$ is the marginal covariance matrix of $\boldsymbol{y}_J$ (obtained from $\boldsymbol{y}$ by selecting the elements of $\boldsymbol{y}$ indexed by $J$). Denote by $\boldsymbol{\Sigma}_{JJ.\bar{J}}$ the conditional covariance matrix $\mathbb{V}\mathrm{ar}(\boldsymbol{y}_J \,|\, \boldsymbol{y}_{\bar{J}})$ of $\boldsymbol{y}_J$ given $\boldsymbol{y}_{\bar{J}}$ (obtained from $\boldsymbol{y}$ by complementary selection). Then

$$\boldsymbol{\Sigma}_{JJ.\bar{J}} = \boldsymbol{\Sigma}_{JJ} - \boldsymbol{\Sigma}_{J\bar{J}}\boldsymbol{\Sigma}_{\bar{J}\bar{J}}^{-1}\boldsymbol{\Sigma}_{\bar{J}J} = \boldsymbol{\Omega}_{JJ}^{-1}, \tag{8}$$

that is, $\boldsymbol{\Sigma}_{JJ.\bar{J}}$ is the *Schur complement* of $\boldsymbol{\Sigma}_{\bar{J}\bar{J}}$ in $\boldsymbol{\Sigma}$, as well as the inverse of $\boldsymbol{\Omega}_{JJ}$.

Formula (8) expresses a relationship between four blocks of $\boldsymbol{\Sigma}$ and a corresponding block of $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega}$. Hence, by switching the roles of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$, we obtain

$$\boldsymbol{\Sigma}_{JJ} = \left(\boldsymbol{\Omega}_{JJ} - \boldsymbol{\Omega}_{J\bar{J}}\boldsymbol{\Omega}_{\bar{J}\bar{J}}^{-1}\boldsymbol{\Omega}_{\bar{J}J}\right)^{-1} = \boldsymbol{\Omega}_{JJ.\bar{J}}^{-1}, \tag{9}$$

where $\boldsymbol{\Omega}^{-1}_{JJ.\bar{J}}$ is to be interpreted as Schur complementation followed by inversion. Therefore, working with covariance matrices, marginalization corresponds to submatrix extraction and conditioning to Schur complementation, whereas, working with precision matrices, marginalization corresponds to Schur complementation and conditioning to submatrix extraction.

Now let $\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{R})$, with $\boldsymbol{R}$ an s.p.d. matrix and $a > q-1$. If $\boldsymbol{\Omega}$ is partitioned as in (7), and $\boldsymbol{R}$ is partitioned accordingly, then

$$\boldsymbol{\Omega}_{JJ.\bar{J}} \quad \sim \quad \mathcal{W}_{|J|}(a - |\bar{J}|, \boldsymbol{R}_{JJ}), \tag{10}$$

independently of $(\boldsymbol{\Omega}_{J\bar{J}}, \boldsymbol{\Omega}_{\bar{J}\bar{J}})$, where of course $|\bar{J}| = q - |J|$; see Lauritzen (1996, prop. C.15) who also gives the distribution of $(\boldsymbol{\Omega}_{J\bar{J}}, \boldsymbol{\Omega}_{\bar{J}\bar{J}})$.

# 3 Gaussian multivariate regression

We return to the scenario discussed in the Introduction, leading to covariate-adjusted graphical model selection, and to the response matrix $\boldsymbol{Y}$ introduced at the beginning of section 2. Denote by $\boldsymbol{Z}$ the $n \times p_\star$ matrix of all possible $p_\star$ predictors. In eQTL analysis $p_\star$ is typically very large, and often much larger than $n$. However, because of sparsity considerations, only models of the type $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ need be taken into consideration, where $\boldsymbol{X}$ is an $n \times (p+1)$ design matrix having the unit vector $\mathbf{1}_n$ as first column and $p \ll p_\star$ predictors selected from $\boldsymbol{Z}$ as remaining columns, while $\boldsymbol{E}$ is an $n \times q$ matrix of error terms with distribution $\mathcal{N}_{n,q}(\mathbf{0}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1})$, $\boldsymbol{B}$ is an unconstrained $(p+1) \times q$ matrix of regression coefficients, $\mathbf{0}$ is the $n \times q$ zero matrix, and $\boldsymbol{\Omega}$ is an unconstrained (s.p.d.) $q \times q$ matrix. Henceforth we will assume $n > p+1$, which is quite a reasonable assumption as illustrated in section 6. Notice that the $p$ predictors to be used will not be known *a priori*, and thus it will be necessary to carry out variable selection together with covariance selection; this will be feasible using the marginal likelihoods corresponding to different design matrices. For simplicity, we will use a single $\boldsymbol{X}$ in our notation (without explicitly conditioning on it).

In section 3.1 we summarize the main features of a standard *conjugate analysis* of the model

$$\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Omega} \sim \mathcal{N}_{n,q}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1}). \tag{11}$$

This is done for completeness and for the benefit of the reader, so that the subsequent sections can be followed more easily; see also Rowe (2003), whose notation is somewhat different from ours. We remark that, because of the theory presented in

section 5.1, we need only consider an *unconstrained* $\boldsymbol{\Omega}$ even when the actual context involves covariance matrices having a graphical structure. This is indeed a major simplification characterizing the approach taken in this paper; we will return to this issue later on. Next, in section 3.2, we derive the marginal data distribution for a subset of variables (selected columns of $\boldsymbol{Y}$) which represents the building block for computing the marginal likelihood of a general DAG model (as detailed in section 5.1).

## 3.1 Conjugate analysis

If we denote by $\hat{\boldsymbol{B}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$ the least squares estimator of $\boldsymbol{B}$, the likelihood function can be written as

$$f(\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Omega}) = \frac{|\boldsymbol{\Omega}|^{\frac{n}{2}}}{(2\pi)^{\frac{nq}{2}}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Omega}\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}\}\right)\right\}, \quad (12)$$

where $\hat{\boldsymbol{E}} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}})$ is the matrix of residuals. Hence, a conjugate prior for $(\boldsymbol{B}, \boldsymbol{\Omega})$ is obtained by letting

$$\begin{aligned} \boldsymbol{B} \mid \boldsymbol{\Omega} &\sim \mathcal{N}_{p+1,q}(\underline{\boldsymbol{B}}, \boldsymbol{C}^{-1}, \boldsymbol{\Omega}^{-1}), \\ \boldsymbol{\Omega} &\sim \mathcal{W}_q(a, \boldsymbol{R}), \end{aligned}$$

which results in the prior density

$$p(\boldsymbol{B}, \boldsymbol{\Omega}) = \frac{|\boldsymbol{\Omega}|^{\frac{(p+1)+(a-q-1)}{2}}}{K(\boldsymbol{C}, \boldsymbol{R}, a)} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Omega}\{(\boldsymbol{B}-\underline{\boldsymbol{B}})^\top\boldsymbol{C}(\boldsymbol{B}-\underline{\boldsymbol{B}}) + \boldsymbol{R}\}\right)\right\}, \quad (13)$$

where

$$K(\boldsymbol{C}, \boldsymbol{R}, a) = \frac{(2\pi)^{\frac{q(p+1)}{2}} 2^{\frac{aq}{2}} \Gamma_q(\frac{a}{2})}{|\boldsymbol{C}|^{\frac{q}{2}} |\boldsymbol{R}|^{\frac{a}{2}}} \quad (14)$$

is the prior normalizing constant. We remark that $\boldsymbol{C}$ is the prior precision matrix of $(\boldsymbol{\Omega}^{-1})_{jj}\boldsymbol{B}_j$, given $\boldsymbol{\Omega}$, for all $j = 1, \ldots, q$. The prior (13) is a matrix normal Wishart.

Some algebraic manipulations show that the posterior distribution of $(\boldsymbol{B}, \boldsymbol{\Omega})$ is

$$\begin{aligned} \boldsymbol{B} \mid \boldsymbol{\Omega}, \boldsymbol{Y} &\sim \mathcal{N}_{p+1,q}(\overline{\boldsymbol{B}}, (\boldsymbol{C} + \boldsymbol{X}^\top\boldsymbol{X})^{-1}, \boldsymbol{\Omega}^{-1}), \\ \boldsymbol{\Omega} \mid \boldsymbol{Y} &\sim \mathcal{W}_q(a+n, \boldsymbol{R} + \hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}} + \boldsymbol{D}), \end{aligned}$$

where $\overline{\boldsymbol{B}} = (\boldsymbol{C} + \boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{C}\underline{\boldsymbol{B}})$ is the posterior expectation (matrix) of $\boldsymbol{B}$, and $\boldsymbol{D} = (\underline{\boldsymbol{B}} - \hat{\boldsymbol{B}})^\top\{\boldsymbol{C}^{-1} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\}^{-1}(\underline{\boldsymbol{B}} - \hat{\boldsymbol{B}})$ is a measure of discrepancy between

$\underline{\boldsymbol{B}}$ and $\hat{\boldsymbol{B}}$ (prior and data). Prior-to-posterior updating thus takes the form

$$\underline{\boldsymbol{B}} \mapsto \overline{\boldsymbol{B}}, \quad \boldsymbol{C} \mapsto \boldsymbol{C} + \boldsymbol{X}^\top \boldsymbol{X}, \quad a \mapsto a + n, \quad \boldsymbol{R} \mapsto \boldsymbol{R} + \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}} + \boldsymbol{D}, \tag{15}$$

and the posterior density $p(\boldsymbol{B}, \boldsymbol{\Omega} \,|\, \boldsymbol{Y})$ is as in (13) with hyper-parameters updated by (15); the posterior normalizing constant will be given by

$$K(\boldsymbol{C} + \boldsymbol{X}^\top \boldsymbol{X}, \boldsymbol{R} + \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}} + \boldsymbol{D}, a + n), \tag{16}$$

with the function $K(\cdot, \cdot, \cdot)$ defined in (14).

## 3.2 Marginal data distribution

The marginal distribution of the matrix $\boldsymbol{Y}$ can be obtained as

$$m(\boldsymbol{Y}) = \frac{f(\boldsymbol{Y} \,|\, \boldsymbol{B}, \boldsymbol{\Omega}) p(\boldsymbol{B}, \boldsymbol{\Omega})}{p(\boldsymbol{B}, \boldsymbol{\Omega} \,|\, \boldsymbol{Y})},$$

which in light of conjugacy gives

$$m(\boldsymbol{Y}) = \frac{K(\boldsymbol{C} + \boldsymbol{X}^\top \boldsymbol{X}, \boldsymbol{R} + \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}} + \boldsymbol{D}, a + n)}{(2\pi)^{\frac{nq}{2}} K(\boldsymbol{C}, \boldsymbol{R}, a)}, \tag{17}$$

that is, up to a multiplicative factor, the ratio of the posterior and prior normalizing constants, (16) and (14), respectively.

