Biometrika (1999), **86,** 4, *pp.* 785–801
© 1999 Biometrika Trust
Printed in Great Britain

Decomposable graphical Gaussian model determination

By PAOLO GIUDICI

Department of Economics and Quantitative Methods, University of Pavia,
Via San Felice n. 5, 27100 Pavia, Italy
pgiudici@eco.unipv.it

AND PETER J. GREEN

Department of Mathematics, University of Bristol, Bristol BS8 1TW, U.K.
p.j.green@bristol.ac.uk

Summary

We propose a methodology for Bayesian model determination in decomposable graphical Gaussian models. To achieve this aim we consider a hyper inverse Wishart prior distribution on the concentration matrix for each given graph. To ensure compatibility across models, such prior distributions are obtained by marginalisation from the prior conditional on the complete graph. We explore alternative structures for the hyperparameters of the latter, and their consequences for the model. Model determination is carried out by implementing a reversible jump Markov chain Monte Carlo sampler. In particular, the dimension-changing move we propose involves adding or dropping an edge from the graph. We characterise the set of moves which preserve the decomposability of the graph, giving a fast algorithm for maintaining the junction tree representation of the graph at each sweep. As state variable, we use the incomplete variance-covariance matrix, containing only the elements for which the corresponding element of the inverse is nonzero. This allows all computations to be performed locally, at the clique level, which is a clear advantage for the analysis of large and complex datasets. Finally, the statistical and computational performance of the procedure is illustrated by mean of both artificial and real datasets.

Some key words: Bayesian model selection; Hyper-Markov distribution; Inverse Wishart distribution; Junction tree; Reversible jump Markov chain Monte Carlo.

1. Bayesian graphical models

1·1. *Introduction*

This paper is concerned with model determination for a random vector X, and in particular with inference about its conditional independence graph g. We focus on the case where g is decomposable and X is multivariate Gaussian, although some of our formulation and analysis applies much more generally.

Our research is related to work on Bayesian model determination for directed graphical models and probabilistic expert systems; see for instance Geiger & Heckerman (1994) and Spiegelhalter et al. (1993). For undirected graphical Gaussian models the main reference is Dawid & Lauritzen (1993), who introduced hyper-Markov priors allowing local computations in Bayesian model determination. Applications of such priors include those of Madigan & Raftery (1994) and Madigan & York (1995), who analyse discrete graphical models according to Occam's razor and using Markov chain Monte Carlo over the graph space. Finally, Dellaportas & Forster (1999) use reversible jump Markov chain Monte Carlo for model determination over undirected discrete graphical models.

All the above papers consider only non-hierarchical and, typically, conjugate priors, with the advantage of allowing the derivation of closed-form expressions of the posterior probabilities. Quantitative learning is, however, limited to quantities having an explicit posterior distribution. Our motivation is that richer information is often to be extracted from the data and, furthermore, that more flexible priors may be better suited for this purpose. Our main contributions are therefore the introduction of a hierarchical Bayesian graphical Gaussian model and the design of a reversible jump Markov chain Monte Carlo algorithm to perform both structural and quantitative learning in a graphical Gaussian model by means of local computations.

After some preliminaries on graphical models we present our proposed Bayesian graphical models in § 2. In § 3 we provide a complete characterisation of the one-edge-at-a-time incremental changes to a graph that preserve its decomposability, and then use this to define our reversible jump Markov chain Monte Carlo scheme for performing Bayesian model determination in graphical models. In § 4 we examine the statistical performance of the proposed methodology, as well as the performance of the Markov chain Monte Carlo sampler.

1·2. *Background on graphical Gaussian models*

Here we briefly review the theory of graphical models relevant for our work following the exposition in Dawid & Lauritzen (1993), to which we refer readers for further details and explanations. For an introduction to graphical models, see Lauritzen (1996).

Let $g=(V, E)$ be an undirected graph, where the vertex set V has p elements. A graph or subgraph is complete if all its vertices are joined by an edge. A complete subgraph that is not contained within another complete subgraph is called a clique. An ordering of the cliques of an undirected graph, $(C_1, \ldots, C_n)$, is said to be perfect if the vertices of each clique C_i also contained in any previous clique $C_1, \ldots, C_{i-1}$ are all members of one previous clique; that is, for $i=2, 3, \ldots, n$,

$$S_i = C_i \cap \bigcup_{j=1}^{i-1} C_j \subseteq C_h$$

for some $h=h(i)\in\{1, 2, \ldots, i-1\}$. The sets S_i are called separators. If an undirected graph admits a perfect ordering it is said to be decomposable. A pair (A, B) of subsets of the vertex set V of an undirected graph g is said to form a decomposition of g if (i) $V=A\cup B$, (ii) $A\cap B$ is complete, and (iii) $A\cap B$ separates A from B.

With each vertex $v\in V$ associate a random variable X_v taking values in a sample space $\mathcal{X}_v$. For $A\subseteq V$ we let $X_A=(X_v)_{v\in A}$ indicate the collection of random variables $\{X_v : v\in A\}$ with values of $\mathcal{X}_A = \times_{v\in A}\mathcal{X}_v$. To ease the notation, let $X=X_V$. By a probability distribution over $A\subseteq V$ we mean a joint distribution for X_A over $\mathcal{X}_A$. If P is the distribution over $U\subseteq V$ and $A, B\subseteq U$, then P_A will denote the marginal distribution of X_A and $P_{B|A}$ the conditional distribution of X_B given $X_A=x_A$. A distribution P over V is Markov with respect to g if, for any decomposition (A, B) of g, $X_A \perp\!\!\!\perp X_B | X_{A\cap B}$, where $\perp\!\!\!\perp$ means 'is independent of', using the notation introduced by Dawid (1979). A graphical model is a family of probability distributions which are Markov with respect to a graph. Henceforth

P is a graphical model with respect to some graph g, which is not fixed, and will be implicit in the notation. We assume that g is decomposable.

A graphical Gaussian model, also known as a covariance selection model (Dempster, 1972), is defined by a p-dimensional multivariate Gaussian distribution, with expected value μ and covariance matrix Σ:

$$P = N_p(\mu, \Sigma).$$

In a graphical Gaussian model, the mean parameter μ is typically set to zero; we shall assume so, and therefore the data we analyse will be expressed as deviations from the sample mean. The matrix Σ is positive definite and such that P is Markov over g. In a graphical Gaussian model, the global, local and pairwise Markov properties are identical (Lauritzen, 1996, pp. 36, 129). The last property is particularly useful for interpretability. Define $K = \Sigma^{-1}$ to be the precision matrix of X. The pairwise Markov property specifies that

$$X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}} \Leftrightarrow k_{ij} = 0. \tag{1}$$

Thus, g constrains Σ by imposing a pattern of zeros on to K. The effect of this constraint on Σ can be better specified using the notation of matrix completion with respect to a graph; see for instance Roverato & Whittaker (1998). Let Γ be a $p \times p$ matrix such that $\gamma_{ij} = \Sigma_{ij}$ if and only if $(i, j) \in E$, and is otherwise unspecified. A completion of Γ with respect to g is a positive definite matrix obtained from Γ by fixing its unspecified elements so that its inverse D satisfies $d_{ij} = 0$, for all $(i, j) \dotplus E$; see Dempster (1972) and Grone et al. (1984) for a proof of the uniqueness and existence of such a matrix. It turns out that Σ is the completion of Γ with respect to g.

