Decomposable graphical gaussian model determination

Paolo Giudici* Università di Pavia.
Peter J. Green† University of Bristol.

18 February 1999

Abstract

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 MCMC 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 propose to 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 data-sets. Finally, the statistical and computational performance of the procedure is illustrated by means of both artificial and real data-sets.

Keywords: Bayesian Model Selection; Hyper Markov distributions; Junction Tree; Inverse Wishart Distribution; Reversible Jump MCMC.

1 Bayesian graphical models

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 in the area of Bayesian model determination for directed graphical models and probabilistic expert systems, see for instance Geiger and Heckerman (1994) and Spiegelhalter *et al* (1993). For undirected graphical gaussian models the main reference is Dawid and Lauritzen (1993), who introduced hyper Markov priors allowing local computations in Bayesian model determination. Applications of such priors include those of Madigan and Raftery (1994) and Madigan and York (1995), who analyse discrete graphical models according

*Dipartimento di Economia Politica e Metodi Quantitativi, Università di Pavia, Via San Felice 5, I-27100, Italy.
Email: pgiudici@eco.unipv.it
†Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
Email: P.J.Green@bristol.ac.uk.

to Occam's razor and using MCMC over the graph space. Finally, Dellaportas and Foster (1996) use reversible jump MCMC 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 it is often the case that richer information is 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 MCMC 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 Section 2. In Section 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 MCMC scheme for performing Bayesian model determination in graphical models. In Section 4 we examine the statistical performance of the proposed methodology, as well as the performance of the MCMC sampler. Finally, Section 5 contains some concluding remarks.

1.1 Background on graphical gaussian models

In this Subsection we briefly review the theory of graphical models relevant for our work following the exposition in Dawid and Lauritzen (1993), hereafter DL, to which we refer 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, $(C_1, \ldots, \mathcal{C}_n)$, 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 $i = 2, 3, \ldots, n$,

$$S_i = C_i \cap \bigcup_{j=1}^{i-1} C_i \subseteq C_h \quad \text{for some} \quad h = h(i) \in \{1, 2, \ldots, i\};$$

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; (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 $X_A = (X_v)_{v \in V}$ indicate the collection of random variables $(X_v : v \in A)$ with values in $\mathcal{X}_A = \times_{v \in A} \mathcal{X}_v$. 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 a 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, $X_A \amalg X_B | X_{A \cap B}$, where $\amalg$ 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 tacit 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:

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

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

$$X_i \amalg X_j | X_{V \setminus \{i,j\}} \Leftrightarrow k_{ij} = 0, \tag{1}$$

Thus, g constrains Σ imposing a pattern of zeros onto K. The effect of this constraint on Σ can be better specified using the notion of matrix completion with respect to a graph (see, for instance, Roverato and 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 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 $\{d_{ij} = 0, \forall (i, j) \notin E\}$. 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, say g, 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 similarly for S. When the graph is decomposable the likelihood of the graphical gaussian model specified by P is:

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

where

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

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

1.2 Prior distributions for graphical gaussian models

Two kind of uncertainties may affect a graphical model: (a) uncertainty about the probability distributions P on X or about the quantities, say θ, 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 both the above uncertainties simultaneously in a Bayesian fashion, using the data and the available expert information (expressed through the prior) to learn about θ (*quantitative learning*) and/or about g (*qualitative or structural learning*). To achieve such an objective, we need to consider formulating a prior distribution on θ and g. Concerning the latter, we shall assume throughout, for simplicity, a uniform prior on the class of decomposable graphs under comparison:

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

with d the number of 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 Section 3. However, 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 number of edges. We remark that, using importance sampling ideas, it is in principle possible to reweight results to replace this assumption by any other desired prior on g.

