 Vol 29: 391–411, 2002

Hyper Inverse Wishart Distribution for Non-decomposable Graphs and its Application to Bayesian Inference for Gaussian Graphical Models

ALBERTO ROVERATO

University of Modena and Reggio Emilia

ABSTRACT. While conjugate Bayesian inference in decomposable Gaussian graphical models is largely solved, the non-decomposable case still poses difficulties concerned with the specification of suitable priors and the evaluation of normalizing constants. In this paper we derive the DY-conjugate prior (Diaconis & Ylvisaker, 1979) for non-decomposable models and show that it can be regarded as a generalization to an arbitrary graph *G* of the hyper inverse Wishart distribution (Dawid & Lauritzen, 1993). In particular, if *G* is an incomplete prime graph it constitutes a non-trivial generalization of the inverse Wishart distribution. Inference based on marginal likelihood requires the evaluation of a normalizing constant and we propose an importance sampling algorithm for its computation. Examples of structural learning involving non-decomposable models are given. In order to deal efficiently with the set of all positive definite matrices with non-decomposable zero-pattern we introduce the operation of triangular completion of an incomplete triangular matrix. Such a device turns out to be extremely useful both in the proof of theoretical results and in the implementation of the Monte Carlo procedure.

Key words: Cholesky decomposition, conjugate distribution, Gaussian graphical model, importance sampling, hyper inverse Wishart distribution, matrix completion, non-decomposable graph, normalizing constant

1. Introduction

Bayesian analysis of undirected Gaussian graphical models has been considered in the literature mainly for the subclass of models with decomposable graph (Dawid & Lauritzen, 1993; Cowell *et al.*, 1999; Giudici & Green, 1999).

An arbitrary graph G can be successively decomposed into its prime components $(P_1, \ldots, P_k)$. Since the prime components of a decomposable graph G are all complete, the analysis of a graphical model with decomposable graph may be reduced to the analysis of saturated marginal models corresponding to the cliques of G (see Lauritzen, 1996).

The inverse Wishart distribution is conjugate for the parameter Σ of the saturated model and, for the analysis of decomposable Gaussian graphical models, Dawid & Lauritzen (1993) introduced a generalization of such a distribution and named it hyper inverse Wishart. The characterizing features of this distribution are the hyper-Markov property and its clique-marginals that are all inverse Wishart, but Dawid & Lauritzen (1993) also showed that it satisfies the strong hyper-Markov property that allows the local computation of several statistical quantities. Furthermore, Bjerg & Nielsen (1993) showed that the hyper inverse Wishart distribution is the DY-conjugate prior for the moment parameter of the model (Diaconis & Ylvisaker, 1979).

For a model selection procedure, however, it is natural to choose candidates from the set of all graphs and a restriction to the subset of decomposable models is somewhat artificial. An example where a non-decomposable graph like that in Fig. 1 might well

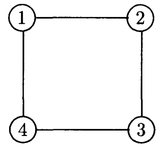

Fig. 1. Chordless four-cycle.

describe the relations between variables is shown in Table 1. These data are reproduced in several texts; for an analysis based on graphical models see Whittaker (1990) and Bjerg & Nielsen (1993).

This paper is concerned with the Bayesian analysis of Gaussian graphical models with non-decomposable graph G. A central role is played by the incomplete prime components of G. The set of all prime graphs on p vertices includes the complete graph but, if $p \geq 4$, it also includes incomplete graphs; the chordless four-cycle in Fig. 1 is the smallest incomplete prime graph (for a discussion on its interpretation see Cox & Wermuth, 2000).

We derive the DY-conjugate prior for an arbitrary graph G and call it a hyper inverse Wishart because it satisfies the strong hyper-Markov property and its marginal distributions corresponding to the complete prime components of G are inverse Wishart. Of interest are the marginal distributions for the incomplete prime components of G that can be regarded as a generalization of the inverse Wishart distribution to the set of all prime graphs.

A marginal likelihood approach to structural learning requires the evaluation of a normalizing constant. We propose an important sampling procedure for its computation in the case where the prior distribution is hyper inverse Wishart. For an efficient implementation of the procedure, a strategy to deal efficiently with the set of all positive definite matrices with non-decomposable zero-pattern is required. We solve this problem by considering the Cholesky decomposition of Σ^{-1} and introducing the operation of triangular completion of an incomplete triangular matrix. This paper relies heavily on this idea, which is also used in the proof of theoretical results. Related works on the triangular decomposition of either the covariance matrix or its inverse include Massam & Neher (1997, 1998), Pourahmadi (1999), Roverato (2000), Wermuth (1980) and Wermuth & Cox (1998, 2000).

The notation, including the definitions of incomplete triangular matrix and triangular completion, is presented in section 2. The general theory relating to Gaussian graphical models and the hyper inverse Wishart distribution is presented in section 3. In section 4 we derive the DY-conjugate prior for non-decomposable models and investigate its properties. The proposed importance distribution and a discussion on the performance of the Monte Carlo procedure is given in section 5. Structural learning for the data in Table 1 is discussed in section 6. Finally, in appendix A we present a detailed description of the technical aspects concerning the triangular completion operation.

Table 1. *Fisher's iris data. Summary statistics about four variables for 50 flowers from iris-Virginica species. Variances (main diagonal), covariances (lower triangle), partial correlations given the other two variables (upper triangle)*

	SL	SW	PL	PW
Sepal length	0.396	0.269	0.838	−0.125
Sepal width	0.092	0.102	−0.076	0.484
Petal length	0.297	0.070	0.298	0.180
Petal width	0.048	0.047	0.048	0.074

2. Notation

Unlike decomposable models, in the non-decomposable case the analysis of graphical models cannot be reduced to the analysis of saturated marginal models but requires handling non-saturated irreducible models explicitly. To deal with the consequent notational difficulties, Roverato & Whittaker (1998) proposed the use of an edge set indexing notation to be applied to Isserlis matrices, incomplete symmetric matrices and matrix completion. In this section we review the main features of such a framework, introduce the graph theory notation used in this paper and extend the idea of incomplete matrix and matrix completion to the set of upper triangular matrices. A full account of graph theory requisite for graphical models can be found in Cowell *et al.* (1999), where definitions of subgraphs, separations, decompositions, decomposable graphs, complete subsets, cliques, perfect sequences and perfect numberings are given.

2.1. Incomplete matrices and matrix completion

Let V be a finite set with $|V| = p$, and let Γ be a $p \times p$ matrix. The rows and columns of Γ are indexed by the elements of V, so that Γ itself is indexed by $V \times V$. If $V = \{1, \ldots, p\}$, $\Gamma = \{\gamma_{ij}\}$ is indexed by row and column numbers.