In the sequel, we will also need the marginal distribution of selected columns of the data matrix $\boldsymbol{Y}$, corresponding to a proper subset $J$ of the full set of $q$ response variables. Let $\boldsymbol{Y}_J$ be the $n \times |J|$ selected data submatrix, and $\boldsymbol{B}_J$ be the corresponding $(p+1) \times |J|$ submatrix of $\boldsymbol{B}$, whose columns contain the regression coefficients for the selected responses. When restricted to the set $J$ of response variables, by the results presented in section 2, the Gaussian multivariate regression model (11) can be written as

$$\boldsymbol{Y}_J \,|\, \boldsymbol{B}_J, \boldsymbol{\Omega}_{JJ.\bar{J}} \sim \mathcal{N}_{n,|J|} \left( \boldsymbol{X} \boldsymbol{B}_J, \boldsymbol{I}_n, \boldsymbol{\Omega}_{JJ.\bar{J}}^{-1} \right),$$

with induced prior

$$\begin{aligned} \boldsymbol{B}_J \,|\, \boldsymbol{\Omega}_{JJ.\bar{J}} &\sim \mathcal{N}_{p+1,|J|} \left( \underline{\boldsymbol{B}}_J, \boldsymbol{C}^{-1}, \boldsymbol{\Omega}_{JJ.\bar{J}}^{-1} \right), \\ \boldsymbol{\Omega}_{JJ.\bar{J}} &\sim \mathcal{W}_{|J|}(a - |\bar{J}|, \boldsymbol{R}_{JJ}), \end{aligned}$$

where $\underline{\boldsymbol{B}}_J$ is the appropriate submatrix of $\underline{\boldsymbol{B}}$.

One readily sees that the formal structure of model and prior for a subset $J$ of response variables is the same as for the full data matrix. As a consequence, the marginal data distribution for the submatrix $\boldsymbol{Y}_J$ is given by (17) with the following substitutions:

$$q \mapsto |J|,\ \boldsymbol{R} \mapsto \boldsymbol{R}_{JJ},\ a \mapsto a - |\bar{J}|,\ \underline{\boldsymbol{B}} \mapsto \underline{\boldsymbol{B}}_J,\ \hat{\boldsymbol{B}} \mapsto \hat{\boldsymbol{B}}_J,\ \hat{\boldsymbol{E}} \mapsto \hat{\boldsymbol{E}}_J,\ \boldsymbol{D} \mapsto \boldsymbol{D}_{JJ},$$

while $n$, $\boldsymbol{C}$ and $\boldsymbol{X}$ remain unchanged.

# 4 Objective analysis

We assume the reader is familiar with the basic concepts of model selection from the Bayesian perspective, as described for instance in O'Hagan & Forster (2004, ch. 7). Here, in section 4.1, we provide some background on *objective Bayes* model selection, focusing in particular on a proposal by O'Hagan (1995). Then, in section 4.2, we give the expression for the marginal data distribution of a generic subset of columns of $\boldsymbol{Y}$ under the prior implied by such proposal; this will be instrumental in the construction of the marginal likelihood of a DAG model given in section 5.1.

## 4.1 Fractional parameter priors

Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be a collection of Bayesian models for the same observable $\boldsymbol{Y}$. Each model $\mathcal{M}_k$, $k = 1, \ldots, K$, consists of a family of sampling densities $f_k(\boldsymbol{Y} \mid \boldsymbol{\theta}_k)$, indexed by a model specific parameter $\boldsymbol{\theta}_k$, and of a prior density $p_k(\boldsymbol{\theta}_k)$ on $\boldsymbol{\theta}_k$, which we assume to be *proper*. We focus on the comparison of $\mathcal{M}_k$ with $\mathcal{M}_{k'}$ through the Bayes factor. The Bayes factor for $\mathcal{M}_k$ against $\mathcal{M}_{k'}$ is defined as $BF_{kk'}(\boldsymbol{Y}) = m_k(\boldsymbol{Y})/m_{k'}(\boldsymbol{Y})$, where $m_k(\boldsymbol{Y}) = \int f_k(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p_k(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k$ is the marginal density of $\boldsymbol{Y}$ under $\mathcal{M}_k$, also known as the marginal likelihood of $\mathcal{M}_k$.

In lack of substantive prior information, we would like to take $p_k(\boldsymbol{\theta}_k) = p_k^D(\boldsymbol{\theta}_k)$ for some objective default (non-informative) parameter prior $p_k^D(\boldsymbol{\theta}_k)$. However, objective priors are often improper and they cannot be naively used to compute Bayes factors, even when the marginal likelihoods $m_k(\boldsymbol{Y})$ are finite and non-zero, because of the presence of arbitrary constants which do not cancel out in their ratios. Pericchi (2005) reviews several proposals put forward to address this issue. In this paper, we take advantage of the fractional Bayes factor originally introduced by O'Hagan (1995); see also O'Hagan & Forster (2004).

Let $b = b(n)$, $0 < b < 1$, be a fraction of the number of observations $n$. Define

$$m_k(\boldsymbol{Y}; b) = \frac{\int f_k(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p_k^D(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k}{\int f_k^b(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p_k^D(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k}, \tag{18}$$

where $f_k^b(\boldsymbol{Y} \mid \boldsymbol{\theta}_k)$ 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. The *fractional marginal likelihood* (18) of model $\mathcal{M}_k$, can be rewritten as

$$m_k(\boldsymbol{Y}; b) = \int f_k^{1-b}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p_k^F(\boldsymbol{\theta}_k \mid b, \boldsymbol{Y}) d\boldsymbol{\theta}_k,$$

where $p_k^F(\boldsymbol{\theta}_k \mid b, \boldsymbol{Y}) \propto f_k^b(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p_k^D(\boldsymbol{\theta}_k)$ is the implied *fractional prior* (actually a "posterior" based on the fractional likelihood and the default prior). The fractional Bayes factor for $\mathcal{M}_k$ against $\mathcal{M}_{k'}$ is then defined as the ratio of $m_k(\boldsymbol{Y}; b)$ to $m_{k'}(\boldsymbol{Y}; b)$. In essence, a fraction of the likelihood is used to obtain a proper prior, which is then applied to the complementary fraction.

Clearly, the fractional prior depends on the choice of $b$. Usually $b$ will be small, so that dependence of the prior on the data will be weak. Consistency is achieved as long as $b \to 0$ for $n \to \infty$. O'Hagan (1995, sect. 4) suggests $b = n_0/n$ as a default choice, where $n_0$ is the minimal (integer) training sample size for which the fractional marginal likelihood is well defined, together with a couple of alternative choices, to be used when robustness is an issue. Moreno (1997) has an argument according to which the default choice is the only valid one, and we stick to this choice in this paper.

## 4.2 Fractional marginal likelihoods

Consider the Gaussian multivariate regression model (11). We start from the improper prior

$$p^D(\boldsymbol{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D - q - 1}{2}}, \tag{19}$$

which is flexible enough to accommodate different choices of default distributions. In particular, $a_D = 0$ gives $p^D(\boldsymbol{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{-(q+1)}{2}}$, equivalently $p^D(\boldsymbol{B}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{-(q+1)}{2}}$ for $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$, because the Jacobian of $(\boldsymbol{B}, \boldsymbol{\Omega}) \mapsto (\boldsymbol{B}, \boldsymbol{\Sigma})$ is proportional to $|\boldsymbol{\Sigma}|^{(q+1)}$. This is the "independence" Jeffreys prior, that is, the prior obtained by multiplying the Jeffreys priors for the two parameters assuming the other one is known; see Press (1982, sect. 3.6.2 and (14.2.7)). Alternatively, $a_D = q - 1$ gives $p^D(\boldsymbol{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{-1}$, or $p^D(\boldsymbol{B}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-q}$. Both these priors are discussed in Geisser & Cornfield (1963),

whereas Geisser (1965) focusses more deeply on the independence Jeffreys prior. Sun & Berger (2007) present further objective priors for the multivariate normal model, but we content ourselves with these two well-established alternatives.

The default prior (19) formally corresponds to the conjugate prior (13) with $\boldsymbol{C} = 0$, $\boldsymbol{R} = 0$, and $a = a_D - p - 1$. Setting the fraction $b$ equal to $n_0/n$, the posterior hyperparameters in (15) are given by $\hat{\boldsymbol{B}}$, $n_0 n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$, $a_D - p - 1 + n_0$, and $n_0 n^{-1}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}$. Hence, the fractional prior for the multivariate regression model (11) is a matrix normal Wishart of the form (13) with

$$\underline{\boldsymbol{B}} = \hat{\boldsymbol{B}}, \quad \boldsymbol{C} = n_0\tilde{\boldsymbol{C}}, \quad a = a_D + n_0 - p - 1, \quad \boldsymbol{R} = n_0\tilde{\boldsymbol{R}},$$

if we define $\tilde{\boldsymbol{C}} = n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ and $\tilde{\boldsymbol{R}} = n^{-1}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}$. In this way, we can write

$$p(\boldsymbol{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D+n_0-q-1}{2}} \exp\left\{-\frac{n_0}{2}\mathrm{tr}\left(\boldsymbol{\Omega}\left\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top\tilde{\boldsymbol{C}}\,(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \tilde{\boldsymbol{R}}\right\}\right)\right\}, \qquad (20)$$

which is proper under two conditions: i) $a_D + n_0 - p > q$, so that $a > q - 1$; ii) $n - p - 1 > q - 1$, so that $\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}$ is (almost surely) positive definite.

Condition ii), which simplifies to $n > p + q$, may not be met in realistic scenarios, but we will be able to relax it in the context of sparse DAG models; see section 5.1. Condition i) becomes $n_0 > p + q$, if $a_D = 0$, or $n_0 > p + 1$, if $a_D = q - 1$. Clearly, a larger $n_0$ is needed in the case $a_D = 0$ (independence Jeffreys prior) with respect to the case $a_D = q - 1$ (Geisser & Cornfield, 1963), especially if $q$ is much larger than 1. Since $n_0$ is intended to be minimal, we recommend setting $a_D = q - 1$, and $n_0 = p + 2$, so that $a = q$. Notice that, for the fraction $b = n_0/n$ to be small with $n_0 = p+2$, we need $p << n$, which is a stronger requirement than assuming $n > p+1$ as in section 3. However, as illustrated in section 6, this requirement will be typically satisfied in our intended application setting.

Posterior updating of the hyper-parameters leads to

$$\overline{\boldsymbol{B}} = \hat{\boldsymbol{B}}, \quad \boldsymbol{C} \mapsto n\tilde{\boldsymbol{C}}, \quad a \mapsto a_D + n - p - 1, \quad \boldsymbol{R} \mapsto n\tilde{\boldsymbol{R}},$$

keeping into account that the fractional prior is to be applied to the likelihood (12) raised to the $(1-b)$-th power, which corresponds to $n - n_0$ observations having the same $\hat{\boldsymbol{B}}$, $\tilde{\boldsymbol{C}}$ and $\tilde{\boldsymbol{R}}$, as the whole dataset. Consequently, using (17), one gets

$$m(\boldsymbol{Y}) = \frac{K(\boldsymbol{X}^\top\boldsymbol{X}, \hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}, a_D + n - p - 1)}{(2\pi)^{\frac{nq}{2}} K(n_0 n^{-1}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}, n_0 n^{-1}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}, a_D + n_0 - p - 1)},$$

which after some simplifications leads to

$$m(\boldsymbol{Y}) = \pi^{-\frac{(n-n_0)q}{2}} \frac{\Gamma_q(\frac{a_D+n-p-1}{2})}{\Gamma_q(\frac{a_D+n_0-p-1}{2})} \left(\frac{n_0}{n}\right)^{\frac{q(a_D+n_0)}{2}} |\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}|^{-\frac{n-n_0}{2}}. \tag{21}$$

In order to apply the method presented in section 5 one also needs the fractional marginal likelihood based on the submatrix $\boldsymbol{Y}_J$ which only contains the columns of $\boldsymbol{Y}$ belonging to the subset $J$, which we write as $m(\boldsymbol{Y}_J)$. This marginal likelihood is germane to our approach, and represents a half-way house towards computing the entire fractional marginal likelihood for a DAG model; see section 5.1. Based on the results presented in section 3.2, it is immediate to conclude that $m(\boldsymbol{Y}_J)$ can be obtained from equation (21) upon making the substitutions

$$q \mapsto |J|, \quad a_D \mapsto a_D - |\bar{J}|, \quad \hat{\boldsymbol{E}} \mapsto \hat{\boldsymbol{E}}_J = (\boldsymbol{Y}_J - \boldsymbol{X}\hat{\boldsymbol{B}}_J).$$

These substitutions lead to

$$m(\boldsymbol{Y}_J) = \pi^{-\frac{(n-n_0)|J|}{2}} \frac{\Gamma_{|J|}\left(\frac{a_D+n-p-1-|\bar{J}|}{2}\right)}{\Gamma_{|J|}\left(\frac{a_D+n_0-p-1-|\bar{J}|}{2}\right)} \left(\frac{n_0}{n}\right)^{\frac{|J|(a_D+n_0-|\bar{J}|)}{2}} |\hat{\boldsymbol{E}}_J^\top \hat{\boldsymbol{E}}_J|^{-\frac{n-n_0}{2}}, \tag{22}$$

which returns (21) upon setting $J = \{1, \ldots, q\}$.