Turning to the parameters, a very general class of priors are the hyper Markov laws introduced in DL. 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 $\theta_{B|A}$ 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 $\mathcal{L}$ on θ is *hyper Markov* over g if, for any decomposition (A, B) of g, $\theta_A \amalg \theta_B | \theta_{A\cap B}$. In order to construct such laws, DL define two distributions $\mathcal{M}$ over θ_A and $\mathcal{N}$ over θ_B as *hyperconsistent* if they induce the same prior law over $\theta_{A\cap B}$. Given the family of sets $\mathcal{C}$ and $\mathcal{S}$, and a collection of pairwise hyperconsistent distributions $(\mathcal{L}_C, C \in \mathcal{C})$, DL show that there exists a *unique* hyper Markov law $\mathcal{L}$ 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 incomplete variance-covariance matrix Γ. Such elements are subject only to symmetry and positive definiteness of the submatrices $\{\Gamma_C = \Sigma_C, C \in \mathcal{C}\}$, 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:

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

A natural (although not necessary) 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), letting (Σ_C) to be distributed as an inverse Wishart distribution, with parameters α and Φ^C. We employ the parametrisation in DL 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 DL.

The construction previously described requires the specification of 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, say A, B, 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, say g and g'. Let Σ and Σ' be the corresponding precision matrices and, finally, $\mathcal{L}$ and $\mathcal{L}'$ be two hyper Markov laws on them. It is quite natural to require that $\mathcal{L}(\Sigma^A) = \mathcal{L}'(\Sigma^A)$, for any clique A common to both g and g'. This notion of compatibility is the same as in DL 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 DL, 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 a $IW(\alpha, \Phi)$ distribution to Σ. Marginalisation from this law will then imply that, for each $C \in \mathcal{C}$, $\mathcal{L}_C(\Sigma_C) = IW(\alpha, \Phi^C)$, 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 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

Our proposed statistical models differ in terms of the proposed prior on Σ, conditionally on g. A first class of models considers Σ to be hyper inverse Wishart with respect to g, with fixed hyperparameters α and Φ. A second class of models considers α and Φ *random* quantities, giving rise to a *hierarchical hyper inverse Wishart.*

2.1 A non-hierarchical model

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

$$\begin{aligned} X|\Sigma, g &\sim N_p(0, \Sigma); \\ \Sigma|g &\sim HIW_g(\alpha, \Phi); \\ p(g) &= d^{-1}, \end{aligned}$$

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 $C \in \mathcal{C}$, and d is the number of decomposable graphs on the vertex set V.

Notice that 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 *intra-class 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. Notice 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 among each pair of random variables. An assumption of common covariance is inevitably asymmetric about zero correlation (the prior correlation is constrained below by $-1/(p-1)$), and this may lead, particularly in large graphs, to an asymmetric evaluation 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 difficulties of prior specification, outlined above, it is desirable to devise a more automatic, yet flexible, method of assigning a prior distribution. A natural choice is letting 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 assigning a prior for α. Notice that α expresses the relative weight of the prior. A reasonable prior for α is a *Gamma* distribution, with mean f and variance fs, namely:

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

where $f > 0$ and $s > 0$ are positive quantities 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 Section 4.

Consider now the assignment of a prior on Φ. The representation adopted for Φ determines the set of random quantities which are to be assigned a prior distribution. We shall consider the following two situations: (a) Φ unstructured; (b) Φ with an intra-class correlation structure.

Unstructured Φ. An unstructured prior for Φ involves the assignment of a prior on $p(p+1)/2$ elements, that is, of p variances and $p(p-1)/2$ covariances. To ease the calculations, one can take a conjugate prior distribution. Notice 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 is 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 $\tau_{ii} = \tau$. The diagonal elements of T should be fixed coherently with the scale of the corresponding random variables. In the absence of such information, they can be taken as equal.

Intra-class Φ. In the intra-class case, as remarked in the last Subsection, all partial correlation coefficients are assumed to be equal *a priori*. A prior on the random elements (τ, ρ) which characterise the intra-class correlation structure can be obtained by restriction from the $W(d, T)$ prior on the unstructured Φ, as follows:

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

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 $t_{ij} = 0, i \neq j$), 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

(a) τ and ρ are independent random variables;

(b) τ has the Gamma distribution;

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

(c)

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

where γ has the Beta distribution

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

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. Notice also that $E(\rho) > 0, \forall p$ 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 MCMC algorithms

Markov chain Monte Carlo methods have considerably enlarged the domain of application of Bayesian inference (see for example Tierney, 1994). In particular, the *reversible jump* MCMC described in Green (1995) is particularly suited to deal with problems where the dimension of the parameter space changes. In our context, the dimension-changing aspect concerns proposing a change in the current graphical structure (say g) to a new structure, say g'. An important point is that, since we are considering *only* decomposable graphs, the proposed moves should consider only members of the latter class as candidate graphical structures.

