A Monte Carlo method for computing the marginal likelihood in nondecomposable graphical Gaussian models

By ALIYE ATAY-KAYIS

Faculty of Economic and Administrative Sciences, Department of Business Administration, Suleyman Demirel University, 32260 Isparta, Turkey

aliye@emk.com.tr

and

HÉLÈNE MASSAM

Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, M3J 1P3, Canada.

massamh@yorku.ca.

SUMMARY

A centred Gaussian model Markov with respect to an undirected graph G is characterised by the parameter set of its precision matrices which is the cone $M^+(G)$ of positive definite matrices with entries corresponding to the missing edges of G constrained to be equal to zero. In a Bayesian framework, the conjugate family for the precision parameter is the distribution with Wishart density with respect to the Lebesgue measure restricted to $M^+(G)$. We call this distribution the G-Wishart. When G is nondecomposable, the normalising constant of the G-Wishart cannot be computed in closed form. In this paper, we give a simple Monte-Carlo method for computing this normalising constant. The main feature of our method is that the sampling distribution is exact and consists of a product of independent univariate standard normal and chi-squared distributions that can be read off the graph G. Computing this normalising constant is necessary for obtaining the posterior distribution of G or the marginal likelihood of the corresponding graphical Gaussian model. Our method also gives a way of sampling from the posterior distribution of the precision matrix.

Some keywords: Estimation in covariance selection models; Exact sampling distribution Wishart; Marginal likelihood; Nondecomposable graphical Gaussian models; Normalising constant.

1 Introduction

In multivariate Gaussian data analysis, the independence and conditional independence relationships between the variables can be represented by means of an undirected graph G with p vertices equal to the number of variables and the notion of a model Markov with respect to G. An undirected graph G is a pair $G = (V, E)$ where $V = \{1, \ldots, p\}$ and E is the set of undirected edges $E = \{(i, j) | i \in V, j \in V\}$. A p-dimensional Gaussian model is said to be Markov with respect to G if, for any edge (i, j) not in E, the ith and jth variables are conditionally independent given all the other variables. Such models are known as covariance selection models (Dempster, 1972) or graphical Gaussian models (see Whittaker, 1990, Ch. 6 or Lauritzen, 1996, Ch. 5). Without loss of generality, we can assume that these models are centred $N_p(0, \Sigma)$, and it is well known that they are characterised by the parameter set of the precision matrices which is the set of positive definite matrices $K = \Sigma^{-1}$ such that $K_{ij} = 0$ whenever the edge (i, j) is not in E. Finding the right model or models underlying the data is therefore equivalent to finding the graph or graphs G which best represent the conditional independences between the different variables.

In a Bayesian framework, the model selection is then made on the basis of the posterior distribution for G or, equivalently, the marginal likelihood (Berger, 1985, Ch. 3) for the model corresponding to G. As we shall see further in §3, if we use the Diaconis & Ylvisaker (1979) conjugate prior distribution for the precision matrix K the marginal likelihood is actually equal to the ratio of normalising constants of two Wishart distributions conditional on having those entries corresponding to the missing edges of G equal to zero. Such Wishart distributions will be called G-Wishart. The normalising constant of the G-Wishart with shape parameter δ and inverse scale parameter D is therefore equal to the integral over the set $M^+(G)$ of positive definite matrices with zero ij entries whenever $(i,j) \notin E$, of a Wishart density

$$g(K) \propto \det(K)^{\frac{\delta-2}{2}} \exp\{-\frac{1}{2} \text{ tr } (KD)\}. \tag{1}$$

This normalising constant is also needed for the computation of the Bayes factors in model comparisons and for any Markov chain or stochastic search on the space of all possible graphs on p vertices. Its efficient and accurate computation is therefore very important.

When G is non decomposable, this normalising constant has to be computed numerically and the aim of this paper is to present a simple Monte Carlo method for its computation. Roverato (2002) and Dellaportas et al. (2003) have already addressed this problem, using an importance sampling method. Our method is a simple Monte Carlo method and therefore without the risk of instability sometimes present in the weights of an importance sampling method (Tanner, 1985, Ch. 3). The distribution from which we sample is exact and can be directly determined from the graph visually, or through the incidence matrix of the edges of the graph G; see §4.

The different methods given in Roverato (2002), Dellaportas et al. (2003) and the present paper follow from three different views of the Wishart distribution when G is complete. Dellaportas et al. (2003) view a Wishart random variable $K \sim W(\alpha, \Sigma)$, for integer values of α, as the sum $\sum_{i=1}^{\alpha} Z^{(i)} Z^{(i)T}$ where $(Z^{(i)}, i = 1, \ldots, \alpha)$ is a sample from the $N_p(0, \Sigma)$ distribution . For arbitrary values of α, Roverato (2002) considers the Choleski decomposition $K = \phi^T\phi$ of K where ϕ is uppertriangular with positive diagonal elements. It is well known that, for G complete or decomposable, the rows of ϕ are independent, the diagonal elements ϕ_{ii} are distributed as chi-squared and the remainder of the ith row, conditional on ϕ_{ii}, is multivariate normal. We also start with the Choleski decomposition $K = \phi^T\phi$ but then immediately use the action of the triangular group on the set of positive definite matrices to obtain a Wishart variate with the identity as its scale parameter. Indeed, if D is the inverse scale parameter as in (1) above and $D = (T^T T)^{-1}$ with T uppertriangular, then $V = (T^{-1})^T K T^{-1}$ is a Wishart variate with scale parameter the identity. Its Choleski decomposition $V = \psi^T\psi$, with $\psi = \phi T^{-1}$, is such that all the entries $\psi_{ij}, 1 \leq i \leq j \leq p$ are mutually independent, the ψ_{ii}'s are chi-squared and the $\psi_{ij}, i < j$, are univariate $N(0, 1)$; see for example Muirhead (1982).

When G is not decomposable, these properties are no longer true but they are the basis for the derivation of the importance sampling distribution in the case of Dellaportas et al. (2003) and Roverato (2002) and for the Monte Carlo method presented here. As we shall show in §4, even when G is nondecomposable, we can still use the transformation $\psi = \phi T^{-1}$ to transform the expression of the normalising constant so that it is expressed as the expected value of a function $f_T(\psi^{\mathcal{V}})$, where $\mathcal{V}$ is the set of pairs (i, j) such that either $i = j, i = 1, \ldots, p$, or (i, j) is an edge of G, where $\psi^{\mathcal{V}}$ is the vector with components $\psi_{ij}, (i, j) \in \mathcal{V}$ and where the expected value is taken with respect to the product of independent chi-squared and univariate standard normal distributions. From this, we will immediately derive our Monte Carlo method. We will also see that this allows us to sample from the distribution of $K = \Sigma^{-1}$.

2 Preliminaries: the cones $M^+(G)$ and $M_*^+(G)$

Let $G = (V, E)$ be a graph as defined in the introduction. Following the notation introduced by Roverato (2000, 2002) we let

$$\begin{aligned} \mathcal{V} &= \{(i,j),\ i \leq j \ \text{ such that either } \ i = j,\ i \in V \text{ or } (i,j) \in E\} \\ \mathcal{W} &= \{(i,j),\ i,j \in V, i \leq j\} \ \text{ and } \ \bar{\mathcal{V}} = \mathcal{W} \setminus \mathcal{V}. \end{aligned} \tag{2}$$

Moreover, if A is any nonempty subset of V, we will write $\mathcal{A} = \mathcal{V} \cap (A \times A)$ for the edge set of the induced graph G_A together with the pairs $(i,i), i \in A$.

Let M^+ denote the cone of $p \times p$ positive definite matrices and let $M^+(G) \subset M^+$ be the cone of $p \times p$ positive definite matrices with ij entry equal to 0 whenever $(i,j) \in \bar{\mathcal{V}}$; that is,

$$M^+(G) = \{X \in M^+ | \quad \text{for } \ (i,j) \in \bar{\mathcal{V}}, X_{ij} = 0\}. \tag{3}$$

For any set $\mathcal{C} \subset \mathcal{W}$, a $\mathcal{C}$-incomplete symmetric matrix $X^{\mathcal{C}}$ is a symmetric $p \times p$ matrix with specified entries $X_{ij}^{\mathcal{C}}$ when $(i,j) \in \mathcal{C}$ and the remaining entries empty. For any given symmetric matrix X, we denote by $X_{\mathcal{V}}$ the projection of X onto the space of $\mathcal{V}$-incomplete matrices obtained by keeping the entries X_{ij} and X_{ji} when $(i,j) \in \mathcal{V}$ and not specifiying the others. For a given $\mathcal{C}$-incomplete matrix $X^{\mathcal{C}}$, any matrix X such that $X_{\mathcal{C}} = X^{\mathcal{C}}$ is called a completion of $X^{\mathcal{V}}$.