Conditionally on a graph, g, say, consider a sample x of size n from P. Let $S = xx'$ denote the observed sum-of-products matrix. For a subset of vertices $A \subset V$, let Σ_A denote the variance-covariance matrix of the variables in X_A, and define S_A similarly. When the graph is decomposable the likelihood of the graphical Gaussian model specified by P is

$$p(x \mid \Sigma, g) = \frac{\prod_{C \in \mathscr{C}} p(x_C \mid \Sigma_C)}{\prod_{S \in \mathscr{S}} p(x_S \mid \Sigma_S)},$$

where $\mathscr{C}$ and $\mathscr{S}$ respectively denote the sets of cliques and separators,

$$p(x_C \mid \Sigma_C) = (2\pi)^{-n|C|/2} \det(\Sigma_C)^{-n/2} \exp[-\tfrac{1}{2} \operatorname{tr}\{S_C(\Sigma_C)^{-1}\}], \tag{2}$$

and similarly for $p(x_S \mid \Sigma_S)$, with $|.|$ denoting cardinality.

1·3. *Prior distributions for graphical Gaussian models*

Two kinds of uncertainty may affect a graphical model: (a) uncertainty about the probability distributions P on X or about the quantities, θ, say, which parameterise such distributions; (b) uncertainty about the graphical structure g, describing the conditional independence relationships among the random variables considered. Our objective is to deal with (a) and (b) simultaneously in a Bayesian fashion. To this end, we must formulate a prior distribution on θ and g. Concerning the latter, we shall assume throughout, for simplicity, a uniform prior,

$$p(g) = d^{-1},$$

on the class of d decomposable graphs with vertex set V. Note that d is actually hard to compute. We can indeed estimate its value, using the algorithm outlined in § 3, but d is not needed in our approach. Note also that the above prior distribution is simple, but not neutral, being concentrated around models that are 'medium-sized' in terms of their numbers of edges. Importance sampling ideas allow us in principle to reweight results to accommodate any other desired prior on g.

For the parameters, a very general class of priors are the hyper-Markov laws introduced in Dawid & Lauritzen (1993). Let θ be a quantity parameterising a graphical model P, for a given undirected decomposable graph $g=(V, E)$. Similarly, for $A, B \subseteq V$ let θ_A parameterise the marginal distribution P_A, Markov with respect to the subgraph g_A, and let $\theta_{B|A}$ parameterise the conditional distribution $P_{B|A}$, with $P_{A\cup B}$ Markov with respect to $g_{A\cup B}$. A hyper-Markov law is then defined by a property which mimics the global Markov property, at the parameter level: a law $\mathscr{L}$ on θ is hyper-Markov over g if, for any decomposition (A, B) of g, $\theta_A \perp\!\!\!\perp \theta_B \,|\, \theta_{A\cap B}$. In order to construct such laws, Dawid & Lauritzen (1993) define two distributions $\mathscr{M}$ over θ_A and $\mathscr{N}$ over θ_B as hyperconsistent if they induce the same prior law over $\theta_{A\cap B}$. Given the family of sets $\mathscr{C}$ and $\mathscr{S}$, and a collection of pairwise hyperconsistent distributions $(\mathscr{L}_C, C \in \mathscr{C})$, they show that there exists a unique hyper-Markov law $\mathscr{L}$ over g, with the assigned marginals, concentrated on the set of parameters such that P is Markov with respect to g.

A hyper-Markov prior for a graphical Gaussian model is a prior on Σ. We can take, as dominating measure, the product of Lebesgue measures on the elements of the incomplete variance-covariance matrix Γ. Such elements are subject only to symmetry and positive definiteness of the submatrices $\{\Gamma_C = \Sigma_C, C \in \mathscr{C}\}$, as consistency restrictions over the corresponding marginal distributions are automatically satisfied. Let l_C and l_S be the densities of a generic clique and separator, with respect to the corresponding product of Lebesgue measures. A hyper-Markov law on Σ can then be obtained from the clique-specific marginal densities as

$$l(\Sigma) = \frac{\prod_{C\in\mathscr{C}} l_C(\Sigma_C)}{\prod_{S\in\mathscr{S}} l_S(\Sigma_S)}.$$

A natural choice for a prior distribution over each clique-specific covariance matrix, and, therefore, for each separator, is to take a prior conjugate to the likelihood in (2), assuming Σ_C to be distributed as inverse Wishart with parameters α and Φ^C. We employ the parameterisation in Dawid & Lauritzen (1993) which implies that, for $\alpha > 2$, $E(\Sigma_C) = (\alpha - 2)^{-1}\Phi^C$. The resulting distribution for Σ has been named the hyper inverse Wishart by Dawid & Lauritzen (1993), denoted by $\textsc{hiw}_g(\alpha, \Phi)$.

This construction involves many hyperparameters, namely the precision parameter α, common to all cliques, and one prior matrix, Φ^C, for each clique; the separator specific priors can be obtained by marginalisation. Furthermore, in order to satisfy hyperconsistency of the clique-specific priors, it is necessary and sufficient that, for each pair of cliques, A, B, say, with intersection $S = A \cap B$, the submatrices of Φ^A and Φ^B corresponding to the elements in S coincide. This requirement is rather stringent, particularly when large graphs are considered.

A further complication in the practical specification of a hyper-Markov law, which is indeed common to all Bayesian model comparison problems, is that of compatibility. The simplest case involves comparison between two graphs, g and g', say. Let Σ and Σ' be the corresponding precision matrices and let $\mathscr{L}$ and $\mathscr{L}'$ be two hyper-Markov laws on them. It is quite natural to require that $\mathscr{L}(\Sigma^A) = \mathscr{L}'(\Sigma^A)$, for any clique A common to both g and g'. This notation of compatibility is the same as in Dawid & Lauritzen (1993) and corresponds to requiring the two prior distributions to be consistent on the common marginals. Given the difficulty of the above specification tasks, especially in large graphs, it becomes desirable to have a 'semi-automatic' method for assigning compatible hyper-Markov distributions. One possibility, suggested in Dawid & Lauritzen (1993), is to consider an 'embedding' graph, g^*, and derive the required marginal distributions, for each $g \subset g^*$, by marginalisation from those of g^*. Note that this use of an embedding graph is not without critics; see Cowell (1996) for an alternative approach.

In the remainder of this work we shall take g^* as the complete graph, for which Σ is not constrained, and assign an $\text{IW}(\alpha, \Phi)$ distribution to Σ. Marginalisation from this law will then imply that, for each $C \in \mathscr{C}$, $\mathscr{L}_C(\Sigma_C) = \text{IW}(\alpha, \Phi^C)$, with $\Phi^C = \Phi_C$, the submatrix of Φ corresponding to the variables indexed by C. The graph g thus determines which collection of submatrices of Φ are to be taken to form a hyper-Markov law on Σ with respect to g. Although the specification task is now reduced, there remains the issue of specifying the matrix Φ. One possibility is to consider an assignment that is a default or uninformative, yet leads to a proper prior on Σ. However, it is difficult to understand what a default setting really means in the present context. A different strategy is to add one further layer of uncertainty, and consider α and Φ as random quantities, regulated by a few hyperparameters. This leads us to consider a hierarchical hyper-Markov law.