3.1 Incremental changes to decomposable graphs

It is well-known (see for instance Frydenberg and Lauritzen, 1989) that the space of all decomposable graphs can be traversed by adding and deleting single edges at a time. Since such changes are convenient for MCMC implementation (in terms of algebraic tractability and computational efficiency) they will form the basis for the sampling algorithm we introduce in the next Subsection. 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. This characterisation may have application outside our MCMC context. While "legal" deletion moves can be characterised using a standard result which will be now quoted, "legal" addition moves still need to be characterised and we shall propose a Theorem for this purpose.

Theorem 1 (see, for instance, Frydenberg and 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* $\mathcal{T}$ 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 $\mathcal{T}$ is a tree and satisfies the *junction property*: for any two cliques $C_i, C_j \in \mathcal{C}$, and any clique C' on the unique path between them in $\mathcal{T}$,

$C_i \cap C_j \subset C'$. A *junction forest* representation of an undirected graph g is a collection of junction trees $\mathcal{T}_i$, each corresponding to a collection of cliques $\mathcal{C}_i$, with $\mathcal{C} = \bigcup \mathcal{C}_i$ and $C_i \cap C_j = \emptyset, i \neq j$. 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.*

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). We note that 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 $q \geq 0$ be the shortest path in g from an element of $R \setminus T$ to one of $T \setminus R$ avoiding vertices in $R \cap T$. No two of the $\{v_i\}$ are adjacent except for (v_i, v_{i+1}), $i = 0, 1, \ldots, q$, 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 $a \to v_0 \to v_1 \to \cdots \to v_{q+1} \to b \to a$ of length $q + 4 \geq 4$ 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 $p \geq 1$. 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 $P = R \setminus S$ and $Q = T \setminus S$. We distinguish 4 cases, according to which of P and Q are empty or non-empty. 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 4 cases, it is easy to see that the junction property is maintained, so g' is decomposable.

Example. Consider the graph in Figure 1: it is characterised by the cliques: (a, b, f), (b, c, f), (c, d, f) and (d, e, f).

Figure 1 about here

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'. This is 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 incremental changes to the edge-set of a decomposable graph. An alternative possibility is to reject such moves by running maximum cardinality search (MCS, see for instance Spiegelhalter *et al.*, 1993) after each graphical update proposal, to check if the proposed graph g' is decomposable. However, while MCS 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 and, furthermore, it constructs the new junction forest, so that the cliques are already constructed ready for use in probability calculations. Often, a and b will be adjacent so that the search will be very fast. In Section 4 we present empirical results that show the better performance of our algorithm.

3.2 Reversible jump MCMC design

We now briefly summarise the main features of reversible jump MCMC methodology, referring to 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 (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 vectorr 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 suitable 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 $y'(\cdot, \cdot)$ 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 $r_m(\cdot)$ is the probability density of a move of type m, evaluated at y and $q(u)$ is the density unction of u.

We now detail the reversible jump MCMC sampler we propose for the models specified in Section 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 variables. In our context, an important choice to be made is on how to represent Σ. Considering the collection $(\Sigma_C, C \in \mathcal{C})$ would be too computationally expensive: 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 simply to draw (set to zero) an element of K previously set to zero (unconstrained). However, notice 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; second, 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' requires inverting each Σ_C. Notice 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 effort of performing 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 to occasionally complete Γ to obtain K and Σ. An important result in this direction is contained, for instance, in DL: it turns out that $\Sigma = K^{-1}$, with

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

where the exponent [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:

(a) adding *or* deleting one edge from the graph g, ensuring that the proposed graph g' is decomposable. Notice that this move involves also making changes to Γ.

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

(c) updating the hyperparameter α.

(d) updating the hyperparameter Φ.