Formula (22) derives from $\boldsymbol{\Omega}_{JJ.\bar{J}} \sim \mathcal{W}_{|J|}(a_J, \boldsymbol{R}_{JJ})$ with $a_J = a_D + n_0 - p - 1 - |\bar{J}|$, which is (almost surely) proper if $n > p + |J|$. The latter condition guarantees positive definiteness of $\boldsymbol{R}_{JJ}$, while $a_J = q - |\bar{J}| = |J|$ using our recommended choices for $a_D$ and $n_0$. Therefore, formula (22) provides us with a valid value for $m(\boldsymbol{Y}_J)$, whenever $|J| < n - p$, even if $n \leq p + q$. We will exploit this fact in section 5.1 to derive the marginal likelihood of a sparse DAG. In the next paragraph we specialize (22) to the simplest regression setup, which is of some interest in its own right.

If the sampling distribution corresponds to i.i.d. observations from a $q$-dimensional Gaussian density with expectation $\boldsymbol{\mu}$ and precision $\boldsymbol{\Omega}$, conditionally on $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, the corresponding marginal data distribution $m(\boldsymbol{Y}_J)$ can be derived from (22) upon setting $p = 0$ (no predictors) and $\hat{\boldsymbol{E}} = \boldsymbol{Y} - \boldsymbol{1}_n \bar{\boldsymbol{y}}^\top$, where $\bar{\boldsymbol{y}}$ is the $q$-dimensional vector of sample means. In this way we obtain

$$m(\boldsymbol{Y}_J) = \pi^{-\frac{(n-n_0)|J|}{2}} \frac{\Gamma_{|J|}\left(\frac{a_D+n-1-|\bar{J}|}{2}\right)}{\Gamma_{|J|}\left(\frac{a_D+n_0-1-|\bar{J}|}{2}\right)} \left(\frac{n_0}{n}\right)^{\frac{|J|(a_D+n_0-|\bar{J}|)}{2}} |\hat{\boldsymbol{E}}_J^\top \hat{\boldsymbol{E}}_J|^{-\frac{n-n_0}{2}}, \tag{23}$$

with $(\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}})_{jj'} = \sum_i (y_{ij} - \bar{y}_j)(y_{ij'} - \bar{y}_{j'})$. Expression (23) complements formula (22) in Consonni & La Rocca (2012), which holds for i.i.d. $q$-dimensional Gaussian observations with zero expectation.

# 5 Covariance selection

So far we have analyzed the Gaussian multivariate regression model (11) under the condition that $\boldsymbol{\Omega}$ is unconstrained. We now assume instead that $\boldsymbol{\Omega}$ is constrained by a DAG, aiming at graphical model (or covariance) selection after having adjusted for the presence of covariates. In section 5.1, we develop an extension of the approach by Geiger & Heckerman (2002) explicitly for the regression setup. An advantage of the method we present is that the computation of the marginal likelihood for each DAG only requires the results established, for an unconstrained $\boldsymbol{\Omega}$, in section 4.2. In section 5.2, taking advantage of the fact that any two Markov equivalent DAGs obtain the same marginal likelihood, we specify our results to the case of Gaussian decomposable graphical models, and relate them to those obtained by Carvalho & Scott (2009) in the i.i.d. case.

## 5.1 Error term with directed acyclic graph structure

Let $\mathcal{D}$ be a DAG with vertex set $\{1, \ldots, q\}$. Denote by $\mathrm{pa}_{\mathcal{D}}(j)$ the *parents* of $j$ in $\mathcal{D}$, that is, the set of all vertices in $\mathcal{D}$ from which an edge points to vertex $j$, and by $\boldsymbol{y}_{i\mathrm{pa}_{\mathcal{D}}(j)}$ the subvector of $\boldsymbol{y}_i$ indexed by $\mathrm{pa}_{\mathcal{D}}(j)$. The Gaussian multivariate regression sampling density of $\boldsymbol{y}_i \,|\, \boldsymbol{B}, \boldsymbol{\Omega}$, assumed Markov with respect to $\mathcal{D}$, can be written as

$$f_{\mathcal{D}}(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_{\mathcal{D}}) = \prod_{j=1}^{q} f_{\mathcal{D}}(y_{ij} \,|\, \boldsymbol{y}_{i\mathrm{pa}_{\mathcal{D}}(j)}; \boldsymbol{\theta}_j), \tag{24}$$

where $\boldsymbol{\theta}_j = (\boldsymbol{\alpha}_j, \boldsymbol{\gamma}_j, \lambda_j)$ is defined by

$$\mathbb{E}(y_{ij} \,|\, \boldsymbol{y}_{i\mathrm{pa}_{\mathcal{D}}(j)}; \boldsymbol{B}, \boldsymbol{\Omega}) = \boldsymbol{x}_i^\top \boldsymbol{\alpha}_j + \boldsymbol{y}_{i\mathrm{pa}_{\mathcal{D}}(j)}^\top \boldsymbol{\gamma}_j, \tag{25}$$

$$\mathbb{V}\mathrm{ar}(y_{ij} \,|\, \boldsymbol{y}_{i\mathrm{pa}_{\mathcal{D}}(j)}; \boldsymbol{B}, \boldsymbol{\Omega}) = \lambda_j^{-1}, \tag{26}$$

and $\boldsymbol{\theta}_{\mathcal{D}} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$ is the collection of all $\boldsymbol{\theta}_j$s; recall that $\boldsymbol{x}_i^\top$ is the $i$-th row of the design matrix $\boldsymbol{X}$, and notice that we drop dependence on $\mathcal{D}$ when we move from $\boldsymbol{\theta}_{\mathcal{D}}$ to its components (to lighten notation). We illustrate below the reparameterization from $(\boldsymbol{B}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega}$ s.p.d., to $\boldsymbol{\theta}_{\mathcal{D}}$, with $\lambda_j > 0$, $j = 1, \ldots, q$, after a remark on (24).

The conditional vertex density $f_{\mathcal{D}}(y_{ij} \,|\, \boldsymbol{y}_{i\mathrm{pa}_{\mathcal{D}}(j)}; \boldsymbol{\theta}_j)$ is a univariate normal density with expectation and variance given by (25) and (26), respectively. It is important to remark that such density depends on $\mathcal{D}$ only through $\mathrm{pa}_{\mathcal{D}}(j)$. In other words, if two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are such that $\mathrm{pa}_{\mathcal{D}_1}(j) = \mathrm{pa}_{\mathcal{D}_2}(j)$, then the vertex-specific parameter $\boldsymbol{\theta}_j$ varies in the same space under $\mathcal{D}_1$ and $\mathcal{D}_2$, because $\boldsymbol{\gamma}_j$ has the same dimension under the two DAGs, and $f_{\mathcal{D}_1}(y_{ij} \,|\, \boldsymbol{y}_{i\mathrm{pa}_{\mathcal{D}_1}(j)}; \boldsymbol{\theta}_j) = f_{\mathcal{D}_2}(y_{ij} \,|\, \boldsymbol{y}_{i\mathrm{pa}_{\mathcal{D}_2}(j)}; \boldsymbol{\theta}_j)$. This property, called *likelihood modularity* by Geiger & Heckerman (2002), represents a condition to be satisfied for the subsequent theory to apply.

Assume (without loss of generality) that the vertices of $\mathcal{D}$ are well-numbered; this means that, if $j'$ is a parent of $j$, then $j' < j$. If $\mathcal{D}$ is *complete*, that is, it has all pairs of vertices joined by an edge, then the parameters indexing the last ($j = q$) conditional vertex density in (24) are: $\boldsymbol{\alpha}_q = \boldsymbol{B}_q + \boldsymbol{B}_{\bar{q}}\boldsymbol{\Omega}_{\bar{q}q}\Omega_{qq}^{-1}$, $\boldsymbol{\gamma}_q = -\boldsymbol{\Omega}_{\bar{q}q}\Omega_{qq}^{-1}$, and $\lambda_q = \Omega_{qq}$, where $\bar{q} = \{1, \ldots, q-1\} = \mathrm{pa}_{\mathcal{D}}(q)$; see the end of section 2.1. Then, since $\boldsymbol{y}_{i\bar{q}} \,|\, \boldsymbol{B}, \boldsymbol{\Omega} \sim \mathcal{N}_{q-1}(\boldsymbol{B}_{\bar{q}}^\top \boldsymbol{x}_i, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}^{-1})$, one can repeat the previous argument and recursively find $\boldsymbol{\theta}_{q-1}, \ldots, \boldsymbol{\theta}_1$. If $\mathcal{D}$ is *incomplete*, its missing edges will impose on $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q$ the constraints $\gamma_{jj'} = 0$, $j' \notin \mathrm{pa}_{\mathcal{D}}(j)$, $j = 1, \ldots, q$, so that a corresponding set of constraints will be imposed on $\boldsymbol{\Omega}$.

We now show that, for complete DAGs, the transformation $(\boldsymbol{B}, \boldsymbol{\Omega}) \mapsto \boldsymbol{\theta}_{\mathcal{D}}$ is a smooth bijection. This fact, which is arguably not new, is reported here because it will be used below for constructing priors under general DAGs. Given the recursive definition of $(\boldsymbol{B}, \boldsymbol{\Omega}) \mapsto \boldsymbol{\theta}_{\mathcal{D}}$, it is enough to show that the transformation from $(\boldsymbol{B}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega}$ s.p.d., to $(\boldsymbol{B}_{\bar{q}}, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}; \boldsymbol{\alpha}_q, \boldsymbol{\gamma}_q, \lambda_q)$, with $\boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}$ s.p.d. and $\lambda_q > 0$, is a smooth bijection. We do this by composing a few simpler reparameterizations. First, we go from $(\boldsymbol{B}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega}$ s.p.d., to $(\boldsymbol{B}, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}, \boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq})$, with $\boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}$ s.p.d. and $\Omega_{qq} > 0$, where the smooth inverse map is provided by $\boldsymbol{\Omega}_{\bar{q}\bar{q}} = \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q} + \boldsymbol{\Omega}_{\bar{q}q}\Omega_{qq}^{-1}\boldsymbol{\Omega}_{\bar{q}q}^\top$, recalling that $\boldsymbol{\Omega}_{q\bar{q}} = \boldsymbol{\Omega}_{\bar{q}q}^\top$ (unconstrained); see for instance Lauritzen (1996, Lemma B.1). Then, we trivially split $\boldsymbol{B}$ as $(\boldsymbol{B}_q, \boldsymbol{B}_{\bar{q}})$, and replace $\boldsymbol{B}_q$ with $\boldsymbol{\alpha}_q$, where the smooth inverse map is given by $\boldsymbol{B}_q = \boldsymbol{\alpha}_q - \boldsymbol{B}_{\bar{q}}\boldsymbol{\Omega}_{\bar{q}q}\Omega_{qq}^{-1}$. Finally, we reparameterize from $\boldsymbol{\Omega}_{\bar{q}q}$ to $\boldsymbol{\gamma}_q$, with smooth inverse map given by $\boldsymbol{\Omega}_{\bar{q}q} = -\Omega_{qq}\boldsymbol{\gamma}_q$, and we rename $\Omega_{qq}$ as $\lambda_q$ (constrained to be positive).

