Objective Bayesian model selection in Gaussian graphical models

By C. M. CARVALHO

Booth School of Business, University of Chicago, Chicago, Illinois 60637, U.S.A.

carlos.carvalho@chicagobooth.edu

AND J. G. SCOTT

Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A.

james@stat.duke.edu

Summary

This paper presents a default model-selection procedure for Gaussian graphical models that involves two new developments. First, we develop a default version of the hyper-inverse Wishart prior for restricted covariance matrices, called the hyper-inverse Wishart *g*-prior, and show how it corresponds to the implied fractional prior for selecting a graph using fractional Bayes factors. Second, we apply a class of priors that automatically handles the problem of multiple hypothesis testing. We demonstrate our methods on a variety of simulated examples, concluding with a real example analyzing covariation in mutual-fund returns. These studies reveal that the combined use of a multiplicity-correction prior on graphs and fractional Bayes factors for computing marginal likelihoods yields better performance than existing Bayesian methods.

Some key words: Bayesian model selection; Fractional Bayes factor; Gaussian graphical model; Hyper-inverse Wishart distribution; Multiple hypothesis testing.

1. Bayesian graphical models

1·1. *Introduction*

Gaussian graphical models are tools for modelling conditional independence relationships, and they offer many practical advantages in high-dimensional problems. They can make computing more efficient by alleviating the need to handle large matrices; they can yield better predictions by fitting sparser models; and they can aid scientific understanding by breaking down a global model into a collection of local models that are easier to parse.

Yet often the graph itself must be inferred from the data. Our approach to this problem is Bayesian, meaning that two quantities must be specified: the prior distribution for Σ under each graph and a prior distribution over different possible graphs.

The first specification is difficult because there is no common covariance matrix shared by all graphs, but rather an entire collection of covariance matrices $\{\Sigma_G\}$ indexed by all possible graphs. Different graphs imply different numbers of free elements in Σ, and so it is not possible to use an improper prior for each Σ_G as one might do for covariance estimation under a fixed graph, since this would leave the resulting model probabilities defined only up to an arbitrary constant. Instead, we must either elicit a subjective prior for each Σ_G, which is clearly intractable in high dimensions, or we must choose some default proper prior that is neither too vague nor too

precise. Regardless, it is clear that any answer will depend on the priors chosen for the various Σ_Gs. To handle this difficulty, we introduce an objective-Bayesian approach using fractional Bayes factors.

The second task, specifying a prior across different graphs, is deceptively easy. Yet there are still pitfalls: graphical model selection poses an implicit problem in multiple hypothesis testing, where a null hypothesis is the exclusion of a single edge from the graph. The seemingly objective choice of assigning all graphs equal prior probability will be shown to flag many false-positive edges, and we develop a class of fully Bayesian edge-selection priors to avoid this problem.

1·2. *Notation and priors for constrained covariance matrices*

An undirected graph is a pair $G = (V, E)$ with vertex set V and edge set $E = \{(i, j)\}$ for some pairs $(i, j) \in V$. Nodes i and j are adjacent, or neighbours, if $(i, j) \in E$. Complete graphs are those having $(i, j) \in E$ for every $i, j \in V$. Maximal complete subgraphs $C \subset V$ are called cliques; two cliques that overlap in a set S are said to have S as a separator. A decomposition is a partitioning of a graph G into subgraphs (A, S, B) such that $V = A \cup B$, $S = A \cap B$ is complete, and any path from a node in A to a node in B goes through the separator S. All graphs in this paper are assumed to be decomposable, thus admitting a recursive decomposition into a sequence of cliques. We denote the set of cliques and separators of a graph by $\mathcal{C}$ and $\mathcal{S}$, respectively. A decomposable graph G can be represented by a perfect ordering of cliques and separators. An ordering of cliques $C_i \in \mathcal{C}$ and separators $S_i \in \mathcal{S}$ is said to be perfect if for every $i = 2, \ldots, |V|$ there exists a $j < i$ such that $S_i = C_i \cap H_{i-1} \subset C_j$, where $H_{i-1} = \bigcup_{j=1}^{i-1} C_j$. Finally, a perfect numbering of vertices in G is obtained by taking first the vertices in C_1, then those in $C_2 \backslash H_1$, $C_3 \backslash H_2$ and so on (Lauritzen, 1996). Taking this order in reverse creates what is called a perfect vertex-elimination scheme of G.

A Gaussian graphical model uses a graphical structure to define a set of pairwise conditional-independence relationships on a p-dimensional zero-mean, normally distributed random vector $x \sim N(0, \Sigma)$. The unknown covariance matrix Σ is restricted by its Markov properties; given $\Omega = \Sigma^{-1}$, elements x_i and x_j of the vector x are conditionally independent, given their neighbours, if and only if $\Omega_{ij} = 0$. If $G = (V, E)$ is an undirected graph describing the joint distribution of x, $\Omega_{ij} = 0$ for all pairs $(i, j) \notin E$. The covariance matrix Σ is in $M^+(G)$, the set of all symmetric positive-definite matrices having elements in Σ^{-1} set to zero for all $(i, j) \notin E$.

The hyper-inverse Wishart distribution is a general class of hyper-Markov laws introduced by Dawid & Lauritzen (1993) for a covariance matrix $\Sigma \in M^+(G)$, where $G = (V, E)$ is a decomposable graph. The notation is $(\Sigma \mid G) \sim \text{HIW}_G(b, D)$, where $b \in \mathbb{R}^+$ is a degrees-of-freedom parameter, and where $D \in M^+(G)$ is a symmetric positive-definite scale matrix. The density of this distribution is defined with respect to the product of Lebesgue measures for the (i, j) elements of Σ for which $(i, j) \in E$, subject to the conditions that Σ_C is symmetric and positive definite for all $C \in \mathcal{C}$.

The density of Σ can be obtained from the clique-specific marginal densities as a ratio of products over cliques and separators

$$p(\Sigma \mid G) = \frac{\prod_{C \in \mathcal{C}} p(\Sigma_C \mid b, D_C)}{\prod_{S \in \mathcal{S}} p(\Sigma_S \mid b, D_S)}, \tag{1}$$

where, for each clique $C \in \mathcal{C}$ (and separators $S \in \mathcal{S}$), $\Sigma_C \sim \text{IW}(b, D_C)$ with density

$$p(\Sigma_C \mid b, D_C) \propto |\Sigma_C|^{-(b/2+|C|)} \exp\left\{-\frac{1}{2}\text{tr}\left(\Sigma_C^{-1} D_C\right)\right\}. \tag{2}$$

The factorization in (1) holds if we assume that if $S = C_1 \cap C_2$ the elements of Σ_S are common in Σ_{C_1} and Σ_{C_2}. For further details and explanations, refer to Dawid & Lauritzen (1993); Giudici & Green (1999) and Letac & Massam (2007).

