Consonni, Guido; La Rocca, Luca

Working Paper

Objective Bayes Factors for Gaussian Directed Acyclic Graphical Models

Quaderni di Dipartimento, No. 141

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

Suggested Citation: Consonni, Guido; La Rocca, Luca (2011) : Objective Bayes Factors for Gaussian Directed Acyclic Graphical Models, Quaderni di Dipartimento, No. 141, Università degli Studi di Pavia, Dipartimento di Economia Politica e Metodi Quantitativi (EPMQ), Pavia

This Version is available at:
https://hdl.handle.net/10419/95330

Quaderni di Dipartimento

Objective Bayes Factors for Gaussian Directed Acyclic Graphical Models

Guido Consonni
(Università di Pavia)

Luca La Rocca
(Università di Modena e Reggio Emilia)

141 (03-11)

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

Marzo 2011

Abstract

We propose an objective Bayesian method for the comparison of all Gaussian directed acyclic graphical models defined on a given set of variables. The method, which is based on the notion of fractional Bayes factor, requires a single default (typically improper) prior on the space of unconstrained covariance matrices, together with a prior sample size hyper-parameter, which can be set to its minimal value. We show that our approach produces genuine Bayes factors. The implied prior on the concentration matrix of any complete graph is a data-dependent Wishart distribution, and this in turn guarantees that Markov equivalent graphs are scored with the same marginal likelihood. We specialize our results to the smaller class of Gaussian decomposable undirected graphical models, and show that in this case they coincide with those recently obtained using limiting versions of hyper-inverse Wishart distributions as priors on the graph-constrained covariance matrices.

Keywords: Bayes factor; Bayesian model selection; Directed acyclic graph; Exponential family; Fractional Bayes factor; Gaussian graphical model; Objective Bayes; Standard conjugate prior; Structural learning.

1 Introduction

Consider a set of variables whose independence structure can be represented by a Directed Acyclic Graph (DAG). A DAG model is a (parametric) family of multivariate distributions which are Markovian with respect to the DAG; see Cowell *et al.* (1999). Different DAGs may define the same DAG model, in which case they are called *Markov equivalent*. Nevertheless, it is often useful to confound a DAG with its model, *if* invariance can be achieved within Markov equivalence classes. For a given DAG, Bayesian inference requires the specification of a prior on the corresponding parameter space, which we call a *DAG-conditional parameter prior*. If a collection of DAGs is entertained, Bayesian model comparison requires: i) the elicitation of a parameter prior for each DAG; ii) a prior distribution on the collection of DAGs. In this paper we focus on the first point.

Geiger & Heckerman (2002) list a set of assumptions on DAG-conditional priors which permit their construction, for all possible DAGs, starting from a single prior associated to a *complete* DAG (a DAG where all pairs of vertices are directly connected). Additionally, their method is such that Markov equivalent DAGs have the same marginal likelihood. This is an important *desideratum* for model comparison, whenever DAGs are regarded as models of *conditional independence*, as opposed to *causal* models; see Lauritzen (2001) and Dawid (2003) for an appreciation of this distinction. For Gaussian DAG models with zero expectation, the method of Geiger & Heckerman (2002) requires a prior distribution on the unconstrained covariance matrix associated to *any* complete DAG (all of them being equivalent). This specification can be very hard, if a purely subjective viewpoint is adopted, especially when many variables are involved. On the other hand, weakly informative (proper) priors do not represent a viable solution for model determination; see Berger & Pericchi (2001). Finally, default noninformative priors, which are typically improper, cannot be used because the Bayes factors would depend on arbitrary constants.

The above remarks suggest the adoption of an *objective* approach, which requires minimal prior inputs, and yet produces meaningful comparisons among models. The

best known objective Bayesian methods for model determination to date are those based on fractional Bayes factors (O'Hagan, 1995), intrinsic Bayes factors (Berger & Pericchi, 1996), intrinsic priors (Moreno, 1997), and expected-posterior priors (Perez & Berger, 2002). Pericchi (2005) provides a comprehensive review.

Recently, Carvalho & Scott (2009) have proposed an objective Bayesian model selection procedure for Gaussian decomposable Undirected Graph (UG) models based on a suitable improper prior and a Fractional Bayes Factor (FBF) approach. They show the superiority of their method in terms of structural learning, relative to some conventional proper priors supposedly believed to be weakly informative.

In this paper we propose an objective methodology based on the FBF to carry out model determination in the class of Gaussian DAG models, which is strictly larger than the class of Gaussian decomposable UG models; see Andersson *et al.* (1997). A key result is that our method satisfies the assumptions of Geiger & Heckerman (2002) and is thus invariant with respect to Markov equivalence. We show this by means of a general interpretation of the FBF within the setup of exponential families and *generalized* (possibly improper) standard conjugate priors. Since any decomposable UG model coincides with some DAG model, and the approach by Geiger & Heckerman (2002) does not discriminate between Markov equivalent DAGs, our method naturally applies to the class of Gaussian decomposable UG models. We adapt our formulas to deal with these models, and show that in this special case we obtain the results presented in Carvalho & Scott (2009).

Our contribution can be seen from two perspectives. On the one hand, we reformulate the procedure by Geiger & Heckerman (2002) in the context of Gaussian models using an objective approach, thus making it more easily applicable; incidentally, we also correct a result in a crucial formula for computing the marginal likelihood of a DAG. On the other hand, we extend the procedure by Carvalho & Scott (2009) to a larger class of Gaussian graphical models. Interestingly, the latter result is achieved using elementary distributional tools, namely *ordinary* Wishart distributions, whereas Carvalho & Scott (2009) have to rely on the more elaborate notion of *hyper-inverse* Wishart law.

The rest of the paper is organized as follows: in section 2 we provide results on marginal data distributions for subsets of multivariate Gaussian variables; section 3 points out a useful interpretation of the FBF for exponential families; section 4 presents our method for comparing Gaussian DAG models, and section 5 applies it to Gaussian decomposable UG models. Finally, Section 6 considers potential implementation of our method for searching a space of DAGs and discusses extensions to non-local priors.

2 Multivariate Gaussian variables

In this section we report some results useful to compute marginal data distributions for subsets of Gaussian variables. These will be needed in the sequel to obtain the marginal likelihood of any DAG. Preliminarily, we set out notation for relevant distributions and discuss the important issue of conditioning and marginalization.

2.1 Distributions

We write $u|\mu, \Omega \sim N_p(\mu, \Omega^{-1})$ to say that the random vector u follows a p-dimensional normal distribution conditionally on the expectation μ and covariance matrix Ω^{-1}; Ω is also referred to as the *precision*, or *concentration*, matrix. Unless explicitly stated otherwise, we assume that the only constraint satisfied by the $p(p+1)/2$ elements of Ω is that Ω be symmetric and positive definite (s.p.d.); occasionally, we will underline this fact by writing that Ω is *unconstrained*. The lack of constraints on Ω characterizes the *complete* Gaussian DAG (or UG for that matter) model.