2. The proposed models

2·1. *A non-hierarchical model*

Consider first the case of fixed hyperparameters. The model we assume specifies that

$$X \mid \Sigma, g \sim N_p(0, \Sigma), \quad \Sigma \mid g \sim \text{HIW}_g(\alpha, \Phi), \quad p(g) = d^{-1},$$

where α is a fixed positive quantity, Φ is a fixed $p \times p$ symmetric positive definite matrix, whose elements satisfy $\Phi_C = \Phi^C$, for all $C \in \mathscr{C}$, and d is the number of decomposable graphs on the vertex set V.

The complete prior specification of the dispersion matrix Φ involves setting $p(p+1)/2$ prior quantities and satisfying the positive definiteness condition, a clearly difficult task, so that one would typically try to simplify the structure of Φ. A reasonable default specification for Φ is to consider an intraclass correlation structure:

$$\Phi = \tau\{\rho J + (1-\rho) I\}, \tag{3}$$

where J is the $p \times p$ matrix of 1's and I the identity matrix of order p. Note that Φ is positive definite if and only if $\tau > 0$ and $\rho \in (-1/(p-1), 1)$.

However, the above parameterisation exhibits some drawbacks: for instance, it may not be reasonable to assume a priori a common correlation between each pair of random variables. An assumption of common covariance is inevitably asymmetric about zero correlation, since the prior correlation is constrained below by $-1/(p-1)$, and this may lead, particularly in large graphs, to an asymmetric evolution of the association signs. Concerning τ, the assumption of a common prior scale is clearly reasonable if the random variables are standardised or on a similar scale.

2·2. *A hierarchical model*

Given the above difficulties of prior specification, it is desirable to devise a more automatic, yet flexible, method of assigning a prior distribution. A natural choice is to let

α and Φ become random quantities, to be assigned a prior distribution. A reasonable assumption is that α, Φ and g are mutually independent.

First consider α. Note that α expresses the relative weight of the prior. A reasonable prior for α is a gamma distribution, with mean f, variance fs, and density

$$\pi(\alpha) \propto \alpha^{(f/s)-1} e^{-\alpha/s},$$

where $f > 0$ and $s > 0$ are to be fixed, A rationale for choosing them is that $\pi(\alpha)$ be as uninformative as possible; sensitivity to the choice will be discussed in § 4.

Now consider Φ. The representation adopted for Φ determines the set of random quantities to which are to be assigned a prior distribution. We shall consider the two situations of Φ unstructured and Φ with an intra-class correlation structure.

Unstructured Φ. Here we assign a prior on $p(p+1)/2$ elements, consisting of p variances and $p(p-1)/2$ covariances. To ease the calculations, one can take a conjugate prior distribution. Note that the prior on Σ can be interpreted as a likelihood for Φ, suggesting that a conjugate prior for Φ is a Wishart distribution with fixed hyperparameters $d > 0$ and T positive definite. Note that, although still difficult, this prior specification is considerably easier than the specification of a hyper inverse Wishart law in the non-hierarchical case. For instance, since Φ is already a prior opinion, a reasonable requirement on the second-stage prior on Φ is that it be not very informative, taking $d = 1$ and embodying a belief of a very simple structure, such as $T = \mathrm{diag}(\tau_{11}, \ldots, \tau_{pp})$, possibly with $\tau_{ii} \equiv \tau$.

Intraclass Φ. In the intraclass case, as remarked in § 2·1, all partial correlation coefficients are assumed to be equal a priori. A prior on the random elements (τ, ρ) which characterise the intraclass correlation structure can be obtained by restriction from the $W(d, T)$ prior on the unstructured Φ, as follows:

$$\begin{aligned}\pi(\tau, \rho) &\propto \pi_\Phi[\tau\{\rho J + (1-\rho)I\}] \\ &\propto [\tau^p(1-\rho)^{p-1}\{1+\rho(p-1)\}]^{(d-2)/2} \exp\left\{-\frac{1}{2}\tau\left(\sum_{i=1}^{p} t_{ii} + \rho \sum_{i \neq j} t_{ij}\right)\right\}. \qquad (4)\end{aligned}$$

Note that the above kernel does not factorise as $\pi(\tau) \times \pi(\rho)$. However, if $\sum_{i \neq j} t_{ij} = 0$, for example, if $t_{ij} = 0$, for $i \neq j$, τ and ρ become independent, as in the following proposition.

PROPOSITION. *Let Φ be a random symmetric matrix of form* (3) *with τ and ρ distributed as* (4), *with $\sum_{i \neq j} t_{ij} = 0$. Suppose that $d > 2 - 2/p$ and let $t_0 = \sum_{i=1}^{p} t_{ii}$. Then τ and ρ are independent random variables*;

$$\tau \sim \mathrm{Ga}\left\{\frac{p(d-2)+2}{2}, \frac{t_0}{2}\right\};$$

$$\rho = -\frac{1}{p-1} + \frac{p}{p-1}\gamma,$$

where

$$\gamma \sim \mathrm{Be}\left\{\frac{d}{2}, \frac{(p-1)(d-2)+2}{2}\right\}.$$

Thus, the prior on τ depends on two hyperparameters; the mean and variance are increasing in d and decreasing in t_0. On the other hand, the prior on ρ depends only on

d, and $E(\rho)$ is non-increasing in d. Note also that $E(\rho)>0$, for all p, and that, as $p\to\infty$, $E(\rho)\to\frac{1}{2}$. It follows that d should be fixed to regulate the prior on ρ, with t_0 adjusting its effect on the prior on τ.

3. Modifying graphs to preserve decomposability, and Markov chain Monte Carlo algorithms

3·1. *Incremental changes to decomposable graphs*

A key aspect of our work concerns proposing a change in the current graphical structure g, say, to a new structure, g'. Since we are considering only decomposable graphs, the proposed moves should consider only members of the latter class as candidate graphical structures.

It is well known, see for instance Frydenberg & Lauritzen (1989), that the space of all decomposable graphs can be traversed by adding and deleting single edges at a time. Such changes will form the basis for the sampling algorithm we introduce in § 3·2. Here we characterise in graph-theoretic terms those incremental changes to a graph's edge set that preserve decomposability, making particular use of a junction forest representation of the graph. While 'legal' deletion moves can be characterised using the standard result in Theorem 1, a new result, Theorem 2, is required to characterise 'legal' addition moves.