In light of the above discussion, all complete DAGs define the same statistical model, in which $\boldsymbol{\Omega}$ is unconstrained, and there is a smooth bijection between their collections of parameters; in the terminology of Geiger & Heckerman (2002) we have *complete model equivalence*, and *regularity*. It follows that any prior on $(\boldsymbol{B}, \boldsymbol{\Omega})$ will induce a prior on $\boldsymbol{\theta}_{\mathcal{D}}$, if $\mathcal{D}$ is complete. We now show that, if we let $(\boldsymbol{B}, \boldsymbol{\Omega})$ follow the conjugate prior (13), then $p_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) = \prod_{j=1}^q p_{\mathcal{D}}(\boldsymbol{\theta}_j)$, so that $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q$ will be *a priori*

independent. This property is called *global parameter independence*, and represents a crucial ingredient in the approach of Geiger & Heckerman (2002); it can be obtained by recursive application of the following result.

**Proposition 1.** *If* $\boldsymbol{B} \,|\, \boldsymbol{\Omega} \sim \mathcal{N}_{(p+1)\times q}(\underline{\boldsymbol{B}}, \boldsymbol{C}^{-1}, \boldsymbol{\Omega}^{-1})$ *and* $\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{R})$*, then the pair* $(\boldsymbol{B}_{\bar{q}}, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q})$ *is independent of the triple* $(\boldsymbol{B}_q + \boldsymbol{B}_{\bar{q}}\boldsymbol{\Omega}_{\bar{q}q}\Omega_{qq}^{-1}, \boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq})$.

*Proof.* Consider the reparameterization in terms of $\boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}$ s.p.d., $\boldsymbol{\Omega}_{\bar{q}q}$, $\Omega_{qq} > 0$, $\boldsymbol{B}_{\bar{q}}$, $\boldsymbol{\alpha}_q = \boldsymbol{B}_q + \boldsymbol{B}_{\bar{q}}\boldsymbol{\Omega}_{\bar{q}q}\Omega_{qq}^{-1}$, and factorize the corresponding joint parameter density as

$$p(\boldsymbol{\alpha}_q \,|\, \boldsymbol{B}_{\bar{q}}, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}, \boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq}) \times p(\boldsymbol{B}_{\bar{q}} \,|\, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}, \boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq}) \times p(\boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}, \boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq}).$$

We know, from our statement following (10), that $\boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}$ is independent of $(\boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq})$ under the assumed distribution for $\boldsymbol{\Omega}$. Moreover, from the law of $\boldsymbol{B} \,|\, \boldsymbol{\Omega}$, we obtain

$$\begin{aligned} \boldsymbol{B}_{\bar{q}} \,|\, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}, \boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq} &\sim \mathcal{N}_{(p+1),(q-1)}(\underline{\boldsymbol{B}}_{\bar{q}}, \boldsymbol{C}^{-1}, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}^{-1}), \\ \boldsymbol{\alpha}_q \,|\, \boldsymbol{B}_{\bar{q}}, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}, \boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq} &\sim \mathcal{N}_{p+1}(\underline{\boldsymbol{B}}_q - \underline{\boldsymbol{B}}_{\bar{q}}\boldsymbol{\Omega}_{\bar{q}q}\Omega_{qq}^{-1}, \Omega_{qq}^{-1}\boldsymbol{C}^{-1}), \end{aligned}$$

first using column marginalization (4), and (9), then using column conditioning (5). Therefore, the joint density of $\boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}$, $\boldsymbol{\Omega}_{\bar{q}q}$, $\Omega_{qq}$, $\boldsymbol{B}_{\bar{q}}$, and $\boldsymbol{\alpha}_q$, factorizes as

$$p(\boldsymbol{\alpha}_q \,|\, \boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq}) \times p(\boldsymbol{B}_{\bar{q}} \,|\, \boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}) \times p(\boldsymbol{\Omega}_{\bar{q}\bar{q}\cdot q}) \times p(\boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq}),$$

which implies the desired result. □

If $\mathcal{D}$ is incomplete, global parameter independence can be guaranteed by letting $p_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) = \prod_{j=1}^{q} p_{\mathcal{C}_j}(\boldsymbol{\theta}_j)$, where $\mathcal{C}_j$ is any complete DAG such that $\mathrm{pa}_{\mathcal{C}_j}(j) = \mathrm{pa}_{\mathcal{D}}(j)$. The actual choice of each $\mathcal{C}_j$ is immaterial, because all $j' \notin \mathrm{pa}_{\mathcal{D}}(j)$, $j' \neq j$, are such that $j \in \mathrm{pa}_{\mathcal{C}_j}(j')$, and therefore they follow $j$ in the well-ordering of $\mathcal{C}_j$, so that $p_{\mathcal{C}_j}(\boldsymbol{\theta}_j)$ is induced by the law of $(\boldsymbol{B}_F, \boldsymbol{\Omega}_{FF\cdot\bar{F}})$, where $F = \mathrm{fa}_{\mathcal{D}}(j) = \mathrm{pa}_{\mathcal{D}}(j) \cup \{j\}$ is the *family* of $j$ in $\mathcal{D}$. Notice that $j$ is the last element of $\mathrm{fa}_{\mathcal{D}}(j)$ in the well-ordering of $\mathcal{C}_j$, and recall that $\boldsymbol{B}_F \,|\, \boldsymbol{\Omega}_{FF\cdot\bar{F}} \sim \mathcal{N}_{(p+1)\times|F|}(\underline{\boldsymbol{B}}_F, \boldsymbol{C}^{-1}, \boldsymbol{\Omega}_{FF\cdot\bar{F}}^{-1})$, by column marginalization, while $\boldsymbol{\Omega}_{FF\cdot\bar{F}} \sim \mathcal{W}_{|F|}(a - |F|, \boldsymbol{R}_{FF})$, as per (10). Assigning parameter priors in this way, we also guarantee *prior modularity*: $p_{\mathcal{D}_1}(\boldsymbol{\theta}_j) = p_{\mathcal{D}_2}(\boldsymbol{\theta}_j)$, if $\mathrm{pa}_{\mathcal{D}_1}(j) = \mathrm{pa}_{\mathcal{D}_2}(j)$. This is the last ingredient required by the method of Geiger & Heckerman (2002) to compute the marginal likelihood of *any* DAG model, based on the assignment of the *single* prior (13). We now detail the computations for our regression setting.

The marginal density of the matrix $\boldsymbol{Y}$ under the DAG $\mathcal{D}$, equivalently the marginal likelihood of $\mathcal{D}$ observing $\boldsymbol{Y}$, can be found as $m_{\mathcal{D}}(\boldsymbol{Y}) = \int f_{\mathcal{D}}(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_{\mathcal{D}}) p_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) d\boldsymbol{\theta}_D$,

where $f_{\mathcal{D}}(\boldsymbol{Y} \mid \boldsymbol{\theta}_{\mathcal{D}}) = \prod_{i=1}^{n} f_{\mathcal{D}}(\boldsymbol{y}_i \mid \boldsymbol{\theta}_{\mathcal{D}})$ with $f_{\mathcal{D}}(\boldsymbol{y}_i \mid \boldsymbol{\theta}_{\mathcal{D}})$ given by (24), and furthermore $p_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) = \prod_{j=1}^{q} p_{\mathcal{D}}(\boldsymbol{\theta}_j)$ by global parameter independence. We can thus write

$$\begin{aligned} m_{\mathcal{D}}(\boldsymbol{Y}) &= \prod_{j=1}^{q} \int p_{\mathcal{D}}(\boldsymbol{\theta}_j) \prod_{i=1}^{n} f_{\mathcal{D}}(y_{ij} \mid \boldsymbol{y}_{i\mathrm{pa}_{\mathcal{D}}(j)}; \boldsymbol{\theta}_j) d\boldsymbol{\theta}_j \\ &= \prod_{j=1}^{q} \int p_{\mathcal{C}_j}(\boldsymbol{\theta}_j) \prod_{i=1}^{n} f_{\mathcal{C}_j}(y_{ij} \mid \boldsymbol{y}_{i\mathrm{pa}_{\mathcal{C}_j}(j)}; \boldsymbol{\theta}_j) d\boldsymbol{\theta}_j \\ &= \prod_{j=1}^{q} \int p_{\mathcal{C}_j}(\boldsymbol{\theta}_j) f_{\mathcal{C}_j}(\boldsymbol{Y}_j \mid \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{C}_j}(j)}; \boldsymbol{\theta}_j) d\boldsymbol{\theta}_j, \end{aligned}$$

where the second equality is based on prior and likelihood modularity. It follows that

$$m_{\mathcal{D}}(\boldsymbol{Y}) = \prod_{j=1}^{q} m_{\mathcal{C}_j}(\boldsymbol{Y}_j \mid \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{C}_j}(j)}) = \prod_{j=1}^{q} \frac{m_{\mathcal{C}_j}(\boldsymbol{Y}_{\mathrm{fa}_{\mathcal{C}_j}(j)})}{m_{\mathcal{C}_j}(\boldsymbol{Y}_{\mathrm{pa}_{\mathcal{C}_j}(j)})} = \prod_{j=1}^{q} \frac{m(\boldsymbol{Y}_{\mathrm{fa}_{\mathcal{D}}(j)})}{m(\boldsymbol{Y}_{\mathrm{pa}_{\mathcal{D}}(j)})}, \tag{27}$$

recalling that $\mathrm{pa}_{\mathcal{C}_j}(j) \equiv \mathrm{pa}_{\mathcal{D}}(j)$, by construction, and $m_{\mathcal{C}_j}(\cdot)$ is nothing else but $m(\cdot)$ under our prior (13), by complete model equivalence and regularity.

The great advantage of (27) is that the computations of the required terms in the rightmost product can be done under the assumption that the precision matrix $\boldsymbol{\Omega}$ is unconstrained, and thus one can use the standard matrix normal Wishart prior (13). Notice that the DAG $\mathcal{D}$ enters (27) only through the specification of the set of parents, $\mathrm{pa}_{\mathcal{D}}(j)$, for each vertex $j$. The expressions for $m(\boldsymbol{Y}_{\mathrm{fa}_{\mathcal{D}}(j)})$ and $m(\boldsymbol{Y}_{\mathrm{pa}_{\mathcal{D}}(j)})$ are available in section 3.2, upon replacing $J$ with $\mathrm{fa}_{\mathcal{D}}(j)$ and $\mathrm{pa}_{\mathcal{D}}(j)$, respectively.