Let U be a $p \times p$ unconstrained s.p.d. random matrix. We write $U \sim W_p(a, A)$ to mean that U follows a Wishart distribution with density

$$p^W(U) = c(p,a)|A|^{\frac{a}{2}}|U|^{\frac{a-p-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(UA)\right\}, \quad U \in \mathcal{U}, \tag{1}$$

and $p^W(U) = 0$, $U \notin \mathcal{U}$, where $\mathcal{U}$ is the set of all unconstrained s.p.d. $p \times p$-matrices, A is a $p \times p$ s.p.d. matrix, a is a scalar strictly greater than $p-1$ and

$$c(p,a) \;=\; \left\{\int_{\mathcal{U}} |A|^{\frac{a}{2}}|U|^{\frac{a-p-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(UA)\right\} dU\right\}^{-1}$$

$$= \left\{ 2^{\frac{ap}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma\left(\frac{a+1-j}{2}\right) \right\}^{-1}, \tag{2}$$

where dU stands for the differential of the $p(p+1)/2$ distinct elements of U, i.e., the Lebesgue measure element on $\mathbb{R}^{p(p+1)/2}$. As for parameter interpretation, it can be shown that, given a and A, $\mathbb{E}[U] = aA^{-1}$. The notation $W_p(a, A)$ for the density (1) is essentially that employed by Geiger & Heckerman (2002), following DeGroot (1970, p. 59); other authors (Press, 1982; Lauritzen, 1996) would instead use A^{-1} in place of A in (1).

2.2 Conditioning and marginalization

Let u be a p-dimensional random vector with covariance matrix Σ (a $p \times p$ s.p.d. matrix). Partition u as

$$u' = [v'\ w'], \tag{3}$$

where v has dimension p_v and w has dimension p_w, with $p_v + p_w = p$; then partition Σ and $\Omega = \Sigma^{-1}$ as

$$\Sigma = \begin{bmatrix} \Sigma_{vv} & \Sigma_{vw} \\ \Sigma_{wv} & \Sigma_{ww} \end{bmatrix}, \quad \Omega = \begin{bmatrix} \Omega_{vv} & \Omega_{vw} \\ \Omega_{wv} & \Omega_{ww} \end{bmatrix}. \tag{4}$$

The block Σ_{vv} is the *marginal* covariance matrix of v. The *partial* covariance matrix of v given w (defined as the residual variance associated to the linear least squares predictor of v from w) is given by

$$\mathbb{V}\mathrm{ar}(v|w) = \Sigma_{vv} - \Sigma_{vw}\Sigma_{ww}^{-1}\Sigma_{wv} \equiv \Sigma_{vv \cdot w} = (\Omega_{vv})^{-1}, \tag{5}$$

where $\Sigma_{vv \cdot w}$ is called the *Schur complement* of Σ_{ww} in Σ. If u follows a multivariate normal distribution, then $\mathbb{V}\mathrm{ar}(v|w)$ coincides with the *conditional* covariance matrix of v given w; see for instance Whittaker (1990, Ch. 5).

Notice that (5) expresses a relationship between four blocks of Σ and a corresponding block of $\Sigma^{-1} = \Omega$. Hence, by switching the roles of Σ and Ω, we obtain

$$\Sigma_{vv} = \left(\Omega_{vv} - \Omega_{vw}\Omega_{ww}^{-1}\Omega_{wv}\right)^{-1}. \tag{6}$$

Since $\Sigma_{vv} = (\Omega^{-1})_{vv}$, we can also write

$$((\Omega^{-1})_{vv})^{-1} = \Omega_{vv} - \Omega_{vw}\Omega_{ww}^{-1}\Omega_{wv} \equiv \Omega_{vv\cdot w}. \tag{7}$$

Thus, 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.

Theorem 2.1 *Let $\Omega \sim W_p(a, A)$, with A an s.p.d. matrix and $a > p - 1$. If Ω is partitioned as in (4), and A is partitioned accordingly, then*

$$\Omega_{vv\cdot w} \sim W_{p_v}(a - p_w, A_{vv}). \tag{8}$$

Proof. See Press (1982, Theorem (5.1.4)), recalling that Press's parameterization for the Wishart differs from ours. More in detail, start with $\Omega \sim \tilde{W}_p(a, A^{-1})$, where the tilde reminds us that $\mathbb{E}[\Omega] = aA^{-1}$, and use the theorem to conclude that $\Omega_{vv\cdot w} \sim \tilde{W}_{p_v}(a - p_w, (A^{-1})_{vv\cdot w})$, whence (8) follows because of (7). See also Lauritzen (1996, Proposition C.15) with similar care for the notation.

2.3 Marginal data distributions

Let $u_1, \ldots, u_n | \Omega \overset{i.i.d.}{\sim} N_p(0, \Omega^{-1})$ and $\Omega \sim W_p(a, A)$, with A an s.p.d. matrix and $a > p - 1$. We want to compute the *marginal density* $m(u_1, \ldots, u_n)$ of the data; when model comparison is the focus this is also called the *marginal likelihood* (of the underlying model). Let $S = \sum_{i=1}^n u_i u_i'$ be the $p \times p$ matrix of sums of squares and products of the coordinates of u_i, $i = 1, \ldots n$. The marginal data density is then

$$\begin{aligned}
m(u_1, \ldots, u_n) &= \int f(u_1, \ldots, u_n | \Omega) p^W(\Omega) d\Omega \\
&= \int (2\pi)^{-\frac{np}{2}} |\Omega|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Omega S)\right\} \\
&\qquad c(p, a)|A|^{\frac{a}{2}} |\Omega|^{\frac{a-p-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Omega A)\right\} d\Omega \\
&= (2\pi)^{-\frac{np}{2}} c(p, a)|A|^{\frac{a}{2}} \int |\Omega|^{\frac{a+n-p-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Omega(S + A))\right\} d\Omega \\
&= (2\pi)^{-\frac{np}{2}} \frac{c(p, a)}{c(p, a + n)} \frac{|A|^{\frac{a}{2}}}{|S + A|^{\frac{a+n}{2}}},
\end{aligned} \tag{9}$$

where $c(p,a)$ is defined in (2) and the integral is over the set of all s.p.d. matrices. Now

$$\frac{c(p,a)}{c(p,a+n)} = \frac{2^{\frac{(a+n)p}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{j=1}^{p}\Gamma\left(\frac{a+n+1-j}{2}\right)}{2^{\frac{ap}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{j=1}^{p}\Gamma\left(\frac{a+1-j}{2}\right)} = 2^{\frac{np}{2}}\frac{\prod_{j=1}^{p}\Gamma\left(\frac{a+n+1-j}{2}\right)}{\prod_{j=1}^{p}\Gamma\left(\frac{a+1-j}{2}\right)},$$

whence

$$m(u_1,\ldots,u_n) = (2\pi)^{-\frac{np}{2}}2^{\frac{np}{2}}\frac{\prod_{j=1}^{p}\Gamma\left(\frac{a+n+1-j}{2}\right)}{\prod_{j=1}^{p}\Gamma\left(\frac{a+1-j}{2}\right)}\frac{|A|^{\frac{a}{2}}}{|S+A|^{\frac{a+n}{2}}}, \tag{10}$$

leaving the factor $(2\pi)^{-np/2}2^{np/2}$ untouched in view of future comparisons.