1·3. *Marginal likelihoods for graphs*

Suppose we observe n samples $(x_1, \ldots, x_n)$ of p-dimensional vectors, where each $x_i \sim N(0, \Sigma)$, with unknown covariance matrix Σ. Let X be the $n \times p$ matrix of samples and let X_j refer to column j of X and x_i to row i; also, let X_C refer to the columns of X corresponding to the nodes in clique C; and assume that $\Sigma \in M^+(G)$ for some unknown decomposable graph G on p nodes. The posterior probability of a graph G is

$$p(G \mid X) \propto p(G) \int_{\Sigma \in M^+(G)} p(X \mid \Sigma, G) \, p(\Sigma \mid G) \, \mathrm{d}\Sigma, \tag{3}$$

where $p(G)$ is the prior probability of the graph, and where the integral is the marginal likelihood of the data under G. If $\Sigma \sim \mathrm{HIW}_G(b, D)$, the integral in (3) is available in closed form using the ratio of the prior and posterior normalizing constants

$$p(X \mid G) = (2\pi)^{-np/2} \frac{h(G, b, D)}{h(G, b^*, D^*)}, \tag{4}$$

where $b^* = b + n$ and $D^* = D + X'X$. The normalizing constant $h(\cdot)$ is

$$h(G, b, D) = \frac{\prod_{C \in \mathcal{C}} \left| \frac{1}{2} D_C \right|^{(b+|C|-1)/2} \Gamma_{|C|} \left(\frac{b+|C|-1}{2} \right)^{-1}}{\prod_{S \in \mathcal{S}} \left| \frac{1}{2} D_S \right|^{(b+|S|-1)/2} \Gamma_{|S|} \left(\frac{b+|S|-1}{2} \right)^{-1}}, \tag{5}$$

where $\Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\{x + (1-j)/2\}$ is the multivariate gamma function.

2. Model-selection priors for restricted covariance matrices

2·1. *Criteria for model-selection priors*

The expression for the marginal likelihood in (3) involves an integral over the prior for Σ under the graph G. This integral will typically be very sensitive to different choices of the prior, which is a general phenomenon in model-selection problems (Berger & Pericchi, 2001; Jones et al., 2005). Unlike in estimation problems, this sensitivity does not diminish as more data are collected, making an intelligent choice of $p(\Sigma \mid G)$ for each decomposable graph G the main difficulty in graphical-model selection. In particular, neither improper priors nor vague proper priors can be used.

In all but the smallest of problems, $p(\Sigma \mid G)$ must be a conjugate hyper-inverse Wishart prior; otherwise it will not be possible to make use of (4) for computing marginal likelihoods. Other priors will require approximating the integrals in (3), and in such a large model space, the need to do so repeatedly will usually pose an insurmountable obstacle. We accept that this practical requirement tethers us to a very restricted class of models, and we seek priors that are as well-behaved as possible under this constraint.

Well-behaved is somewhat difficult to judge, since distributions over the space of constrained covariance matrices are not very accessible by intuition. But since every graph implies a complete set of univariate conditional-independence relationships, graphical-model selection can be viewed as simultaneously performing variable selection for all of these conditionals. It is typically easier to assess priors for graphically constrained covariance matrices by studying the properties of the priors they induce on all implied conditional regression models. This point will be developed in § 4.

2·2. *A conventional proper prior*

The most popular choice of prior for use in Gaussian graphical models is $\Sigma \sim \mathrm{HIW}_G(\delta, \tau I)$, where the scale matrix $D = \tau I$ is proportional to the identity matrix. We call this the conventional proper prior, following Berger & Pericchi (2001). Examples of these or similar priors being used to compute marginal likelihoods can be found in Giudici (1996), Giudici & Green (1999), Dobra et al. (2004), Jones et al. (2005), Atay-Kayis & Massam (2005) and Carvalho & West (2007), among others. The scale parameter τ must be chosen to match the expected scale of the data; if τ is too large, the prior for Σ under each graph will wash out the likelihood, and there will be no basis for discriminating among competing graphs.

The now-standard notation of the conventional proper prior can be confusing: a single scale matrix D may be used to specify $p(\Sigma \mid G)$ for all graphs, but this scale matrix means different things under each graph, since different graphs imply different configurations of free elements in Σ. Under G, only the (i, j) elements of D for which the edge $(i, j) \in G$ are relevant in determining the distribution of Σ; other, nonfree, elements of Σ must be filled in using the deterministic completion operation described in Massam & Neher (1998); see also Dawid & Lauritzen (1993) and Carvalho et al. (2007). Hence the notation $\Sigma \sim \mathrm{HIW}_G(b, D)$ must be taken as convenient shorthand for the statement that Σ depends upon the free elements of D implied by the graph G.

2·3. *The hyper-inverse Wishart g-prior*

Given the practical need for conjugacy, we suggest another possible form of the hyper-inverse Wishart distribution, one where D involves the cross-product matrix: $(\Sigma \mid G) \sim \mathrm{HIW}_G(\delta, gX'X)$, where g is some suitably small fraction such as $1/n$.

We call this the hyper-inverse Wishart g-prior by analogy with Zellner's g-prior in linear regression (Zellner, 1986), and we recommend it as an alternative to the conventional $\mathrm{HIW}(\delta, \tau I)$ for use in graphical model-selection. The similarity to the g-prior in regression is more than superficial, since this prior will be shown to induce a g-like prior for the univariate conditional regression models implied by G. This, along with other theoretical and methodological properties, will be examined in §§ 4 and 5.

3. Fractional Bayes factors for Gaussian graphical models

3·1. *Motivation for the hyper-inverse Wishart g-prior*

Direct use of the hyper-inverse Wishart g-prior for selecting graphs is incoherent, since it involves a double use of the data. We now show, however, that the prior arises very naturally through the use of fractional Bayes factors, and that reusing the data can be avoided.

Fractional Bayes factors were proposed by O'Hagan (1995) as a default Bayesian model-selection technique for use when prior information is weak. The idea is to train a noninformative prior for each model using a small fractional power g of the likelihood function. This is done simultaneously for all models being considered, converting all noninformative priors into proper priors that are then used to select a model with the remainder of the likelihood.

Choose $g \in (0, 1)$ and let $p_N(\Sigma \mid G)$ be a noninformative, typically improper, prior for Σ under a decomposable graph G. The fractional Bayes factor for graphs G_1 and G_2 is then $\mathrm{FBF}_g(G_1, G_2) = Q_g(X \mid G_1)/Q_g(X \mid G_2)$, where

$$Q_g(X \mid G) = \frac{\int p_N(\Sigma \mid G) p(X \mid \Sigma, G) \, \mathrm{d}\Sigma}{\int p_N(\Sigma \mid G) p(X \mid \Sigma, G)^g \, \mathrm{d}\Sigma}. \tag{6}$$

Then $p^*(\Sigma \mid G) \propto p_N(\Sigma \mid G) \, p(X \mid \Sigma, G)^g$ is called the implied fractional prior, where the constant of proportionality is the integral of the given expression, and $p(X \mid \Sigma, G)^{1-g}$ is called the implied fractional likelihood.