Definition 1 . *For a given arbitrary undirected graph G, a $\mathcal{V}$-incomplete symmetric matrix $X^{\mathcal{V}}$ is said to be G-partially positive definite if all block submatrices of $X^{\mathcal{C}}$*

corresponding to the cliques of G are positive definite. The set of G-partially positive definite matrices $X^{\mathcal{V}}$ such that there exists a completion $X \in M^+$ with $X^{-1} \in M^+(G)$ is denoted $M_*^+(G)$. When such an X exists, we say that X is a PD-completion of $X^{\mathcal{V}}$.

Example 1 . *If G is the graph given in Figure 1 (a), then $V = \{1,2,3,4\}$, $\mathcal{C} = \{(11),(12),(13),(22),(24),(33),(34)\}$ and the $\mathcal{C}$-symmetric incomplete matrix $X^{\mathcal{C}}$ is*

$$X^{\mathcal{C}} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & * \\ x_{21} & x_{22} & * & x_{24} \\ x_{31} & * & x_{33} & x_{34} \\ * & x_{42} & x_{43} & x_{44} \end{pmatrix},$$

where asterisks denote empty entries. Such a matrix will be said to be G partially positive definite if

$$\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}, \begin{pmatrix} x_{11} & x_{13} \\ x_{31} & x_{33} \end{pmatrix}, \begin{pmatrix} x_{22} & x_{24} \\ x_{42} & x_{44} \end{pmatrix}, \begin{pmatrix} x_{33} & x_{34} \\ x_{43} & x_{44} \end{pmatrix}$$

are all positive definite.

Proposition 1 . *(Grőne et al., 1984) If G is decomposable, then, given any $\mathcal{V}$-incomplete partially positive definite matrix $X^{\mathcal{V}} \in M_*^+(G)$, there exists a unique complete positive definite matrix X such that $(X)_{\mathcal{V}} = X^{\mathcal{V}}$ and X^{-1} belongs to $M^+(G)$. If G is nondecomposable, a PD-completion of $X^{\mathcal{V}}$ does not necessarily exist but, if it does, it is unique and we can therefore talk about the PD-completion of $X^{\mathcal{V}}$.*

This defines a bijection

$$\gamma \;:\; Y \in M^+(G) \mapsto X^{\mathcal{V}} = (Y^{-1})_{\mathcal{V}} = \gamma(Y) \in M_*^+(G)\ . \tag{4}$$

According to Proposition 1, the completion of $X^{\mathcal{V}}$ is unique when it exists and therefore the elements $X_{ij}, (i,j) \in \bar{\mathcal{V}}$ are functions of the elements $X_{ij}, (i,j) \in \mathcal{V}$, which will be called the 'free' elements of X.

Any positive definite matrix $x \in M^+$ can be written as $x = \phi^T \phi$ where ϕ is in the set $M^{\triangleleft}$ of uppertriangular matrices with positive diagonal elements; this is the Choleski decomposition of X. The following proposition shows that, if $x \in M^+(G)$, the free elements of ϕ are the $\phi_{ij}, (i,j) \in \mathcal{V}$. Its proof is trivial and has already been given in Roverato (2002) but we still give it here since the expression of $\phi_{ij}, (i,j) \in \mathcal{V}$ is essential to the development of our method.

Proposition 2 . *Let x be in $M^+(G)$ and let $x = \phi^T \phi$ be its Choleski decomposition. Then the entries $\phi_{ij}, (i,j) \in \mathcal{V}$ are such that*

$$\phi_{ij} = \frac{x_{ij} - \sum_{k=1}^{i-1} \phi_{ki}\phi_{kj}}{\phi_{ii}}. \tag{5}$$

For $(i,j) \in \bar{\mathcal{V}}$, that is, when $x_{ij} = 0$,

$$\phi_{1k} = 0, \quad \textit{for} \quad i = 1 \textit{ and } k = 2, \ldots, p \tag{6}$$

$$\phi_{ij} = -\frac{\sum_{k=1}^{i-1} \phi_{ki}\phi_{kj}}{\phi_{ii}} \quad \textit{for} \quad 1 < i \leq j \leq p\, . \tag{7}$$

Proof. The entries $\phi_{ij},\ i \leq j$ are determined in a unique way by the equations

$$x_{ij} = \sum_{k=1}^{i} \phi_{ki}\phi_{kj}, \quad \phi_{ii} > 0\, . \tag{8}$$

This immediately yields (5) and it follows that, for $(i,j) \in \mathcal{V}$, the ϕ_{ij} are free entries of ϕ since the corresponding x_{ij} are free entries of x.

When $(i,j) \in \bar{\mathcal{V}}$, (8) yields (6) and (7). Ordering the elements of ϕ according to the lexicographic order, we note that for $(i,j) \in \bar{\mathcal{V}}, \phi_{ij}$ is function only of ϕ_{rs} with $(r,s) < (i,j)$. □

From Proposition 2, we see that $X = \phi^T \phi$ is in $M^+(G)$ if and only if the entries $\phi_{ij}, (ij) \in \mathcal{V}$ are free elements and the entries $\phi_{ij}, (ij) \in \bar{\mathcal{V}}$, satisfy (6) and (7). Thus, denoting by $M_*^{\triangleleft}(G)$ the set of $\mathcal{V}$-incomplete uppertriangular matrices $\phi^{\mathcal{V}}$ with positive diagonal elements and completion $\phi \in M^{\triangleleft}$ and letting $\phi_{\mathcal{V}}$ denote the projection of $\phi \in M^{\triangleleft}$ onto $M_*^{\triangleleft}(G)$, we see that the mapping π, defined by

$$x = \phi^T \phi \in M^+(G) \mapsto \pi(x) = \phi^{\mathcal{V}} = \phi_{\mathcal{V}} \in M_*^{\triangleleft}(G),$$

is a bijection. The matrix obtained by completing any $\mathcal{V}$-incomplete upper triangular matrix $\phi^{\mathcal{V}}$ according to (6) and (7) is called the T-completion of $\phi^{\mathcal{V}}$.

The results presented in this section allow us to say that, if the precision matrix K is in $M^+(G)$, then the covariance parameter of interest is $\Sigma^{\mathcal{V}} = \Sigma_{\mathcal{V}}$, where $\Sigma = K^{-1}$. Conversely, if an incomplete covariance matrix $\Sigma^{\mathcal{V}} \in M_*^+(G)$ is given, it is of course G-partially positive definite. When G is decomposable, its PD-completion Σ always exists and is unique. If G is nondecomposable, its completion does not necessarily exist but when it exists, it is unique. When the Choleski decomposition $K = \phi^T \phi$ of $K \in M^+(G)$ is considered, the parameter of interest is $\phi^{\mathcal{V}}$ as defined above and it is uniquely defined.

3 The normalising constant

In this section, we recall how, in a Bayesian framework, the model selection problem leads to that of the computation of the normalising constant of the G-Wishart already mentioned in the introduction and formally defined below. We also recall that the normalising constant need only be computed prime component by prime component, thus reducing the dimensionality of the problem.

Recall that a prime component G_A of an arbitrary undirected graph G is a subgraph induced by $A \subset E$ such that G_A does not admit a proper decomposition and is maximal with respect to inclusion among all subgraphs which do not admit a proper decomposition. We have already used, and will use throughout the paper, graphical model notions such as decomposability, proper decomposability, cliques, minimal separators or the perfect numbering of cliques and vertices; see Lauritzen (1996, Ch. 2) for detailed definitions and studies of these concepts. Less common notions are redefined here.

Consider a Gaussian model

$$\{N_p(0, \Sigma), \Sigma \in M^+\}.$$

Let $G = (V, E)$ be an arbitrary given graph. By definition, if $Z \sim N_p(0, \Sigma)$ and its distribution is Markov with respect to G then the ith and jth components of Z are conditionally independent given all the other components of Z; that is,

$$Z_i \perp\!\!\!\perp Z_j | Z_{V \setminus \{i,j\}}, \tag{9}$$

whenever $(i,j) \notin E$. It is is well known that, for a Gaussian model, (9) is equivalent to

$$K_{ij} = 0 \text{ for } (i,j) \notin \mathcal{V} \, ;$$

that is, $K \in M^+(G)$.

Therefore, for a given graph G, the graphical Gaussian model Markov with respect to G can now be defined as

$$\mathcal{M}_G = \{N_p(0,\Sigma) : \; K = \Sigma^{-1} \in M^+(G)\}. \tag{10}$$

Let $\mathcal{G}$ denote the set of all possible graphs with vertex set V. The model selection problem is, given the data, to choose between all the models $\{\mathcal{M}_G, G \in \mathcal{G}\}$, that is, between all G in $\mathcal{G}$. This usually involves finding the quantitative as well as the qualitative aspects of the models; that is the covariance matrix, or its inverse, as well as the underlying graph. With high-dimensional data such as those obtained from gene expression experiments, the primary objective is to show how different variables affect each other. In such cases, it is most important to find the various conditional independences between variables and therefore the object of primary interest is the underlying graph G. Finding the 'best' graphs is done through model selection. An estimate of the precision matrix can then be obtained by various means, one of which is to sample from the posterior distribution of K corresponding to the 'best' graphs.