Theorem 1 (*Frydenberg & Lauritzen*, 1989). *Let g and g' be two undirected decomposable graphs, with the same vertex set V and with $E'\subseteq E$, with g having exactly one more edge than g'. Such an edge must then be contained in exactly one clique of g.*

A junction tree $\mathscr{T}$ representation of a connected undirected graph g is a graph whose vertex set is the set of cliques of g, and whose edge set is such that $\mathscr{T}$ is a tree and satisfies the junction property: for any two cliques C_i, $C_j\in\mathscr{C}$ and any clique C' on the unique path between them in $\mathscr{T}$, $C_i\cap C_j\subset C'$. The junction property is evidently necessary and sufficient for the existence of a perfect ordering, and hence for the decomposability of g. A junction forest representation of an undirected graph g is a collection of junction trees $\{\mathscr{T}_r\}$, each $\mathscr{T}_r$ corresponding to a collection $\mathscr{C}_r$ of cliques with $\mathscr{C}=\cup\mathscr{C}_r$ and $\mathscr{C}_r\cap\mathscr{C}_s=\emptyset$, for $r\neq s$. Finally, for each $v\in V$ let $[v]$ indicate the connectivity component of V, that is the set of all vertices which are connected to v.

Theorem 2. *Let $g=(V,E)$ be an undirected decomposable graph in which vertices a and b are not adjacent, and let g' denote the graph modified by the addition of edge (a,b). Then g' is decomposable if and only if either*

(i) $[a]\neq[b]$,

or

(ii) $[a]=[b]$ *and there exist $R, T\subset V$ such that $a\cup R$ and $b\cup T$ are cliques, and $S=R\cap T$ is a separator on the path between $a\cup R$ and $b\cup T$ in a junction forest representation of the graph g.*

The proof of Theorem 2 is given in the Appendix.

As a simple example consider the graph in Fig. 1, which is characterised by the cliques (a,b,f), (b,c,f), (c,d,f) and (d,e,f). The separators are (b,f), (c,f) and (d,f). By Theorem 1, the edges (b,f), (c,f) and (d,f) cannot be deleted. On the other hand, the pairs (a,e), (a,d) and (b,e) cannot be joined in g', because, for all such pairs, $R\cap T=\{f\}$ but $\{f\}$ is not a separator.

Remark. Theorems 1 and 2 can be employed to characterise completely the legitimate

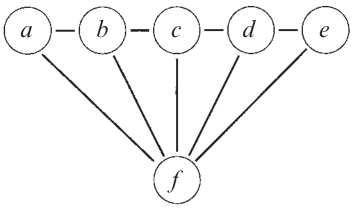

Fig. 1. Graph illustrating the reversibility condition.

incremental changes to the edge set of a decomposable graph. An alternative possibility is to reject illegitimate moves by running maximum cardinality search, see for instance Spiegelhalter et al. (1993), after each graphical update proposal, to check if the proposed graph g' is decomposable. However, while maximum cardinality search tests for decomposability by means of a global search through the whole of the junction forest, without building the new clique organisation, our method only requires searching through a section of the junction forest, corresponding to the shortest path between cliques containing a and b. Furthermore, it constructs the new junction forest so that the cliques are already constructed ready for use in probability calculations. Sometimes a and b will be adjacent so that the search will be very fast. In § 4 we present an empirical comparison between the two algorithms.

3·2. *Reversible jump Markov chain Monte Carlo design*

We now briefly summarise the main features of reversible jump Markov chain Monte Carlo methodology, which is particularly suited to problems where the dimension of the parameter space changes; see Green (1995) for further details. Let y denote a state variable. For instance, in our hierarchical Bayesian graphical Gaussian model, y is the complete set of unknowns $(g, \Sigma, \alpha, \Phi)$. Let $\pi(dy)$ be the target probability measure of interest, here the posterior distribution. When the current state is y we propose a move of type m that would take the chain to the destination y', with probability $q_m(y, dy')$. It is then accepted with probability given by

$$\alpha_m(y, y') = \min\left\{1, \frac{\pi(dy')q_m(y', dy)}{\pi(dy)q_m(y, dy')}\right\}, \tag{5}$$

which ensures that detailed balance is achieved within each move type.

For an 'ordinary' move type, that is, a move which does not change the dimension of the parameter vector, expression (5) reduces to the usual Metropolis–Hastings acceptance probability, using an ordinary ratio of densities with respect to a measure on the underlying fixed parameter subspace. For dimension-changing moves, Green (1995) shows that expression (5) can be interpreted as a ratio of Radon–Nikodym derivatives with respect to a suitably chosen common dominating measure. Suppose that a move from y to y' is proposed, with y' lying in a higher dimensional space. Then the method can be implemented by drawing a vector of continuous random variables u, independent of y, and setting $y' = y'(y, u)$, with $y'(.,.)$ an invertible deterministic function. Correspondingly, the reverse move can be achieved by the inverse transformation, in a deterministic fashion. Then expression (5) simplifies to

$$\alpha_m(y, y') = \min\left\{1, \frac{\pi(y')}{\pi(y)} \times \frac{r_m(y')}{r_m(y)q(u)} \times \left|\frac{\partial y'}{\partial(y, u)}\right|\right\}, \tag{6}$$

where $r_m(y)$ is the probability of a move of type m, evaluated at y, and $q(u)$ is the density function of u.

We now detail the reversible jump Markov chain Monte Carlo sampler we propose for the models specified in § 2. In the exposition we refer to the more general hierarchical model. An important issue in the design of the algorithm is the choice of the state variable. In our context, an important choice to be made is how to represent Σ. It would be too computationally expensive to consider the collection $(\Sigma_C, C \in \mathscr{C})$; for instance, a change in g would require changing most of the possibly overlapping clique-specific variances Σ_C. On the other hand, it seems that using the precision matrix K is a good choice; because of (1), adding (deleting) an edge requires one simply to draw (set to zero) an element of K previously set to zero (unconstrained). However, note that the hyper inverse Wishart model considered means that several time-consuming operations have to be performed: first K has to be inverted, to obtain Σ; secondly, the collection of submatrices Σ_C is to be extracted from Σ; finally, both the likelihood and the prior contribution to the Metropolis–Hastings acceptance ratio for g' require inversion of each Σ_C. Note also that the inversion from K to Σ prevents local computation of the ratio.

A more efficient representation for Σ is to consider as state variable the incomplete version of Σ, Γ. This has the advantage of avoiding the inversion of K into Σ, thus leading to local Metropolis–Hastings computations. Note that, since it will be important to draw inferences on functions of K or Σ, such as the partial correlation coefficients, we may want occasionally to complete Γ to obtain K and Σ. A related important result is contained in Dawid & Lauritzen (1993); it turns out that $\Sigma = K^{-1}$, with

$$K = \sum_{\mathscr{C}} (\Sigma_C^{-1})^{[0]} - \sum_{\mathscr{S}} (\Sigma_S^{-1})^{[0]},$$

where [0] means that the corresponding matrix is filled with zeros to match dimensions.

Thus, for our hierarchical Bayesian graphical model we shall consider a systematic scan over the following move types.

Move (a). Add or delete one edge from the graph g ensuring that the proposed graph g' is decomposable. Note that this move also involves making changes to Γ.

Move (b). Update the incomplete covariance matrix Γ and, correspondingly, Σ.

Move (c). Update the hyperparameter α.

Move (d). Update the hyperparameter Φ.

The only randomness in the above scan is the choice between adding and deleting an edge in Move (a). An update of $(g, \Sigma, \alpha, \Phi)$ is complete when all of the above move types are completed.