Prior (13) requires to specify the hyper-parameters $\underline{\boldsymbol{B}}$, $\boldsymbol{C}$, $a$, and $\boldsymbol{R}$. This can be problematic, especially when the dimension of the problem is large, and we know that marginal likelihoods are quite sensitive to changes in the hyper-parameters; see O'Hagan & Forster (2004, Ch. 7). We therefore suggest an objective choice, based on the fractional matrix normal Wishart prior (20) applied to the Gaussian likelihood (12) with $(n - n_0)$ observations and the same $\hat{\boldsymbol{B}}$, $\tilde{\boldsymbol{C}}$ and $\tilde{\boldsymbol{R}}$ as the whole data. With this choice, the terms $m(\boldsymbol{Y}_{\mathrm{fa}_{\mathcal{D}}(j)})$ and $m(\boldsymbol{Y}_{\mathrm{pa}_{\mathcal{D}}(j)})$ in formula (27) can be computed from (22) provided that the condition $|\mathrm{fa}_{\mathcal{D}}(j)| = |\mathrm{pa}_{\mathcal{D}}(j)| + 1 < n - p$ is satisfied. This condition guarantees a valid value for $m(\boldsymbol{Y}_j | \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{D}}(j)}) = m(\boldsymbol{Y}_{\mathrm{fa}_{\mathcal{D}}(j)})/m(\boldsymbol{Y}_{\mathrm{pa}_{\mathcal{D}}(j)})$ by granting a proper distribution to the marginal precision matrix $\boldsymbol{\Omega}_{\mathrm{fa}_{\mathcal{D}}(j)\mathrm{fa}_{\mathcal{D}}(j)\cdot\overline{\mathrm{fa}_{\mathcal{D}}(j)}}$; see section 4.2. In this way, formula (27) provides a marginal likelihood for every DAG $\mathcal{D}$ whose parent sets have size smaller than the number of observations minus the number of columns in the design matrix $\boldsymbol{X}$ (number of predictors in the model plus

one). The latter is a sparsity condition on the structure of the DAG, involving the maximal number of parents across vertices, which is quite reasonable in our intended application setting (eQTL analysis) as discussed in the Introduction.

## 5.2 Error term with decomposable graph structure

It is often appropriate to model the conditional independence structure of a set of variables in terms of an undirected graph; see Lauritzen (1996) for an authoritative exposition. This is for instance the approach followed in Cai *et al.* (2013) and Chen *et al.* (2016) for the analysis of genetical genomics data. With reference to the Gaussian multivariate regression model (11), this means that the precision matrix $\boldsymbol{\Omega}$ of the response vector $\boldsymbol{y}_i$ is constrained by an undirected graph $\mathcal{G}$: if an edge is missing between $j$ and $j'$ in $\mathcal{G}$, then $\boldsymbol{\Omega}_{jj'} = 0$. Equivalently, $\boldsymbol{y}_i$ is Markov with respect to $\mathcal{G}$, that is, if $j$ and $j'$ are not joined by an edge in $\mathcal{G}$, the responses $y_{ij}$ and $y_{ij'}$ are conditionally independent, under the sampling distribution, given all remaining responses; in symbols $y_{ij} \perp\!\!\!\perp y_{ij'} \mid \boldsymbol{y}_{i(\{1,\ldots,q\}\setminus\{j,j'\})}, \boldsymbol{B}, \boldsymbol{\Omega}$ (Drton & Perlman, 2004).

To enhance tractability, the undirected graph $\mathcal{G}$ is often assumed to satisfy some conditions, such as *decomposability*; see for instance Bhadra & Mallick (2013). Recall that $\mathcal{G}$ is decomposable when all cycles in $\mathcal{G}$ admit a *chord*, that is, an edge joining two non-consecutive vertices of the cycle (Cowell *et al.*, 1999, sect. 4.2). It is well known that a decomposable $\mathcal{G}$ is Markov equivalent to some DAG (Andersson *et al.*, 1997). Specifically, one can always well-number the vertices of $\mathcal{G}$ and construct a directed version $\mathcal{G}^{<}$, which is a DAG Markov equivalent to $\mathcal{G}$; see Lauritzen (1996, p. 18). It follows that the methodology developed in section 5.1 can also be applied to decomposable graphs, because the marginal likelihoods given by such methodology are invariant with respect to Markov equivalence. Indeed, the proof of Theorem 4 in Geiger & Heckerman (2002) directly carries over into our regression setting.

In practice, the marginal likelihood of the model defined by the decomposable graph $\mathcal{G}$, $m_{\mathcal{G}}(\boldsymbol{Y}) = m_{\mathcal{G}^{<}}(\boldsymbol{Y})$, will be given by (27) with $\mathcal{D} = \mathcal{G}^{<}$. Since the parameter prior used to compute (27) satisfies global parameter independence, $m_{\mathcal{G}^{<}}(\boldsymbol{Y})$ is readily seen to be $\mathcal{G}^{<}$-Markov; see for instance Cowell *et al.* (1999, sect. 9.4). Then $m_{\mathcal{G}}(\boldsymbol{Y})$ is also $\mathcal{G}$-Markov, and thus it admits the representation

$$m_{\mathcal{G}}(\boldsymbol{Y}) = \frac{\prod_{C \in \mathcal{C}} m(\boldsymbol{Y}_C)}{\prod_{S \in \mathcal{S}} m(\boldsymbol{Y}_S)}, \tag{28}$$

where $\mathcal{C}$ is the set of cliques (inclusion maximal complete subgraphs) and $\mathcal{S}$ the set

of separators in a perfect ordering of $\mathcal{C}$; see Lauritzen (1996, sect. 2.1.3). The explicit expression of each factor in (28) can be deduced from (17) as explained in section 3.2.

In particular, when using the fractional matrix normal Wishart prior (20), one computes $m(\boldsymbol{Y}_C)$ and $m(\boldsymbol{Y}_S)$ in (28) by means of (22), with $J = C$ and $J = S$, respectively, assuming $|C| < n-p$ (hence $|S| < n-p$) whenever $C$ is a clique ($S \subseteq C$ a separator) of $\mathcal{G}$. In this way, we cope with decomposable graphs whose clique sizes are smaller than the number of observations minus the number of predictors in the model. This is again a sparsity assumption on the graph, well-suited to our intended application setting, which grants a proper distribution to $\boldsymbol{\Omega}_{CC\cdot\bar{C}}$ (hence to $\boldsymbol{\Omega}_{SS\cdot\bar{S}}$); see section 4.2. We remark that formulae (28) and (22) generalize to the multivariate regression setup the results established by Carvalho & Scott (2009) for i.i.d. Gaussian observations with zero expectation. As a special case, formulae (28) and (22) also cover the i.i.d. Gaussian setup with unknown expectation.

# 6 Experimental studies

In the present section the proposed methodology is applied to a problem of joint variable and graphical model selection. Different simulated scenarios are discussed, and the results are compared with state-of-the-art competing approaches. To this aim, the theoretical results developed in the previous sections are operationalized in a Markov chain Monte Carlo (MCMC) algorithm that follows the structure of the sampling algorithm proposed in Bhadra & Mallick (2013). All codes were written for parallel computing in `R` (R Core Team, 2016) and are available upon request.

Given $p^\star$ available predictors (or variables) we assume a regression model which includes only a subset of the variables. Specifically, let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p^\star})$ denote the vector of binary indicators which identifies the predictors present in the model: $\gamma_i = 1$ if the $i$-th predictor is present, $\gamma_i = 0$ otherwise. Let $\sum_{i=1}^{p^\star} \gamma_i = p_\gamma$ denote the dimension of the regression model. For simplicity we identify the regression model with $\boldsymbol{\gamma}$. Additionally, given a graph $\mathcal{G}$, we assume that the precision matrix of the $q$ response variables $\boldsymbol{\Omega}_{\mathcal{G}}$ is Markov with respect to $\mathcal{G}$.

This leads to the following graphical Gaussian multivariate regression model

$$\boldsymbol{Y} \mid \boldsymbol{B}_{\boldsymbol{\gamma}}, \boldsymbol{\Omega}_{\mathcal{G}}, \mathcal{G}, \boldsymbol{\gamma} \sim \mathcal{N}_{n,q}\left(\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{B}_{\boldsymbol{\gamma}}, \boldsymbol{I}_n, \boldsymbol{\Omega}_{\mathcal{G}}^{-1}\right), \tag{29}$$

where $\boldsymbol{X}_\gamma$ is an $n \times (p_\gamma + 1)$ matrix of selected predictors, $\boldsymbol{B}_{\boldsymbol{\gamma}}$ is a $(p_\gamma + 1) \times q$ matrix of regression coefficients, and $\boldsymbol{\Omega}_{\mathcal{G}}$ is a $q \times q$ precision matrix constrained by $\mathcal{G}$. Clearly,

neither $\boldsymbol{\gamma}$ not $\mathcal{G}$ are known, and the goal is to perform the *joint* task of variable and graph selection. This is reflected in the notation used in (29) which indexes parameters by either $\boldsymbol{\gamma}$ or $\mathcal{G}$. Notice that this differs from the notation employed in the more general section 5 where a fixed regression structure was tacitly understood.

As discussed at the beginning of section 3, in a typical scenario for genetical genomics applications the total number of predictors $p^\star$ is comparable to, or larger than, the number of observations, but interest lies in sparse models. For instance, the two simulations considered by Bhadra & Mallick (2013) have: i) $p^\star = 498$, $q = 300$, and $n = 120$, with $p_\gamma = 11$ for the actual data generating distribution; ii) $p^\star = 498$, $q = 100$, and $n = 120$, with $p_\gamma = 3$ for the actual data generating distribution. Similarly, their real data analysis (eQTL Analysis on Publicly Available Human Data) has $p^\star = 3125$, $q = 100$, and $n = 60$, with $p_\gamma = 1$ or $p_\gamma = 2$ identified as the most likely values. Accordingly we restrict our simulation studies to scenarios wherein $p_\gamma \ll n$.

Now consider a Gaussian multivariate regression model $\boldsymbol{\gamma}$ having linear predictor $\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{B}_{\boldsymbol{\gamma}}$, and an unconstrained precision matrix $\boldsymbol{\Omega}$. In order to compare our results with those obtained using alternative methods, we assume that $\mathcal{G}$ is an undirected decomposable graph, and let $\mathcal{G}_j$ denote the indicator of the $j$-th off-diagonal element of the lower triangular part of the adjacency matrix of $\mathcal{G}$. We assign to $(\boldsymbol{B}_\gamma, \boldsymbol{\Omega})$ the fractional prior shown in (20), while the remaining prior specifications are standard and follow Bhadra & Mallick (2013):

$$\begin{aligned}
\gamma_i &\sim \text{Ber}(\pi_\gamma), \quad i = 1, \ldots, p^\star, \\
\mathcal{G}_j &\sim \text{Ber}(\pi_\mathcal{G}), \quad j = 1, \ldots, q(q-1)/2, \\
\pi_\gamma, \pi_G &\sim \text{Unif}(0, 1),
\end{aligned}$$

all independently.

Our MCMC procedure is a collapsed Metropolis-Hastings algorithm, because the marginal data distribution, after marginalization of $\boldsymbol{B}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\Omega}$, but conditionally on $\boldsymbol{\gamma}$ and $\mathcal{G}$, is available in closed form from formula (22) and factorization (28), thus permitting draws from the full conditionals of $\boldsymbol{\gamma}$ and $\mathcal{G}$. In this way the parameter space investigated by the sampler is significantly reduced, with substantial computational gains. The sampler iteratively extracts instances of $\boldsymbol{\gamma}$ and $\mathcal{G}$ from their conditional posteriors; we omit the details of the algorithm and we refer the interested reader to sections 2.3 and 2.4 of Bhadra & Mallick (2013). It should be noted however that we depart from the latter authors in a few significant directions. First of all, they employ weakly informative parameter priors, which are not tailor cut for model selection in

the same way as our fractional priors are. Next, they resort to hyper-inverse Wishart priors on the constrained covariance matrices $\boldsymbol{\Sigma}_{\mathcal{G}} = \boldsymbol{\Omega}_{\mathcal{G}}^{-1}$, whereas we only need a single distribution on the unconstrained precision matrix $\boldsymbol{\Omega}$ which turns out to be, in the implied fractional prior, a standard Wishart distribution with data dependent hyperparameters; see (20). As a consequence, the data distribution conditional on $\boldsymbol{\gamma}$ and $\mathcal{G}$ will be different in the two approaches. Finally, the moves in the space of decomposable graphical models are implemented in our approach following a most recent theoretical contribution, as we detail shortly below.