We denote an undirected graph by $G = (V, \mathscr{V})$, where V is the vertex set and $\mathscr{V}$ is the set of edges. We use the convention that for all $i \in V$ the pair (i, i) is included in the edge set $\mathscr{V}$ and that if $(i, j) \in \mathscr{V}$ then $i \leq j$. Thus, $\mathscr{W} = \{(i, j) : i, j \in V, i \leq j\}$ is the edge set of the complete graph and we denote by $\bar{\mathscr{V}} = \mathscr{W} \backslash \mathscr{V}$ the set of edges not in G. Similarly, for a non-empty subset A of V the edge set of the induced subgraph G_A is denoted by $\mathscr{A} = \mathscr{V} \cap (A \times A)$.

Example 1. For the graph $G = (V, \mathscr{V})$ in Fig. 1, the vertex and edge sets are $V = \{1, 2, 3, 4\}$ and $\mathscr{V} = \{(1, 1), (2, 2), (3, 3), (4, 4), (1, 2), (1, 4), (2, 3), (3, 4)\}$ respectively, while $\bar{\mathscr{V}} = \{(1, 3), \ (2, 4)\}$. For $A = \{1, 3\}$, $\mathscr{A} = \{(1, 1), (3, 3)\}$ and for $B = \{1, 2\}$, G_B is complete so that $\mathscr{B} = \{(1, 1), (2, 2), (1, 2)\}$.

Vertices i and j, with $i < j$, are said to be neighbours if $(i, j) \in \mathscr{V}$. The boundary bd(A) of a subset A of V is the set of vertices in $V \backslash A$ that are neighbours of vertices in A. A central notion here is that of collapsibility: a graph $G = (V, \mathscr{V})$ is said to be collapsible onto $B \subseteq V$ if every connected component of $V \backslash B$ has complete boundary in G.

Let M^+ denote the set of all $|V| \times |V|$ positive definite matrices. For a set $\mathscr{C} \subseteq \mathscr{W}$ the $\mathscr{C}$-incomplete symmetric matrix $\Gamma^{\mathscr{C}}$ is defined as the symmetrized matrix indexed by $V \times V$ with elements $\{\gamma_{ij}\}$ for all $(i, j) \in \mathscr{C}$, and with the remaining elements unspecified.

Example 2. For the graph $G = (V, \mathscr{V})$ in Fig. 1, the $\mathscr{V}$-incomplete symmetric matrix and its submatrix induced by $A = \{1, 3\}$ are respectively

$$\Gamma^{\mathscr{V}} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & * & \gamma_{14} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & * \\ * & \gamma_{32} & \gamma_{33} & \gamma_{34} \\ \gamma_{41} & * & \gamma_{43} & \gamma_{44} \end{pmatrix} \quad \text{and} \quad \Gamma^{\mathscr{V}}_{AA} = \Gamma^{\mathscr{A}}_{AA} = \begin{pmatrix} \gamma_{11} & * \\ * & \gamma_{33} \end{pmatrix},$$

where asterisks denote unspecified elements.

If it is possible to fill an incomplete symmetric matrix $\Gamma^{\mathscr{C}}$ to obtain a (full) positive definite matrix we say that $\Gamma^{\mathscr{C}}$ admits a positive completion. For an undirected graph $G = (V, \mathscr{V})$, we denote by $M_*^+(G)$ the set of all $\mathscr{V}$-incomplete symmetric matrices $\Gamma^{\mathscr{V}}$ that admit positive completion and by $M^+(G)$ the subset of all positive definite matrices Γ^{-1} satisfying $\{\Gamma^{-1}\}_{ij} = 0$ whenever $(i, j) \in \bar{\mathscr{V}}$. We say that Γ is the PD-completion of $\Gamma^{\mathscr{V}} \in M_*^+(G)$ if it is the unique positive definite matrix such that $(\Gamma)^{\mathscr{V}} = \Gamma^{\mathscr{V}}$ and $\Gamma^{-1} \in M^{+}(G)$. Grone *et al.* (1984) proved the existence and uniqueness of such matrix, showing in this way that the set $M_{*}^{+}(G)$ is isomorphic with the set $M^{+}(G)$ through the PD-completion operation:

$$\begin{aligned} &M_{*}^{+}(G) \leftrightarrow M^{+}(G) \\ &\Gamma^{\mathscr{V}} \leftrightarrow \Gamma^{-1}. \end{aligned}$$

The Isserlis matrix of a positive definite matrix Γ, $\mathrm{Iss}(\Gamma)$, (Isserlis, 1918; Roverato & Whittaker, 1998) is the symmetric matrix indexed by $\mathscr{W} \times \mathscr{W}$ with elements $\{\mathrm{Iss}(\Gamma)\}_{(i,j),(r,s)} = \gamma_{ir}\gamma_{js} + \gamma_{is}\gamma_{jr}$. The determinant of the Isserlis matrix is $|\mathrm{Iss}(\Gamma)| = 2^{p}|\Gamma|^{p+1}$. The edge set of the undirected graph $G = (V, \mathscr{V})$ identifies the submatrix $\mathrm{Iss}(\Gamma)_{\mathscr{V}\mathscr{V}}$ of $\mathrm{Iss}(\Gamma)$ and the non-empty subset $A \subseteq V$ identifies the submatrix $\mathrm{Iss}(\Gamma)_{\mathscr{A}\mathscr{A}}$ through the edge set of the corresponding induced subgraph.

For a graph $G = (V, \mathscr{V})$, the PD-completion Γ of $\Gamma^{\mathscr{V}} \in M_{*}^{+}(G)$ satisfies useful properties related with the decompositions of G (see Lauritzen, 1996, lem. 5.5) and Roverato & Whittaker (1998) showed that similar properties hold for $\mathrm{Iss}(\Gamma)_{\mathscr{V}\mathscr{V}}$.

We consider now the set $M^{\triangleleft}$ of all $|V| \times |V|$ upper triangular matrices with positive diagonal and, in parallel with symmetric incomplete matrices and positive definite completion, we define triangular incomplete matrices and introduce the concept of triangular completion.

For a set $\mathscr{C} \subseteq \mathscr{W}$ we define the $\mathscr{C}$-incomplete triangular matrix $\Phi^{\mathscr{C}}$ as the upper triangular matrix indexed by $V \times V$ with elements $\{\phi_{ij}\}$ for all $(i,j) \in \mathscr{C}$, and with the remaining elements unspecified.