Equation (6) clearly depends upon the choice of a noninformative prior for Σ. The obvious choice given our need for conjugacy is to simply define, for $\Sigma \in M^+(G)$,

$$p_N(\Sigma \mid G) \propto \frac{\prod_{C \in \mathcal{C}} |\Sigma_C|^{-|C|}}{\prod_{S \in \mathcal{S}} |\Sigma_S|^{-|S|}}, \tag{7}$$

an improper prior that makes use of the same factorization over cliques and separators, and is defined with respect to the same measure, as the hyper-inverse Wishart distribution. Interestingly, this prior has clique marginals that correspond to one of many versions of Fisher's fiducial prior (Sun & Berger, 2007), so called because it yields Fisher's fiducial distribution for marginal variances. These clique marginals, $p(\Sigma_C) \propto |\Sigma_C|^{-|C|}$, also have the same form as the priors that yield exact frequentist matching for means and variances (Geisser & Cornfield, 1963) when used for covariance estimation. We make no such claims regarding frequentist matching for the prior in (7), and use it only because it is a usefully conjugate generalization of a familiar prior to the space of covariance matrices with graphical structure.

We can now state the main result of this section:

THEOREM 1. *The hyper-inverse Wishart g-prior* $(\Sigma \mid G) \sim \text{HIW}_G(gn, gX'X)$ *is the implied fractional prior for* Σ *corresponding to the prior in* (7), *where* $0 < g < 1$ *is the fraction of the likelihood used for training and* $n > \max_{c \in \mathcal{C}} |C|$ *is the sample size. The fractional marginal likelihood is*

$$Q_g(X \mid G) = (2\pi)^{-np/2} \frac{h(G, gn, gX'X)}{h(G, n, X'X)}, \tag{8}$$

with $h(G, b, D)$ *defined as in* (5).

Proof. This follows immediately from the conjugacy of the hyper-inverse Wishart prior with the normal likelihood, and from (4). □

We emphasize that this procedure does not merely specify a single prior distribution, but rather a whole cohort of objective prior distributions for all $\Sigma_G \in M^+(G)$. One necessary condition for the marginal likelihoods to be defined is that n be greater than the maximum clique size for all graphs under consideration, which we will assume for the remainder of this paper.

3·2. *Choice of g*

An obvious issue with the use of this methodology is the choice of g. If the hyper-inverse Wishart g-prior were used as a real prior, it would be possible to place a hyperprior on g, and not tie ourselves down to a specific value; indeed, Liang et al. (2008) recommend exactly this approach towards g-priors in linear-model selection.

Yet when interpreting the hyper-inverse Wishart g-prior in terms of fractional Bayes factors, it is no longer possible to put a prior on g. This is because g is not a model parameter about which there is information in the likelihood, but rather the fractional power of the likelihood itself used for training the noninformative prior in (7). This fraction must be chosen outright in order for the fractional marginal likelihoods in (8) to be well-defined.

Several criteria help to guide this choice. First, there is an established tradition of using minimal training sample sizes to calibrate default Bayes factors; see, for example, O'Hagan (1995) and Berger & Pericchi (1996, 2001). A minimal training sample is the smallest sample size needed

to convert all improper priors such as (7) into proper priors. The intuition is that as much of the data as possible should be held back to choose between models. It is easy to see from (1) and (2) that the minimal training sample size is 1, suggesting that g be $1/n$.

Second, it is clear that g must be $O(1/n)$ in order for the implied fractional prior to correspond asymptotically to the gold standard of a carefully elicited subjective prior distribution. If g decreases too slowly as a function of n, the implied fractional prior will asymptotically overwhelm the likelihood; if it decreases too fast, the prior will become arbitrarily diffuse. Neither behaviour could possibly result from the choices of a careful elicitee making intelligent decisions about each $p(\Sigma \mid G)$. We find the hypothetical careful elicitation a useful ideal to keep in mind, even if dimensionality makes this ideal impossible to attain.

Finally, the implied fractional prior for Σ should have heavy tails. Choosing $gn = 1$ implies that the vector x is marginally Cauchy, and that each $p(\Sigma \mid G)$ is heavy-tailed without being too vague. This choice dovetails with the advice given by Liang et al. (2008), who themselves generalize the recommendations of Jeffreys (1961) and Zellner & Siow (1980).

As a default choice, we recommend setting $g = 1/n$, though this is not a hard rule, and other choices that decay like $1/n$ may be reasonable and can be judged by the reasonableness of their effect on the implied fractional prior. Robustness to these choices should be considered, just as it should be in subjective analyses.

4. PROPERTIES OF THE HYPER-INVERSE WISHART g-PRIOR

4·1. *Information consistency*

The consistency of fractional Bayes factors as $n \to \infty$ is a well-known result from O'Hagan (1995). We instead consider a second notion of consistency, often called information consistency or finite-sample consistency, that describes how a Bayes factor behaves for fixed n with respect to a test statistic that would be used to perform a classical test of significance on the same problem.

The canonical example of an information-inconsistent procedure is model selection in linear regression using fixed-g versions of Zellner's g-prior (Zellner & Siow, 1980; Liang et al., 2008). Imagine testing a specific regression model M_A having k possible covariates against the null model M_0 having only an intercept term. If the usual F statistic for testing M_A against M_0 goes to infinity for fixed n and $k < n - 1$, the evidence against M_0 is overwhelming, and one would expect the Bayes factor $\mathrm{BF}(M_A : M_0)$ to diverge.

But under the standard g-prior, this Bayes factor instead converges to the fixed constant $(1 + g)^{(n-k-1)/2}$. This gives an intrinsic limitation, one that is not shared by the F statistic, to how strongly the Bayes factor may support the bigger model. Such behaviour is intuitively unappealing, since $\mathrm{pr}(F > C \mid M_0) \to 0$ as $C \to \infty$.

A natural question is whether the fractional Bayes factors defined above exhibit a similar information paradox. We show that they do not, in two related senses.