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

Updating g. Consider first moves of type (a), which are the only ones 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, say (i, j) 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}. 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 centered at the previous state, as in: $\gamma'_{ij} = \sigma_{ij} + u$. In our computations we prefer to perform only 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 ratio can be calculated as in (6). First note that the Jacobian of the transformation is equal to 1, and 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 adopting the proposal just described, based on the Γ representation, it turns out that the posterior ratio localises to the four subsets S, $S \cup i$, $S \cup j$, $S \cup i \cup j$ (abbreviated below as S, Si, Sj and Sij):

$$R_{\mathrm{post}} = \frac{\pi(y')}{\pi(y)} = \frac{h(\Sigma_S)h(\Sigma'_{Sij})}{h(\Sigma_{Si})h(\Sigma_{Sj})},$$

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:

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

When $S = \emptyset$, $h(\Sigma_S) = 1$. Notice that the requirement of positive definiteness of Σ constrains γ'_{ij}: if Σ'_{Sij} is not positive definite then $h(\Sigma'_{Sij}) = 0$, 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, $r_m(y) = r_m(y') = 1/\binom{n}{2}$. Finally, when (i,j) is added, γ'_{ij} is drawn from a gaussian distribution, with zero mean and standard deviation σ_G, so that

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

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

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

Notice that the calculation of R_{a} completely localises to *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 $R_{\mathrm{d}} = 1/R_{\mathrm{a}}$.

Updating Σ. Our strategy consists of perturbing *each* element of the corresponding incomplete matrix Γ with independent gaussian random walk proposal, centered 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:

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

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, which discourage implementation.

We now calculate the acceptance probability for our proposed updating of Σ to a new covariance matrix, say Σ', by means of a perturbation of 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:

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

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

Updating α. We perturb α with a gaussian random walk proposal, centered around the current value, namely: $q(\alpha'|\alpha) = N(\alpha, \sigma_\alpha)$, where σ_α is a spread parameter, to be appropriately chosen. Consequently, the proposal ratio is equal to 1. On the other hand, the posterior ratio is equal to:

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

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, namely: $\phi'_{ij} = N(\phi_{ij}, \nu_{ij})$, where the ν_{ij}'s are spread parameters, to be suitably chosen. The acceptance probability of the move will be equal to $\min(1, R_\Phi)$, 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:

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

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

Notice 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 data-sets, in order of increasing difficulty.

(a) Fret's heads data-set, with $p = 4$; there are 64 possible graphs, of which 3 are not decomposable. This is a small but challenging data-set, 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.

(b) Fowl bones data-set, with $p = 6$; there are 32768 possible graphs, of which 80% are decomposable. This is a more complex problem, but less multimodal than the previous one.

(c) An artificial data-set, with $p = 16$; there are 2^{16} possible models, of which 45% are decomposable. Data is 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 data-set will illustrate, besides the process of learning the true simulated data, how a mixture of non-decomposable models can approximate the true non-decomposable model.

The analysis of the three data-sets will be presented simultaneously, considering the following aspects: (i) prior settings; (ii) posterior distributions of main quantities of interest; (iii) sensitivity to prior specification; (iv) performance of the MCMC sampler.

Prior setting. For all data-sets 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 intra-class correlation structure for Φ, with $f = p + 1$, $s = 0.1$, $T = I$, and $d = 2$. Concerning the parameter Σ, it is important to understand see what such a prior specification corresponds to in terms of the prior expected partial correlation coefficients. This can be done by MCMC 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 MCMC sweeps after $10,000$ burn-in, with $p = 4$, gives essentially an identity matrix.

Posterior distributions. We now present results on the three data-sets considered. 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.

Figure 2 about here

Notice first that the posterior distribution of g is dispersed, as expected. For instance, the most probable graph receives only about 15% of the posterior probability and, in order to obtain 80% of the posterior probability, at least 10 structures have to be considered. These results are similar to those in Giudici (1996) who performed an exact non-informative Bayesian analysis on the same data-set using a non-hierarchical model.

It is often of interest to evaluate not only if an edge is present, but the strength of the association described by the edge itself. This can be done looking at the posterior distribution of the partial correlation coefficients, which cannot be derived analytically, but can be easily obtained from the MCMC output. Figure 3 reproduces the posterior distributions of the partial correlation coefficients for Fret's data.