In each step of the MCMC algorithm, acceptance of proposed moves are subject to the verification of the conditions outlined in the previous sections for the validity of formula (22) and factorization (28). We also verify that local perturbations of the graph at the current iteration result in a new graph which is still decomposable: this can be done by accepting only those moves which satisfy two conditions outlined in Green & Thomas (2013) on the junction tree representation of the proposed graph.

Following the simulation settings in Bhadra & Mallick (2013) and Chen *et al.* (2016), we explore the performance of our method and other competing procedures in two scenarios: *sparse block* and *magnified block* settings (both described below).

In the *sparse block* setting, we let $\mathcal{G}_j$=0 for $j \leq q(q-1)/2 - 10$ and $\mathcal{G}_j$=1 for $j > q(q-1)/2 - 10$, so that the $q \times q$ adjacency matrix of $\mathcal{G}$ has a sparse right-below block of active edges, where the sparsity of $\mathcal{G}$ increases with $q$. Given $\mathcal{G}$, we extract $\boldsymbol{\Omega}_{\mathcal{G}}$ from the $\mathcal{G}$-Wishart distribution (Roverato, 2002; Letac & Massam, 2007) with degrees of freedom and scale matrix parameters respectively equal to 10 and the identity matrix (as in Bhadra & Mallick 2013). The actual number of covariates is also very sparse: out of $p^\star = 100$ potential covariates, the true model $\boldsymbol{\gamma}$ assumes only two predictors (plus the intercept), namely the first and the third. Then, conditionally on $\boldsymbol{\Omega}_{\mathcal{G}}$, we sample $\boldsymbol{B}_\gamma$ from $\mathcal{N}_{3,q}\left(\boldsymbol{0}_{3,q}, 0.3^2\boldsymbol{I}_3, \boldsymbol{\Omega}_{\mathcal{G}}^{-1}\right)$, where again the hyperparameters are set as in Bhadra & Mallick (2013). Given the true $\boldsymbol{\gamma}$ and $\mathcal{G}$, and the sampled $\boldsymbol{\Omega}_{\mathcal{G}}$ and $\boldsymbol{B}_\gamma$, we generate 60 repetitions of $\boldsymbol{X}_\gamma$ and $\boldsymbol{Y}$, for a sample size $n = 50$: beyond its first column of ones, all the elements of $\boldsymbol{X}_\gamma$ are drawn from $\mathcal{N}(10, 1)$, whilst $\boldsymbol{Y}$ is drawn from (29). The experiment is then replicated for different values of $q \in \{30, 60, 120\}$, so that in all scenarios the number of potential predictors is greater than the sample size ($p^\star > n$), and in two of the three scenarios the number of vertices in the graph is greater than the sample size ($q > n$).

In the *magnified block* setting, we first create a graph $\mathcal{G}^I$ by fixing a $50 \times 50$ adjacency matrix as in the previous scenario. Given $\mathcal{G}^I$, we extract $\boldsymbol{\Omega}_{\mathcal{G}^I}$ from the $\mathcal{G}^I$-Wishart distribution with the hyperparameters fixed as in the sparse block setting.

The whole graph $\mathcal{G}$ is the disjoint union of three copies of $\mathcal{G}^I$, so that $q = 150$, and $\boldsymbol{\Omega}_{\mathcal{G}}$ is a block diagonal matrix having on its diagonal $\boldsymbol{\Omega}_{\mathcal{G}^I}$, $5\boldsymbol{\Omega}_{\mathcal{G}^I}$ and $10\boldsymbol{\Omega}_{\mathcal{G}^I}$. This results in a precision matrix $\boldsymbol{\Omega}_{\mathcal{G}}$ with sequentially magnified signals. The active predictors are randomly chosen, each with probability 0.05, among $p^\star = 100$ potential predictors producing a true $\boldsymbol{\gamma}$. Once $\boldsymbol{B}_{\boldsymbol{\gamma}}$ is generated as in the previous setting, we simulate 60 replicates of $\boldsymbol{X}_{\boldsymbol{\gamma}}$ and $\boldsymbol{Y}$ for a sample size $n = 50$ as described above.

We compare our method, which we name OBFBF (Objective Bayes Fractional Bayes Factor) for easier reference in the sequel, with five alternative procedures: the two-step ANTAC (Asymptotically Normal with Thresholding after Adjusting Covariates) estimator of Chen *et al.* (2016), the GLASSO (graphical lasso) of Friedman *et al.* (2008), the HYPERT (hyper-matrix t) method of Bhadra & Mallick (2013), the CONDIT (Sparse Gaussian Conditional) method of Wytock & Kolter (2013), and the LOWRANK (Low Rank plus Sparse) methodology of Chandrasekaran *et al.* (2012). For GLASSO, the precision matrix is estimated without taking into account the effects of $\boldsymbol{X}_{\boldsymbol{\gamma}}$; its tuning parameter is selected with five-fold cross-validation by maximizing the log-likelihood function. For ANTAC, the penalty parameters are set to their theoretically optimal values, and so is the theoretical bound of edge inclusion. The hyperparameters in HYPERT are fixed to be the same as the ones used by Bhadra & Mallick (2013) in their simulation study. Finally, for the comparison with Wytock & Kolter (2013) and Chandrasekaran *et al.* (2012), we adopt the optimization routines of Nesterov (2005), as implemented in Frot *et al.* (2016).

For each procedure we evaluate its performance in learning the graphical structure (graph selection) in terms of misspecification rate, specificity, sensitivity, precision and Matthews correlation coefficient, defined as

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

where $TP$, $TN$, $FP$, $FN$ are the numbers of true positives, true negatives, false positives and false negatives (respectively).

The results in the different simulated settings are summarized in Table 1: save for MISR, better performances correspond to higher indicators. It appears that our method OBFBF performs equally or better than the competitors for all measures and scenarios. The exception is ANTAC, which provides better results for scenarios with low number of vertices, but the difference with our method significantly reduces in settings with high $q$, corresponding to big networks. It should be emphasized, how-

ever, that ANTAC is finely tuned to graph selection, with variable selection being of lesser concern; see further comments below. On the other hand, OBFBF takes care of both variable and graph selection. Furthermore OBFBF, which like HYPERT is Bayesian, returns a richer output, namely a posterior distribution on $\boldsymbol{\gamma}$ (indexing the space of regression models) and $\mathcal{G}$ (indexing the space of graphs) thus fully accounting for model uncertainty. The above considerations should be borne in mind also when evaluating the computational time for a given task, which is admittedly lower for ANTAC. GLASSO, CONDIT and LOWRANK also have lower computational costs than the MCMC-based Bayesian methods. However, their performances are hardly impressive; in particular, the bad behaviour of GLASSO, which disregards the regression structure, highlights the relevance of including covariates in the problem of learning a graphical structure.

[Table 1 ABOUT HERE]

Moving from lower to higher $q$, that is, increasing the number of vertices in the graph, does not necessarily increase the computational times of MCMC-based algorithms (OBFBF and HYPERT). As $q$ increases, the space of decomposable graphs spans a decreasing portion of the whole space which can be explored more efficiently provided one uses, as we do, an efficient method to screen out MCMC proposals which fall outside the admissible subspace of graphs. The comparison of the two Bayesian competitors in terms of computational time shows an advantage of HYPERT over OBFBF only for the low-dimensional settings with $q = 30$. In the remaining cases OBFB prevails, because HYPERT needs to go through a rather elaborate derivation of the marginal likelihood using the hyper-Inverse Wishart distribution on the constrained covariance matrix, leading to an hyper-matrix t density; see their formula (10). On the other hand, we only require standard Wishart distributions on the unconstrained precision matrix which results in faster computations.

We report in Figure 1 run time sensitivity to different sample sizes and different numbers of potential regressors. The results summarize the computational times of 10 000 accepted MCMC moves in stationary regime, for samples of size from 50 to 250, and for a number of potential regressors between 50 and 250, assuming as reference equal to 1 the time of the algorithm for 5 nodes, sample size 50 and 50 potential regressors. There is a clear increase in time as the two dimensions increase, but such an increase does not seem problematic, because there is a difference of roughly 7% between the scenarios of smallest and highest dimensions.

[Figure 1 ABOUT HERE]

Figure 2 reports a plot showing the ability of the methodologies under comparison in recovering the structure of the graph. For the *sparse block* simulation setting and $(n, p^\star, q) = (200, 100, 30)$, it is apparent that ANTAC and OBFBF perform similarly and better than the alternatives in recovering the true graph structure; both identify the same edges in the bottom-right corner of the adjacency matrix, but ANTAC incorrectly finds one more edge outside the true set of edges. GLASSO, HYPERT and CONDIT are not able, for the given number of MCMC/optimization iterations used with OBFBF, to identify the correct graphical structure, with GLASSO performing worst, because it neglects covariates, followed by CONDIT. LOWRANK recognizes the high sparsity of the graph, but it does not recover the bulk of connected edges.

[Figure 2 ABOUT HERE]

Table 1 and Figure 2 together show that OBFBF and ANTAC outperform their competitors; moreover OBFBF can improve on ANTAC in sparse settings with sizeable sample sizes, and importantly it is a valid alternative to ANTAC in scenarios featuring a high number of vertices together with a modest sample size, corresponding to highly relevant scientific benchmarks.

Both OBFBF and HYPERT provide as output a posterior probability of edge inclusion. In our summary results we decided to report the presence of an edge in the adjacency matrix whenever this probability is higher than the fixed threshold 0.5. This threshold is a natural default choice, and could be changed in applications where context may suggest a different value. However, it turns out that our results are quite robust with respect to this choice: for instance, with reference to Figure 2, the matrix of estimated edge posterior probabilities is virtually indistinguishable from the reported 0/1 adjacency matrix; among all edges identified as missing, the maximum posterior probability is estimated as 0.0094, whilst among those identified as present, the minimum estimated posterior probability is 0.9152.

In order to highlight that our simulation results are not dependent on the chosen threshold, we compare in Figure 3(b) the two Bayesian methods on the basis of their receiver operating characteristic (ROC) curve, which plots $1-\text{SPE}$ *versus* SEN as the threshold for the edge inclusion probability is varied. Figure 3(b) confirms that, using the same number of iterations, OBFBF is able to estimate the graphical structure better than HYPERT.

[Figure 3 ABOUT HERE]

The performance of OBFBF on the variable selection part of the problem is compared only with HYPERT: the GLASSO procedure does not address the issue of variable selection, and in ANTAC the selection of variables is performed in the first step only, as an intermediate result deliberately meant to be good enough to contribute to the primary goal of precision matrix estimation, which is pursued in the second step; finally in CONDIT and LOWRANK regressors enter into the analysis through the conditioning of the graphical structure, but these methods do not report estimates of significant regression coefficients.