We will also need the marginal density of the random sample $v_1,\ldots,v_n$ corresponding to the *subvector* v in the partition of u defined in (3). To this aim, we first note that $v_1,\ldots,v_n|\Omega \overset{i.i.d.}{\sim} N_p(0,(\Omega_{vv\cdot w})^{-1})$. Then, we obtain from Theorem 2.1 that $\Omega_{vv\cdot w} \sim W_{p_v}(a-p_w, A_{vv})$. Hence, we just need to make the following substitutions in (9) and (10): $p \to p_v$; $a \to a-p_w$; $A \to A_{vv}$; $S \to S_{vv}$. We conclude that

$$m(v_1,\ldots,v_n) = (2\pi)^{-\frac{np_v}{2}}\frac{c(p_v,a-p_w)}{c(p_v,a-p_w+n)}\frac{|A_{vv}|^{\frac{a-p_w}{2}}}{|S_{vv}+A_{vv}|^{\frac{a-p_w+n}{2}}} \tag{11}$$

$$= (2\pi)^{-\frac{np_v}{2}}2^{\frac{np_v}{2}}\frac{\prod_{j=1}^{p_v}\Gamma\left(\frac{a-p_w+n+1-j}{2}\right)}{\prod_{j=1}^{p_v}\Gamma\left(\frac{a-p_w+1-j}{2}\right)}\frac{|A_{vv}|^{(a-p_w)/2}}{|S_{vv}+A_{vv}|^{(a-p_w+n)/2}}. \tag{12}$$

Since the sampling distribution of the unconstrained normal model belongs to an exponential family (Lauritzen, 1996, Sect. 5.1.2), we can derive (10) and (12) as a special case of a more general expression which holds for exponential families paired with standard conjugate priors. This setting is especially useful in view of the FBF interpretation which we present in section 3.

A statistical model for the random sample of size n, $y \in \mathcal{Y}$, is an exponential family if the sampling density of y can be written as

$$f(y|\theta) = h_n(y)\exp\left\{\langle\theta,s\rangle - nM(\theta)\right\}, \quad y \in \mathcal{Y}, \tag{13}$$

where $s \equiv s(y)$ is the *canonical statistic* belonging to a real Euclidean vector space endowed with inner product $\langle\cdot,\cdot\rangle$, θ is the corresponding *canonical parameter*, and $e^{nM(\theta)}$ is, for each given θ, the normalizing constant; we do not absorb the leading

factor $h_n(y) = \prod_{i=1}^n h_1(y_i)$ into the dominating measure, because we like the latter to be a product of either Lebesgue or counting measures.

The standard conjugate prior density on θ with respect to the Lebesgue measure is given by

$$p^C(\theta) \;=\; K(n_\bullet, s_\bullet) \exp\left\{\langle \theta, s_\bullet \rangle - n_\bullet M(\theta)\right\}, \tag{14}$$

where $s_\bullet$ is a prior guess of s and $n_\bullet$ is a prior sample size, while $K(n_\bullet, s_\bullet)$ is the corresponding normalizing constant, assuming it exists. An alternative term for the prior (14) is DY-prior after Diaconis & Ylvisaker (1979); see also Consonni & Veronese (1992) and Gutiérrez-Peña & Smith (1995) for extensive discussions on conjugate prior families. The corresponding posterior density is

$$p^C(\theta|y) \;=\; K(n_\bullet + n, s_\bullet + s) \exp\left\{\langle \theta, s_\bullet + s \rangle - (n_\bullet + n) M(\theta)\right\}. \tag{15}$$

We can easily specialize the above setup to multivariate normal data with zero mean. Let the random sample be $u_i|\Omega \sim N_p(0, \Omega^{-1})$ independently for $i = 1, \ldots, n$. Then, $s = -S/2$, where $S = \sum_{i=1}^n u_i u_i'$ belongs to the vector space of matrices with inner product $\langle A, B \rangle = \mathrm{tr}(A'B)$, and $\theta = \Omega$ is the canonical parameter. Additionally $M(\Omega) = -\log|\Omega|/2$, while $h_n(u_1, \ldots, u_n) = (2\pi)^{-np/2}$. The standard conjugate prior family on Ω is Wishart. In particular, if we set $n_\bullet = a - p - 1$ and $s_\bullet = -A/2$, then we recover our original $W_p(a, A)$ formulation, so that we can write with a slight abuse of notation $K(a, A) = c(p, a)|A|^{a/2}$.

We can now derive an expression for $m^C(y) = \int f(y|\theta) p^C(\theta) d\theta$, the marginal data density, in the exponential-conjugate family setup. In fact, from $m^C(y) = f(y|\theta)p^C(\theta)/p^C(\theta|y)$, we get

$$\begin{aligned} m^C(y) \;&=\; h_n(y) \frac{\exp\left\{\langle \theta, s\rangle - nM(\theta)\right\} K(n_\bullet, s_\bullet) \exp\left\{\langle \theta, s_\bullet \rangle - n_\bullet M(\theta)\right\}}{K(n_\bullet + n, s_\bullet + s) \exp\left\{\langle \theta, s_\bullet + s \rangle - (n_\bullet + n) M(\theta)\right\}} \\ &=\; h_n(y) \frac{K(n_\bullet, s_\bullet)}{K(n_\bullet + n, s_\bullet + s)}. \end{aligned} \tag{16}$$

When the data are multivariate normal with zero mean, it is immediate to realize that (16) specializes to (9), and hence to (10). With obvious modifications, the marginal density of the data subset $v_1, \ldots, v_n$ in (11) and (12) can also be derived in the same way (thanks to Theorem 2.1).