Figure 3 about here

From Figure 3 notice that only the partial correlation between (1,2) and that between (3,4) have a relatively small posterior probability around zero, so that the evidence supports strongly the presence of such two edges.

Consider now analysis of the Fowl bones data-set, 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): 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 between (1,3,4,6) in Whittaker's.

We now finally comment on results on the spatial lattice data-set, obtained with a run of $100,000$ sweeps after $10,000$ burn-in. Our aim here is to show that, although the considered model space is very large, and does not contain the true model, MCMC learning can still give sensible answers. Figure 4 reproduces the trace and the corresponding histogram of two sampled partial correlations: the left trace plots a partial correlation which is equal to 0.2 in the true model; the right one a partial correlation which is zero in the true model.

Figure 4 about here

Notice 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. We remark that this difference is general; for presentation purposes 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 of times that edge is in the simulated model is greater than .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 require 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 we have obtained, with a run of $n = 1,000,000$ iterations is equal to 42, corresponding to a rate of 14%.

On the other hand, we remark that Bayesian structural learning is a very difficult task for this problem and, more generally, for large data-sets. Our results show that this is indeed possible with MCMC, although slow and requiring a considerable amount of diagnostic checking, in order to ascertain the validity of the results.

Sensitivity to the prior. Fret's data-set is useful for evaluating the sensitivity of results to the prior distribution, because of its highly correlated structure, leading 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 about here

Inference on partial correlation coefficients is quite robust, this seems an advantage of model averaging. Table 2 shows such a robustness of the inference on 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 a conclusion on this. Similar results can be obtained with a hierarchical prior.

Table 2 about here

The hierarchical prior has a smaller impact on the posterior over graphs (compare Table 3 with Table 1). On the other hand, the hierarchical model seem to select models with a higher number of edges. These results are confirmed by the analysis of the two other data-sets considered.

Table 3 about here

Performance of the MCMC sampler. First of all we remark that the correctness of our program implementing the sampler has been partially validated for all of our models by running them without any data, and with the likelihood term omitted: that is, we have simulated from the prior distribution. We have derived analytically the exact distribution of some marginal quantities, such as g and the number of edges present, and checked whether the MCMC output was in agreement with such theoretical results. Our algorithms gave very good agreement between the exact and simulated prior marginals.

A requirement in designing any Metropolis-Hastings sampler is the appropriate choice of the spread parameters of the proposal distribution, in order to ensure satisfactory mixing of the chain. In our case, after a number of pilot runs, we have decided to take $\sigma_G = 0.5n/p$; $\sigma_{ij} = 0.1/p$; $\sigma_\alpha = 1.0$; $\nu_{ij} = 1.0/p$. We remark that, concerning the proposal on g, the choice of centering the proposal at σ_{ij} leads to better performances compared to a "blind" proposal centered 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 on g, Γ_g, Φ and α for the three previously described simulations. We also report the corresponding total computation times obtained on a SPARC4 workstation.

Table 4 about here

In order to evaluate the computational times of our method relative to MCS, we conducted a small trial with the two methods in parallel, timing just the graph manipulation part of the procedures (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 MCS for these graph operations were respectively 0.63, 1.21 and 3.49 times those with our method. This is a necessarily incomplete comparison, and it is difficult to be absolutely objective as times depend on details of the coding. However, the comparisons are in one sense favourable to MCS, 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. It can be monitored by looking at an appropriate summary measure of g, such as the number of edges present, which describes the graph complexity. For all data-sets trace and autocorrelation plots on 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 Fowl bones data-sets $n = 100,000$ iterations are sufficient, whereas for the spatial lattice model a 10 times longer run is needed. We have also assessed performance of the MCMC dimension-jumping move more formally: for each of the three considered data-sets we have evaluated Gelman and Rubin's convergence diagnostic for the trace of the number of edges in the simulated graphs, according to the iterated graphical approach suggested by Brooks and Gelman (1998). Our result indicate that each of the simulated parallel chains is close to the target distribution.