Example 3. For the graph $G = (V, \mathscr{V})$ in Fig. 1, the $\mathscr{V}$-incomplete triangular matrix and its submatrix induced by $A = \{1, 3\}$ are respectively

$$\Phi^{\mathscr{V}} = \begin{pmatrix} \phi_{11} & \phi_{12} & * & \phi_{14} \\ 0 & \phi_{22} & \phi_{23} & * \\ 0 & 0 & \phi_{33} & \phi_{34} \\ 0 & 0 & 0 & \phi_{44} \end{pmatrix} \quad \text{and} \quad \Phi_{AA}^{\mathscr{V}} = \Phi_{AA}^{\mathscr{A}} = \begin{pmatrix} \phi_{11} & * \\ 0 & \phi_{33} \end{pmatrix}.$$

We denote by $M_{*}^{\triangleleft}(G)$ the set of all $\mathscr{V}$-incomplete triangular matrices with positive diagonal. It is straightforward to see that any completion Φ of $\Phi^{\mathscr{V}} \in M_{*}^{\triangleleft}(G)$ satisfies the relation $\Phi^{\top}\Phi \in M^{+}$.

Assume now that the ordering of the elements of V is fixed and that the rows and columns of Γ are ordered accordingly. The factorization $\Gamma^{-1} = \Phi^{\top}\Phi$ with $\Gamma^{-1} \in M^{+}$ and $\Phi \in M^{\triangleleft}$ is called the Cholesky decomposition of Γ^{-1} and defines a bijective mapping from M^{+} to $M^{\triangleleft}$; $\pi\colon M^{+} \rightarrow M^{\triangleleft}$. For an undirected graph $G = (V, \mathscr{V})$, the mapping $\pi(\cdot)$ identifies the subset $M^{\triangleleft}(G) = \pi\{M^{+}(G)\}$ of all $\Phi \in M^{\triangleleft}$ such that $\Phi^{\top}\Phi \in M^{+}(G)$. We say that Φ is the T-completion of $\Phi^{\mathscr{V}} \in M_{*}^{\triangleleft}(G)$ if it is the unique upper triangular matrix such that $(\Phi)^{\mathscr{V}} = \Phi^{\mathscr{V}}$ and $\Phi \in M^{\triangleleft}(G)$. For any $\Phi^{\mathscr{V}} \in M_{*}^{\triangleleft}(G)$ the T-completion exists and is unique (see appendix A). As a consequence, similarly to PD-completion, the following correspondence holds:

$$\begin{aligned} &M_{*}^{\triangleleft}(G) \leftrightarrow M^{+}(G) \\ &\Phi^{\mathscr{V}} \leftrightarrow \Gamma^{-1} = \Phi^{\top}\Phi. \end{aligned}$$

In our perspective, both PD- and T-completion are computational devices for the generation of elements from the set $M^{+}(G)$. In Gaussian graphical models $M^{+}(G)$ represents the canonical parameter space and PD-completion is a well-known technique because it is used in the computation of the maximum likelihood estimate $\hat{\Sigma}$ of the covariance matrix Σ (Speed & Kiivery, 1986). However, when G is non-decomposable PD-completion cannot be performed efficiently (see Lauritzen, 1996, p. 134) and this proscribes its use in Bayesian computational intensive procedures where an efficient exploration of the parameter space $M^{+}(G)$ is required. T-completion cannot be applied to the computation of maximum likelihood estimates but turns out to be very useful in Bayesian applications. The next section is concerned with an ordering of the vertex set V that allows an efficient computation of T-completion.

2.2. Vertex ordering for efficient triangular completion

Any undirected graph $G = (V, \mathscr{V})$ can be successively decomposed into its prime components $P_1, \ldots, P_k$. A prime component is a subset of V that induces a, possibly incomplete, maximal subgraph without a complete separator. Any non-decomposable graph has at least one incomplete prime component. Efficient algorithms have been developed to find all prime components of any undirected graph (Leimer, 1993).

The prime components of an undirected graph can be ordered to form a perfect sequence $(P_1, \ldots, P_k)$ of subsets of V and, following standard notation, we denote by $H_j = P_1 \cup \cdots \cup P_j$ the histories, by $R_j = P_j \backslash H_{j-1}$ the residuals and by $S_j = H_{j-1} \cap P_j$ the separators of the sequence. Recall that for all $j = 2, \ldots, k$, the pair (H_{j-1}, P_j) forms a decomposition of G_{H_j} (Lauritzen, 1996, lem. 2.11) and that, for a decomposable graph $G = (V, \mathscr{V})$ a perfect sequence of prime components of G is a perfect sequence $(C_1, \ldots, C_k)$ of cliques of G.

If the vertices V are enumerated by taking first the vertices in R_k, then those in $R_{k-1}, \ldots, R_2, P_1$, we say that the vertices are enumerated along a perfect sequence of prime components of G. Assume now the vertex ordering $V = \{1, 2, \ldots, p\}$ fixed and, for $i = 1, \ldots, p$, let $\{i] = \{1, \ldots, i\}$ and $\langle i\} = \{i+1, \ldots, p\}$ so that $\{p] = V$ and $\langle p\} = \emptyset$. We denote by $v_i = |\text{bd}(i) \cap \langle i\}|$ the number of vertices adjacent to i following i in the given ordering.

In appendix A it is shown that, for an arbitrary graph G, the T-completion Φ of $\Phi^{\mathscr{V}} \in M_*^{\triangleleft}(G)$ can be computed in closed form. Furthermore, if the vertices V are enumerated along a perfect sequence of prime components $(P_1, \ldots, P_k)$ of G, then in T-completion a large amount of local and parallel computations are possible. In particular, the unspecified elements on the same row of $\Phi^{\mathscr{V}}$ can be computed in parallel and, more importantly, the T-completion of $\Phi^{\mathscr{V}}$ only involves the $\mathscr{P}_j$-incomplete triangular submatrices $\Phi^{\mathscr{P}_j}_{P_jP_j}$ corresponding to the incomplete prime components of G. These may be completed locally with respect to G_{P_j} and in parallel (see theorem 1). Throughout this paper, for an undirected graph $G = (V, \mathscr{V})$, we always consider the vertices V enumerated along a perfect sequence of prime components of G and the rows and columns of Γ ordered accordingly. Furthermore, we write $\Gamma^{-1} = \Phi^{\top}\Phi$ to define the Cholesky decomposition of Γ^{-1}.

3. Hyper inverse Wishart distribution

In this section we define Gaussian graphical models and present the theory related to conjugate priors, hyper-Markov properties and hyper-inverse Wishart distribution required in this paper. We refer to Lauritzen (1996) for a full account of Gaussian graphical model theory and to Dawid & Lauritzen (1993) for the theory of hyper-Markov distributions. For a review of standard and DY-conjugate families see Gutiérrez-Peña & Smith (1997).