3 Fractional Bayes factors

We first recall the definition of FBF, then we cast it into the exponential family-conjugate prior setting, and finally we focus on the case of multivariate normal data with zero mean.

3.1 Definition

Consider a collection of models M_k, $k = 1, \ldots, K$, for the same observables y. Let $f_k(y|\theta_k)$ denote the sampling distribution of y under M_k, and let $p_k(\theta_k)$ be the corresponding prior density, which we assume proper. We focus on the comparison of M_k with M_j through the Bayes Factor (BF)

$$B_{kj}(y) = \frac{m_k(y)}{m_j(y)},$$

where $m_k(y) = \int f_k(y|\theta_k)p_k(\theta_k)d\theta_k$ is the marginal likelihood of M_k.

In lack of specific prior information, we would like to take $p_k(\theta_k) = p_k^D(\theta_k)$ for some default, noninformative, objective prior. However, objective priors are often improper and they cannot be naively used to compute BFs, even when the marginal likelihoods $m_k(y)$ are finite, because of the presence of arbitrary constants which do not cancel out when taking their ratios. Several proposals to overcome this difficulty have been put forward; see Pericchi (2005). In this paper we focus on the FBF introduced by O'Hagan (1995).

Let $0 < b < 1$ be a quantity depending on the sample size n, and define

$$m_k(y; b) = \frac{\int f_k(y|\theta_k)p^D(\theta_k)d\theta_k}{\int f_k^b(y|\theta_k)p^D(\theta_k)d\theta_k}, \tag{17}$$

where $f_k^b(y|\theta_k)$ is the sampling density of model M_k raised to the b-th power, and the integrals are assumed to be finite and nonzero. We refer to $m_k(y; b)$ as the *fractional marginal likelihood* for the k-th model. For later purposes it is useful to rewrite (17) as

$$m_k(y; b) = \int f_k^{1-b}(y|\theta_k)p^F(\theta_k|b, y)d\theta_k, \tag{18}$$

where $p^F(\theta_k|b,y) \propto f_k^b(y|\theta_k)p^D(\theta_k)$ is the implied *fractional prior* (actually a "posterior" based on the b-fractional likelihood). The FBF of M_k against M_j is then defined as

$$FBF_{kj}(y;b) = \frac{m_k(y;b)}{m_j(y;b)}.$$

Clearly, the FBF depends on the choice of b. Usually b will be small, so that the dependence on the data of the prior will be weak. Consistency is achieved as long as $b \to 0$ for $n \to \infty$, and O'Hagan (1995, Sect. 4) suggests three possible choices: i) $b = n_0/n$, where n_0 is the minimal (integer) training sample size for which the fractional marginal likelihood is well defined; ii) $b = \max\{n_0, \sqrt{n}\}/n$; iii) $b = \max\{n_0, \log n\}/n$. Choice i) is suggested as the standard option, when robustness issues are of little concern, while ii) is recommended when robustness is a serious concern, and iii) represents an intermediate option. Moreno (1997) has an argument for i) being the only valid choice, and we stick to this choice in this paper.

3.2 Interpretation for exponential families

For exponential families and conjugate priors, the FBF admits a simple and intuitive interpretation, which both puts it on firmer grounds and makes its computation straightforward. We detail this interpretation below, and exploit it in section 4 for Gaussian DAG model comparison.

Suppose the sampling density can be written as in (13). Furthermore, suppose the default prior has a conjugate form

$$p^D(\theta|n_\bullet^D, s_\bullet^D) \propto \exp\{\langle\theta, s_\bullet^D\rangle - n_\bullet^D M(\theta)\}, \tag{19}$$

where we allow $n_\bullet^D$ and $s_\bullet^D$ to be such that (19) is an improper prior. In this way (19) includes Jeffreys's prior on Ω in the multivariate normal family, and more generally Jeffreys's prior for the canonical parameter of exponential families having a simple quadratic variance function; see Gutiérrez-Peña & Smith (1995). Now write the fraction b as $b = n_0/n$ for some $0 < n_0 < n$, as suggested at the end of the previous subsection. The fractional likelihood becomes

$$f(y|\theta)^{\frac{n_0}{n}} = h_n(y)^{\frac{n_0}{n}} \exp\{\langle\theta, n_0\bar{s}\rangle - n_0 M(\theta)\},$$

where $\bar{s} = s/n$ is the average value of the canonical statistic. By writing $h_n(y) = \bar{h}^n$, where $\bar{h} \equiv \bar{h}(y)$ is the geometric mean of $\{h_1(y_i),\, i = 1, \ldots, n\}$, the fractional (n_0/n)-likelihood can be interpreted as an ordinary likelihood based on n_0 observations, canonical statistic $n_0\bar{s}$ and leading factor $\bar{h}^{n_0}$. The same can be said for the fractional likelihood $f(y|\theta)^{(n-n_0)/n}$ appearing in (18), with canonical statistic $(n - n_0)\bar{s}$, sample size $n - n_0$ and leading factor $\bar{h}^{n-n_0}$, which will be paired with the fractional prior

$$p^F(\theta|n_0) \propto \exp\{\langle \theta, n_0\bar{s} + s^D_\bullet \rangle - (n_0 + n^D_\bullet)M(\theta)\} \tag{20}$$

to compute the fractional marginal likelihood. Specifically, assuming n_0 is such that $p^F(\theta|n_0, \bar{s})$ is proper, the fractional marginal likelihood $m(y; b) \equiv m(y; n_0)$ will be written, using (18) without the unnecessary subscript k and (16), as

$$m(y; n_0) \;\; = \;\; \bar{h}^{n-n_0} \frac{K(n^D_\bullet + n_0, s^D_\bullet + n_0\bar{s})}{K(n^D_\bullet + n, s^D_\bullet + \bar{s})}.$$

Notice that in the p-dimensional normal case $\bar{h} = (2\pi)^{-p/2}$ independently of y.

We have thus shown that, when the model is an exponential family and the default prior belongs to the family (19), the fractional marginal likelihood corresponds to an ordinary conjugate marginal likelihood based on a particular data-dependent prior. Specifically, the sample size is split as $n = n_0 + (n - n_0)$, with n_0 usually much smaller than n. Notice that $\bar{s}$ is used both as data and as prior information, which we believe is a sensible choice, because it reduces prior-likelihood conflicts lying at the heart of many difficulties surrounding the comparison of nested models. We believe this discussion lends support to the use of the FBF in the setting we have described, because it adheres to the principle laid out in Berger & Pericchi (2001, Sect. 3), namely that "Testing and model selection methods should correspond, in some sense, to actual Bayes factors, arising from reasonable default prior distributions".