4·2. *Tests against the null graph*

Let G_0 denote the null graph having no edges, and let G_A denote the graph to be compared with the null. The Bayes factor for comparing these two models is

$$
\mathrm{BF}(G_0 : G_A) = K \; \frac{\prod_{j=1}^{p} \left| \frac{g}{2} X_j' X_j \right|^{\frac{gn}{2}}}{\prod_{j=1}^{p} \left| \frac{1}{2} X_j' X_j \right|^{\frac{n}{2}}} \; \frac{\prod_{S \in \mathcal{S}} \left| \frac{g}{2} X_S' X_S \right|^{\frac{gn+|S|-1}{2}}}{\prod_{C \in \mathcal{C}} \left| \frac{g}{2} X_C' X_C \right|^{\frac{gn+|C|-1}{2}}} \times \frac{\prod_{C \in \mathcal{C}} \left| \frac{1}{2} X_C' X_C \right|^{\frac{n+|C|-1}{2}}}{\prod_{S \in \mathcal{S}} \left| \frac{1}{2} X_S' X_S \right|^{\frac{n+|S|-1}{2}}}, \quad (9)
$$

where g is fixed, $\mathcal{C}$ and $\mathcal{S}$ are the cliques and separators of G_A and the leading term K is

$$K = \left\{ \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{gn}{2}\right)} \right\}^p \frac{\prod_{S\in\mathcal{S}} \Gamma_{|S|}\left(\frac{n+|S|-1}{2}\right)}{\prod_{C\in\mathcal{C}} \Gamma_{|C|}\left(\frac{n+|C|-1}{2}\right)} \frac{\prod_{C\in\mathcal{C}} \Gamma_{|C|}\left(\frac{gn+|C|-1}{2}\right)}{\prod_{S\in\mathcal{S}} \Gamma_{|S|}\left(\frac{gn+|S|-1}{2}\right)}.$$

It remains to define a suitable test statistic b as a basis for assessing information consistency. Following Lauritzen (1996), let $\hat{\Omega}_0$ be the maximum likelihood estimate for the precision matrix under G_0, and let $\hat{\Omega}_A$ be the maximum likelihood estimate under G_A. Then there is a nested sequence $G_0 \subset \cdots \subset G_d = G_A$ of decomposable graphs that differ only by a single edge. Let e_i denote the edge in G_i but not in G_{i-1}, and let C_i be the (unique) clique of G_i containing e_i. Then we have the following proposition, proven in Lauritzen (1996).

PROPOSITION 1. *The test of significance for G_A against the null graph G_0 can be performed by rejecting G_0 for sufficiently small values of $b = |\hat{\Omega}_0|/|\hat{\Omega}_A|$. Under G_0, b is distributed as the product of independent beta random variables $B_1 \cdots B_d$, with $B_i \sim \mathrm{BE}\{(n-|C_i|)/2, 1/2\}$.*

This defines the relevant test statistic b. We now give a precise statement of information consistency for the fractional Bayes factors in Theorem 1.

THEOREM 2. *Let G_A be a decomposable graph having cliques $\mathcal{C}$, let G_0 be the null graph, and let $FBF_g(G_A : G_0)$ be the fractional Bayes factor, given data X, corresponding to the noninformative prior in (7). For any $0 < g < 1$, $FBF_g(G_0 : G_A) \to 0$ as $b \to 0$.*

Proof. The Bayes factor in (9) simplifies to

$$K \left(\frac{1}{g}\right)^{(S_1-gnp)/2} \left(\frac{1}{2}\right)^{(n-gn)(S_2-p)/2} \prod_{j=1}^p |X_j'X_j|^{-(n-gn)/2} \frac{\prod_{C\in\mathcal{C}} |X_C'X_C|^{(n-gn)/2}}{\prod_{S\in\mathcal{S}} |X_S'X_S|^{(n-gn)/2}},$$

where the exponent terms S_1 and S_2 are

$$S_1 = \sum_{C\in\mathcal{C}} |C|\,(gn+|C|-1) - \sum_{S\in\mathcal{S}} |S|\,(gn+|S|-1),$$

$$S_2 = \sum_{C\in\mathcal{C}} |C| - \sum_{S\in\mathcal{S}} |S|.$$

Now apply the formula of Lauritzen (1996) for the determinant of $\hat{\Omega}_G$, which will exist due to the previously stated restriction that $n > \max_{C\in\mathcal{C}} |C|$:

$$|\hat{\Omega}_G| = n^p \frac{\prod_{S\in\mathcal{S}} |X_S'X_S|}{\prod_{C\in\mathcal{C}} |X_C'X_C|}.$$

This gives

$$\mathrm{BF}(G_0 : G_A) = C \left(\frac{|\hat{\Omega}_0|}{|\hat{\Omega}_A|}\right)^{(n-gn)/2},$$

where C is a fixed, finite term involving g, p, n and the structure of the graph G_A. The proof of information consistency now follows immediately by plugging the test statistic b into the above equation, and noticing that for fixed $0 < g < 1$, $\mathrm{BF}(G_0 : G_A) \to 0$ as $b \to 0$. □

4·3. *Tests for an implied conditional regression model*

Information consistency is important in a second sense: nonzero entries in a precision matrix imply a set of nonzero conditional regression coefficients for each element of x upon the other

elements, and Bayes factors for model selection in Gaussian graphical models perform variable selection on all of these implied regressions simultaneously. We find that observing the behaviour of these implied conditional regression models is a useful window on the behaviour of $p(\Sigma \mid G)$, which is far harder to understand intuitively.

The following lemma provides a characterization of the implied conditionals.

LEMMA 1. *Suppose $(x \mid \Sigma) \sim N(0, \Sigma)$ and $\Sigma \sim \text{HIW}_G(b, D)$ for some decomposable graph G. Suppose $x = (z, y)'$, where z is a scalar and the vertices are numbered following the perfect vertex elimination scheme of G, with z as the first vertex. Let Σ and D be partitioned as*

$$\Sigma = \begin{pmatrix} \Sigma_{zz} & \Sigma_{zy} \\ \Sigma_{yz} & \Sigma_{yy} \end{pmatrix}, \quad D = \begin{pmatrix} D_{zz} & D_{zy} \\ D_{yz} & D_{yy} \end{pmatrix}.$$

Then

(i) $\Sigma_{z|y}^{-1} = (\Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz})^{-1} \sim \text{Ga}(\frac{b+k}{2}, \frac{D_{z|y}}{2})$,

(ii) $(\Sigma_{zy}\Sigma_{yy}^{-1} \mid \Sigma_{z|y}) \sim N(D_{zy}D_{yy}^{-1}, \Sigma_{z|y}D_{yy}^{-1})$,

with k representing the number of neighbours of z under G.

One proof of this lemma is presented in the Appendix. The same result could be obtained using the fact that the neighbourhood of z, as the first element in a perfect elimination scheme, is complete. Therefore, the marginal covariance of z and its neighbours is a standard inverse-Wishart law and the above result follows from Lauritzen (1996).