For convenience, from now on, we will denote the inner product $\operatorname{tr}(ab^T)$ of a and b in the space of $p \times p$ matrices by $\langle a, b \rangle$. This implies of course that, for a and b symmetric, $\langle a, b \rangle = \operatorname{tr}(ab)$. Let $(Z^{(1)}, \ldots, Z^{(n)})$ be a sample from $\mathcal{M}_G$ with G unknown and with joint density

$$p(z^{(1)},\ldots,z^{(n)}|K,G)=\frac{|K|^{\frac{n}{2}}}{(2\pi)^{\frac{np}{2}}}\exp-\frac{1}{2}\langle K,u\rangle\ ,$$

where $K\in M^+(G)$ and $u=\sum_{i=1}^{n}z^{(i)}z^{(i)t}$. For graphical Gaussian model selection in a Bayesian framework, we now have to choose a prior distribution for the parameter K. Since the family of distributions for U when K varies is a natural exponential family a widely accepted family of distributions is the conjugate family as defined by Diaconis & Ylvisaker (1979): the conjugate family for the canonical parameter $K\in M^+(G)$ has density

$$f(K|G)\sim|K|^{\frac{\delta-2}{2}}\exp-\frac{1}{2}\langle K,D\rangle\ , \tag{11}$$

where the parameters $\delta\in\mathbb{R}$ and $D\in M^+$ are such that

$$I_G(\delta,D)=\int_{M^+(G)}|K|^{\frac{\delta-2}{2}}\exp-\frac{1}{2}\langle K,D\rangle dK<\infty. \tag{12}$$

A sufficient condition for the convergence of $I_G(\delta,D)$ (Diaconis & Ylvisaker, 1979, Theorem 1) is $\delta>2$ and $D^{-1}\in M^+(G)$.

Therefore, for $\delta>2$ and $D^{-1}\in M^+(G)$, the distribution with density, with respect to the Lebesgue measure on $M^+(G)$, equal to

$$f(K|G)=\frac{1}{I_G(\delta,D)}|K|^{\frac{\delta-2}{2}}\exp-\frac{1}{2}\langle K,D\rangle \tag{13}$$

will be called the G-Wishart distribution $W_G(\delta,D)$.

We next have to choose a prior distribution on the set $\mathcal{G}$ of all possible graphs on p vertices. Without loss of generality for our problem and for the sake of simplicity, we can assume that the prior on $\mathcal{G}$ is the discrete uniform distribution with density

$\pi(G) = \frac{1}{|\mathcal{G}|}$. When the prior distribution on K, given G, is taken to be $W_G(\delta, D)$ the joint density of $(Z^{(1)}, \ldots, Z^{(n)}, K, G)$ is then

$$f((Z^{(1)}, \ldots, Z^{(n)}, K, G) = \frac{1}{(2\pi)^{\frac{np}{2}} |\mathcal{G}|} \frac{1}{I_G(\delta, D)} |K|^{\frac{\delta+n-2}{2}} \exp -\frac{1}{2} \langle K, D+U \rangle \ , \tag{14}$$

the marginal likelihood is

$$p(Z^{(1)}, \ldots, Z^{(n)} | G) = \frac{1}{(2\pi)^{\frac{np}{2}}} \frac{I_G(\delta + n, D + U)}{I_G(\delta, D)} \ , \tag{15}$$

and the posterior density of G given $Z^{(1)}, \ldots, Z^{(n)}$ is

$$p(G | Z^{(1)}, \ldots, Z^{(n)}) = \frac{J_G(\delta, n, D, U)}{\sum_{G' \in \mathcal{G}} J_{G'}(\delta, n, D, U)} \ , \tag{16}$$

where

$$J_G(\delta, n, D, U) = \frac{I_G(\delta + n, D + U)}{I_G(\delta, D)} \ .$$

The problem of computing the marginal likelihood or the posterior distribution is therefore reduced to the problem of computing normalising constants of the type $I_G(\delta, D)$ as given in (12), with $\delta > 2$ and $D^{-1} \in M^+(G)$.

In practice, the number of variables p is large and it is desirable to perform computations locally that is within relatively small subgroups of variables whenever possible. Fortunately this is the case. To prove the decomposition of $I_G(\delta, D)$ into smaller components, it is necessary to go through the conjugate prior for Σ corresponding to the conjugate prior (13) on K. As we have seen in §2 and more precisely in Proposition 1, if $K \in M^+(G)$, the covariance parameter of interest is not Σ but the incomplete matrix $\Sigma^{\mathcal{V}}$. We now give a quick review of the conjugate prior for $\Sigma^{\mathcal{V}}$ and its properties, in particular the fact that the normalising constant $I_G(\delta, D)$ need only be computed on each prime component of G. While (12) and (13) are valid for any undirected graph G, the form of the corresponding conjugate family for $\Sigma^{\mathcal{V}}$ changes with G.

When G is complete, for $\delta > 0$ and $D \in M^+$, (13) is the density of the Wishart distribution $W(\delta, D)$ where

$$I_G(\delta, D) = \frac{2^{\frac{np}{2}} \Gamma_p(\frac{\delta+p-1}{2})}{|D|^{\frac{\delta+p-1}{2}}},$$

and where, in general, for $a > \frac{p-1}{2}$,

$$\Gamma_p(a) = \pi^{\frac{p(p-1)}{4}} \prod_{i=0}^{p-1} \Gamma(a - \frac{i}{2}).$$

The covariance parameter is the full covariance matrix $\Sigma = K^{-1}$ which follows the inverse Wishart distribution $IW(\delta, D)$ with density

$$IW(\Sigma|\delta, D) = I_G(\delta, D)^{-1} |\Sigma|^{-\frac{\delta+2p}{2}} \exp -\frac{1}{2}\langle \Sigma^{-1}, D\rangle \ . \qquad (17)$$

For G decomposable, let $(C_j, j = 1, \ldots, k)$ and $(S_j, j = 2, \ldots, k)$ denote a perfect ordering of the set of cliques of G and its corresponding set of separators (Lauritzen, 1996, Ch. 2) and let $c_j = |C_j|, j = 1, \ldots, k$, and $s_j = |S_j|, j = 2, \ldots, k$, denote the cardinalities of the cliques and separators respectively. In this case, K belongs to $M^+(G)$. From Proposition 1, Σ always exists and is unique and the moment parameter of interest is the incomplete matrix $\Sigma^{\mathcal{V}}$ with entries the entries of $(\Sigma_{C_j}, j = 1, \ldots, k)$. Dawid & Lauritzen (1993) defined the distribution for $\Sigma^{\mathcal{V}}$ called the hyper inverse Wishart distribution with density a Markov combination of inverse Wisharts for $\Sigma_{C_j}, j = 1, \ldots, k$, and $\Sigma_{S_j}, j = 2, \ldots, k$; that is

$$HIW_G(\Sigma^{\mathcal{V}}|\delta, D) d\Sigma^{\mathcal{V}} = \frac{\prod_{j=1}^{k} IW(\Sigma_{C_j}|\delta, D_{C_j})}{\prod_{j=2}^{j=k} IW(\Sigma_{S_j}|\delta, D_{S_j})} d\Sigma^{\mathcal{V}}, \qquad (18)$$

with $\delta > 0$ and D such that $D^{-1} \in M^+(G)$. Roverato (2000) proved that the inverse of the hyper inverse Wishart distribution for δ and D as above is the $W_G(\delta, D)$ distribution defined on $M^+(G)$ and with normalising constant

$$I_G(\delta, D) = \frac{\prod_{j=1}^k I_{G_{C_j}}(\delta, D_{C_j})}{\prod_{j=2}^k I_{G_{S_j}}(\delta, D_{S_j})} \; ; \tag{19}$$

see also an unpublished 1993 Aalborg University M.Sc. thesis by A.M. Bjerg and T.W. Nielsen. The ratio (19) is, of course, also the normalising constant of the $HIW_G(\delta, D)$.

For G not necessarily decomposable, we have similar properties; that is, $K \in M^+(G)$ and from Proposition 1 the unique corresponding moment parameter of interest is again $\Sigma^{\mathcal{V}}$. The Jacobian $J(K \mapsto \Sigma^{\mathcal{V}})$ for the change of variable $K \mapsto \Sigma^{\mathcal{V}}$ has been computed in Roverato (2000) and the density for $\Sigma^{\mathcal{V}}$, obtained by multiplying (13) by $J(K \mapsto \Sigma^{\mathcal{V}})$ has been studied in Roverato (2002) and named the hyper inverse Wishart distribution $HIW_G(\delta, D)$ with shape parameter δ and scale parameter D. The density with respect to the measure $d\Sigma^{\mathcal{V}}$ is of course

$$HIW_G(\Sigma^{\mathcal{V}}|\delta, D) = \{I_G(\delta, D)\}^{-1}|\Sigma|^{-\frac{\delta-2}{2}} J(K \mapsto \Sigma^{\mathcal{V}}) \exp -\frac{1}{2}\langle \Sigma^{-1}, D\rangle, \tag{20}$$

with $\delta > 2$ and $D^{-1} \in M^+(G)$.