3.3 Marginal distributions for normal data

Assume $u_1, \ldots, u_n|\Omega \overset{i.i.d.}{\sim} N_p(0, \Omega^{-1})$ with Ω unconstrained (s.p.d.), so that the joint density is

$$f(u_1, \ldots, u_n|\Omega) = (2\pi)^{-\frac{np}{2}} |\Omega|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Omega S)\right\},$$

where $S = \sum_i u_i u_i'$ is the matrix of sums of squares and products of coordinates. We shall write $\bar{S}$ to denote $n^{-1}S$.

Let

$$p^D(\Omega) \propto |\Omega|^{\frac{a_\Omega - p - 1}{2}}, \tag{21}$$

be a default improper prior. Notice that if $a_\Omega = 0$ we recover Jeffreys's prior. Indeed, Jeffreys's prior on Σ is given by $p^J(\Sigma) \propto |\Sigma|^{-\frac{p+1}{2}}$ (Press, 1982, p. 76). Consider now the transformation $\Omega = \Sigma^{-1}$. The Jacobian of this transformation is $J(\Sigma \to \Omega) = |\Omega|^{-(p+1)}$ (Press, 1982, p. 47 (2.15.8)), whence $p^J(\Omega) \propto |\Omega|^{\frac{p+1}{2}} |\Omega|^{-(p+1)} = |\Omega|^{-\frac{p+1}{2}}$, which coincides with (21) if $a_\Omega = 0$. On the other hand, if $a_\Omega = p - 1$, then $p^D(\Omega) \propto |\Omega|^{-1}$. Since $J(\Omega \to \Sigma) = |\Sigma|^{-(p+1)}$, we find $p^D(\Sigma) \propto |\Sigma|^{-p}$. We will use this prior later on for comparison with existing results.

Since $p^D(\Omega)$ in (21) can be written as in (19), with $n_\bullet^D = a_\Omega - p - 1$ and $s_\bullet^D = 0$, we obtain from (20) that the fractional prior for Ω is $W_p(a_\Omega + n_0, n_0 \bar{S})$. If we set $a_\Omega = p - 1$ then the fractional prior is proper provided $n_0 > 0$, so that the minimal training sample size is $n_0 = 1$ and the corresponding fraction becomes $b = n_0/n = 1/n$. Thus, the fractional marginal likelihood for data $v_1, \ldots, v_n$ relative to the subvector v can be deduced from (12) by replacing the quantities which appear therein according to the scheme below:

$$a \to a_\Omega + n_0; \quad A \to n_0 \bar{S}; \quad n \to (n - n_0); \quad S \to (n - n_0)\bar{S}.$$

The result is

$$m(v_1, \ldots, v_n; n_0) = \frac{2^{\frac{(n-n_0)p_v}{2}}}{(2\pi)^{\frac{(n-n_0)p_v}{2}}} \frac{\prod_{j=1}^{p_v} \Gamma\left(\frac{a_\Omega - p_w + n + 1 - j}{2}\right)}{\prod_{j=1}^{p_v} \Gamma\left(\frac{a_\Omega - p_w + n_0 + 1 - j}{2}\right)} \frac{n_0^{\frac{(a_\Omega - p_w + n_0)p_v}{2}}}{n^{\frac{(a_\Omega - p_w + n_0)p_v}{2}} |S_{vv}|^{\frac{(n-n_0)}{2}}}. \tag{22}$$

Notice that (22) is valid provided $|S_{vv}| > 0$; in particular this implies that $n \geq p_v$.

4 Objective priors for Gaussian DAG models

We first review the approach by Geiger & Heckerman (2002, henceforth G&H); then we discuss their formulas for the Gaussian case, and finally we present our proposal.

Although we aim at giving a self-contained presentation, for reasons of space we must assume the reader is familiar with the basics of graphical modeling theory and notation; see for instance Cowell *et al.* (1999), Lauritzen (1996), Whittaker (1990).

4.1 Geiger and Heckerman's approach

With the aim of comparing DAG models using marginal likelihoods (or equivalently BFs), G&H propose a method for the construction of (DAG-conditional) parameter priors on *all* DAG models with given vertex set, which is particularly attractive because of its simplicity and because it satisfies a natural compatibility requirement for Markov equivalent DAGs.

G&H lay down five assumptions which must be satisfied by their procedure. The first three concern regularity conditions on the sampling distribution of the data, which are naturally fulfilled in the Gaussian case (as detailed by the authors). The next two concern structural properties of the prior and represent the cornerstone of their approach. Recall that in the model specified by a DAG $\mathcal{D}$ the joint density of the p-dimensional random vector $(u(1), \ldots, u(p))'$, where coordinate $u(j)$ is the variable associated to vertex j of $\mathcal{D}$, can be written as

$$f_{\mathcal{D}}(u(1), \ldots, u(p)|\theta) = \prod_{j=1}^{p} f(u(j)|u(\mathrm{pa}_{\mathcal{D}}(j)); \theta_j), \tag{23}$$

where $\mathrm{pa}_{\mathcal{D}}(j)$ denotes the *parents* of j in $\mathcal{D}$, i.e., all nodes in $\mathcal{D}$ from which a directed edge points to j, while θ is the collection of all θ_js, and $u(\mathrm{pa}_{\mathcal{D}}(j))$ is the collection of variables belonging to the vertex set $\mathrm{pa}_{\mathcal{D}}(j)$. Assumption 4, called *prior modularity*, requires that, given two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ prescribing the same set of parents for node j, the parameter prior on θ_j should be the same for the two corresponding models. Assumption 5, called *global parameter independence*, states that for every DAG model the parameters θ_js should be *a priori* independent; equivalently the joint density $p^{\mathcal{D}}(\theta)$ should factorize as $\prod_j p_j^{\mathcal{D}}(\theta_j)$. Both these assumptions had already been used in earlier works containing Bayesian analyses of DAG models, as recounted in Cowell *et al.* (1999, Sect. 9.2 & 9.4), and they are feasible within the Gaussian setting because parameters pertaining to distinct local structures are variation independent.