We now apply the lemma to give the following theorem, which is intended to be understood as if the implied fractional prior were a true prior, and the implied fractional likelihood were a true likelihood. Since the fractional prior has the exact functional form of a proper hyper-inverse Wishart prior, and the fractional likelihood has the functional form of a Gaussian likelihood, there is no mathematical ambiguity in operationally defining prior and likelihood this way, even if the interpretation is difficult from a pure subjectivist viewpoint.

THEOREM 3. *Let $X = (z\ Y)$ be the $n \times p$ matrix of observed data having rows $x_i = (z_i, y_i)$. Let G_A be a decomposable graph having cliques $\mathcal{C}$, and let M_A be the conditional regression model for z in terms of Y implied by the neighbours of z in G_A, and let M_0 be the null regression model for z. Assume further, that the vertices are listed in the perfect vertex elimination scheme of G_A, with z as the first vertex. Let F be the usual F-statistic for testing M_A against M_0, and let $BF_g(M_A : M_0)$ be the likelihood ratio $p(z \mid Y, M_A)/p(z \mid Y, M_0)$, where these marginals are defined under the fractional prior and the fractional likelihood in (6) for a fixed g. For any $0 < g < 1$, $F \to \infty$ implies $BF_g(M_A : M_0) \to \infty$.*

Proof. Let Y_z denote the columns of the matrix X corresponding to the neighbours of z under M_A. Applying Lemma 1 and recalling the global assumption that $n > \max_{C \in \mathcal{C}} |C|$, the hyper-inverse Wishart g-prior gives the following regression relationship

$$z = Y_z f + \epsilon, \quad \epsilon \sim N(0, \phi^{-1}I) \tag{10}$$

$$(f \mid \phi, M_A) \sim N\{\hat{f}, (g\phi)^{-1}(Y_z'Y_z)^{-1}\} \tag{11}$$

$$(\phi \mid M_A) \sim \text{Ga}\left(\frac{gn+k}{2}, \frac{gr}{2}\right), \tag{12}$$

where ϕ is the conditional precision or $\Sigma_{z|Y_z}^{-1}$, I is the $n \times n$ identity matrix, $\hat{f}' = (Y_z'Y_z)^{-1}Y_z'z$ is the traditional least-squares estimate for f and $r = z'(I - P_{Y_z})z$ with P_{Y_z} denoting the

perpendicular projection matrix onto the column space of Y_z. Hence r is the residual sum of squares after regressing z upon Y_z.

Marginalizing over f and ϕ, taking care to use the fractional likelihood rather than the full likelihood, gives

$$P(z \mid Y, M_A) = (\pi)^{-n/2} g^{\frac{gn+2k}{2}} (1-g)^{n/2} \frac{\Gamma\left(\frac{n+gn+k}{2}\right)}{\Gamma\left(\frac{gn+k}{2}\right)} \, r^{-n/2}. \qquad (13)$$

Assuming $0 < g < 1$, the relevant Bayes factor can then be computed by recognizing the null model M_0 as a special case of (13) with $k = 0$ and $r = z'z$:

$$\mathrm{BF}(M_A : M_0) = C\left(1 - R^2_{M_A}\right)^{-n/2}, \qquad (14)$$

where C is a fixed term involving g, n and k, and where $R^2_{M_A}$ is the usual coefficient of determination for model M_A. As the F-statistic grows without bound, $R^2_{M_A} \to 1$, and the Bayes factor in (14) clearly diverges. □

The hierarchical model in (10), (11) and (12) is immediately recognizable as a modified form of Zellner's g-prior for the vector of conditional regression coefficients. Unlike the g-prior, however, this procedure avoids the information paradox.

4·4. *Remarks*

One possible source of alarm is that the priors in equations (11)–(12) are centred at their maximum-likelihood estimates. These priors, however, are fractional priors, not real priors. They are only used in conjunction with a diminished fractional power of the likelihood function, and so the fact that they are centred should not be viewed as an improper double-use of the data, at least in the way that double-use is normally understood. This notion can be formalized by solving a particular set of equations to yield the intrinsic prior, which is the real, non-likelihood-centred prior to which a default Bayes factor corresponds asymptotically; see Berger & Pericchi (2001) for further discussion.

Theorem 2 applies to any true decomposable graph, and so offers a universal guarantee of information consistency with respect to the b-statistic. Theorem 3, however, does not apply to all possible univariate conditional regression models; it applies only to those in which the variable z to be predicted comes first in a perfect vertex elimination scheme. There are multiple such schemes, meaning that the theorem will apply to multiple models for any given true graph. But without a more general distributional result analogous to Lemma 1, we cannot show information consistency of all univariate regression models. Nonetheless, the theorem does demonstrate information consistency with respect to the F-statistic for a useful, albeit restricted, subclass of these conditional sparse regression models. This proof-of-concept result, while not completely general, is still enough to demonstrate that the hyper-inverse Wishart g-prior has certain desirable properties that the conventional proper prior lacks, and that these properties relate to precisely the concept of information consistency familiar from linear models.

5. Fractional marginal likelihoods: a simulation study

We now study the behaviour of the fractional marginal likelihoods through simulations that compare models of differing complexity. The baseline for comparison will be the conventional alternative, the $\mathrm{HIW}_G(\delta, \tau I)$ prior. We hope to illustrate that the casual use of the conventional prior
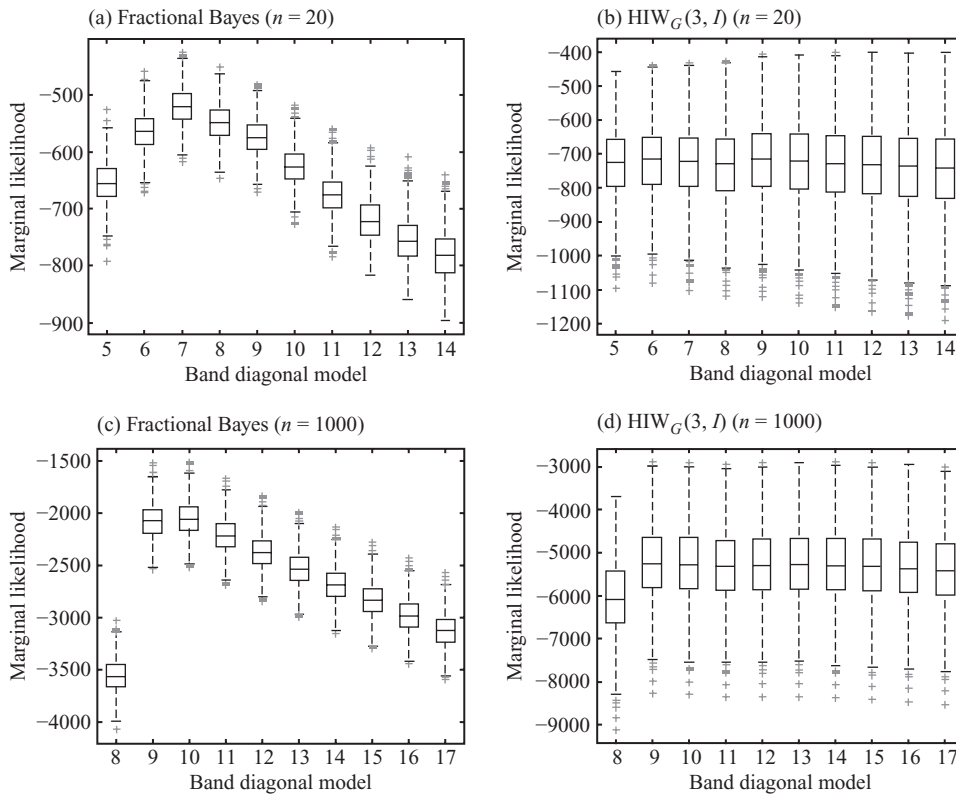
Fig. 1. Boxplots of realized marginal likelihoods in the simulation study. The x-axis is the bandwidth of the precision matrix; the y-axis is marginal loglikelihood.