This density can also be factorised according to the prime components of G and their separators. Let $(P_1, \ldots, P_k)$ be a perfect sequence of prime components of G and let $(S_2, \ldots, S_k)$ be the corresponding set of minimal separators. Roverato (2002) has proved that the $HIW_G(\delta, D)$ density can be written as

$$HIW_G(\Sigma^{\mathcal{V}}|\delta, D) d\Sigma^{\mathcal{V}} = \frac{\prod_{j=1}^k HIW_{G_{P_j}}(\Sigma_{P_j}^{\mathcal{P}_j}|\delta, D_{P_j})}{\prod_{j=2}^k HIW_{G_{S_j}}(\Sigma_{S_j}|\delta, D_{S_j})} d\Sigma^{\mathcal{V}}, \tag{21}$$

where according to the notation introduced in §2, $\Sigma_{P_j}^{\mathcal{P}_j}$ denotes the $\mathcal{P}_j$-incomplete submatrix of $\Sigma^{\mathcal{V}}$ corresponding to the induced graph G_{P_j}; that is $\Sigma_{P_j}^{\mathcal{P}_j}$ is obtained by taking the P_j submatrix Σ_{P_j} of Σ and deleting the entries $(\Sigma_{P_j})_{rs}$ such that (r, s) is not an edge of G_{P_j}. This implies that the normalising constant for the density $HIW_G(\Sigma^{\mathcal{V}}|\delta, D)$ is equal to

$$I_G(\delta, D) = \frac{\prod_{j=1}^{k} I_{G_{P_j}}(\delta, D_{P_j})}{\prod_{j=2}^{k} I_{G_{S_j}}(\delta, D_{S_j})} \,. \tag{22}$$

The S_j are complete, so that $I_{G_{S_j}}(\delta, D_{S_j})$ can be computed explicitly, as in (17). Therefore, to compute $I_G(\delta, D)$ when G is nondecomposable, it is sufficient to compute each $I_{G_{P_j}}(\delta, D_{P_j})$ for P_j a prime component of G.

The task of computing the normalising constant for the $HIW_G(\delta, D)$ distribution is therefore reduced to that of computing

$$I_G(\delta, D) = \int_{M^+(G)} |K|^{\frac{\delta-2}{2}} \exp -\frac{1}{2}\langle K, D\rangle dK, \tag{23}$$

where G is a prime graph. This is the objective of the next section.

4 A Monte Carlo method for computing $I_G(\delta, D)$

4.1 The method

Let $G = (V, E)$ be a prime graph and let $|V| = p$. We will transform the expression of $I_G(\delta, D)$ as given in (23) according to a series of four steps. As in Roverato (2002) we first consider the Choleski decomposition $K = \phi^T \phi$ of $K \in M^+(G)$. Following this, we rely on the properties of the Choleski decomposition of a Wishart variate with the identity as its scale parameter (Muirhead, 1982, Th. 3.2.14). As a result, we do not have to choose and compute an importance sampling distribution, and the integral $I_G(\delta, D)$ can be expressed as an expectation, where the sampling distribution is read off the graph G or equivalently from its incidence matrix and consists of a product of independent chi-squared and standard univariate normal distributions. This makes the implementation simple and it allows us to sample from the posterior distribution of K for a given G, thus allowing us to obtain estimates of the precision matrix.

Step 1. Let $K = \phi^T \phi, \phi \in M^{\triangleleft}$ be the Choleski decomposition of K. Make the change of variable

$$K \in M^+(G) \mapsto \phi^{\mathcal{V}} = \phi_{\mathcal{V}} \in M_*^{\triangleleft}(G) \tag{24}$$

with Jacobian given in Lemma 1 below.

Step 2. Let $D = (T^T T)^{-1}$, $T \in M^{\triangleleft}$ where $T^T T$ is the Choleski decomposition of D^{-1}. Make the change of variable

$$\phi^{\mathcal{V}} \in M_*^{\triangleleft}(G) \mapsto \psi^{\mathcal{V}} \in M_*^{\triangleleft}(G) \tag{25}$$

where $\psi = \phi T^{-1}$ and $\psi^{\mathcal{V}} = \psi_{\mathcal{V}}$, with Jacobian given in Lemma 3 below.

Step 3. Give the expression of $\psi_{ij}, (i,j) \in \bar{\mathcal{V}}$ in terms of $\psi_{ij}, (i,j) \in \mathcal{V}$; see Lemma 2 below.

Step 4. Express $I_G(\delta, D)$ as the expectation of a function of $\psi_{ij}, (i,j) \in \mathcal{V}$, where the sampling distribution is a product of independent chi-squareds and standard normals.

From Proposition 2, we know that the free entries of ϕ, where $K = \phi^T\phi$, are the entries $\phi_{ij}, (i,j) \in \mathcal{V}$, and that the change of variable in (24) is well defined and one to one. Let $ne(i) = \{j \in V : \ (i,j) \in E\}$ denote the set of neighbours of i and let $\nu_i = |ne(i) \cap \{i+1, \ldots, p\}|$. The Jacobian for (24) is given in the following lemma.

Lemma 1 . *(Roverato, 2002) Let K be an element of $M^+(G)$ and let $K = \phi^T\phi$ be its Choleski decomposition. Then the Jacobian of the change of variable (24) is*

$$J_1 = 2^p \prod_{i=1}^{p} \phi_{ii}^{\nu_i + 1} \tag{26}$$

Let $D^{-1} = T^T T$ be the Choleski decomposition of the parameter D^{-1}. From Lemma 1, the equality $|K| = |\phi|^2 = \prod_{i=1}^{i=p} \phi_{ii}^2$ and the fact that the $\phi_{ij}, (i,j) \in \bar{\mathcal{V}}$ are functions of the $\phi_{ij}, (i,j) \in \mathcal{V}$, we immediately obtain that

$$\begin{aligned}
& I_G(\delta, D) \\
&= \int_{M^+(G)} |K|^{\frac{\delta-2}{2}} \exp -\frac{1}{2}\langle K, D\rangle dK \\
&= 2^p \int_{M_*^{\triangleleft}(G)} \prod_{i=1}^{p} (\phi_{ii}^2)^{\frac{\delta+\nu_i-1}{2}} \exp -\frac{1}{2}\langle \phi^T\phi, (T^tT)^{-1}\rangle d\phi^{\mathcal{V}} \\
&= 2^p \int \prod_{i=1}^{p} (\phi_{ii}^2)^{\frac{\delta+\nu_i-1}{2}} \exp -\frac{1}{2}\langle (\phi T^{-1})^T, \phi T^{-1}\rangle \prod_{i=1}^{p} d\phi_{ii} \prod_{i\neq j, (i,j)\in\mathcal{V}} d\phi_{ij},
\end{aligned} \tag{27}$$

where the third integration is over $(\mathbb{R}^+)^p \times \mathbb{R}^{|\mathcal{V}-p|}$.

We now consider the change of variable (25). The following lemma shows that this change of variable is well defined and is one to one. For any upper triangular matrix $T = (t_{ij})_{1\leq i\leq j\leq p}$, we use the notation

$$t_{\langle ij]} = \frac{t_{ij}}{t_{jj}}. \tag{28}$$

Lemma 2 . *Let $\phi^{\mathcal{V}} \in M_*^{\triangleleft}(G)$ and let ϕ be its completion. Let T be a given matrix in $M^{\triangleleft}$. Then $\psi = \phi T^{-1}$ is such that $\psi^{\mathcal{V}} = \psi_{\mathcal{V}}$ belongs to $M_*^{\triangleleft}(G)$ and the mapping given in (25) is a bijection. We have*

for $(r,s) \in \mathcal{W}$, that

$$\psi_{rs} = \sum_{j=r}^{s-1} -\psi_{rj} t_{\langle js]} + \frac{\phi_{rs}}{t_{ss}}. \tag{29}$$

In particular, for $r = s$,

$$\psi_{ss} = \frac{\phi_{ss}}{t_{ss}}. \tag{30}$$

For $(rs) \in \bar{\mathcal{V}}$ and $r < s$,

$$\psi_{rs} = \sum_{j=r}^{s-1} -\psi_{rj} t_{\langle js]} - \sum_{i=1}^{r-1} \Big(\frac{\psi_{ir} + \sum_{j=i}^{r-1} \psi_{ij} t_{\langle jr]}}{\psi_{rr}}\Big)\Big(\psi_{is} + \sum_{j=i}^{s-1} \psi_{ij} t_{\langle js]}\Big). \tag{31}$$

In particular, for $r = 1$ and $(1s) \in \bar{\mathcal{V}}, 1 < s$,

$$\psi_{1s} = \sum_{j=1}^{s-1} -\psi_{1j} t_{\langle js]}. \tag{32}$$

Proof. By straightforward matrix multiplication, we obtain $\phi_{rs} = \sum_{i=r}^{s} \psi_{ri} t_{is}$, from which it follows that

$$\psi_{rs} = \frac{1}{t_{ss}}(\phi_{rs} - \sum_{l=r}^{s-1} \psi_{rl} t_{ls}). \tag{33}$$

We see that the ψ_{rs} are determined recursively, following the lexicographic order, and that ψ_{rs} depends only on ϕ_{rs} and the preceding $\psi_{rj}, j = 1, \ldots, s-1$. Therefore, if $(r,s) \in \mathcal{V}$, that is, if ϕ_{rs} is a free variable, then so is ψ_{rs}. If $(r,s) \in \bar{\mathcal{V}}$, ϕ_{rs} is not a free variable and therefore ψ_{rs} is a function of $\psi_{ij}, (i,j) < (r,s)$, and ψ_{rs} is not free. We have therefore proved equalities (29) and (30) and the fact that $\psi^{\mathcal{V}}$ belongs to $M_*^{\triangleleft}(G)$.