A first basic result of G&H is reported in their Theorem 1: under Assumptions 1–5, the parameter prior for *any* DAG model is determined by a specified parameter prior for an arbitrary *complete* DAG model. As a consequence, once we specify the parameter prior for one complete DAG model, all other priors can be generated automatically. A crucial implication of this result concerns the computation of the marginal likelihood for a general DAG model. This is contained in Theorem 2 of G&H. Assume $u_1, \ldots, u_n$ is a sample of complete (no missing) data, where $u_i' = (u_i(1), \ldots, u_i(p))$, $i = 1, \ldots, n$. For any complete DAG $\mathcal{D}_c$ and any DAG $\mathcal{D}$, the marginal data density of $(u_1, \ldots, u_n)$ given $\mathcal{D}$, equivalently the marginal likelihood for $\mathcal{D}$, can be written as

$$m_{\mathcal{D}}(u_1, \ldots, u_n) = \prod_{j=1}^{p} \frac{m_{\mathcal{D}_c}(u_1(\mathrm{fa}_{\mathcal{D}}(j)), \ldots, u_n(\mathrm{fa}_{\mathcal{D}}(j)))}{m_{\mathcal{D}_c}(u_1(\mathrm{pa}_{\mathcal{D}}(j)), \ldots, u_n(\mathrm{pa}_{\mathcal{D}}(j)))}, \tag{24}$$

where $\mathrm{fa}_{\mathcal{D}}(j) = \mathrm{pa}_{\mathcal{D}}(j) \cup \{j\}$ is the *family* of j in $\mathcal{D}$ and $u_i(S)$ represents the i-th observation on the collection of variables indexed by set S. The great advantage of (24) is that we simply need to compute the required factors for a *single complete* DAG. The particular features of the DAG structure $\mathcal{D}$ under consideration enter (24) only through the specification of the set of parents for each node j.

Another consequence of Assumptions 1-5 is that every two Markov equivalent DAGs have the same marginal likelihood; see Theorem 4 of G&H. As recalled in the Introduction, this feature is clearly attractive whenever DAGs are regarded purely as models of *conditional independence*, as opposed to *causal* models. Notice that, as a consequence of this feature, we can always consider $\mathcal{D}$ as a subgraph of $\mathcal{D}_c$, in (24), because all complete DAGs are Markov equivalent.

4.2 Geiger and Heckerman for Gaussian DAG models

G&H also address the specific issue of constructing parameter priors for the comparison of Gaussian DAG models. Let the underlying p-dimensional distribution be $N_p(\mu, \Omega^{-1})$. Assuming a complete DAG model, equivalently that Ω be unconstrained (s.p.d.), G&H deduce from Assumptions 1–5 that the prior for (μ, Ω) must be Normal-Wishart: μ normally distributed conditionally on Ω and $\Omega \sim W_p(a_\Omega, A)$.

More specifically, Theorem 10 of G&H establishes that, if $p \geq 3$, global parameter independence holds if and only if the prior on (μ, Ω) is Normal-Wishart. The other assumption on the prior, namely prior modularity, is automatically satisfied whenever the prior on the parameter θ_j of each conditional distribution appearing in the factorization (23) is derived through a unique prior on the parameter indexing the joint distribution. If $\mu = 0$, as in our simplified setup, which is often adopted when dealing with graphical models, global parameter independence holds if and only if the prior on Ω is Wishart; see Theorem 7 of G&H.

G&H produce the marginal data density for a *subset* of the p-variables, which is necessary to implement formula (24). As already remarked, their discussion is slightly more general than ours because they allow for $\mu \neq 0$. Their result on the marginal data density must however specialize to ours upon choosing a prior for $\mu|\Omega$ that is degenerate on zero. This does *not* occur: their result, reported in Geiger & Heckerman (2002, p. 1425, formula (18)), when adapted to our zero-expectation context, *disagrees* with our formula (11). We believe that formula (18) of G&H' is incorrect. A detailed explanation of our claim is provided in the Appendix.

4.3 Fractional Bayes factors for Gaussian DAG models

The approach by Geiger & Heckerman (2002) requires to specify (m, a_μ, a_Ω, A) or, when $\mu = 0$, simply (a_Ω, A). This can be problematic, especially when the dimension of the problem is large. Even when substantial prior information on (a_Ω, A) is available, special care must be exercised because, as recalled in the Introduction, the BF is typically quite sensitive to the choice of these inputs. For these reasons we find it advisable to proceed using an objective method, at least as a way to provide a benchmark result. In the following we develop a proposal based on the notion of FBF.

The key to our proposal is the interpretation of the FBF pointed out in subsection 3.2, which holds true in particular for the normal model, so that the fractional marginal likelihood corresponds to an actual marginal likelihood (with a reduced sample size) based on a data-dependent proper conjugate prior. Hence, the FBF can be

accommodated within the approach by Geiger & Heckerman (2002). Specifically, for a complete DAG $\mathcal{D}_c$, we will use the prior $\Omega \sim W_p(a_\Omega + n_0, n_0\bar{S})$ and pair it to an ordinary Gaussian likelihood with $(n - n_0)$ observations and mean canonical statistic $\bar{S}$. Then, formula (24) will give us the marginal likelihood of any DAG $\mathcal{D}$.

As for the specific expression of $m_{\mathcal{D}}(u_1, \ldots, u_n)$, using (22) the j-th term in the numerator of (24) becomes

$$m_{\mathcal{D}_c}(u_1(\mathrm{fa}_{\mathcal{D}}(j)), \ldots, u_n(\mathrm{fa}_{\mathcal{D}}(j))) = (2\pi)^{-\frac{(n-n_0)p_j}{2}} 2^{\frac{(n-n_0)p_j}{2}} |S_{\mathrm{fa}_{\mathcal{D}}(j)\mathrm{fa}_{\mathcal{D}}(j)}|^{-\frac{(n-n_0)}{2}} \cdot \frac{\prod_{j=1}^{p_j} \Gamma\left(\frac{a_\Omega+n-p+p_j+1-j}{2}\right)}{\prod_{j=1}^{p_j} \Gamma\left(\frac{a_\Omega+n_0-p+p_j+1-j}{2}\right)} \frac{n_0^{\frac{(a_\Omega+n_0-p+p_j)p_j}{2}}}{n^{\frac{(a_\Omega+n_0-p+p_j)p_j}{2}}}, \quad (25)$$

where $p_j \equiv |\mathrm{fa}_{\mathcal{D}}(j)|$ is the cardinality of the set $\mathrm{fa}_{\mathcal{D}}(j)$. To understand (25) simply apply the following substitution into (22): $v \to \mathrm{fa}_{\mathcal{D}}(j)$, $p_v \to p_j$, $p_w \equiv p - p_v \to p - p_j$. On the other hand, the j-th term in the denominator of (24) is exactly as in (25), but with $\mathrm{fa}_{\mathcal{D}}(j)$ replaced by $\mathrm{pa}_{\mathcal{D}}(j)$. Clearly, for (25) to exist $S_{\mathrm{fa}_{\mathcal{D}}(j)\mathrm{fa}_{\mathcal{D}}(j)}$ must be positive definite, which requires $n \geq p_j$. Since this condition must hold for all $j = 1, \ldots, p$, it is necessary that $n \geq \max\{p_j,\, j = 1, \ldots, p\}$.