Let $X \equiv X_V$ be a $|V|$-variate normal random vector with mean equal to zero and covariance matrix $\Sigma = \{\sigma_{ij}\}$. The covariance selection model, or Gaussian graphical model, for X with graph $G = (V, \mathscr{V})$ is specified by assuming that $K = \Sigma^{-1}$ belongs to $M^{+}(G)$ (Dempster, 1972; Wermuth, 1976). A model whose graph is decomposable is itself called decomposable. For a subset $B \subset V$, the submatrix Σ_{BB} is the parameter of the marginal distribution of X_B and, if $A = V \backslash B$,

$$\Sigma_{AA|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \quad \text{and} \quad \Gamma_{A|B} = \Sigma_{AB}\Sigma_{BB}^{-1}$$

are the parameters relative to the conditional distribution of X_A given X_B or, more compactly, of $X_A|X_B$.

In Bayesian analysis of Gaussian graphical models the parameter Σ has a probability distribution associated with it. Let $G = (V, \mathcal{V})$ be a decomposable graph. A distribution for Σ is said to satisfy the weak hyper Markov property with respect to G if for any decomposition (A, B) of G, Σ_{AA} is conditionally independent of Σ_{BB} given $\Sigma_{A\cap B, A\cap B}$; shortly $\Sigma_{AA} \perp\!\!\!\perp \Sigma_{BB}|\Sigma_{A\cap B,A\cap B}$. If the distribution of Σ also satisfies the stronger property $(\Sigma_{AA|B}, \Gamma_{A|B}) \perp\!\!\!\perp \Sigma_{BB}$ then it is said to be strong hyper Markov with respect to G (Dawid & Lauritzen, 1993, defs. 3.5, 3.12).

In an exponential family context, the standard conjugate family of distributions for the parameter θ is the family of distributions with density function $q(\theta)$ proportional to the likelihood function of the model, $q(\theta) \propto L(\theta)$ (see Gutiérrez-Peña & Smith, 1997, def. 3.1). The Gaussian graphical model with graph $G = (V, \mathcal{V})$ is a regular exponential family (Lauritzen, 1996, p. 132) and the density function of the standard conjugate prior for the canonical parameter $K = \{\kappa_{ij}\}$ can be written as

$$q_G(K|\delta, D) \propto |K|^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(KD)\right\}, \quad K \in M^+(G). \tag{1}$$

In the saturated case G is the complete graph and it can be easily checked from (1) that the standard conjugate prior has a Wishart distribution, $K \sim W(\delta + |V| - 1, Q)$, where $\delta > 0$ and $Q = D^{-1}$ is a positive definite matrix. For an arbitrary graph G, (1) is the density of a Wishart distribution conditioned on the event $\{K \in M^+(G)\}$. Such a distribution was introduced by Bjerg & Nielsen (1993) who named it Markov Wishart (see also Giudici, 1996); here, following Roverato (2000) we call it a G-conditional Wishart and write $K \sim W_G(\delta + |V| - 1, Q)$. Note that the matrix hyperparameter of the distribution is an incomplete matrix $D^{\mathcal{V}} \in M_*^+(G)$. This can be seen by recalling that $\text{tr}(KD) = \sum_{i=1}^p \sum_{j=1}^p \kappa_{ij} d_{ij}$ so that, since $\kappa_{ij} = \kappa_{ji} = 0$ for all $(i, j) \in \bar{\mathcal{V}}$, only the specified elements of $D^{\mathcal{V}}$ enter into the specification of the density. For this reason, in the following we can assume without loss of generality that D is a PD-completed matrix so that $Q \in M^+(G)$.

Diaconis & Ylvisaker (1979) defined the DY-conjugate family of distributions relative to the parameter θ of the model as the conjugate family induced on θ by the standard conjugate family of the canonical parameter (see Gutiérrez-Peña & Smith, 1997, def. 3.2). Hence, in the saturated case the DY-conjugate prior for the moment parameter Σ of the Gaussian graphical model has an inverse Wishart distribution. Here, using the parameterization of Dawid (1981), we write $\Sigma \sim IW(\delta, D)$ with density

$$f_V(\Sigma|\delta, D) = h_V(\delta) \frac{|\Sigma|^{-(\delta+2|V|)/2}}{|D|^{-(\delta+|V|-1)/2}} \exp\left\{-\frac{1}{2}\text{tr}(\Sigma^{-1}D)\right\},$$

where

$$h_V(\delta) = \frac{2^{-|V|(\delta+|V|-1)/2}}{\Gamma_{|V|}\{\frac{1}{2}(\delta + |V| - 1)\}} \tag{2}$$

is the normalizing constant (Muirhead, 1982, p. 113). We recall that the multivariate Gamma function has form $\Gamma_{|V|}(y) = \pi^{|V|(|V|-1)/4} \prod_{i=1}^{|V|} \Gamma\{y - (i-1)/2\}$.

For a decomposable graph $G = (V, \mathcal{V})$, Bjerg & Nielsen (1993) showed that the DY-conjugate prior for $\Sigma^{\mathcal{V}}$ is the hyper inverse Wishart distribution, $\Sigma^{\mathcal{V}} \sim \text{HIW}_G(\delta, D^{\mathcal{V}})$ (see also Roverato, 2000). The hyper inverse Wishart distribution was first introduced by

Dawid & Lauritzen (1993) who defined it as the unique hyper-Markov distribution for $\Sigma^{\mathcal{V}}$ with marginals $\Sigma_{C_jC_j} \sim \mathrm{IW}(\delta, D_{C_jC_j})$, for all $j = 1, \ldots, k$. The hyper inverse Wishart distribution can be obtained as hyper-Markov combination of the marginal inverse Wishart distributions for the cliques of G so that its density factorizes as

$$f_G(\Sigma^{\mathcal{V}}|\delta, D^{\mathcal{V}}) = \frac{\prod_{j=1}^{k} f_{C_j}(\Sigma_{C_jC_j}|\delta, D_{C_jC_j})}{\prod_{j=2}^{k} f_{S_j}(\Sigma_{S_jS_j}|\delta, D_{S_jS_j})}. \tag{3}$$

We now summarize some known, although scattered, properties of the hyper inverse Wishart distribution. For a decomposable graph $G = (V, \mathcal{V})$ assume $\Sigma^{\mathcal{V}} \sim \mathrm{HIW}_G(\delta, D^{\mathcal{V}})$ and let $K = \Sigma^{-1}$ where Σ is the PD-completion of $\Sigma^{\mathcal{V}}$. Then, for every pair of subsets B and $A = V \backslash B$ of V such that G is collapsible onto B,

R.1 $\Sigma_{BB}^{\mathcal{B}} \sim \mathrm{HIW}_{G_B}(\delta, D_{BB}^{\mathcal{B}})$, or equivalently $\Sigma_{BB}^{-1} = K_{BB|A} \sim W_{G_B}(\delta + |B| - 1, Q_{BB|A})$;
R.2 $\Sigma_{BB} \perp\!\!\!\perp (\Sigma_{AA|B}, \Gamma_{A|B})$, or equivalently $K_{BB|A} \perp\!\!\!\perp (K_{AA}, K_{AB})$.