It remains to prove (31) and (32). If $(rs) \in \bar{\mathcal{V}}, K_{rs} = 0 = \sum_{i=1}^{i=r} \phi_{ir}\phi_{is}$ and therefore

$$\frac{\phi_{rs}}{t_{ss}} = -\sum_{i=1}^{r-1} \frac{\frac{\phi_{ir}}{t_{rr}}}{\frac{\phi_{rr}}{t_{rr}}} \frac{\phi_{is}}{t_{ss}} = -\sum_{i=1}^{r-1} \Big(\frac{\psi_{ir} + \sum_{j=i}^{r-1} \psi_{ij} t_{\langle jr]}}{\psi_{rr}}\Big)\Big(\psi_{is} + \sum_{j=i}^{s-1} \psi_{ij} t_{\langle js]}\Big)$$

which proves (31). Then (32) follows from the fact that $\phi_{1s} = 0$ for $(1s) \in \bar{\mathcal{V}}$. $\square$

Let us now give the Jacobian of the change of variables (25). Let k_i be the number of vertices preceding i in the given order of the vertices.

Lemma 3 . *The Jacobian of the change of variable $\phi^{\mathcal{V}} \mapsto \psi^{\mathcal{V}}$ as given in (25) is*

$$J_2 = \prod_{i=1}^{p} t_{ii}^{k_i+1} \; . \tag{34}$$

Proof. Order the elements of both matrices ϕ and ψ according to the lexicographic order. Since the transformation from ϕ to ψ is linear, the Jacobian matrix is the matrix representative of the mapping $\phi^{\mathcal{V}} \mapsto \psi^{\mathcal{V}} = (\phi T^{-1})^{\mathcal{V}}$ and is of dimension $|\mathcal{V}| \times |\mathcal{V}|$. Since $\phi = \psi T$, we have that $\phi_{rs} = \sum_{i=r}^{s} \psi_{ri} t_{is}$ and it is clear that ϕ_{rs} is a function only of $\psi_{ij}, (i,j) \in \mathcal{V}$, with $i = r, j = r, \cdots, s$, i.e. of elements ψ_{ij} such that $(ij) < (rs)$. Therefore, the Jacobian matrix is an uppertriangular matrix and its determinant is the product of the diagonal elements. The (r,s) diagonal element for $(r,s) \in \mathcal{V}$ is the coefficient of ψ_{rs} in the expression of ϕ_{rs}, that is, t_{ss}. Thus

$$J_2 = \prod_{(rs)\in\mathcal{V}} t_{ss}. \tag{35}$$

For a given $s = 1, \ldots, p$, the number of edges $(rs) \in \mathcal{V}$ is equal to k_s and therefore $J_2 = \prod_{i=1}^{i=p} t_{ii}^{k_i+1}$, which is the desired result. $\square$

Since $\langle (\phi T^{-1})^T, \phi T^{-1} \rangle = \langle \psi^T, \psi \rangle = \sum_{i=1}^{i=p} \psi_{ii}^2 + \sum_{(i,j)\in\mathcal{V}, i\neq j} \psi_{ij}^2 + \sum_{(i,j)\in\bar{\mathcal{V}}} \psi_{ij}^2$, if we use also (30), $I_G(\delta, D)$ in (27) becomes

$$\begin{aligned} &I_G(\delta, D) \\ &= \; 2^p \prod_{i=1}^{p} (t_{ii}^2)^{\frac{\delta+\nu_i-1+k_i+1}{2}} \quad \times \end{aligned} \tag{36}$$

$$\int \prod_{i=1}^{p} (\psi_{ii}^2)^{\frac{\delta+\nu_i-1}{2}} \exp -\frac{1}{2}\left(\sum_{i=1}^{p} \psi_{ii}^2 + \sum_{(i,j)\in\mathcal{V}, i\neq j} \psi_{ij}^2 + \sum_{(i,j)\in\bar{\mathcal{V}}} \psi_{ij}^2\right) \prod_{i=1}^{p} d\psi_{ii} \prod_{(i,j)\in\mathcal{V}, i\neq j} d\psi_{ij}.$$

Let $b_i = \nu_i + k_i + 1$. Then

$$\begin{aligned} I_G(\delta, D) \quad = \quad & 2^p \prod_{i=1}^{p} (t_{ii}^2)^{\frac{\delta+b_i-1}{2}} \int \exp -\frac{1}{2} \sum_{\bar{\mathcal{V}}} \psi_{ij}^2 \prod_{i=1}^{p} (\psi_{ii}^2)^{\frac{\delta+\nu_i-1}{2}} \exp -\frac{1}{2} \sum_{i=1}^{p} \psi_{ii}^2 \\ & \times \exp -\frac{1}{2} \sum_{(i,j)\in\mathcal{V}, i\neq j} \psi_{ij}^2 \prod_{i=1}^{p} d\psi_{ii} \prod_{(i,j)\in\mathcal{V}, i\neq j} d\psi_{ij}. \end{aligned} \tag{37}$$

Since $d\psi_{ii} = \frac{1}{2}\psi_{ii}^{-1} d(\psi_{ii}^2)$, (37) becomes

$$\begin{aligned} I_G(\delta, D) \quad = \quad & \prod_{i=1}^{p} 2^{\frac{\delta+\nu_i}{2}} (2\pi)^{\frac{\nu_i}{2}} \Gamma(\frac{\delta+\nu_i}{2}) \prod_{i=1}^{p} (t_{ii}^2)^{\frac{\delta+b_i-1}{2}} \times \\ \times \quad & \int \exp -\frac{1}{2} \sum_{(i,j)\in\bar{\mathcal{V}}} \psi_{ij}^2 \quad \prod_{i=1}^{p} \frac{1}{\Gamma(\frac{\delta+\nu_i}{2})} \left(\frac{\psi_{ii}^2}{2}\right)^{\frac{\delta+\nu_i}{2}-1} \exp -\frac{1}{2}\psi_{ii}^2 \\ & \prod_{(i,j)\in\mathcal{V}, i\neq j} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}\psi_{ij}^2 \quad \prod_{i=1}^{p} d(\psi_{ii})^2 \prod_{(i,j)\in\mathcal{V}, i\neq j} d\psi_{ij}. \end{aligned} \tag{38}$$

We are now ready to state our main result.

Theorem 1 . *Let G be an arbitrary undirected graph and let $I_G(\delta, D)$ be the normalising constant of the G-Wishart distribution $w_G(\delta, D)$ as defined in (23). Then*

$$I_G(\delta, D) = \prod_{i=1}^{p} 2^{\frac{\delta+\nu_i}{2}} (2\pi)^{\frac{\nu_i}{2}} \Gamma(\frac{\delta+\nu_i}{2}) (t_{ii}^2)^{\frac{\delta+b_i-1}{2}} E\{f_T(\psi^{\mathcal{V}})\}, \tag{39}$$

where

$$f_T(\psi^{\mathcal{V}}) = \exp -\frac{1}{2} \sum_{(i,j)\in\bar{\mathcal{V}}} \psi_{ij}^2, \tag{40}$$

and the $\psi_{i,j}, (i,j) \in \bar{\mathcal{V}}$, are well-defined functions of $\psi_{ij}, (i,j) \in \mathcal{V}$, as given in (31) and (32), and where the expectation in (39) is taken with respect to the distribution with density equal to the product of independent chi-squared distributions with $\delta + \nu_i$ degrees *of freedom and standard normal distributions. More precisely*

$$\psi_{ii} \sim \sqrt{\chi^2}_{\delta+\nu_i}, i = 1, \ldots, p, \tag{41}$$

$$\psi_{ij} \sim N(0,1), (i,j) \in \mathcal{V}, i \neq j \ , \tag{42}$$

and where the $\psi_{ii}, i = 1, \ldots, p$ and the $\psi_{ij}, (ij) \in \mathcal{V}$ are mutually independent.