Our method and HYPERT provide a posterior probability of inclusion for each variable, so that a varying inclusion threshold creates a ROC curve for each method; this is shown in Figure 3(a). With reference to the *sparse block* simulation setting with $(n, p^\star, q) = (200, 100, 30)$, the curves show that OBFBF performs extremely well in selecting the correct variables. This is also clear from Figure 3(c) and Figure 3(d): in Figure 3(c) we represent graphically the situation in which, out of $p^\star = 100$ potential regressors, only two predictors (plus intercept) are actually used to generate the data, and in Figure 3(d) the posterior probability of inclusion for each potential regressor is reported under the OBFBF procedure; they are correctly equal to 1 (0) for each predictor actually included in (excluded from) the model. The corresponding results for HYPERT (not reported) exhibit lower accuracy.

# 7 Discussion

Motivated by covariate-adjusted graphical model selection under sparsity, this paper proposes an objective Bayes method for computing the marginal likelihood of a graphical Gaussian multivariate regression model whose covariance matrix is constrained by a DAG. This calculation represents an essential ingredient to obtain a posterior probability over the space of DAG models, after having adjusted for the effect of relevant predictors. Since the proposed method is invariant with respect to Markov equivalence of DAGs, it can also be used to select covariate-adjusted decomposable undirected graphical models. Furthermore, by adding an extra standard layer to our modeling setup, it can successfully address joint variable and graphical selection.

The simulation studies we report in section 6 show that our method is quite effective even when the sample size is small to moderate, and both the regression and covariance structure exhibit sparsity. More specifically, with regard to graph selection, our procedure is highly competitive with the best method for covariance selection available in our comparative exercise.

A natural alternative to our method is represented by Bhadra & Mallick (2013) who derive their results for regression models whose error term is Markov with respect to a decomposable graph. They employ weakly informative priors that need be specified by the user; in particular they assign a hyper-inverse Wishart distribution on the graph-constrained covariance matrix depending on a global shrinkage parameter which can highly influence the results. In the experiments we carried out our results are significantly better both with regard to variable and graphical selection.

Although we implemented our methodology in an objective Bayes setup, our approach can be seamlessly applied also with a subjectively specified matrix normal Wishart prior under any complete DAG model, and then applying the general results of section 3.2 in the context of DAG models as described in section 5.1, or with regard to decomposable models as illustrated in 5.2. In either case, the sparsity conditions relating the sample size $n$, the number of predictors $p$ and the maximal size of the cliques, which we had to impose to make our objective Bayes analysis feasible, could be relaxed.

Our method does not deal with covariate adjusted selection of general undirected graphical models, where the main challenge is finding an efficient method for computing the marginal likelihood; see Carvalho *et al.* (2007), Wang & Carvalho (2010), Lenkoski (2013). However, working within the class of decomposable graphs can still be very effective, even when the true graph is not decomposable; see Fitch *et al.* (2014) for asymptotic results on the posterior model probabilities, and for a high performing stochastic search of the model space.

Finally, it would be useful to consider an extension of our methodology to regression models having latent variables; a recent contribution in this direction has been given by Frot *et al.* (2016) from a penalized optimization point of view.

# Acknowledgements

Work partially supported by a D1-grant from Università Cattolica del Sacro Cuore. The authors are grateful to Alberto Roverato for pointing out a useful reference, and to the Associate Editor and an anonymous Reviewer for helpful comments.

# Supporting information

Additional information for this article is available online including a brief summary of our graph terminology.

# References


Andersson, S. A., Madigan, D. & Perlman, M. D. (1997). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scand. J. Statist.* **24**, 81–102.

Bhadra, A. & Mallick, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics* **69**, 447–457.

Brem, R. B. & Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 1572–1577.

Cai, T. T., Li, H., Liu, W. & Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100**, 139–156.

Carvalho, C. & Scott, J. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96**, 497–512.

Carvalho, C. M., Massam, H. & West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94**, 647–659.

Chandrasekaran, V., Parrilo, P. & Willsky, A. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40**, 1935–1967.

Chen, M., Ren, Z., Zhao, H. & Zhou, H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *J. Amer. Statist. Assoc.* **111**, 394–406.

Consonni, G. & La Rocca, L. (2012). Objective Bayes factors for Gaussian directed acyclic graphical models. *Scand. J. Statist.* **39**, 743–756.

Cowell, R. G., Dawid, P. A., Lauritzen, S. L. & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems.* Springer, New York.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68**, 265–274.

Dawid, A. P. (2003). Causal inference using influence diagrams: the problem of partial compliance. In P. Green, N. L. Hjort & S. Richardson, eds., *Highly structured stochastic systems.* Oxford University Press, Oxford, pp. 45–81.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–1317.

DeGroot, M. H. (1970). *Optimal statistical decisions.* McGraw-Hill, New York.

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

Fitch, A. M., Jones, M. B. & Massam, H. (2014). The performance of covariance selection methods that consider decomposable models only. *Bayesian Anal.* **9**, 659–684.

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

Frot, B., Jostins, L. & McVean, G. (2016). Latent variable model selection for Gaussian conditional random fields. *ArXiv e-print* **1512.06412v2**.

Geiger, D. & Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30**, 1412–1440.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.* **36**, 150–159.

Geisser, S. & Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. Ser. B* **25**, 368–376.

Green, P. & Thomas, A. (2013). Sampling decomposable graphs using a markov chain on junction trees. *Biometrika* **100**, 91–110.

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

Heckerman, D., Geiger, D. & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20**, 197–243.

Kuipers, J., Moffa, G. & Heckerman, D. (2014). Addendum on the scoring of Gaussian directed acyclic graphical models. *Ann. Statist.* **42**, 1689–1691.

Lauritzen, S. L. (1996). *Graphical models.* Oxford University Press, Oxford.

Lauritzen, S. L. (2001). Causal inference from graphical models. In *Complex stochastic systems (Eindhoven, 1999)*, vol. 87 of *Monogr. Statist. Appl. Probab.* Chapman & Hall/CRC, Boca Raton, FL, pp. 63–107.

Lenkoski, A. (2013). A direct sampler for G-Wishart variates. *Stat* **2**, 119–128.

Letac, G. & Massam, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35**, 1278–1323.

Madigan, D., Andersson, S. A., Perlman, M. D. & Volinsky, C. T. (1996). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm. Statist. Theory Methods* **25**, 2493–2519.

Moreno, E. (1997). Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness. In Y. Dodge, ed., *$L_1$-statistical procedures and related topics.* Institute of Mathematical Statistics, pp. 257–270.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming* **103**, 127–152.

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

O'Hagan, A. & Forster, J. (2004). *Kendall's advanced theory of statistics. Vol. 2B. Bayesian inference.* John Wiley & Sons, Chichester.

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

Press, S. J. (1982). *Applied multivariate analysis: Using Bayesian and frequentist methods of inference.* Krieger Publishing Company, Malabar, FL.

R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rothman, A. J., Levina, E. & Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.* **19**, 947–962.

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Stat.* **29**, 391–411.

Rowe, D. B. (2003). *Multivariate Bayesian statistics.* Chapman & Hall/CRC, Boca Raton, FL.

Sohn, K.-A. & Kim, S. (2012). Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS).* pp. 1081–1089. JMLR: W&CP Volume 22.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. & Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statist. Sci.* **8**, 219–283. With comments and a rejoinder by the authors.

Sun, D. & Berger, J. O. (2007). Objective priors for the multivariate normal model. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. Smith & M. West, eds., *Bayesian Statistics 8 – Proceedings of the Eigth Valencia International Meeting.* Oxford University Press, pp. 525–554.

Wang, H. & Carvalho, C. M. (2010). Simulation of hyper-inverse Wishart distributions for non-decomposable graphs. *Electron. J. Stat.* **4**, 1470–1475.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics.* John Wiley & Sons, New York.

Wytock, M. & Kolter, J. Z. (2013). Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *Proceedings of the 30th International Conference on Machine Learning (ICML).* pp. 1265–1273. JMLR: W&CP Volume 28.

Yin, J. & Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5**, 2630–2650.

Zhang, L. & Kim, S. (2014). Learning gene networks under SNP perturbations using eQTL datasets. *PLoS Comput. Biol.* **10**, e1003420. Corrected on March 21, 2014.

Address: *Luca La Rocca, University of Modena and Reggio Emilia, Department of Physics, Informatics and Mathematics, Edificio Matematica, Via Campi 213/b, 41125 Modena, Italy.*

Email: `luca.larocca@unimore.it`

Table 1: Graph selection results. In the different simulation settings described in section 6, for different sample size $n$, potential number of regressors $p^*$ and number of responses $q$, our method OBFBF is compared with ANTAC (Chen *et al.* 2016), GLASSO (Friedman *et al.* 2008), HYPERT (Bhadra & Mallick 2013), CONDIT (Wytock & Kolter 2013) and LOWRANK (Chandrasekaran *et al.* 2012). Performances are measured in terms of misspecification rate (MISR), specificity (SPE), sensitivity (SEN), precision (PRE) and Matthews correlation coefficient (MCC). We report, for each method and setting, the average performance (in percentage points, for each indicator) over 60 simulated datasets, with corresponding standard deviation in brackets. The last column reports the average computational time (in seconds) for each method and setting.

| Setting | $(n, p^*, q)$ | Method | MISR | SPE | SEN | PRE | MCC | Time |
|---|---|---|---|---|---|---|---|---|
| Sparse | $(50, 100, 30)$ | OBFBF | 9(1) | 92(1) | 74(3) | 32(9) | 47(5) | 4769 |
| | | HYPERT | 10(1) | 91(1) | 74(4) | 29(2) | 46(2) | 4270 |
| | | ANTAC | 1(0) | 100(0) | 72(1) | 100(1) | 84(1) | 34 |
| | | GLASSO | 83(5) | 17(5) | 86(4) | 5(0) | 15(2) | 8 |
| | | CONDIT | 52(11) | 48(11) | 90(7) | 8(2) | 21(4) | 99 |
| | | LOWRANK | 49(14) | 50(14) | 91(7) | 9(2) | 22(5) | 75 |
| Sparse | $(50, 100, 60)$ | OBFBF | 3(2) | 97(2) | 84(1) | 49(31) | 60(19) | 5550 |
| | | HYPERT | 5(0) | 95(0) | 84(2) | 28(1) | 47(1) | 5990 |
| | | ANTAC | 0(0) | 100(0) | 83(1) | 100(1) | 91(0) | 109 |
| | | GLASSO | 59(5) | 41(5) | 93(2) | 3(0) | 12(1) | 57 |
| | | CONDIT | 27(19) | 73(20) | 89(4) | 8(3) | 24(6) | 268 |
| | | LOWRANK | 81(3) | 18(3) | 97(3) | 2(0) | 7(1) | 236 |
| Sparse | $(50, 100, 120)$ | OBFBF | 0(0) | 100(0) | 100(0) | 99(9) | 95(5) | 3745 |
| | | HYPERT | 2(0) | 98(0) | 91(1) | 32(2) | 54(1) | 5941 |
| | | ANTAC | 0(0) | 100(0) | 91(0) | 100(0) | 95(0) | 676 |
| | | GLASSO | 36(4) | 64(4) | 95(1) | 2(0) | 12(1) | 547 |
| | | CONDIT | 48(25) | 52(25) | 96(2) | 2(1) | 11(5) | 861 |
| | | LOWRANK | 94(1) | 6(1) | 99(1) | 1(0) | 3(0) | 1002 |
| Magnified | $(50, 100, 150)$ | OBFBF | 0(0) | 100(0) | 93(0) | 94(20) | 92(12) | 5498 |
| | | HYPERT | 2(0) | 99(0) | 93(0) | 32(2) | 54(1) | 6770 |
| | | ANTAC | 0(0) | 100(0) | 93(0) | 99(1) | 96(0) | 1971 |
| | | GLASSO | 78(5) | 22(5) | 97(1) | 1(0) | 5(1) | 4570 |
| | | CONDIT | 96(3) | 4(3) | 100(1) | 1(0) | 2(1) | 3517 |
| | | LOWRANK | 98(0) | 2(0) | 100(0) | 1(0) | 1(0) | 5452 |

Figure 1: Run times of OBFBF for different sample sizes and potential number of regressors, for $q = 5$. The run time of the algorithm at $n = p^* = 50$ is fixed to 1 and taken as reference for the other scenarios.