Result R.1 derives from the constructing procedure of the hyper inverse Wishart distribution. Property R.2 gives, in an alternative formulation, the strong hyper-Markov property for the distribution of $\Sigma^{\mathcal{V}}$ (Dawid & Lauritzen, 1993, prop. 3.18).

The upper triangular matrix Φ defined by $K = \Phi^{\mathrm{T}}\Phi$ constitutes a useful re-parameterization of the model into variation independent terms (see also Consonni & Veronese, 2001). For the distribution induced by $\Sigma^{\mathcal{V}}$ on Φ the following results are available.

R.3 For every ordering of the vertices V such that $B = \{b, b+1, \ldots, p\}$ the first $|A|$ rows of Φ are independent of the remaining $|B|$ rows; that is $(\Phi_{AA}, \Phi_{AB}) \perp\!\!\!\perp \Phi_{BB}$.

Assume now the vertices V enumerated along a perfect sequence of prime components $(P_1, \ldots, P_k)$ of G. We remark that here we assume G decomposable so that $(P_1, \ldots, P_k)$ is a perfect sequence of cliques of G, but we use this more general notation because it allows an easier comparison with the non-decomposable case in the next section.

R.4 Φ is made up of zeros but for the row-blocks

$$\Phi_{P_1P_1}, \quad (\Phi_{R_2R_2}\Phi_{R_2S_2}), \ldots, (\Phi_{R_kR_k}\Phi_{R_kS_k});$$

R.5 the row-block submatrices in R.4 are mutually independent;
R.6 the rows of every row-block submatrix in R.4 are mutually independent;
R.7 the distribution of the rth row of Φ can be computed locally as follows. The rth row of Φ belongs to one and only one of the row-blocks in R.4. To this corresponds one of the prime components, say P, of G. Thus the rth row of Φ is the ith row of Φ_{PP} and can be partitioned into a zero lower-triangular part, a diagonal element ϕ_{ii} and an upper-triangular part $\phi_{\langle i \rangle} := \phi_{i,\langle i \rangle}^{\mathrm{T}} \equiv (\phi_{i,i+1}, \ldots, \phi_{i,|P|})^{\mathrm{T}}$. Then, if $T = \{t_{ij}\}$ is defined by $D_{PP}^{-1} = T^{\mathrm{T}}T$

$$\phi_{ii} \sim t_{ii}\sqrt{\chi^2_{\delta+\nu_i}} \quad \text{and} \quad \phi_{\langle i \rangle}|\phi_{ii} \sim N\left(\phi_{ii}t_{\langle i \rangle}/t_{ii}, T_{\langle i \rangle \langle i \rangle}^{\mathrm{T}}T_{\langle i \rangle \langle i \rangle}\right),$$

where χ^2 denotes a chi-squared random variable and $t_{\langle i \rangle} := (t_{i,i+1}, \ldots, t_{i,|P|})^{\mathrm{T}}$. We remark that here both the set $\langle i \rangle$ and the index ν_i are considered with respect to G_P.

Result R.3 is equivalent to R.2 by (17). Result R.4 was shown by Wermuth (1980); see also proposition 3 in appendix A. A proof of properties R.5 and R.6 can be found in Massam & Neher (1997); see also Roverato (2000). R.7 can be derived by applying standard results of the Wishart distribution (see Muirhead, 1982, th. 3.2.10) because by (15) and (16) the ith row of Φ_{PP} can be obtained by the Cholesky decomposition of $\Sigma_{PP}^{-1} \sim W(\delta - |P| - 1, D_{PP}^{-1})$.

4. DY-conjugate prior for non-decomposable models

Introducing the hyper inverse Wishart distribution as the DY-conjugate prior for the Gaussian graphical model with decomposable graph G provides a natural way to generalize such a distribution to the non-decomposable case. Here we derive the DY-conjugate prior for the moment parameter of a Gaussian graphical model with arbitrary undirected graph $G = (V, \mathscr{V})$ and consider the extension to the non-decomposable case of its properties. An interesting feature of this distribution is that when G is an incomplete prime graph, it can be regarded as a non-trivial generalization of the inverse Wishart distribution. Otherwise it can be obtained by hyper-Markov combination of the DY-conjugate priors for the prime components of G.

Proposition 1

Let $G = (V, \mathscr{V})$ be an arbitrary undirected graph. The density function of the DY-conjugate prior for the moment parameter $\Sigma^{\mathscr{V}}$ of the Gaussian graphical model with graph G can be written as

$$
\begin{aligned}
f_G(\Sigma^{\mathscr{V}}|\delta, D^{\mathscr{V}}) &= h_G(\delta, D^{\mathscr{V}})2^{p/2}|\mathrm{Iss}(D)_{\mathscr{V}\mathscr{V}}|^{1/2} \\
&\quad \times |\mathrm{Iss}(\Sigma)_{\mathscr{V}\mathscr{V}}|^{-1}\frac{|\Sigma|^{-(\delta-2)/2}}{|D|^{-(\delta-2)/2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}D)\right\}
\end{aligned}
\tag{4}
$$

where Σ and D are the PD-completions of $\Sigma^{\mathscr{V}}$ and $D^{\mathscr{V}}$ respectively, and $h_G(\delta, D^{\mathscr{V}})$ is a normalizing constant.

Proof. By direct application of the definition of DY-conjugate prior, we have to show that $f_G(\Sigma^{\mathscr{V}}|\delta, D^{\mathscr{V}}) \propto q_G(\Sigma^{-1}|\delta, D^{\mathscr{V}})|J(K \to \Sigma^{\mathscr{V}})|$; namely (4) is proportional to the density obtained from (1) via the change of variables between the non-zero entries of K and the specified entries of $\Sigma^{\mathscr{V}}$. The result follows because the Jacobian of such a transformation is $|J(K \to \Sigma^{\mathscr{V}})| = 2^{-p}|\mathrm{Iss}(\Sigma^{-1})_{\mathscr{V}\mathscr{V}|\bar{\mathscr{V}}}| = 2^{p}|\mathrm{Iss}(\Sigma)_{\mathscr{V}\mathscr{V}}|^{-1}$ (Roverato & Whittaker, 1998, eqns (10) and (11)).