Proof. The proof follows immediately from the expression (38) of $I_G(\delta, D)$. □

According to the Law of Large Numbers, we are going to estimate $I_G(\delta, D)$ by

$$\frac{1}{N}\sum_{l=1}^{N} f(\psi_l^{\mathcal{V}}), \tag{43}$$

where N is a large integer and $(\psi_l^{\mathcal{V}}, l = 1, \ldots, N)$ is a random sample of values of $\psi^{\mathcal{V}}$ from the distribution given in (41) and (42) above.

Our method will be illustrated by an example and compared to Roverato's method in §4.4 below.

4.2 The algorithm

Given an arbitrary graph G and given δ and D, in order to compute the normalising constant of the G-Wishart distribution $W_G(\delta, D)$ or equivalently the generalised inverse Wishart distribution $HIW_G(\delta, D)$, we must first identify the prime components $(P_1, \ldots, P_k)$ of G. Then, for each $G_{P_j}, j = 1, \ldots, k$, consider D_{P_j} and find the Choleski decomposition $D_{P_j}^{-1} = T^T T$. For G_{P_j} given, denote by p the number of vertices and use the same notation as in the previous sections.

Step 1. Create a $p \times p$ triangular matrix $A = (a_{ij})$ such that $a_{ij} = 0$, if $(i,j) \in \bar{\mathcal{V}}$ or if $i = j$ and $a_{ij} = 1$ otherwise.

Step 2. Using A, find ν_i = number of 1's in the ith row of A and k_i = number of 1 's in the ith column of A.

Define $t_{\langle i,j]} = t_{ij}/t_{jj}$. Choose the sample size N and for $k = 1, \dots N$, go through the following steps.

Step 3. Sample the free variables $\psi_{ij}^k,\ (i,j) \in \mathcal{V}$ as follows:

for $i = 1$ to p, $\psi_{ii}^k = \sqrt{U_i}$, where $U_i \sim \chi^2_{\delta+\nu_i}$,

then, for $i = 1$ to $(p-1)$, $j = (i+1)$ to p and $a_{ij} = 1$, $\psi_{ij}^k \sim V_{ij}$ where $V_{ij} \sim N(0,1)$.

Step 4. Evaluate $\psi_{ij}^k,\ (i,j) \in \bar{\mathcal{V}}$, as follows, for $i = 1$ to $(p-1)$ and for $j = (i+1)$ to p:

if $i = 1$ and $a_{ij} = 0$, then $\psi_{ij}^k = -\sum_{k=i}^{j-1} \psi_{ik} t_{\langle kj]}$;

otherwise, $i > 1$ and $a_{ij} = 0$, then

$\psi_{ij}^k = -\sum_{k=i}^{j-1} \psi_{ik} t_{\langle kj]} - \sum_{r=1}^{i-1} \left(\frac{\psi_{ri} + \sum_{l=r}^{i-1} \psi_{rl} t_{\langle li]}}{\psi_{ii}} \right) \left(\psi_{rj} + \sum_{l=r}^{j-1} \psi_{rl} t_{\langle lj]} \right)$.

Note that, in Step 4, the values $\psi_{ij}^k,\ (i,j) \in \bar{\mathcal{V}}$ are computed line by line and that therefore, for a given (i,j), all values $\psi_{rs}^k, (r,s) < (i,j)$ are available for computing ψ_{ij}^k.

Step 5. Compute $\exp -\frac{1}{2} \sum_{(i,j)\in\bar{\mathcal{V}}} (\psi_{ij}^k)^2$.

Step 6. Compute

$$\hat{J}_{\delta,T}^{mc} = \frac{1}{N} \sum_{k=1}^{N} \left[\exp -\{ \frac{1}{2} \sum_{(i,j)\in\bar{\mathcal{V}}} (\psi_{ij}^k)^2 \} \right] , \tag{44}$$

and multiply it by

$$C_{\delta,T} = \prod_{i=1}^{p} (2\pi)^{\frac{\nu_i}{2}} 2^{\frac{\delta+\nu_i}{2}} \Gamma(\frac{\delta+\nu_i}{2}) t_{ii}^{\delta+b_i-1} \tag{45}$$

to obtain $\hat{I}_{G_{P_j}}(\delta, D)$.

4.3 Sampling from the posterior distribution of K

From the joint distribution (14) of $(Z^{(1)}, \ldots, Z^{(n)}, K, G)$, it is immediate to see that, for a fixed graph G, the posterior density of K given the data is equal to

$$f(K|Z^{(1)}, \ldots, Z^{(n)}, G) = \frac{|K|^{\frac{\delta+n-2}{2}}}{I_G(\delta+n, D+U)} \exp -\frac{1}{2}\langle K, U+D \rangle. \tag{46}$$

From Theorem 1 and the expression of $I_G(\delta, D)$ given in (38), it is clear that to obtain a sample $K^k, k = 1, \ldots, N$ from the posterior distribution of K, we need only follow the steps below:

Step 1. Follow Steps 1, 2, 3 and 4 of the algorithm in §4.2 above but replacing D by $D+U$; that is $(D+U)^{-1} = T^T T$.

Step 2. Write the matrix ψ^k made up of the free elements $\psi^k_{ij}(i,j) \in \mathcal{V}$, and the non-free elements $\psi^k_{ij}(i,j) \in \bar{\mathcal{V}}$, calculated as in Step 4 of §4.2.

Step 3. Compute the matrix $\phi^k = \psi^k T$.

Step 4. Compute the matrix $K^k = (\phi^k)^T \phi^k$.

The desired sample $K^k, k = 1, \ldots, N$, from the posterior distribution of K is thus obtained and we can use the average of these matrices to obtain an estimate of the precision matrix for the corresponding graphical Gaussian model.

This sampling method is natural in the sense that it parallels what is done when G is complete. Indeed, when G is complete, to generate a Wishart variable from the $W(\delta, D+U)$ distribution, we can generate independent $\psi_{ii}^2, i = 1, \ldots, p$ from the $\chi^2_{\delta-i+1}, i = 1, \ldots, p$, distributions respectively and independent standard normal $\psi_{ij}, 1 \leq i < j \leq p$. Let ψ be the $p \times p$ upper triangular matrix with entries $\psi_{ij}, 1 \leq i \leq j \leq p$. For $(D+U)^{-1} = T^T T$, we then compute $\phi = \psi T$ and $K = \phi^T \phi$ to obtain a value of K from the Wishart $W(\delta, D+U)$ distribution. The process used above to generate a value from the $W_G(\delta, D)$ distribution uses the triangular matrix T in the same way but, of course, only after the calculation of the non free variables $\psi_{ij}, (i,j) \in \bar{\mathcal{V}}$ to be able to complete the matrix ψ.

5 Some numerical results

5.1 The two approximations

In this section, we shall first compute the normalising constant (38) for three prime nondecomposable graphs using our method and Roverato's method. We will then use our method for a model selection problem. According to our method, $I_G(\delta, D)$ is viewed as

$$I_G(\delta, D) = C_{\delta,T} E\{f_T(\psi^{\mathcal{V}})\} , \tag{47}$$

where $C_{\delta,T}$ and $f_T(\psi^{\mathcal{V}})$ are as defined in (45) and (40) respectively and where the expectation is taken with respect to (41) and (42). Our numerical method will therefore give $\widehat{J}^{mc}_{\delta,T}$ as given in (44). We will give this estimate for three different values of T and two different values of δ. We will give it separately from the values of $C_{\delta,T}$ in order to be able better to judge the accuracy of the computation. Indeed, typically, $\widehat{J}^{mc}_{\delta,T}$ is a number between 0 and 1 while $C_{\delta,T}$ is a huge number, as we can see in the tables below.

According to Roverato's method, $I_G(\delta, D)$ is viewed as

$$I_G(\delta, D) = E\{\frac{\prod(\phi_{ii}^2)^{\frac{\nu_i+\delta-1}{2}}}{h(\phi^{\mathcal{V}})} \exp -\frac{1}{2}\langle \phi^T \phi, D\rangle\} = C_{\delta,T} E\{f_D(\phi^{\mathcal{V}})\} \tag{48}$$

where $h(\phi^{\mathcal{V}})$ is the chosen importance sampling distribution. Since the importance sampling distribution is a product of chi-square and $|\nu_i|$-dimensional normal distribution, the constant $C_{\delta,T}$ also appears naturally when applying Roverato's method. As a consequence, we will again give $C_{\delta,T}$ separately from the estimate

$$\widehat{J}^{is}_{\delta,D} = \frac{1}{C_{\delta,T}} \frac{1}{N} \sum_{l=1}^{N} \frac{\prod\{(\phi_{ii}^2)^l\}^{\frac{\nu_i+\delta-1}{2}}}{h\{(\phi^{\mathcal{V}})^l\}} \exp -\frac{1}{2}\langle (\phi^l)^T(\phi^l), D\rangle, \tag{49}$$

where $(\phi)^l, l = 1, \ldots, N$ are the different values obtained for ϕ by sampling from the importance sampling distribution $h(\phi^{\mathcal{V}})$.