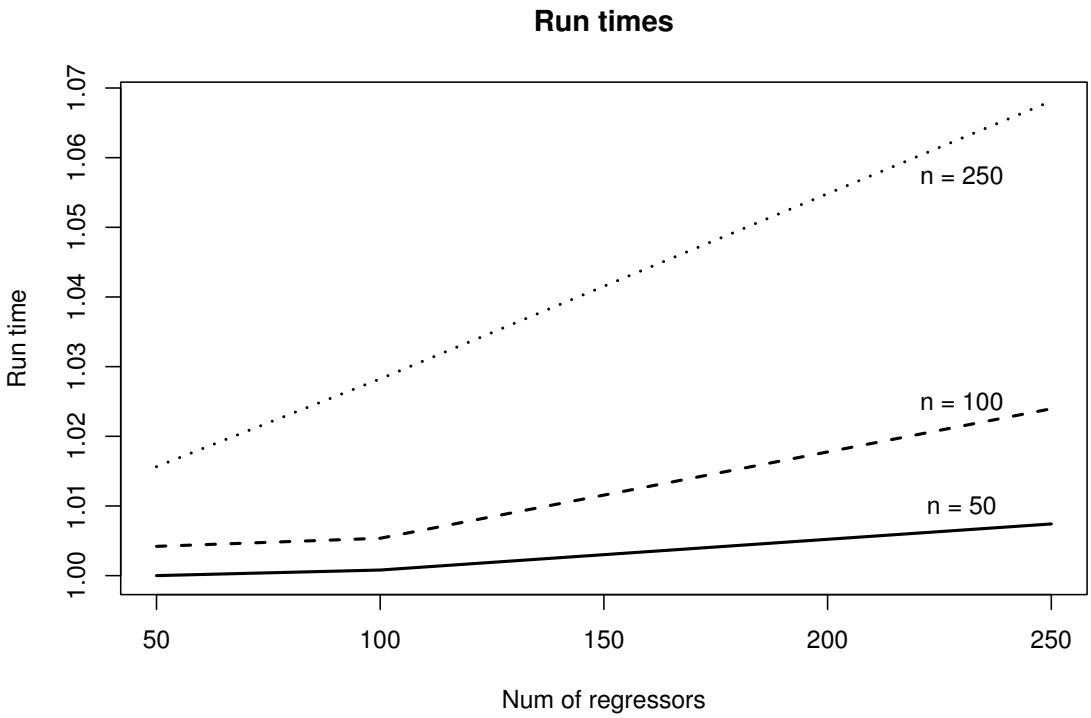

Figure 2: Graph selection results. In the *sparse* simulation setting described in section 6, for sample size $n = 200$, potential number of regressors $p^* = 100$ and number of responses $q = 30$, we report the true adjacency matrix used to generate the data, and we compare it with adjacency matrices estimated by OBFBF, ANTAC, GLASSO, HYPERT, CONDIT and LOWRANK. Red and yellow cells correspond, respectively, to present or absent edges.

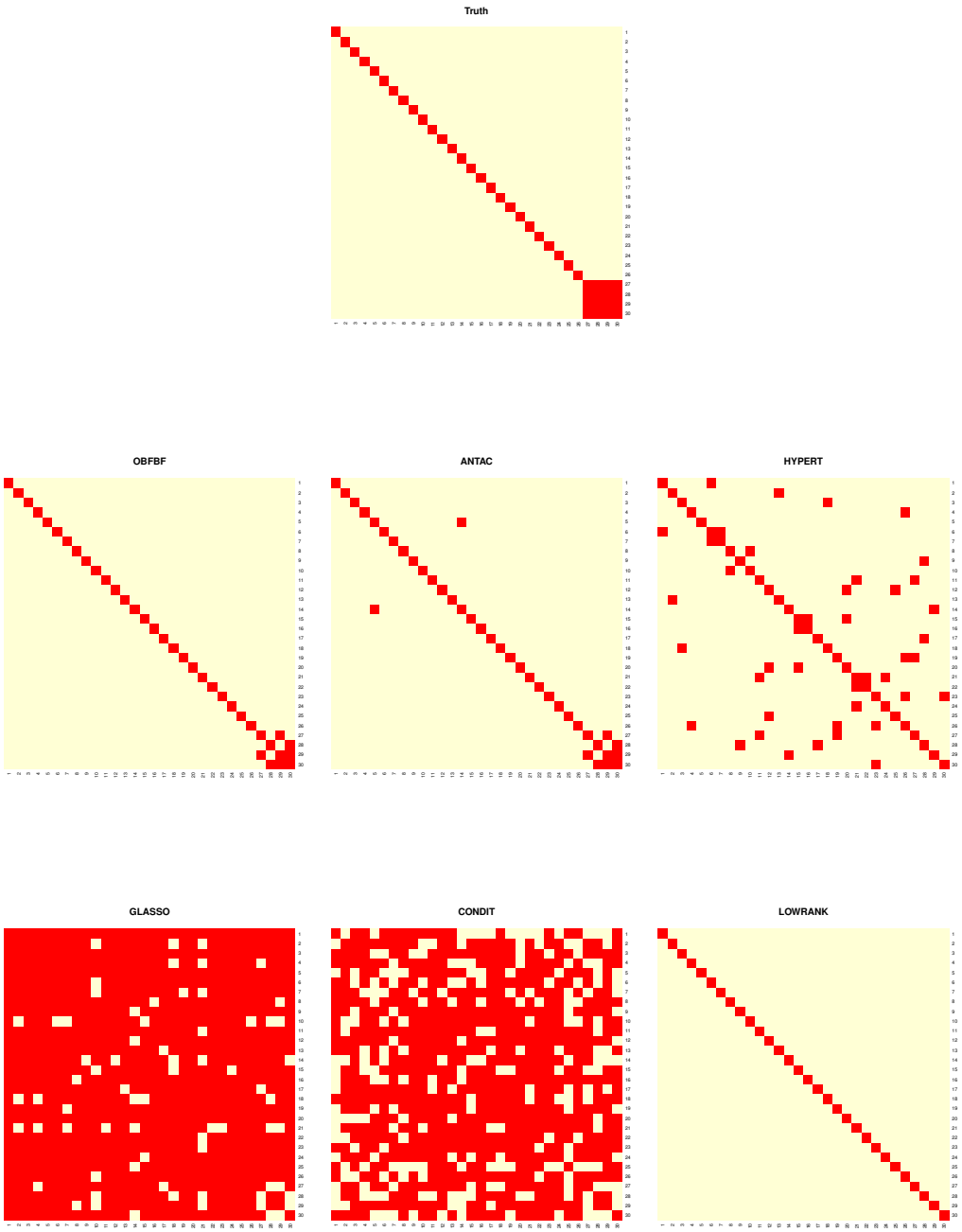
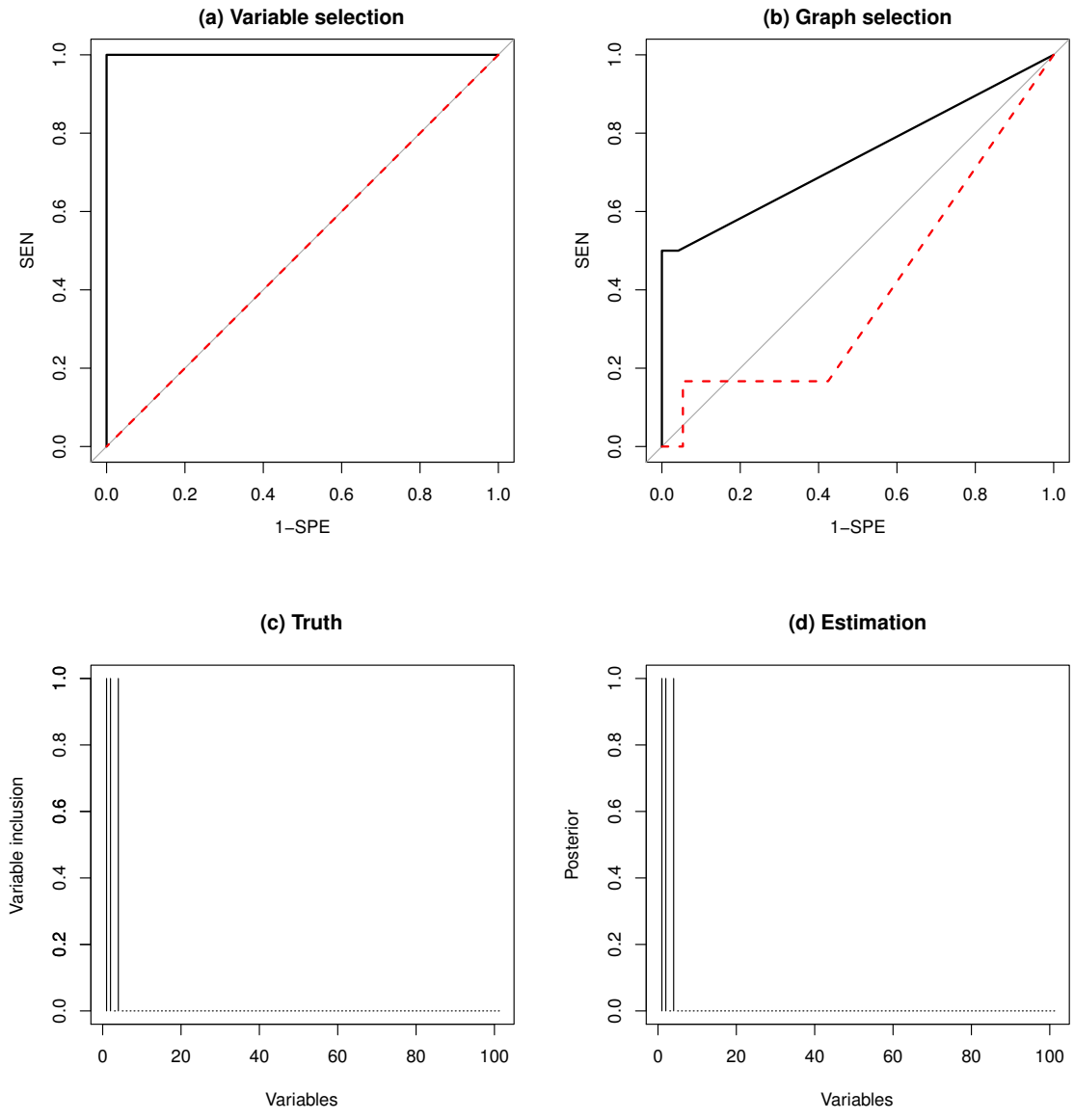
Figure 3: Graph and variable selection results. (a) ROC curve for variable selection, as we vary the threshold of variable inclusion in the model; the black continuous line represents our method OBFBF, the red dashed line is HYPERT. (b) ROC curve for graph selection, as we vary the threshold of edge inclusion in the model; the black continuous line represents OBFBF, the red dashed line is HYPERT. (c) Regressors included in and excluded from the model are depicted as vertical lines of height 1 and 0, respectively, out of $p^* = 100$ potential regressors. (d) Posterior probability of variable inclusion in the model under OBFBF.