It should be noticed that the term $2^{p/2}|\mathrm{Iss}(D)_{\mathscr{V}\mathscr{V}}|^{1/2}/|D|^{-(\delta-2)/2}$ in (4) does not depend on $\Sigma^{\mathscr{V}}$ and could therefore be included into the normalizing constant $h_G(\delta, D^{\mathscr{V}})$. However, in this way, when G is complete $h_G(\delta, D^{\mathscr{V}}) = h_V(\delta)$ in (2). More generally, when G is decomposable, the normalizing constant does not depend on $D^{\mathscr{V}}$, $h_G(\delta, D^{\mathscr{V}}) = h_G(\delta)$, and can be computed by (2) and (3); see also Roverato (2000, prop. 2). In the non-decomposable case the normalizing constant is not known and the next section is concerned with its numerical evaluation.

Proposition 2

If $(P_1, \ldots, P_k)$ is a perfect sequence of prime components of G, then the density (4) can be factorized as

$$
f_G(\Sigma^{\mathscr{V}}|\delta, D^{\mathscr{V}}) = \frac{\prod_{j=1}^{k} f_{G_{P_j}}(\Sigma_{P_jP_j}^{\mathscr{P}_j}|\delta, D_{P_jP_j}^{\mathscr{P}_j})}{\prod_{j=2}^{k} f_{G_{S_j}}(\Sigma_{S_jS_j}|\delta, D_{S_jS_j})}.
\tag{5}
$$

Proof. This factorization can be obtained recursively from (4), by decomposing G_{H_j} with respect to (H_{j-1}, P_j) for $j = k, k-1, \ldots, 2$ and, at every step, factorizing $|\Sigma|^{-(\delta-2)/2}$ and decomposing Σ^{-1} in the trace as in Lauritzen (1996, lem. 5.5) and factorizing both $|\mathrm{Iss}(D)_{\mathcal{V}\mathcal{V}}|$ and $|\mathrm{Iss}(\Sigma)_{\mathcal{V}\mathcal{V}}|$ as in Roverato & Whittaker (1998, eq. (15)).

Proposition 2 above generalizes (3) and shows that the DY-conjugate prior can be constructed by hyper-Markov combination of the DY-conjugate priors for the prime components of the graph that, when complete, have inverse Wishart distribution. For this reason, for an arbitrary graph $G = (V, \mathcal{V})$, we write $\Sigma^{\mathcal{V}} \sim \mathrm{HIW}_G(\delta, D^{\mathcal{V}})$ to say that the distribution of $\Sigma^{\mathcal{V}}$ has density (4), extending in this way the notation of Dawid & Lauritzen (1993).

We now consider results R.1–R.7 given in section 3 for the decomposable case and show that R.1–R.5 hold for an arbitrary graph G. We first give the main result with respect to the G-conditional Wishart distribution and then derive its implications for the hyper inverse Wishart distribution.

Lemma 1

Suppose $K \sim W_G(\delta + |V| - 1, Q)$, where $G = (V, \mathcal{V})$ is an arbitrary undirected graph. For every pair of subsets $B \subseteq V$ and $A = V \backslash B$ of V such that G is collapsible onto B

(i) $K_{BB|A} \sim W_{G_B}(\delta + |B| - 1, Q_{BB|A})$;
(ii) $K_{BB|A} \perp\!\!\!\perp (K_{AA}, K_{AB})$;
(iii) *for every ordering of the vertices V such that $B = \{b, b+1, \ldots, p\}$ it holds that $(\Phi_{AA}, \Phi_{AB}) \perp\!\!\!\perp \Phi_{BB}$ where Φ is defined by $K = \Phi^{\mathrm{T}}\Phi$.*

Proof. See appendix B.

As an immediate consequence of lemma 1 we obtain the generalization to an arbitrary graph G of results R.1 and R.2.

Corollary 1

Let $G = (V, \mathcal{V})$ be an arbitrary undirected graph and $\Sigma^{\mathcal{V}} \sim \mathrm{HIW}_G(\delta, D^{\mathcal{V}})$. Whenever G is collapsible onto $B \subseteq V$

(i) $\Sigma^{\mathcal{B}}_{BB} \sim \mathrm{HIW}_{G_B}(\delta, D^{\mathcal{B}}_{BB})$;
(ii) $(\Gamma_{A|B}, \Sigma_{AA|B}) \perp\!\!\!\perp \Sigma_{BB}$;

where $A = V \backslash B$.

An undirected graph G is collapsible onto any complete subset of V, and in this case result (i) of corollary 1 can be reformulated as follows.

Corollary 2

Let $G = (V, \mathcal{V})$ be an arbitrary undirected graph and $\Sigma^{\mathcal{V}} \sim \mathrm{HIW}_G(\delta, D^{\mathcal{V}})$. For every complete subset $C \subseteq V$, $\Sigma_{CC} \sim \mathrm{IW}(\delta, D_{CC})$.

Note that corollary 1 (ii) can be considered as an extension to the non-decomposable case of the strong hyper-Markov property of Σ. In particular, corollary 1 makes clear that, as well as for the decomposable case, also in our more general formulation the hyper inverse Wishart distribution is characterized by the strong hyper-Markov property and its prime-component marginal distributions. However, in this case it is important to make a distinction between complete and incomplete prime components of G. For every complete prime component C of

G the submatrix Σ_{CC} has an inverse Wishart distribution whereas the distribution of $\Sigma^{\mathscr{P}}_{PP}$, where P is an incomplete prime component, is $\mathrm{HIW}_{G_P}(\delta, D^{\mathscr{P}}_{PP})$ and cannot be further simplified. In fact, the hyper inverse Wishart distribution for incomplete prime graphs is a truly novel distribution. Interestingly, it shares the relevant properties of the inverse Wishart distribution: by corollary 2 all its clique-marginals have inverse Wishart distribution and, by corollary 1, it satisfies R.2. Such a property plays a key role in this context: Dawid & Lauritzen (1993, prop. 3.16) showed that, in the decomposable case, a necessary and sufficient condition for a hyper-Markov distribution for $\Sigma^{\mathscr{V}}$ to be strong hyper-Markov is that all the prime-component marginal distributions of $\Sigma^{\mathscr{V}}$ satisfy R.2. Furthermore, Geiger & Heckerman (2000) showed that, for G complete and $|V| > 2$, R.2 characterizes the inverse Wishart distribution.