We finally remark that the MCMC performance is not greatly affected by the choice of hyperparameter values. However, mixing is sensitive to the strength of the observed interaction effects between the variables in the graph: the higher this is, the slower the convergence.

5 Discussion

We have proposed a methodology for Bayesian model determination in undirected decomposable graphical models and, in particular, for graphical gaussian models. Our proposal is based on hierarchical prior distributions. We believe hierarchical priors have two main advantages with respect to non-hierarchical priors: on one hand, they are easier to specify, and can thus constitute an "automatic" default choice, especially for highly complex problems; on the other hand, they seem to lead to inferences less sensitive to the prior, as they allow "borrowing strength" of sample information between different clique domains.

The second original contribution of our paper is the implementation of a reversible jump MCMC algorithm for Bayesian model determination. This allows the extraction of posterior inference on *any* quantity of interest, in *both* the hierarchical and the non-hierarchical model. For instance, posterior estimates of the partial correlation coefficients, giving the strengths of the associations, can be easily obtained.

Our algorithm is fully based on local computations: the adding/deleting characterisation and the positive completion results have led to the construction of a fast algorithm to rearrange the junction tree and the associated parameters as g varies. On the other hand, a possible disadvantage is that we are restricted to decomposable graphical models. However, as we have also shown in the application section, quantitative learning in non-decomposable models can be reasonably well approximated by learning from mixtures of decomposable models.

Another important weakness of the methodology is that it becomes slow for very large domains, as the dimension of the model space increases more than exponentially with the number of vertices. Research is needed in the design of proposal moves which can improve the speed of convergence as well as on the related issue of monitoring the convergence of the algorithm.

We finally remark that our proposed methodology is quite general, and can be extended to other families of graphical models.

Acknowledgements

The first author acknowledges support from the European Science Foundation (ESF) and the Italian National Research Council (CNR). Computing facilities were supported in part by the EPSRC Stochastic modelling in science and technology initiative. We also acknowledge comments and suggestions from the referees.

References

Brooks, S.P. & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Cowell (1996). On compatible priors for Bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**, 901–911.

Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *J. R. Statist. Soc. B*, **57**, 473–484.

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

Dellaportas, P. & Forster, J.J. (1996). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. To appear in: *Biometrika*.

Dempster, A.P. (1972). Covariance selection. *Biometrics*, **28**, 157–175.

Frydenberg, M. & Lauritzen, S.L. (1989). Decomposition of maximum likelihood in mixed interaction models. *Biometrika* **76**, 539–555.

Geiger & Heckerman (1994). Learning gaussian networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence* **10**, 235-243. New York: Morgan and Kaufmann.

Giudici, P. (1996). Learning in graphical gaussian models. In *Bayesian Statistics* **5** (eds. J. Berger, J.M. Bernardo, A.P. Dawid, and A.F.M. Smith), 621–628. Oxford: Clarendon Press.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711–732.

Grone, R., Johnson, C.R., Sa, E.M. & Wolkowicz, H. (1984). Positive definite completions of partial hermitian matrices. *Linear Algebra and its applications*, **58**, 109–124.

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

Madigan, D. & Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.*, **89**, 1535–1546.

Madigan, D. & York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–232.

Roverato, A. & Whittaker, J. (1998). The Isserlis matrix and its application to non-decomposable graphical gaussian models. *Biometrika*, **85**, 711-725.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., and Cowell, R.J. (1993). Bayesian analysis in expert systems. *Statistical Science*, **8**, 219–283.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics,* **22**, 1701–1762.

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

Table 1: Sensitivity of structural learning with respect to the prior, for a non-hierarchical model.

	$\alpha = p+1$			$\alpha = 2p$		
	$\rho = -.3$	$\rho = 0$	$\rho = .9$	$\rho = -.3$	$\rho = 0$	$\rho = .9$
g_0	100011	111001	100011	110001	111011	110001
$p(g_0\|x)$	.1165	.1267	.2645	.1415	.1171	.2142
$E(\text{n.edges}\|x)$	3.63	4.16	3.13	3.52	4.21	3.04

Table 2: Sensitivity of model averaged inference on partial correlation coefficients with respect to the prior, for a non-hierarchical model.