The values of $\widehat{J}^{mc}_{\delta,D}$ and $\widehat{J}^{is}_{\delta,D}$, the estimates of $E\{f_T(\psi^{\mathcal{V}})\}$ and $E\{f_D(\phi^{\mathcal{V}})\}$ in (47) and (48) respectively, along with their estimated standard errors are given for each example, each value of T and each value of δ.

5.2 Three examples

The prime graphs considered for Examples 1, 2 and 3 are given in Figure 1 and the corresponding results are given in Table 1.

The computations are done in C using 15,000 sample points, for two different values of $\delta, \delta = 3$ and $\delta = 10$, and for three different matrices T such that $D^{-1} = T^T T$. The matrices are given in §7.

For the graphs considered, the numerical results found with the two methods, Monte

Carlo and importance sampling, are the same with an accuracy of 10^{-3}, except in three cases, and the maximum difference in these three cases is less than 5×10^{-3}. The estimated standard deviations are also of the same order for either method.

The times for the computations performed on an IBM RS/6000 (model 39H, Power2) Unix server running at 67 MHz, with 128MB of memory, are given in Table 2. Since these times vary only slightly with the different matrices T_1, T_2 or T_3, we give the average time for the three different computations. We should point out that the *is* method requires the computation of T_i^{-1}. The times given in Table 2 do not include the time spent on this computation since it is done once only. The *mc* method does not require T_i^{-1}. The computation times for the *mc* method are consistently shorter than for the *is* method with the difference being greater in the case $p = 5$ than in the case $p = 4$ or $p = 8$. This is easily explained by the fact that in the case $p = 4$ and $p = 8$ the graphs are both cycles and for cycles, the difference between the two methods is the generation of one two-dimensional normal distribution only. In the case $p = 5$, for the *is* method, we must generate one three-dimensional, two two-dimensional and one one-dimensional normals with various means and variances to be computed while for the *mc* methods we need only generate eight one-dimensional standard normals.

Jones et al. (2004) who have used our method on an example with 150 vertices, see §6, kindly monitored the computing times, using our method, for each $I_G(\delta, D)$ for 10653 prime graphs G obtained from a graph found as 'most likely' by successively adding or removing an edge. Among the 10653 graphs, 10385 are cycles. The computations have been performed on a Dual processor AMD AthalonMP CPU with 1024 MB RAM, a much more powerful computer than the one used for the calculations with times given in Table 2. Times vary from 0 to 6 seconds for graphs with p varying roughly from 2 to 25 vertices and are represented in Fig. 2. The curve of these times can be well fitted by a cubic. The computer code in C for our method can be found at http://www.math.yorku.ca/∼massamh. The computer code in C^{++} can be found in the code given by Jones et al. at http://www.isds.duke.edu/∼mw/ggm.html for their search algorithms.

5.3 Model Selection for Fisher's iris Virginica data set

We now apply our computational method for $I_G(\delta, D)$ to perform a model search for part of Fisher's iris virginica dataset. This data set has been analysed in the context of graphical Gaussian models, in Whittaker (1990), Berg & Nielsen (1993) and Roverato (2002). It consists of $n = 50$ four-dimensional vectors $Z_i, i = 1, \ldots, 50$, giving the measurements in millimetres of the sepal length, sepal width, petal length and petal width of 50 flowers. The matrix $U = \sum_{i=1}^{50} Z^{(i)}(Z^{(i)})^T$ is the following symmetric matrix

$$U = \begin{pmatrix} 19.8 & * & * & * \\ 4.6 & 5.1 & * & * \\ 14.85 & 3.5 & 14.9 & * \\ 2.4 & 2.35 & 2.4 & 3.7 \end{pmatrix}.$$

Since there are four vertices, there are 64 possible graphical Gaussian models. The posterior probability for $G_j, j = 1, \ldots, 64$, given the data U are as given in (16).

Of the sixty four posterior probabilities, sixteen are nonnegligible and are given, together with the corresponding graphs, in Fig. 3. All computations have been done using S-plus with a random sample of size 15,000 in order to obtain an estimated standard error for $J^{mc}_{\delta,T}$ of approximately 10^{-3}. The probability values obtained with our method for the most likely models with nonnegligible positive probability are, within some differences of the order of 10^{-3} the same as those found by Roverato. The most likely model is that corresponding to the nondecomposable four-cycle.

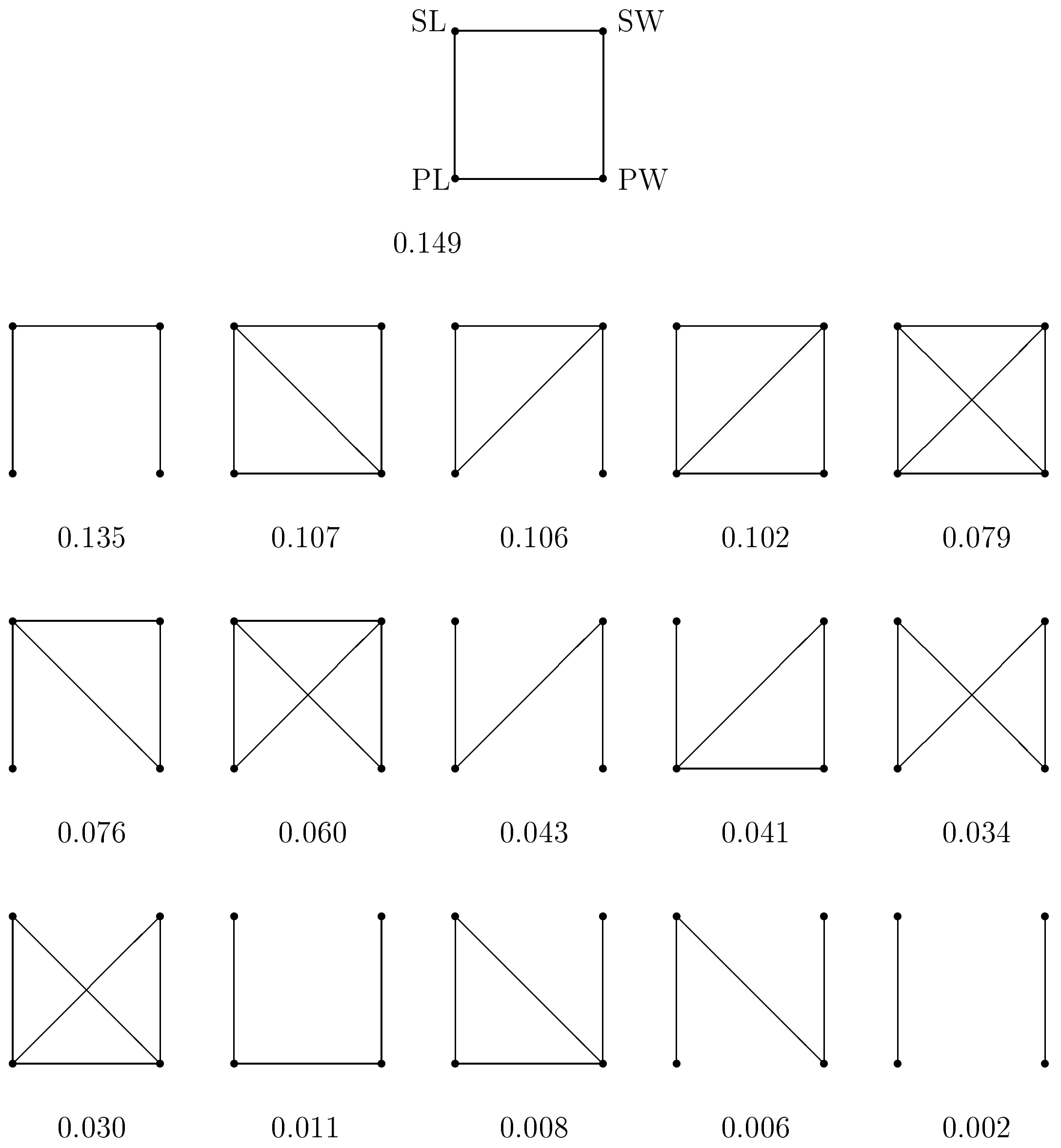

Figure 3: Iris Virginica data. Sixteen most likely models with their associated probabilities $p(G_j|U)$.

6 Applications to high-dimensional data analysis

The method presented in this paper gives a 'natural' way to sample from the G-Wishart distribution and an exact sampling distribution for the computation of $I_G(\delta, D)$. Jones et al. (2004) have used it to analyse a large dataset coming from a gene expression experiment with 150 variables, in preference to other methods precisely because the integral is evaluated by direct Monte Carlo sampling; indeed the variance of the estimation of $I_G(\delta, D)$ is one of their main concerns and according to a private communication from B. Jones, the Monte Carlo method will avoid the variation problems that can occur if the importance sampling distribution does not have significant mass in all locations where the natural target does. Roverato's importance sampling method was not designed specifically for variance reduction of the estimate and might therefore encounter this sort of problem.