undermines our ability to choose between competing models, and that the objective procedure presented so far is well-behaved.

Datasets with $n \in \{20, 50, 100, 500, 1000\}$ were simulated from the true model: each $x_i \sim N(0, \Sigma)$, where Σ was the 50-dimensional correlation matrix of a stationary Gaussian AR(10) process. This represents a Gaussian graphical model due to the band-diagonal form of the precision matrix. The number of nodes $p = 50$ is the length of each observation x_i, but n is the number of such independent draws observed from the true model.

An appropriate choice for the conventional proper prior is $\Sigma \sim \mathrm{HIW}_G(3, I)$, where our choice of $\delta = 3$ reflects the standard advice to give $p(\Sigma \mid G)$ a finite first moment (Jones et al., 2005). We have chosen $\tau = 1$, since Σ is known to be a correlation matrix. Typically, τ is very hard to specify without looking at the data, making this choice, if anything, overly favourable to the conventional prior.

For each sample size, we simulated 1000 datasets from the true model and computed marginal likelihoods for 21 different candidate graphs corresponding to band-diagonal precision matrices of bandwidth 0 through 20, with the true model having bandwidth 10. Figure 1 gives the frequency distributions of marginal loglikelihoods for each candidate model, which show substantially better separation under the fractional Bayes approach. Results are given for both the smallest and largest sample sizes, although the same pattern emerged for all sample sizes.

For each dataset, we also recorded the model with the highest marginal likelihood. This gives a Monte Carlo estimate for the frequency distribution of the preferred band-diagonal size under
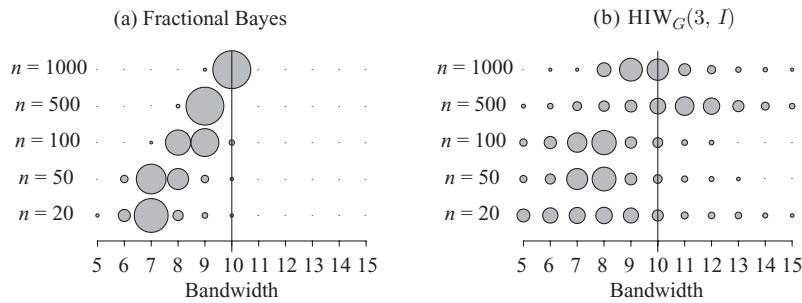
Fig. 2. The empirical distribution of the chosen bandwidth under repeated sampling from the true model in the simulation study. The area of each circle represents the fraction out of 1000 independent datasets in which each model had the highest marginal likelihood under the given prior.

the two priors, as in Fig. 2. As n grows, the fractional Bayes factors favour the true model with increasing accuracy, whereas the results from the $\mathrm{HIW}_G(3, I)$ are highly erratic.

It is clear that, unlike the conventional prior, our objective procedure prefers sparse models in the absence of enough data to justify extra edges. Yet as the sample size increases, the fractional approach favours more complex models, and eventually chooses the true one almost every time. The conventional prior does not exhibit this tendency nearly as strongly, displaying an unintuitively high level of variation in the choice of model.

One might imagine that the conventional prior, by shrinking the covariance structure towards the identity matrix with its strong pattern of off-diagonal zeros, would yield systematically smaller models. This expectation is not confirmed by our simulation study, indicating that intuitions about shrinkage gleaned from covariance estimation do not necessarily apply to model selection. This can be understood in terms of the Ockham's-razor property of Bayesian marginal likelihoods (Jefferys & Berger, 1992). The conventional prior spreads its mass out quite broadly, which is an advantage in estimation problems but crippling in model selection due to the lack of predictions sharp enough to yield any posterior separation of models.

The results from this simulation are consistent with the theoretical results developed so far and strengthen our claim that the fractional approach is indeed an appropriate default procedure for Gaussian-graphical-model selection. The conventional prior induces a set of ridge-regression priors on the complete conditional models considered in § 4·3, as can be shown through a straightforward application of Lemma 1. The problems with ridge-regression priors for variable selection are well understood (Zellner & Siow, 1980; Liang et al., 2008), and give some intuition as to why the conventional $\mathrm{HIW}_G(\delta, \tau I)$ prior is suboptimal for model selection.

6. Priors over graphs

There are two simple approaches to assigning prior probabilities to graphs themselves. The first is to give every graph the same prior probability κ^{-1}, where κ is the number of decomposable graphs on p nodes. As Giudici & Green (1999) note, this prior is quite heavily concentrated on graphs of middling size due to combinatorial explosion.

The second alternative, rapidly becoming the standard, is to imagine a sequence of edge inclusions as having a binomial distribution with success probability r (Dobra et al., 2004; Jones et al., 2005). This yields priors of the form

$$p(G) \propto r^k (1 - r)^{m-k} \tag{15}$$

for decomposable graphs, with nondecomposable graphs given prior probability 0 by construction. Recall that k is the number of edges and $m = p(p-1)/2$ is the maximum number of possible edges. The constant of proportionality is hard to compute since it is typically unknown how many decomposable graphs there are of a given size, but this constant is the same for all models and can thus be ignored.

If the expected fraction of included edges is known quite precisely, this framework may be attractive. Yet often this fraction is not known, making an arbitrary choice of r seem heavy-handed. Instead, we recommend treating r as a model parameter to be estimated from the data rather than as a fixed tuning constant. This shrinks the graph size to a data-determined value of r, which is estimated from the prevailing edge-inclusion fraction of models with favourable marginal likelihoods.

One possibility is to estimate r by empirical-Bayes methods. George & Foster (2000) give an expectation-maximization procedure that can be used to compute a maximum-likelihood estimate $\hat{r}$ in linear regression models, but note that it can prove computationally intractable in high dimensions, making it even less suited to graphical model spaces. This issue is also discussed in a technical report by Scott and Berger available from Duke University.