5 Application to decomposable UG models

It is well known that a decomposable UG is Markov equivalent to some DAG; see Andersson *et al.* (1997). It follows that the methodology developed in subsection 4.3 can be applied to perform Bayesian model determination for decomposable UGs. Notice that the decomposability of an UG $\mathcal{G}$ is equivalent to: i) the cliques of $\mathcal{G}$ can be ordered to form a perfect sequence; ii) the vertices of $\mathcal{G}$ admit a perfect numbering; see Lauritzen (1996, Proposition 2.17).

Now let $\mathcal{G}$ be a decomposable UG. Let $C_1, \ldots, C_K$ be a perfect sequence of cliques. For $k = 2, \ldots, K$, define $H_k = C_1 \cup \ldots \cup C_k$; $S_k = C_k \cap H_{k-1}$; $R_k = C_k \setminus H_{k-1}$. Call H_k the history, S_k the separator and R_k the residual. Notice that $C_1 \cup R_2 \cup \ldots \cup R_K = V$, where V is the vertex set of $\mathcal{G}$. Additionally $R_k \cap R_{k'} = \emptyset$, $k \neq k'$. Let the vertices of $\mathcal{G}$ be numbered with first those in C_1, then those in R_2, R_3, and so on. The numbering

so obtained is perfect; see Lauritzen (1996, Lemma 2.12). Given a perfect numbering of the vertices in $\mathcal{G}$, we can construct its perfect directed version $\mathcal{G}^<$, which is a DAG Markov equivalent to $\mathcal{G}$, simply by directing its edges from lower to higher numbered vertices; see Lauritzen (1996, p. 18).

Now recall the fundamental factorization of a density $f_{\mathcal{G}}(u)$ which is Markovian with respect to the UG $\mathcal{G}$:

$$f_{\mathcal{G}}(u) = \frac{\prod_{C\in\mathcal{C}} f(u(C))}{\prod_{S\in\mathcal{S}} f(u(S))}, \tag{26}$$

where $\mathcal{C}$ is the set of cliques, $\mathcal{S}$ the set of separators, $u(C)$ is the collection of $u(j)$ with $j \in C$, and similarly for $u(S)$. We can write (26) as

$$f_{\mathcal{G}}(u) = \prod_{k=1}^{K} f(u(R_k)|u(S_k)), \tag{27}$$

with the understanding that $R_1 \equiv C_1$ and $S_1 = \emptyset$. Since the vertices of $\mathcal{G}$ are perfectly numbered, one could further decompose $f(u(R_k)|u(S_k))$ into a product of univariate terms (one for each node) thus making it clear that the joint density also factorizes according to the perfect DAG $\mathcal{G}^<$; we omit details. Since the prior satisfies global parameter independence, the marginal data density is also Markovian with respect to $\mathcal{G}$ and we can write

$$m_{\mathcal{G}}(u_1,\ldots,u_n) \;=\; \frac{\prod_{k=1}^{K} m(u_1(C_k),\ldots,u_n(C_k))}{\prod_{k=1}^{K} m(u_1(S_k),\ldots,u_n(S_k))}. \tag{28}$$

Notice that C_k and S_k are complete sets. Hence the k-th factor in the numerator and denominator of (28) is formally equivalent to (25). Consider in particular $m(u_1(C_k),\ldots,u_n(C_k))$. Omitting the subscript k we get for the generic clique C

$$\begin{aligned} m(u_1(C),\ldots,u_n(C)) \;&=\; \frac{2^{\frac{(n-n_0)p_C}{2}}}{(2\pi)^{\frac{(n-n_0)p_C}{2}}} \frac{\prod_{j=1}^{p_C}\Gamma\left(\frac{a_\Omega+n-p+p_C+1-j}{2}\right)}{\prod_{j=1}^{p_C}\Gamma\left(\frac{a_\Omega+n_0-p+p_C+1-j}{2}\right)} \\ &\cdot\; \frac{n_0^{\frac{(a_\Omega+n_0-p+p_C)p_C}{2}}}{n^{\frac{(a_\Omega+n_0-p+p_C)p_C}{2}}} |S_{CC}|^{-\frac{(n-n_0)}{2}}, \end{aligned} \tag{29}$$

where $p_C \equiv |C|$ is the cardinality of the clique C. An analogous expression holds for the denominator of (28): simply replace C with S; this will create the somewhat cumbersome expression S_{SS}, which of course represents the matrix of sums of squares and

products of the data whose vertices belong to the separator S. Formula (28) will be well defined provided each term (29) exists; a necessary condition for this is therefore $n \geq \max\{p_C\}$, where C runs over the set of all cliques. We conclude that the marginal likelihood of the decomposable UG $\mathcal{G}$ is given by (28), where $m(u_1(C_k), \ldots, u_n(C_k))$ is defined in (29) and $m(u_1(S_k), \ldots, u_n(S_k))$ has an analogous expression.

5.1 Relationship to the work by Carvalho and Scott

Carvalho & Scott (2009) consider Bayesian model determination for Gaussian decomposable UG models using an FBF approach. Let $\mathcal{G}$ be a decomposable UG and $\Sigma \in M^+(\mathcal{G})$ be the corresponding covariance matrix, where $M^+(\mathcal{G})$ is the set of all constrained s.p.d. matrices such that $(\Sigma^{-1})_{ij} \equiv \Omega_{ij} = 0$ for all $i \neq j$ not joined by an edge in $\mathcal{G}$. Carvalho & Scott (2009) assign to Σ the default noninformative prior

$$p^D(\Sigma) \propto \frac{\prod_{C \in \mathcal{C}} |\Sigma_{CC}|^{p_C}}{\prod_{S \in \mathcal{S}} |\Sigma_{SS}|^{p_S}}, \; \Sigma \in M^+(\mathcal{G}),$$

which is a limiting form of hyper-inverse Wishart distribution; see Dawid & Lauritzen (1993) and Letac & Massam (2007). When $\mathcal{G}$ is complete, the latter noninformative prior specializes to $p^D(\Sigma) \propto |\Sigma|^{-p}$, or equivalently $p^D(\Omega) \propto |\Omega|^{-1}$, which corresponds to $a_\Omega = p - 1$ in our general notation of subsection 3.3. If we substitute this value into (29) and its analogous expression for $m(u_1(S), \ldots, u_n(S))$, we obtain from (28), after some algebra, the expression for the fractional marginal likelihood of Carvalho & Scott (2009, Theorem 1); notice that there is a typo in their formula (1), where $(2\pi)^{-np/2}$ should be replaced by $(2\pi)^{-(n-n_0)p/2}$.