	$\alpha = p+1$			$\alpha = 2p$		
	$\rho = -.3$	$\rho = 0$	$\rho = .9$	$\rho = -.3$	$\rho = 0$	$\rho = .9$
$E(\rho_{12}\|x)$	0.204	0.207	0.241	0.204	0.208	0.239

Table 3: Sensitivity of structural learning with respect to the prior, for a hierarchical model.

	$f = p+1$			$f = 2p$		
	$d = 2$	$d = p$	$d = 2p$	$d = 2$	$d = p$	$d = 2p$
g_0	111011	110111	110111	111011	111011	111011
$p(g_0\|x)$	.1527	.1304	.1422	.1317	.1412	.1517
$E(\text{n.edges}\|x)$	4.41	4.40	4.46	4.45	4.54	4.53

Table 4: Performance of the MCMC samplers, for the data-sets considered: rejection fractions and computation times (in minutes and seconds, for 100 000 sweeps).

Move type	Fret's	Fowl bones	Spatial lattice
g	.022	.001	.002
Σ	.573	.016	.379
Φ	.566	.595	.642
α	.518	.476	.577
Time	2:16	3:57	31:03

Figure 1: Graph illustrating the reversibility condition

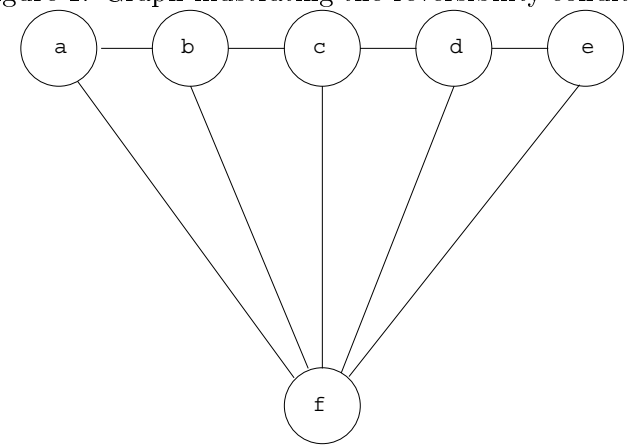

Figure 2: Most probable graphs for Fret's data-set

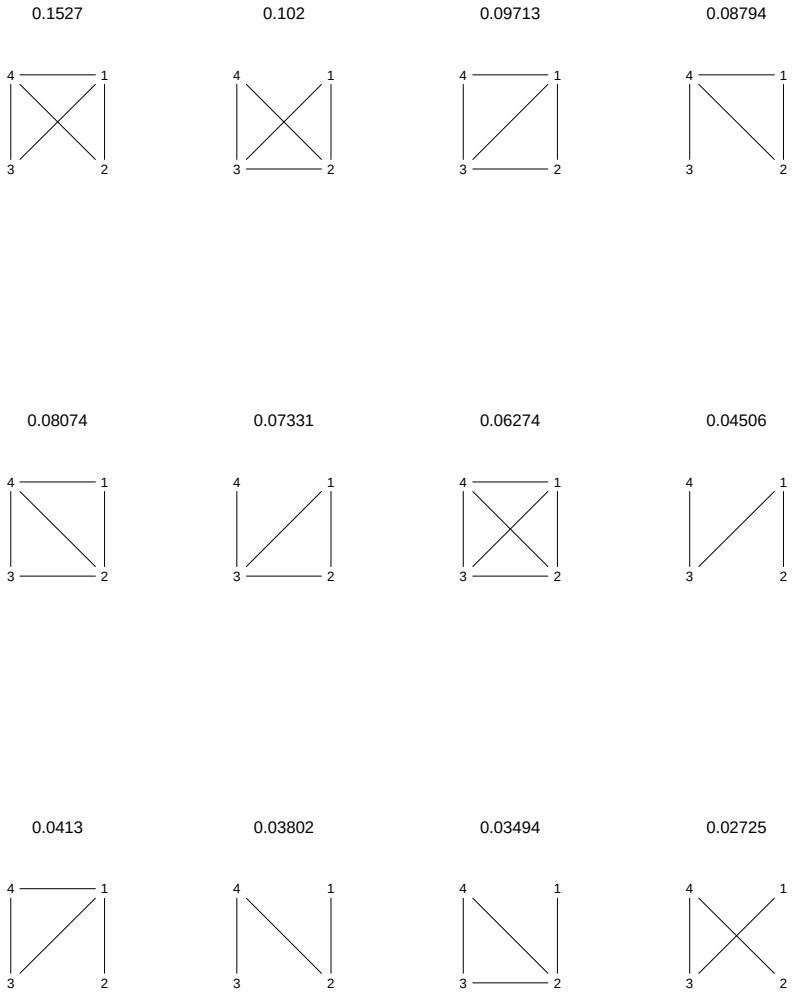

Figure 3: Posterior distribution of the partial correlation coefficients for Fret's data-set

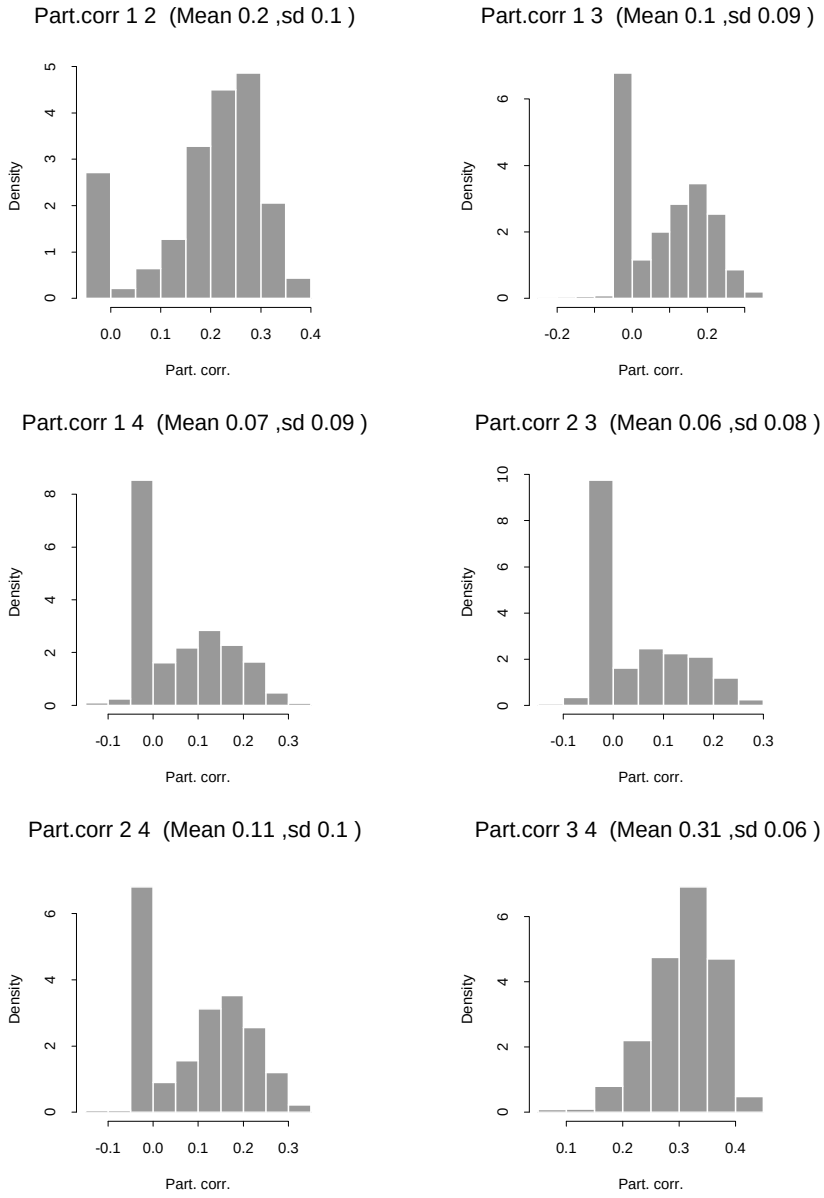

Figure 4: Partial correlation plots for the simulated spatial lattice data