Example 4. Let $\Sigma^{\mathscr{V}} \sim \mathrm{HIW}_G(\delta, D^{\mathscr{V}})$ where $G = (V, \mathscr{V})$ is the incomplete prime graph in Fig. 1. G is collapsible onto every subset $\{i,j\}$ such that $(i,j) \in \mathscr{V}$. Hence, for every $(i,j) \in \mathscr{V}$ with $i \neq j$, the parameter of $X_{\{i,j\}}$, $\Sigma_{\{i,j\}\{i,j\}}$, has inverse Wishart distribution and is independent of the parameter of $X_{V\setminus\{i,j\}}|X_{\{i,j\}}$. In other words, the distribution of $\Sigma^{\mathscr{V}}$ is made up of four inverse Wishart marginals combined together so as that the resulting joint distribution satisfies four independence properties; we conjecture that these elements characterize the distribution.

We now consider properties R.3 to R.7. These concern the distribution of Φ and are useful in stochastic simulation to deal efficiently with the hyper inverse Wishart distribution. The generalization to the non-decomposable case of R.3 is given in lemma 1 (iii) (for a connection between row-blocks of $\Phi^{\mathscr{V}}$ and the parameters of the distributions of X_B and $X_A|X_B$, see corollary 3 in appendix A). R.4 derives from proposition 3 and is shown in appendix A. That R.5 holds in the non-decomposable case can be seen by recursively applying lemma 1 (iii) with $A = R_j$ and $B = H_{j-1}$ for $j = k, \ldots, 2$. Properties R.6 and R.7 do not hold in general. They are clearly true for the row-block submatrices corresponding to the complete prime components of G but it is not obvious how they can be generalized to an incomplete prime component, say P, of G because in this case $\Phi_{PP} \neq \Phi^{\mathscr{P}}_{PP}$.

5. Computing the normalizing constant

The normalizing constant of the hyper inverse Wishart distribution is required for the computation of important statistical quantities. However, it is known in closed form only when G is decomposable and its computation by direct Monte Carlo integration is not possible because a procedure for sampling from $\Sigma^{\mathscr{V}} \sim \mathrm{HIW}_G(\delta, D^{\mathscr{V}})$, when $G = (V, \mathscr{V})$ is non-decomposable, is not available. In this section we propose an importance sampler for the numerical evaluation of $h_G(\delta, D^{\mathscr{V}})$. In the simulation we carried out the normalizing constant seemed not to depend on $D^{\mathscr{V}}$.

5.1. Importance sampler

Importance sampling evaluates a constant $C = \int g(y)\,dy$ by drawing N samples $\{y_1, \ldots, y_N\}$ from an importance distribution with density $I(y)$ and computing $\hat{C} = N^{-1}\sum_{i=1}^{N}[g(y_i)/I(y_i)]$; for instance see Tanner (1996, p. 54). The difficulties related to the implementation of this algorithm to our problem are twofold. First, to deal directly with $\Sigma^{\mathscr{V}}$ is inefficient because the evaluation of $f_G(\cdot)$ in (4) at a given point $\Sigma^{\mathscr{V}}$ requires the PD-completion of $\Sigma^{\mathscr{V}}$. Secondly, the shape of the required distribution over its support needs to be closely tracked by the sampler as the high dimensional spaces quickly lead to numerical inaccuracy.

We solve the first problem by implementing the algorithm with respect to the distribution induced by $\Sigma^{\mathscr{V}} \sim \mathrm{HIW}_G(\delta, D^{\mathscr{V}})$ on $\Phi^{\mathscr{V}}$. The advantage is that the evaluation of the density $g_G(\cdot)$ of $\Phi^{\mathscr{V}}$ only requires an efficient T-completion operation. Note that the distribution of $\Phi^{\mathscr{V}}$, for an arbitrary graph $G = (V, \mathscr{V})$, can be derived from that of $K \sim W_G(\delta + |V| - 1, Q)$ by means of a change of variables. Moreover, it can be easily checked from (1) that if $G^* = (V, \mathscr{V}^*)$ is any undirected graph with $\mathscr{V} \subseteq \mathscr{V}^*$, then the distribution of $K \sim W_G(\delta + |V| - 1, Q)$ can be obtained by conditioning $K \sim W_{G^*}(\delta + |V| - 1, Q)$ on the event $\{K \in M^+(G)\}$. Therefore, the distribution of $\Phi^{\mathscr{V}}$ can be derived by means of the following Condition–Reparameterize (hereafter C–R) procedure:

C–R Procedure

(1) Assume $K \sim W_{G^*}(\delta + |V| - 1, Q)$ where $G^* = (V, \mathscr{V}^*)$ is a decomposable graph such that $\mathscr{V} \subseteq \mathscr{V}^*$ and enumerate the vertices V along a perfect sequence of prime components of G^*;

(2a) for $G = (V, \mathscr{V})$ condition K on the event $\{K \in M^+(G)\}$ so that $K \sim W_G(\delta + |V| - 1, Q)$;

(3a) perform the change of variables to $K \to \Phi^{\mathscr{V}}$. (The density $g_G(\cdot)$ of $\Phi^{\mathscr{V}}$ is obtained by replacing K by $\Phi^{\mathrm{T}}\Phi$ in (1) and multiplying it by the Jacobian in proposition 5).

A solution to the second problem is provided by any distribution for $\Phi^{\mathscr{V}} \in M_*^{\triangleleft}(G)$ such that the importance sampling procedure converges efficiently to the required constant. The rate of convergence depends on how closely the importance density $I_G(\cdot)$ mimics $g_G(\cdot)$ and, for this reason, we propose an importance distribution motivated by a constructive procedure that resembles the C–R procedure:

R–C Procedure

(1) Assume $K \sim W_{G^*}(\delta + |V| - 1, Q)$ where $G^* = (V, \mathscr{V}^*)$ is a decomposable graph such that $\mathscr{V} \subseteq \mathscr{V}^*$ and enumerate the vertices V along a perfect sequence of prime components of G^*;

(2b) perform the change of variables to $K \to \Phi^{\mathscr{V}^*}$ (the distribution of $\Phi^{\mathscr{V}^*}$ is given in R.5–R.7);

(3b) for $G = (V, \mathscr{V})$, $I_G(\cdot)$ is the density of $\Phi^{\mathscr{V}}$ obtained by recursively conditioning the rows of $\Phi^{\mathscr{V}^*}$ on the event $\{\Phi^{\mathscr{V}^*} \in M^{\triangleleft}(G)\}$.

Hence, the C–R and the R–C procedures start from a common distribution and perform on this, in reverse order, one conditioning and one reparameterizing step.