Updating g. Consider first Move (a), which is the only one involving a change in the dimensionality of the parameter space. To accomplish this move we draw randomly a pair of distinct vertices. If such pair, (i, j), say, is in E we propose deleting the edge (i, j); otherwise, if $(i, j) \notin E$, we propose adding (i, j) to the graph.

If (i, j) is proposed for insertion, the dimensionality of the parameter space increases by one; this is expressed by an extra free element of Σ. This requires specifying a new element of Γ, γ'_{ij}, and this is done by drawing a random variable u from a $N(0, \sigma_G^2)$ distribution, with σ_G a scale parameter to be properly chosen, and then letting $\gamma'_{ij} = u$. This is a blind proposal, which does not take into account the previous, constrained state of Σ_{ij}. As an alternative, with the extra computational cost of completing Γ, the proposal can be centred at the previous state, for example $\gamma'_{ij} = \Sigma_{ij} + u$. We prefer local computations and, therefore, we employ the former proposal.

Let R_a indicate the Metropolis–Hastings ratio when the proposed move consists of adding (i, j) to g, leading to g'. Such a ratio can be calculated as in (6). First note that the Jacobian of the transformation is equal to 1. This is not surprising, since we are making proposals on the natural scale. The proposed move can be seen as a change in the appropriate section of the junction tree, possibly after some permutations, as illustrated in the proof of Theorem 2. According to the proposed model, and if we adopt the proposal just described, based on the Γ representation, it turns out that the posterior ratio involves only the four subsets S, $S \cup i$, $S \cup j$ and $S \cup i \cup j$, abbreviated below as S, S_i, S_j and S_{ij}:

$$R_{\mathrm{post}} = \frac{\pi(y')}{\pi(y)} = \frac{h(\Sigma_S)h(\Sigma'_{S_{ij}})}{h(\Sigma_{S_i})h(\Sigma_{S_j})},$$

where each of the above four terms is obtained as the product of the prior and the likelihood of the appropriate submatrix of Σ. For instance, for S,

$$h(\Sigma_S) = \mathrm{IW}(\Sigma_S;\, \alpha, \Phi_S) \times N(x_S;\, \Sigma_S).$$

When $S = \emptyset$, $h(\Sigma_S) = 1$. Note that the requirement of positive definiteness of Σ constrains γ'_{ij}: if $\Sigma'_{S_{ij}}$ is not positive definite then $h(\Sigma'_{S_{ij}}) = 0$, so $R_{\mathrm{post}} = 0$ and the move is rejected.

Consider now the proposal ratio $r_m(y')/\{r_m(y)q(u)\}$. Since the graphs specified by y and y' differ in exactly one edge, $r_m(y)$ and $r_m(y')$ are simply the probabilities of choosing that edge for addition or deletion. Since all edges are chosen with equal probability,

$$r_m(y) = r_m(y') = 1 \bigg/ \binom{n}{2}.$$

Finally, when (i, j) is added, γ'_{ij} is drawn from a Gaussian distribution, with zero mean and standard deviation σ_G, so that

$$q(u) = \frac{1}{\sigma_G \sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{u^2}{\sigma_G^2}\right).$$

Thus the proposal ratio is $1/q(u)$. Putting together the different terms, we obtain that

$$R_a = R_{\mathrm{post}} \times q(u)^{-1}.$$

Note that the calculation of R_a involves at most four cliques.

So far we have considered a move which involves adding an edge to g. When (i, j) is proposed for deletion, we leave γ_{ij} unspecified; it is indeed of no use in the new model. We follow the reverse of the analysis above, and the acceptance ratio R_d is finally obtained as $R_d = 1/R_a$.

Updating Σ. Our strategy consists of perturbing each element of the corresponding incomplete matrix Γ with an independent Gaussian random walk proposal, centred around the current value. More formally, for all (i, j) such that $i = j$ or i and j are adjacent in the current graph g,

$$\gamma'_{ij} \sim N(\gamma_{ij}, \sigma_{ij}^2),$$

where the σ_{ij}'s are spread parameters, to be chosen.

We remark that a more complicated strategy could have been taken, for example updating only one clique-specific block of Σ at a time and exploiting the junction tree representation to construct Gibbs steps. However, the advantages of this do not seem to compensate for the increased complexity of the sampler and the extra computational effort.

We now calculate the acceptance probability for our proposed updating of Σ to a new covariance matrix, Σ', say, by perturbing its specified elements in Γ. As in the ordinary Metropolis–Hastings algorithm, such a probability is equal to $\min(1, R_\Sigma)$, where R_Σ indicates the acceptance ratio of the move and is the product of two terms, the posterior ratio R_{post} and the proposal ratio R_{prop}. The former can be calculated locally, through the junction forest of the graph:

$$R_{\text{post}} = \frac{p(\Sigma' \mid \alpha, \Phi, g)}{p(\Sigma \mid \alpha, \Phi, g)} \frac{p(x \mid \Sigma', g)}{p(x \mid \Sigma, g)},$$

that is, the ratio of two hyper inverse Wishart kernels. Note that, if any of the Σ_C, for $C \in \mathscr{C}$, is not positive definite, the move is rejected, as otherwise we would obtain a Σ' that is not positive definite. Finally, since the proposal distribution is symmetric, R_{prop} is equal to 1.

Updating α. We perturb α with a Gaussian random walk proposal, centred around the current value, that is $q(\alpha' \mid \alpha) = N(\alpha, \sigma_\alpha^2)$, where σ_α is to be appropriately chosen. Consequently, the proposal ratio is equal to 1. On the other hand, the posterior ratio is

$$R_{\text{post}} = \frac{p(\Sigma \mid \alpha', \Phi, g)}{p(\Sigma \mid \alpha, \Phi, g)} \frac{p(\alpha')}{p(\alpha)}.$$

Updating Φ. When Φ is unstructured, it will be updated similarly to Σ. That is, a proposal for Φ will be obtained by perturbing each element of Φ with a random walk proposal, that is $\phi'_{ij} = (\phi_{ij}, v_{ij})$, where the v_{ij}'s are to be suitably chosen. The acceptance probability of the move is $\min(1, R_\Phi)$, with $R_\Phi = R_{\text{post}} R_{\text{prop}}$, as usual. Given the symmetry of the adopted proposals, $R_{\text{prop}} = 1$. On the other hand

$$R_{\text{post}} = \frac{p(\Sigma \mid \alpha, \Phi', g)}{p(\Sigma \mid \alpha, \Phi, g)} \frac{p(\Phi')}{p(\Phi)}.$$

If Φ' is not positive definite, the proposed move is rejected. Note the generality of the above expression, which holds for all of the structures considered for Φ, because of the conditional derivation of the priors. Clearly, more complicated proposals for α and Φ can be considered but, in our experience, such changes do not materially affect the performance of the method.

4. Statistical performance of the methodology

Note first that the data x can be sufficiently summarised by the sample size n and the sample variance-covariance matrix $S = xx'$. We have considered three datasets, in order of increasing difficulty.