Our method allows a model search over a large space of graphical Gaussian models with underlying graphs G that are decomposable or not, using the conjugate prior for the precision parameter. Giudici & Green (1999) developed a reversible jump Markov chain Monte Carlo method for searching over the restricted space of decomposable models only, also using the conjugate prior. More recently Wong et al. (2003) gave a Markov chain Monte Carlo method to estimate the covariance matrix for decomposable or nondecomposable Gaussian models. They do not use the conjugate prior on K. They parameterise the precision matrix by the partial correlation matrix C and a diagonal matrix T such that $K = TCT$ but they are then forced into approximations for some normalising constants.

7 Appendix. Matrices used in §5

The matrices used in §5.1 for $p = 4$ are

$$T_1 = \begin{pmatrix} 8 & 6 & 8 & 0 \\ 0 & 3 & -16 & 2 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}, T_2 = \begin{pmatrix} 4 & 4 & 6 & 0 \\ 0 & 4 & -6 & 6 \\ 0 & 0 & 1 & 7 \\ 0 & 0 & 0 & 2 \end{pmatrix}, T_3 = \begin{pmatrix} 6 & 9 & 4 & 0 \\ 0 & 6 & -6 & 10 \\ 0 & 0 & 7 & 8 \\ 0 & 0 & 0 & 10 \end{pmatrix}$$

For $p = 5$, the matrices are

$$T_1 = \begin{pmatrix} 5 & 10 & 6 & 0 & 7 \\ 0 & 4 & -15 & -1 & 3 \\ 0 & 0 & 10 & 1 & 3 \\ 0 & 0 & 0 & 10 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, T_2 = \begin{pmatrix} 9 & 9 & 7 & 0 & 9 \\ 0 & 3 & -21 & 7 & 4 \\ 0 & 0 & 10 & 10 & 5 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}, T_3 = \begin{pmatrix} 10 & 2 & 1 & 0 & 3 \\ 0 & 2 & -1 & 1 & 4 \\ 0 & 0 & 5 & 2 & 4 \\ 0 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

For $p = 8$, it is more convenient to give the matrices D_i:

$$D_1 = \begin{pmatrix} 6 & 4 & 1 & 0 & 0 & 0 & 0 & 0 \\ 4 & 17 & 0 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 10 & 0 & 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 15 & 0 & 10 & 0 & 0 \\ 0 & 0 & 2 & 0 & 12 & 0 & 9 & 0 \\ 0 & 0 & 0 & 10 & 0 & 17 & 0 & 5 \\ 0 & 0 & 0 & 0 & 9 & 0 & 16 & 6 \\ 0 & 0 & 0 & 0 & 0 & 5 & 6 & 7 \end{pmatrix}, D_2 = \begin{pmatrix} 19 & 7 & 6 & 0 & 0 & 0 & 0 & 0 \\ 7 & 6 & 0 & 2 & 0 & 0 & 0 & 0 \\ 6 & 0 & 11 & 0 & 4 & 0 & 0 & 0 \\ 0 & 2 & 0 & 7 & 0 & 9 & 0 & 0 \\ 0 & 0 & 4 & 0 & 14 & 0 & 3 & 0 \\ 0 & 0 & 0 & 9 & 0 & 20 & 0 & 4 \\ 0 & 0 & 0 & 0 & 3 & 0 & 10 & 1 \\ 0 & 0 & 0 & 0 & 0 & 4 & 1 & 11 \end{pmatrix}, D_3 = \begin{pmatrix} 13 & 4 & 8 & 0 & 0 & 0 & 0 & 0 \\ 4 & 7 & 0 & 1 & 0 & 0 & 0 & 0 \\ 8 & 0 & 8 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 11 & 0 & 6 & 0 & 0 \\ 0 & 0 & 1 & 0 & 11 & 0 & 3 & 0 \\ 0 & 0 & 0 & 6 & 0 & 15 & 0 & 5 \\ 0 & 0 & 0 & 0 & 3 & 0 & 11 & 4 \\ 0 & 0 & 0 & 0 & 0 & 5 & 4 & 11 \end{pmatrix}$$

8 References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag.

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., Giudici, P. & Roberts, G. (2003). An algorithm for sampling from conditional Wishart with application to non-decomposable Gaussian graphical models. *Sankhyā* A **1**, 1-15.

Dempster, A. P. (1972). Covariance selection. *Biometrika* **28**, 157-75.

Diaconis, P. & Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269-81.

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

Giudici, P. & Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785-801.

Jones, B., Carvalho, C., Dobra, A., Hans, C. & West, M. (2003). Stochastic computation for high dimensional graphical models. *submitted.*

Available at ftp://ftp.isds.duke.edu/pub/WorkingPapers/04-01.pdf.

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

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* New York: Wiley.

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

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

Tanner, M.A. (1993), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* New York: Springer-Verlag.

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

Wong, F., Carter, C.K. & Kohn, R. (2004). Efficient estimation of covariance selection models. *Biometrika* **90**, 809-30.

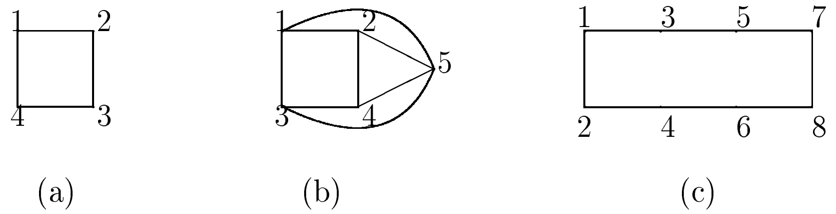

Figure 1: (a) a 4-cycle, (b) a nondecomposable graph on 5 vertices, (c) an 8-cycle

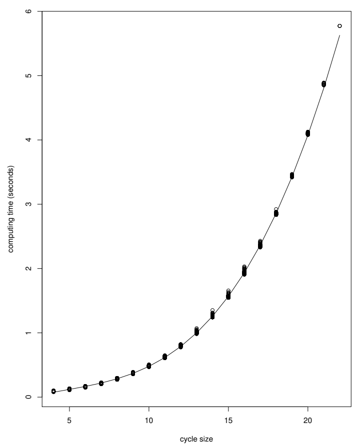

Figure 2: Computation time in C^{++} versus p, the number of vertices.

	$\delta = 3$			$\delta = 10$		
	$\widehat{J}^{mc}_{\delta,D}$	$\widehat{J}^{is}_{\delta,D}$	$C_{\delta,T_i}, i = 1,2,3$	$\widehat{J}^{mc}_{\delta,D}$	$\widehat{J}^{is}_{\delta,D}$	$C_{\delta,T_i}, i = 1,2,3$
Example 1						
T_1	0.11976	0.12183	5.098909×10^{16}	0.12215	0.12162	2.705185×10^{45}
	(0.00197)	(0.00199)		(0.00198)	(0.00197)	
T_2	0.01696	0.01667	3.995128×10^{11}	0.01648	0.01917	1.506348×10^{33}
	(0.00076)	(0.00073)		(0.00073)	(0.00080)	
T_3	0.22239	0.22478	1.209995×10^{21}	0.22933	0.23090	8.56904×10^{55}
	(0.00229)	(0.00231)		(0.00230)	(0.00230)	
Example 2						
T_1	0.18562	0.18562	9.048816×10^{26}	0.19683	0.19610	3.56968×10^{64}
	(0.00254)	(0.00258)		(0.00260)	(0.00260)	
T_2	0.04747	0.04628	1.402279×10^{30}	0.04745	0.04838	5.78653×10^{70}
	(0.00123)	(0.00120)		(0.00122)	(0.00124)	
T_3	0.62453	0.63007	1.643295×10^{28}	0.65021	0.65140	5.297726×10^{66}
	(0.00277)	(0.00276)		(0.00259)	(0.00258)	
Example 3						
T_1	0.01672	0.01729	1.715533×10^{25}	0.01908	0.01898	2.400031×10^{71}
	(0.00049)	(0.00051)		(0.00053)	(0.00054)	
T_2	0.07823	0.07792	7.692845×10^{25}	0.08433	0.08529	8.795647×10^{72}
	(0.00112)	(0.00116)		(0.00110)	(0.00116)	
T_3	0.18666	0.18748	9.982586×10^{25}	0.21026	0.20903	1.643775×10^{73}
	(0.00184)	(0.00191)		(0.00186)	(0.00194)	

Table 1: Examples 1, 2 and 3. Estimates $\widehat{J}^{mc}_{\delta,D}$ and $\widehat{J}^{is}_{\delta,D}$ with their estimated standard errors in parentheses and the constant C_{δ,T_i}, for $\delta = 3$ and $\delta = 10$

	$\delta = 3$		$\delta = 10$	
	mc	is	mc	is
p=4	24.45	26.25	58.18	60.62
p=5	37.92	43.50	80.07	85.40
p=8	51.67	51.92	118.69	122.11

Table 2: Computing times in seconds for the three examples, averaged over $T_i, i = 1, 2, 3$