6 Discussion

Our procedure has been developed under the assumption that the observables to be modeled have zero mean, as it is customary in the analysis of graphical models, which is focussed on the covariance, or precision, matrix. We could extend our procedure to cover the case $N_p(\mu, \Omega^{-1})$ with little extra computations. The implied fractional

prior would then belong to the Normal-Wishart family, as in the approach of Geiger & Heckerman (2002), and the analysis would go through in a very similar fashion.

Our paper is focused on priors for an objective approach to Gaussian DAG model selection. Actual implementation of our method would require setting up a search algorithm over a model space, as for instance the one based on feature inclusion stochastic search implemented in Scott & Carvalho (2009): this would effectively generalize their method to the larger class of DAG models. In this connection efficiency considerations related to exploring only equivalence classes of DAGs should be taken into consideration; see Andersson *et al.* (1997). When the number of variables is very large and the sample size relatively small, as in some current applications to genomic data, searching for the highest probability models may be hopeless. Hence one could resort to learning only some features of the DAG, such as the presence of an edge; see for instance Friedman & Koller (2003).

The parameter priors used in this paper are *local*, in that when comparing nested models the prior under the larger model does not vanish on the subspace corresponding to the smaller model. This is current practice in Bayesian hypothesis testing and model selection, but Johnson & Rossell (2010) have recently advocated the use of *non-local* priors in order to accelerate the rate of learning about the smaller model (when this is actually true) in any pairwise comparison between two nested models. We believe that the rationale behind non-local priors is sound and promising, but in the setting of this paper their use would imply modifying the implied fractional prior, which would no longer be a Wishart. As a consequence, marginal likelihoods would not be invariant within Markov equivalence classes, because this property characterizes the Wishart family, as shown by Geiger & Heckerman (2002). On the other hand, when a fixed ordering of the variables is available, Markov equivalence is not an issue, and an application of non-local priors to Gaussian DAG model comparison can be found in Consonni & La Rocca (2011).

Acknowledgements

This work was partially supported by PRIN grant 2007XECZ7L_001 (MIUR-Italy) and by the University of Pavia. We acknowledge discussion with Alberto Roverato on the issues presented in subsection 2.2.

Appendix

We provide a detailed explanation of the reason why formula (18) on p. 1425 of Geiger & Heckerman (2002, G&H) is incorrect. For the sake of clarity we first summarize G&H's distributional assumptions using *their* notation (but omitting boldface fonts) and specialize their result to the $\mu = 0$ case.

The sampling distribution is $N_n(\mu, W^{-1})$; the prior is $\mu|W \sim N_n(\nu, (a_\mu W)^{-1})$ and $W \sim W_n(a_w, T)$. G&H represent the variables as $(X_1, \ldots, X_n)$ and denote by d a random sample of N complete cases, with x_i the i-th n-dimensional observation of $(X_1, \ldots, X_n)$. In order to specialize G&H's results to the $\mu = 0$ case, it is enough to set $\nu = 0$ and let $a_\mu \to \infty$. If we do so, the following two expressions, appearing on p. 1424 and 1425 of G&H, become

$$\begin{aligned} \frac{a_\mu}{a_\mu + N} &\to 1 \\ R \equiv T + S_N + \frac{a_\mu N}{a_\mu + N}(\nu - \bar{x}_N)(\nu - \bar{x}_N)' &\to T + S_N + N\bar{x}_N\bar{x}_N' = T + S, \end{aligned}$$

where $S_N \equiv \sum_i (x_i - \bar{x}_N)(x_i - \bar{x}_N)'$ and $S \equiv \sum_i x_i x_i'$. Table 1 exhibits the correspondence between our notation and the notation employed by G&H. In this way we can compare formula (18) of G&H with our corresponding formula (11). One realizes that everything agrees *except* for the definition of T_Y and R_Y. According to G&H it should be $T_Y = ((T^{-1})_{YY})^{-1}$ and analogously $R_Y = ((R^{-1})_{YY})^{-1}$, whereas according to our calculations it should be $T_Y = T_{YY}$ and $R_Y = R_{YY}$. Indeed, only in the latter case does the term T_Y match our term A_{vv}, and R_Y match our term $S_{vv} + A_{vv}$.

We believe that the source of error be the following. In the first line of p. 1425 G&H state that, according to their Theorem 5, the distribution of $((W^{-1})_{YY})^{-1}$ is

Geiger & Heckerman	Consonni & La Rocca
X_j	$u(j)$
n	p
N	n
W	Ω
a_w	a
T	A
l	p_v
x_i	u_i
Y	v
$a'_w \equiv a_w - n + l$	$a - p + p_v \equiv a - p_w$

Table 1: Correspondence between G&H's notation and our notation.

$W(a'_w, T_Y \equiv ((T^{-1})_{YY})^{-1})$. Theorem 5 is correct, but the derivation of the distribution of $((W^{-1})_{YY})^{-1}$ is incorrect. Indeed, rephrasing Theorem 5 in the notation of formula (18) on p. 1425, one concludes that $W_{YY \cdot X \setminus Y}$ has distribution $W(a'_w, T_{YY})$; this agrees with our result (8). Consequently, since $((W^{-1})_{YY})^{-1}$ is precisely $W_{YY \cdot X \setminus Y}$, as we have argued in (7), $((W^{-1})_{YY})^{-1}$ must have distribution $W(a'_w, T_{YY})$. If this had been realized, then formula (18) of G&H would have been correctly written. As a double check, notice that T is a covariance-type matrix, because $\mathbb{E}[W] = a_w T^{-1}$, and thus marginalization on Y necessarily corresponds to extracting the submatrix T_{YY}; see subsection 2.2.

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.

Berger, J. & Pericchi, L. (1996). The intrinsic Bayes factor for model selection and

prediction. *Journal of the American Statistical Association* **91**, pp. 109–122.

Berger, J. O. & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison. In *Model selection*, vol. 38 of *IMS Lecture Notes Monogr. Ser.* Inst. Math. Statist., Beachwood, OH, pp. 135–207.

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

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

Consonni, G. & Veronese, P. (1992). Conjugate priors for exponential families having quadratic variance functions. *J. Amer. Statist. Assoc.* **87**, 1123–1127.

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

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 Univ. Press, Oxford, pp. 45–81.

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

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

Diaconis, P. & Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269–281.

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

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

Gutiérrez-Peña, E. & Smith, A. F. M. (1995). Conjugate parameterizations for natural exponential families. *J. Amer. Statist. Assoc.* **90**, 1347–1356. Erratum ibidem 91 (1996), page 1757.

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

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.

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

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.

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

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

Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and Bayes factors. In 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, Inc., Malabar, FL.

Scott, J. & Carvalho, C. (2009). Feature-inclusion stochastic search for gaussian graphical models. *J. Comp. Graph. Stat.* **17**, 790–808.

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