Example 1: *Fret's heads dataset* (*Whittaker*, 1990, p. 225). Here $p = 4$ and there are 64 possible graphs, of which 3 are not decomposable. This is a small but challenging dataset, since all variables appear highly correlated marginally, and there is no evident pattern in the sample precision matrix, resulting in a highly multimodal posterior distribution on the graphical structures.

Example 2: *Fowl bones dataset* (*Whittaker*, 1990, p. 266). Here $p = 6$ and there are 32 768 possible graphs, of which 80% are decomposable. This is a more complex problem, but less multimodal than the previous one.

Example 3: *An artificial dataset.* Here $p = 16$ and there are 2^{16} possible models, of which 45% are decomposable. Data are actually simulated from a non-decomposable model, namely a first-order Gaussian Markov process on a regular 4×4 spatial lattice. We have set equal to 0·2 all the partial correlations not constrained to zero by the graph. This dataset will illustrate, besides the process of learning the true simulated data, how a mixture of decomposable models can approximate the true non-decomposable model.

The analysis of the examples will be presented simultaneously, in terms of four aspects: prior settings; posterior distributions of main quantities of interest; sensitivity to prior specification; and performance of the Markov chain Monte Carlo sampler.

Prior setting. For all datasets we have considered both hierarchical and non-hierarchical models, with several hyperparameter specifications. In the paper we shall report results for only one such prior assessment, namely a hierarchical prior with an intraclass correlation structure for Φ, with $f = p + 1$, $s = 0{\cdot}1$, $T = I$ and $d = 2$. Concerning the parameter Σ, it is important to understand what such a prior specification corresponds to in terms of the prior expected partial correlation coefficients. This can be done by simulation from the assumed mixture of hyper inverse Wisharts prior. For instance, the empirical average of the output obtained from $n = 100\,000$ reversible jump sweeps after 10 000 burn-in, with $p = 4$, gives essentially an identity matrix.

Posterior distributions. Figure 2 reproduces the most plausible graphs, according to the posterior distribution of g, for Fret's data, obtained with a run of $n = 100\,000$ sweeps and $n = 10\,000$ of burn-in.

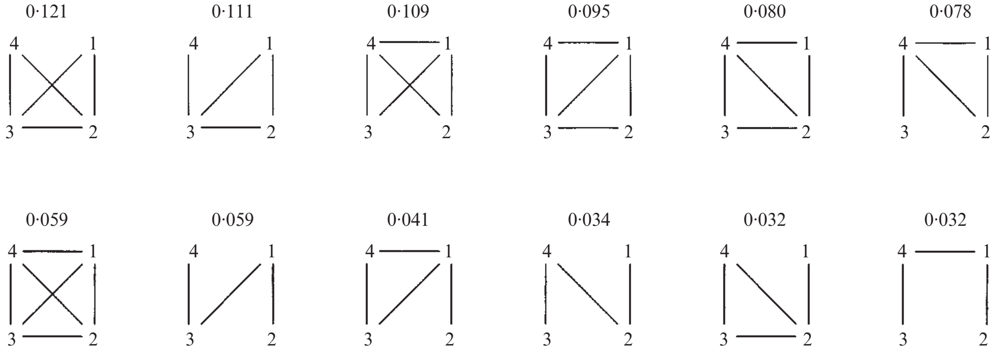

Fig. 2. Most probable graphs for Fret's dataset, together with the associated probabilities.

Note first that the posterior distribution of g is dispersed, as expected. For instance, the most probable graph receives only about 12% of the posterior probability and, in order to obtain 80% of the posterior probability, at least 10 structures have to be considered. The results are similar to those in Giudici (1996), who performed a non-informative Bayesian analysis on the same dataset using a non-hierarchical model.

It is often of interest to assess not only if an edge is present, but the strength of the association described by the edge itself. This can be done by looking at the posterior distribution of the partial correlation coefficients, which cannot be derived analytically, but can be easily obtained from the Markov chain Monte Carlo output. Figure 3 reproduces the posterior distributions of the partial correlation coefficients for Fret's data. Note that only the partial correlations between (1, 2) and between (3, 4) have relatively small

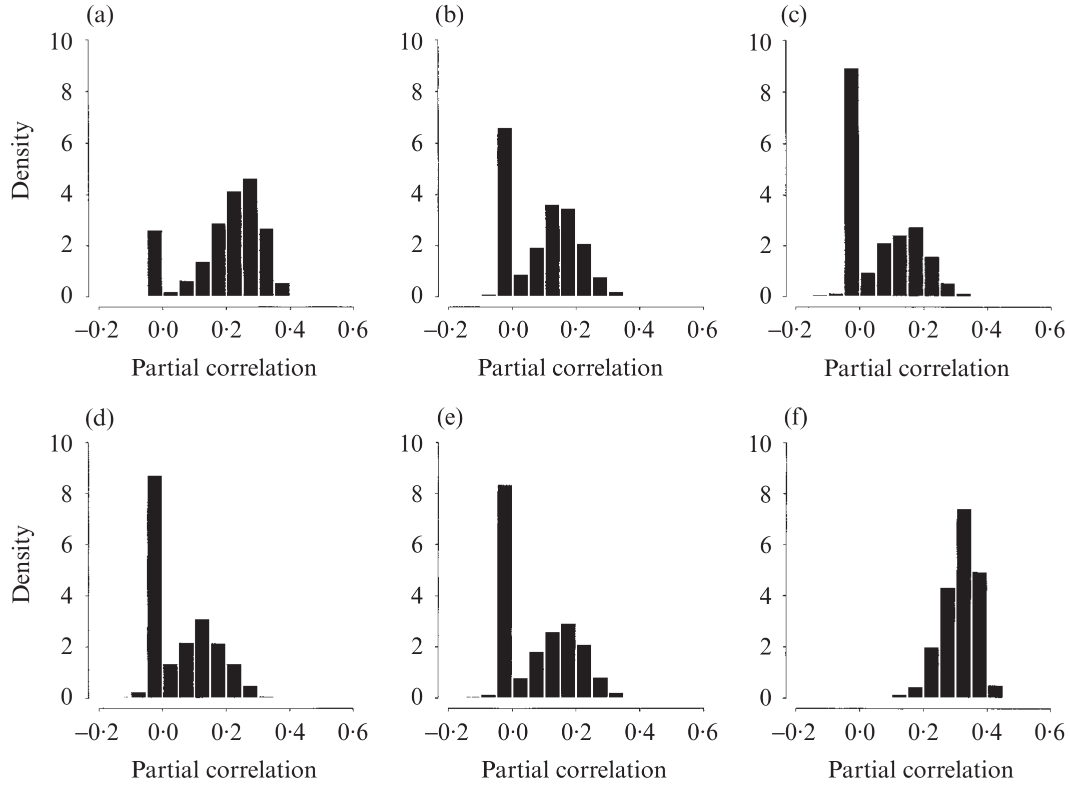

Fig. 3. Posterior distributions of the partial correlation coefficients for Fret's dataset. (a) For variables (1, 2), (b) for (1, 3), (c) for (1, 4), (d) for (2, 3), (e) for (2, 4), (f) for (3, 4).

posterior probabilities around zero, which supports strongly the presence of such two edges.