A second possibility is to place a prior on r, which turns out to be the easier and more attractive option. Assuming the conjugate beta prior $r \sim \text{BE}(a, b)$ allows an explicit marginalization,

$$p(G) \propto \int_0^1 p(G \mid r)p(r)\, \mathrm{d}r \propto \frac{\beta(a+k, b+m-k)}{\beta(a, b)},$$

where $\beta(\cdot, \cdot)$ is the beta function. For the default choice of $a = b = 1$, implying a uniform prior on r, this becomes

$$p(G) \propto \frac{(k)!(m-k)!}{(m+1)(m!)} = \frac{1}{m+1}\binom{m}{k}^{-1}. \tag{16}$$

Scott & Berger (2006) call these multiplicity correction priors in the context of testing exchangeable normal means: if the true k is held fixed while m grows, the expected number of false positives will remain constant. In our context, this will mean an automatic penalty for testing irrelevant edges.

We demonstrate through simulation that the same effect holds here. Beginning with a correlation matrix corresponding to the ten-node graph from Fig. 1 of Scott & Carvalho (2008), we added progressively more noise nodes: nodes unconnected both from the true graph and from each other, but that lead to combinatorial explosion in the number of edges that must be tested for inclusion. The numbers of noise nodes chosen were 5, 15 and 40, which in addition to the 10 connected nodes in the true graph imply 60, 300 and 1225 separate hypothesis tests, respectively. In all cases, the number of true hypotheses remained fixed at 22, one for each edge in the 10-node graph.

Three sets of tests were performed under each of three different priors on models. These results are summarized in Table 1. Here, Fully Bayes uses the model probabilities from (16), while Oracle Bayes involves plugging the true value of r into (15) to compute prior graph probabilities. All marginal likelihoods were computed using fractional Bayes factors, with $g = 1/n$.

The fully Bayesian multiplicity-correction prior in (16) squelches false positives very effectively. This phenomenon is well understood in linear models, but ours is the first demonstration of such an automatic-correction effect in graphical models. The difference between corrected and uncorrected versions is substantial; in the 50-node example, giving all models the same prior probability yields 40 false positives with inclusion probability greater than 50%, whereas

Table 1. *Estimated inclusion probabilities for specific edges* (*left-most column*) *as the number of unconnected noise nodes grows from* 5 *to* 15 *to* 40. *The last row of the table shows the number of falsely positive flags,* FPs, *which are other, nonenumerated edges that have edge-inclusion probability greater than* 0·5. *All probabilities were calculated using five million iterations of the feature-inclusion stochastic search algorithm* (Scott & Carvalho, 2008)

| | Number of noise nodes | | | | | | | | |
| | No correction | | | Oracle Bayes | | | Fully Bayes | | |
Edge	5	15	40	5	15	40	5	15	40
(1,6)	99	99	99	99	99	99	99	99	99
(3,4)	16	2	0	1	0	3	1	1	1
(3,6)	99	99	99	99	99	97	99	99	99
(3,8)	18	0	0	10	2	0	6	1	0
(3,9)	31	3	0	0	0	0	0	0	0
(4,6)	99	99	99	99	99	99	99	99	99
(4,9)	99	99	99	99	99	99	99	99	99
(5,6)	16	5	0	22	21	18	24	23	22
(5,9)	36	14	31	31	30	34	31	31	33
(6,7)	58	76	74	14	2	0	8	3	0
(6,9)	22	3	0	1	0	0	1	0	0
(8,9)	43	4	7	14	2	0	7	1	0
(9,10)	89	95	99	71	69	76	60	60	69
FPs:	6	11	40	1	1	1	1	1	0

imposing the multiplicity-correction prior yields none. Importantly, this approach does not depend upon the choice of an arbitrary hyperparameter r, although if subjective inputs are required, they can be accommodated through a different beta prior while still retaining closed-form answers.

Other differences from standard priors also emerge. Consider, for example, edges $(5, 9)$ and $(6, 7)$. If all models are given equal prior probabilities, adding more noise edges makes $(6, 7)$ appear stronger and $(5, 9)$ appear slightly weaker. Yet the opposite happens using the multiplicity-correction priors: the addition of more noise nodes makes $(6, 7)$ disappear entirely and yet retain $(5, 9)$ at close to its original strength.

This behaviour suggests that (16) does not merely shrink edge-inclusion probabilities to 0 uniformly as k remains fixed and m grows. Rather, it differentially rewards edges that participate in more parsimonious models, suggesting a fundamental difference from (15) in the way mass is apportioned across model space.

7. EXAMPLE: MUTUAL FUNDS

A crucial input for many dynamic portfolio-selection problems is the estimated covariance matrix Σ for a collection of asset returns. Naïve procedures can often yield unstable estimates of Σ, but as Carvalho & West (2007) show, graphical models offer a potent tool for shrinkage.

We now illustrate the proposed methodology on a set of 86 monthly returns for 59 mutual funds from a variety of different sectors: 13 U.S. bond bunds, 30 U.S. stock funds, seven balanced funds investing in both U.S. stocks and bonds and nine international stock funds. The monthly returns were split into a 60-month training set and a 26-month prediction set. We study the predictive performance of each of the different graphical-modelling regimes considered here, incorporating different combinations of priors for Σ and priors over graphs.

The last of these regimes involved expanding the prior specification in (16) to include several different parameters for the prior inclusion probabilities for edges in each of the different fund sectors. This reflects our prior understanding that different patterns of covariation might prevail in different asset classes, and that it may be suboptimal to force edges participating in these different covariational patterns to shrink towards some common inclusion probability. Instead, we allow block-by-block patterns of possibly differing sparseness to emerge. Global shrinkage, of course, could still be accomplished via a hierarchical prior; here we have simply chosen to give each inclusion probability an independent uniform prior.

For each prior specification, we used feature-inclusion stochastic search (Scott & Carvalho, 2008) on the 60-month training set to search for good models and compute posterior means $\hat{\Sigma}$. These means were then used to predict observations in the 26-month validation set, where in each month the 56 observed returns predicted the remaining three missing returns, which are known in reality. We repeated this process for all of the 32 509 combinations of three unobserved columns, and then computed the total squared-error in imputing the missing values.

When all models were given equal prior probability, the fractional approach yielded a 5·5 % improvement in total mean squared error under the top models discovered, to 642·9 from 679·6. Applying the multiplicity-correction priors in (16) reduced the squared error to 623·6, while the block-by-block prior yielded a further reduction to 554·7. These may seem like small improvements, but are actually quite substantial given the highly unpredictable nature of financial markets.