Consider now analysis of the Fowl bones dataset, obtained with a run of $n = 100\,000$ sweeps and $n = 10\,000$ of burn-in. Compared to Fret's data, the posterior distribution of g turns out to be more concentrated, with just two graphs accounting for about 33% of the posterior probability, with the others less important. The results can be compared with the deviance-based analysis in Whittaker (1990, p. 267): while the graph selected by Whittaker contains 8 edges and is not decomposable, our most probable graph contains all edges in Whittaker's as well as two more edges, (3, 6) and (4, 5), thus breaking the cycle involving (1, 3, 4, 6) in Whittaker's graph.

Results for the spatial lattice dataset were obtained with a run of 100 000 sweeps after 10 000 burn-in. Our aim here is to show that, although the model space considered is very large and does not contain the true model, Markov chain Monte Carlo learning can still give sensible answers. Figure 4 reproduces the cumulative average of two sampled partial correlations. Figure 4(a) plots a partial correlation which is equal to 0·2 in the true model, whereas Fig. 4(b) plots a partial correlation which is zero in the true model. Note how well the simulation acknowledges the difference between the two correlations, although the simulation length is certainly short compared to the number of candidate models. This difference is typical; for brevity we have presented only two representative edges.

We have also evaluated the number of edges which are misclassified by the simulation, using a simple binary discriminant function which signals edge presence if the proportion

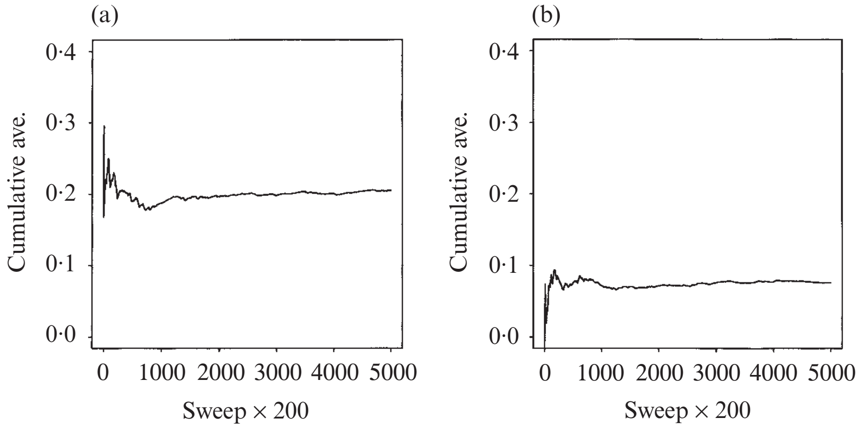

Fig. 4. Cumulative average for the simulated spatial lattice data. (a) Cumulative average of a present partial correlation, (b) cumulative average of an absent partial correlation.

of times that edge is in the simulated model is greater than 0·5 and edge absence otherwise. The total number of misclassified edges is equal to 13, corresponding to a rate of 11% and similar to the number of edges required to make the graph decomposable. Our results seem to be maintained with analysis of an even larger 5 × 5 spatial lattice, although longer runs are needed to achieve the same stability of the output. For instance, the number of misclassifications obtained with a run of $n = 1\,000\,000$ iterations is 42, corresponding to a rate of 14%.

However, Bayesian structural learning is a very difficult task for this problem and for large datasets in general. Our results show that this is indeed possible with Markov chain Monte Carlo, although slow and requiring a considerable amount of diagnostic checking of the validity of the results.

Sensitivity to the prior. Fret's dataset is useful for evaluating the sensitivity of results to the prior distribution, because its highly correlated structure leads to a multimodal posterior distribution over the graph space. Let g_0 denote the graph with the maximum posterior probability; each such graph will be described by a list of binary indicators for edge presence, with edges ordered in lexicographic order of the two vertices. Finally, let 'n.edges' indicate the number of edges of g.

When a non-hierarchical prior is used, the posterior over graphs depends on both α and ρ, particularly on the latter; see Table 1. The support for more complex graphs is lower for larger ρ. As expected, the influence of the prior grows with α.

Table 1. *Sensitivity of structural learning, with respect to the prior, using Fret's data, for a non-hierarchical model*

	$\alpha = p + 1$			$\alpha = 2p$		
	$\rho = -0{\cdot}3$	$\rho = 0$	$\rho = 0{\cdot}9$	$\rho = -0{\cdot}3$	$\rho = 0$	$\rho = 0{\cdot}9$
g_0	100011	111001	100011	110001	111011	110001
$p(g_0 \mid x)$	0·1165	0·1267	0·2645	0·1415	0·1171	0·2142
$E(\text{n.edges} \mid x)$	3·63	4·16	3·13	3·52	4·21	3·04

Inference on partial correlation coefficients is quite robust, which seems to be an advantage of model averaging. Table 2 shows such robustness of the inference about the partial correlation coefficient between X_1 and X_2, with a non-hierarchical prior. From previous analyses of Fret's data, it is quite difficult to draw such a conclusion. Similar results can be obtained with a hierarchical prior.

Table 2. *Sensitivity of model averaged inference on a partial correlation coefficient with respect to the prior, for a non-hierarchical model, using Fret's data*

	$\alpha = p+1$			$\alpha = 2p$		
	$\rho = -0{\cdot}3$	$\rho = 0$	$\rho = 0{\cdot}9$	$\rho = -0{\cdot}3$	$\rho = 0$	$\rho = 0{\cdot}9$
$E(\rho_{12} \mid x)$	0·204	0·207	0·241	0·204	0·208	0·239

The hierarchical prior has less impact on the posterior over graphs; compare Table 3 with Table 1. On the other hand, the hierarchical model seems to select models with more edges. These results are confirmed in the analysis of the two other datasets.

Table 3. *Sensitivity of structural learning with respect to the prior, for a hierarchical model, using Fret's data*

	$f = p+1$			$f = 2p$		
	$d = 2$	$d = p$	$d = 2p$	$d = 2$	$d = p$	$d = 2p$
g_0	110111	111011	110111	111011	111011	111011
$p(g_0 \mid x)$	0·1383	0·1304	0·1422	0·1317	0·1412	0·1517
$E(\text{n.edges} \mid x)$	4·41	4·40	4·46	4·45	4·54	4·53

Performance of the Markov chain Monte Carlo sampler. The correctness of our program was partially validated for all of our models by using it to simulate from the prior distribution. Our algorithms gave very good agreement between the exact and simulated prior marginals of certain marginal distributions that could be calculated analytically.

The spread parameters of the proposal distribution must be chosen so as to ensure satisfactory mixing of the chain. After a number of pilot runs, we took $\sigma_G = 0{\cdot}5n/p$, $\sigma_{ij} = 0{\cdot}1/p$, $\sigma_\alpha = 1{\cdot}0$ and $v_{ij} = 1{\cdot}0/p$. Concerning the proposal on g, centring the proposal at Σ_{ij} leads to better performances than a 'blind' proposal centred at 0. However, the completion of Γ is computationally expensive. In a typical run with $p = 10$ and a hierarchical model this takes about 40% of the CPU time. This percentage increases with p and the number of edges present in the graph.

Table 4 reports the accept/reject rates for g, Γ_g, Φ and α for our three simulations, along with total computation times on a SPARC4 workstation.

Table 4. *Performance of the Markov chain Monte Carlo samplers: rejection fractions and computation times, in minutes and seconds for* 100 000 *sweeps*

Move type	Fret's	Fowl bones	Spatial lattice
g	0·022	0·001	0·002
Σ	0·573	0·016	0·379
Φ	0·566	0·595	0·642
α	0·518	0·476	0·577
Time	2:16	3:57	22:03

We compared the computational times of our method with those for maximum cardinality search, using a small trial with the two methods in parallel, timing just the graph manipulation part of the procedures, i.e. testing for decomposability and constructing the new cliques and separators. For uniformly random decomposable graphs on 6, 10 and 20 vertices, the times to run maximum cardinality search for these graph operations were respectively 0·63, 1·21 and 3·49 times those with our method. Although this evidence is limited, the comparisons are in one sense biased in favour of maximum cardinality search, since in applications with data many graph moves are rejected, and our method then gains an additional advantage through rejecting at an early stage.

The most challenging aspect of the simulation is mixing over g, which can be monitored through a summary measure of g, such as the number of edges present, which describes the graph complexity. For all datasets trace and autocorrelation plots of the number of edges show good performance. However, the number of iterations required to achieve such stability increases with p; for both Fret's and the Fowl bones datasets 100 000 iterations are sufficient, whereas for the spatial lattice model a 10 times longer run is needed. We also assessed performance of the dimension-jumping move more formally. For each of the three datasets we evaluated the Gelman–Rubin convergence diagnostic for the trace of the number of edges in the simulated graphs, according to the iterated graphical approach suggested by Brooks & Gelman (1998). Our results indicate that each of the simulated parallel chains is close to the target distribution.

We finally remark that the sampler's performance is affected little by the choice of hyperparameter values. However, mixing is sensitive to the strength of the observed iterations effects between the variables in the graph; the higher this is, the slower the convergence.

Acknowledgement

P. Giudici acknowledges support from the European Science Foundation and the Italian National Research Council. Computing facilities were supported in part by the UK Engineering and Physical Sciences Research Council Stochastic Modelling in Science and Technology initiative. We also acknowledge comments and suggestions from the editor and the referees.

Appendix

Proof of Theorem 2

The case (i) where the vertices a and b are in different connected components is rather trivial; we can simply add a clique $a \cup b$ to the junction forest, linked to arbitrary existing cliques $a \cup R$ and $b \cup T$. The junction property clearly continues to hold for the modified junction forest.

Turning to the connected case (ii), we first prove the necessity of the condition. Suppose for a contradiction that there are no R, T such that (ii) holds. Let $a \cup R$, $b \cup T$ be the cliques containing a and b that have the shortest connecting path in the junction forest among all such cliques. By assumption $R \cap T$ is not a separator; it may be empty. The connected component containing a and b will remain connected when any vertices in $R \cap T$ are deleted, along with all incident edges. Let $v_0 = r, v_1, \ldots, v_q, v_{q+1} = t$ for some $q \geqslant 0$ be the shortest path in g from an element of $R \backslash T$ to one of $T \backslash R$ avoiding vertices in $R \cap T$. No two of the $\{v_i\}$ are adjacent except for (v_i, v_{i+1}), for $i = 0, 1, \ldots, q$, since it is a shortest path, and a and b are only adjacent to v_0 and v_{q+1} respectively, by definition of R and T. Thus, inserting the edge (a, b) would create a cycle

$a \to v_0 \to v_1 \to \ldots \to v_{q+1} \to b \to a$ of length $q+4 \geqslant 4$ that is chordless. The graph g' would thus not be decomposable, completing the contradiction.

Now we prove the sufficiency of the condition. Given (ii), we can suppose that the cliques $a \cup R$ and $b \cup T$ are adjacent in the junction forest, for, if not, the forest can be manipulated so that this is so, while remaining a valid representation of the graph. To see this, let $C_0 = a \cup R$, $C_1, \ldots, C_p$, $C_{p+1} = b \cup T$ be the path between the cliques, with $p \geqslant 1$. By assumption, $C_i \cap C_{i+1} = S$ for some $i = 0, 1, \ldots, p$. We can delete the edge (C_i, C_{i+1}) from the junction forest and insert (C_0, C_{p+1}) instead. The only pairs of cliques $\{C_+, C_-\}$ for which the path connecting them has any additional cliques as a result of the modification are those for which the original path included both C_i and C_{i+1}; hence $C_+ \cap C_- \subset C_i \cap C_{i+1} = S$. The additional cliques in the modified paths must be some of $\{C_i,\, i = 0, 1, \ldots, p+1\}$, all of which contain S. Thus the junction property remains true for the junction forest as modified.

Thus, without loss of generality, $a \cup R$ and $b \cup T$ are adjacent cliques. Let $P = R \backslash S$ and $Q = T \backslash S$. We distinguish four cases, according to which of P and Q are empty or nonempty. If both are empty, then we simply amalgamate the cliques to form a new clique $a \cup b \cup S$, adding junction forest edges to those cliques adjacent to either of the original cliques. If $P \neq \emptyset = Q$, we replace clique $b \cup T = b \cup S$ by $a \cup b \cup S$, leaving adjacencies unchanged, and similarly by symmetry if $P = \emptyset \neq Q$. Finally, if neither is empty, we insert a new clique $a \cup b \cup S$ in the junction forest, linked to $a \cup R$ and $b \cup T$. In all four cases, it is easy to see that the junction property is maintained, so g' is decomposable.

References

Brooks, S. P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Statist.* **7**, 434–55.

Cowell, R. J. (1996). On compatible priors for Bayesian networks. *IEEE Trans. Pat. Anal. Mach. Intel.* **18**, 901–11.

Dawid, A. P. (1979). Conditional independence in statistical theory (with Discussion). *J. R. Statist. Soc.* B **41**, 1–31.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper Markov Laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–317.

Dellaportas, P. & Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–33.

Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–75.

Frydenberg, M. & Lauritzen, S. L. (1989). Decomposition of maximum likelihood in mixed interaction models. *Biometrika* **76**, 539–55.

Geiger, D. & Heckerman, D. (1994). Learning Gaussian networks. In *Proc. Conf. Uncertainty in Artificial Intelligence* **10**, Ed. R. Lopez de Mantaras and D. Poole, pp. 235–43. New York: Morgan Kaufmann.

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

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

Grone, R., Johnson, C. R., Sa, E. M. & Wolkowicz, H. (1984). Positive definite completions of partial hermitian matrices. *Lin. Algeb. Applic.* **58**, 109–24.

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

Madigan, D. & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Statist. Assoc.* **89**, 1535–46.

Madigan, D. & York, J. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.* **63**, 215–32.

Roverato, A. & Whittaker, J. (1998). The Isserlis matrix and its application to non-decomposable graphical Gaussian models. *Biometrika* **85**, 711–25.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. & Cowell, R. J. (1993). Bayesian analysis in expert systems (with Discussion). *Statist. Sci.* **8**, 219–83.

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

[*Received June* 1998. *Revised April* 1999]