These results demonstrate that each of the methodological advancements considered here can improve predictive accuracy in model selection problems. It is especially interesting to note the strength of the sector-specific multiplicity correction, suggesting that the ease with which our prior specification accommodates structural information can be a real advantage in practice.

APPENDIX

Proof of Lemma 1. Let $\Omega = \Sigma^{-1} = \Phi'\Phi$ be partitioned as

$$\begin{pmatrix} \Omega_{zz} & \Omega_{zy} \\ \Omega_{yz} & \Omega_{yy} \end{pmatrix} = \begin{pmatrix} \Phi'_{zz} & 0 \\ \Phi_{yz} & \Phi'_{yy} \end{pmatrix} \begin{pmatrix} \Phi_{zz} & \Phi_{zy} \\ 0 & \Phi_{yy} \end{pmatrix}.$$

Recall that $(\Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz})^{-1} = \Sigma_{z|y}^{-1} = \Omega_{zz} = \Phi'_{zz}\Phi_{zz}$ and $\Sigma_{zy}\Sigma_{yy}^{-1} = -\Phi_{zz}^{-1}\Phi_{zy} = \Gamma_{z|y}$. Theorem 1 of Roverato (2000) (see also Paulsen et al., 1989; Wermuth, 1980) shows that if G is decomposable and the vertices are listed in a perfect vertex elimination scheme, the pattern of zeros of Ω are preserved in Φ. Now, if $(\Sigma \mid G) \sim \text{HIW}_G(b, D)$, properties of the Cholesky decomposition of the hyper-inverse Wishart as defined in Theorem 3 of Roverato (2000) (see also equation (27) of Atay-Kayis & Massam, 2005) allow us to write $\Psi = \Phi T^{-1}$, where $D^{-1} = T'T$ with

$$\Psi_{zz}^2 \sim \text{GA}\left(\frac{b+k}{2}, \frac{1}{2}\right) \quad \text{and} \quad \Psi_{zy_i} \sim N(0, 1), \tag{A1}$$

for all y_i in the neighbourhood of z. For the non-neighbours of z in y, $\Phi_{zy_i} = \Psi_{zy_i} = 0$.

It is now straightforward to see from ([A1](#)) that

$$
\Omega_{zz} = \Phi'_{zz}\Phi_{zz} = \Phi^2_{zz} = (\Psi_{zz}T_{zz})^2 \sim \mathrm{GA}\left(\frac{b+k}{2}, \frac{T^{-2}_{zz}}{2}\right),
$$

so that

$$
\Sigma^{-1}_{z|y} \sim \mathrm{GA}\left(\frac{b+k}{2}, \frac{D_{z|y}}{2}\right),
$$

which proves part (i) of the lemma. Turning the focus to the form of $\Gamma_{z|y} = \Sigma_{zy}\Sigma^{-1}_{yy} = -\Phi^{-1}_{zz}\Phi_{zy}$ and writing it as a function of Ψ and T, we get

$$
\gamma_i = -\left(\frac{T_{y_iz}}{T_{zz}} + \frac{1}{\Psi_{zz}T_{zz}}\sum_{j=1}^{i}\Psi_{zy_j}T_{y_jy_i}\right).
$$

Given Φ_{zz}, this is just a linear combination of independent standard normals, so that

$$
(\Gamma|\Phi_{zz}) \sim N\left(-T^{-1}_{zz}T_{zy}, \frac{1}{\Phi^2_{zz}}T'_{yy}T_{yy}\right),
$$

$$
\left(\Sigma_{zy}\Sigma^{-1}_{yy}|\Sigma^{-1}_{z|y}\right) \sim N\left(D_{zy}D^{-1}_{yy}, \Sigma_{z|y}D^{-1}_{yy}\right),
$$

proving part (ii) of the lemma. $\qquad\square$

References

ATAY-KAYIS, A. & MASSAM, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* **92**, 317–35.

BERGER, J. O. & PERICCHI, L. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Am. Statist. Assoc.* **91**, 109–22.

BERGER, J. O. & PERICCHI, L. (2001). Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection*, pp. 135–207. Institute of Mathematical Statistics Lecture Notes – Monograph Series 38. Beachwood, OH: Institute of Mathematical Statistics.

CARVALHO, C., MASSAM, H. & WEST, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94**, 647–59.

CARVALHO, C. & WEST, M. (2007). Dynamic matrix-variate graphical models. *Bayesian Anal.* **2**, 69–96.

DAWID, A. P. & LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–317.

DOBRA, A., JONES, B., HANS, C., NEVINS, J. & WEST, M. (2004). Sparse graphical models for exploring gene expression data. *J. Mult. Anal.* **90**, 196–212.

GEISSER, S. & CORNFIELD, J. (1963). Posterior distributions for multivariate normal parameters. *J. R. Statist. Soc.* B **25**, 368–76.

GEORGE, E. I. & FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–47.

GIUDICI, P. (1996). Learning in graphical Gaussian models. In *Bayesian Statistics 5*, Ed. J. Berger, J. Bernardo, A. Dawid and A. Smith, pp. 621–8. New York: Oxford University Press.

GIUDICI, P. & GREEN, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785–801.

JEFFERYS, W. & BERGER, J. (1992). Ockham's razor and Bayesian analysis. *Am. Sci.* **80**, 64–72.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford: Oxford University Press.

JONES, B., CARVALHO, C., DOBRA, A., HANS, C., CARTER, C. & WEST, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.* **20**, 388–400.

LAURITZEN, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

LETAC, G. & MASSAM, H. (2007). Wishart distributions on decomposable graphs. *Ann. Statist.* **35**, 1278–1323.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. & BERGER, J. (2008). Mixtures of *g*-priors for Bayesian variable selection. *J. Am. Statist. Assoc.* **103**, 410–23.

MASSAM, H. & NEHER, E. (1998). Estimation and testing for lattice conditional independence models on euclidean jordan algebras. *Ann. Statist.* **26**, 1051–82.

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

PAULSEN, V., POWER, S. & SMITH, R. (1989). Schur products and matrix completions. *J. Funct. Anal.* **85**, 151–78.

ROVERATO, A. (2000). Cholesky decomposition of a hyper-inverse Wishart matrix. *Biometrika* **87**, 99–112.

SCOTT, J. G. & BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plan. Infer.* **136**, 2144–62.

SCOTT, J. G. & CARVALHO, C. M. (2009). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comp. Graph. Stat.* **17**, 790–808.

SUN, D. & BERGER, J. O. (2009). Objective priors for the multivariate normal model. In *Bayesian Statistics 8*, Ed. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West. Oxford: Oxford University Press.

WERMUTH, N. (1980). Linear recursive equations, covariance selection and path analysis. *J. Am. Statist. Assoc.* **75**, 963–72.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 233–43. Amsterdam: North-Holland.

ZELLNER, A. & SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia*. Valencia: University Press.