We now give a detailed description of the R–C procedure. Since G^* is decomposable and its vertices are enumerated along a perfect sequence of prime components of G^*, it follows that $\Phi^{\mathscr{V}^*}$ in 2b is made up of independent rows whose distribution is known and given in R.7. The density of $\Phi^{\mathscr{V}^*}$ can thus be written as the product of the marginal densities of its rows, and then by further factorizing the density of any row with respect to its diagonal and off-diagonal part as in R.7. The terms of such factorization are then considered one at a time, from the first to the last row of $\Phi^{\mathscr{V}^*}$, and conditioned on the event $\{\Phi^{\mathscr{V}^*} \in M^{\triangleleft}(G)\}$. More precisely, let $(\phi_{rr}, \phi_{\langle r\}})$ and $(\phi^*_{rr}, \phi^*_{\langle r\}})$ denote the rth row of $\Phi^{\mathscr{V}}$ and $\Phi^{\mathscr{V}^*}$ respectively. If the set $\mathscr{V}^* \backslash \mathscr{V}$ does not contain any edge (i, j) with $i = r$, then $(\phi_{rr}, \phi_{\langle r\}}) = (\phi^*_{rr}, \phi^*_{\langle r\}})$ in distribution. Otherwise, $\phi_{\langle r\}}$ has more constrained elements than $\phi^*_{\langle r\}}$, the values of which can be computed by means of the T-completion operation $\Phi_{\{r]\{p]} = \Phi_{\{r]\{p]}(\Phi^{\mathscr{V}}_{\{r]\{p]})$ in (11). In this case, by the conditioning step of the procedure, $\phi_{rr} = \phi^*_{rr}$ in distribution whereas $\phi_{\langle r\}} | \phi_{rr}$ has the normal distribution derived by conditioning $\phi^*_{\langle r\}} | \phi^*_{rr}$ on the given values of the constrained elements.

Hence, the R–C procedure leads to an importance distribution that, exploiting the recursive computation of T-completion jointly with the independence of the rows of $\Phi^{\mathscr{V}^*}$, allows an efficient implementation of the importance sampling procedure. We remark, however, that the used "recursive" conditioning, acting locally on the rows of $\Phi^{\mathscr{V}^*}$, is different from conditioning $\Phi^{\mathscr{V}^*}$ on the event $\{\Phi^{\mathscr{V}^*} \in M^{\triangleleft}(G)\}$ because the latter considers all the rows of $\Phi^{\mathscr{V}^*}$ simultaneously.

The effectiveness of an importance distribution is measured in terms of rate of convergence of the associated importance sampling procedure. On the basis of our experience, the performance of the procedure can be improved adding to the basic R–C procedure one step concerning the diagonal of $\Phi^{\mathscr{V}}$. The conditioning step 3b leaves the distribution of the diagonal of $\Phi^{\mathscr{V}}$ identical to that of $\Phi^{\mathscr{V}^*}$; that is the distribution of the diagonal depends on G^* rather then G. This is clearly a drawback, so that we introduce the following diagonal correction that, in the simulations we carried out, led to a substantial efficiency improvement.

(4b) Correct the degrees of freedom of the chi-squared random variables on the main diagonal of $\Phi^{\mathscr{V}}$ by computing the indexes v_i with respect to G rather than G^*.

The evaluation of the performance of the proposed importance distribution is considered in the next section.

The R–C procedure does not specify how G^* has to be chosen. It is always possible to put G^* equal to the complete graph $(V, \mathscr{W})$; however, we suggest that G^* should be chosen so as that the importance distribution $I_G(\cdot)$ be as close as possible to the required distribution $g_G(\cdot)$. For instance, if G^* is the fill-in decomposable graph described in appendix A, then the importance distribution differs from $g_G(\cdot)$ only on the row-blocks of $\Phi^{\mathscr{V}}$ corresponding to the incomplete prime components of G. Another possible strategy is that of exploiting (5) to independently compute the normalizing constants for each marginal distribution corresponding to every incomplete prime component P with respect to $G_P = (P, \mathscr{P})$ and to choose $G_P^* = (P, \mathscr{P}^*)$ such that $|\mathscr{P}^* \backslash \mathscr{P}|$ is minimized; for a discussion on the triangulation of G_P see Cowell *et al.* (1999, p. 57). In model search, this has the advantage that incomplete prime components common to different graphs need to be considered only once.

We close this section with a remark concerning the hyperparameter space of the hyper inverse Wishart distribution. In the decomposable case the normalizing constant $h_G(\delta)$ is finite for all $\delta > 0$ but, for arbitrary graph $G = (V, \mathscr{V})$, the hyperparameter space $H(G) = \{(\delta, D^{\mathscr{V}}) | \delta > 0, D^{\mathscr{V}} \in M_*^+(G), h_G(\delta, D^{\mathscr{V}}) < \infty\}$ has not been fully identified. Bjerg & Nielsen (1993, p. 87) provided a partial result by showing that $H(G)$ is a convex set. Furthermore, they noticed that a general result for the regular exponential family (Diaconis & Ylvisaker, 1979, th. 1) implies that $(2, +\infty) \times M_*^+(G) \subseteq H(G)$. Hence, in our applications we always set $\delta > 2$.

5.2. Simulations

The performance of the proposed importance sampler is illustrated here with respect to the two graphs in Fig. 2. They are both incomplete prime graphs and the associated hyper inverse Wishart distributions have dimension 16 for graph (a) and 13 for graph (b). Fig. 2 also shows the fill-in decomposable graphs G^* used in the R–C procedures.

The estimated constants are compared with the asymptotic approximation based on Laplace's formulae (see Tierney & Kadane, 1986),

$$\widetilde{h}_G(\delta, D^{\mathscr{V}}) = \exp\left\{-\frac{1}{2}\left[|\mathscr{V}| \log(2\pi\delta') + |V|\{\delta' \log(\delta') + \log(2) - \delta'\}\right]\right\} \tag{6}$$

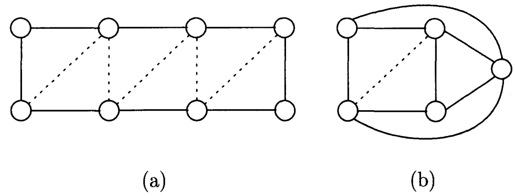

Fig. 2. Non-decomposable graphs and their triangulations. Dashed lines: fill-in edges.

Table 2. *Estimated* $\log\{h_G(\delta, D^{\mathscr{V}})\}$ *for the graph in Fig. 2 (a) (16 parameters). LA: Laplace approximation. I: identity matrix.* D_i *for* $i = 1, \ldots, 4$: *randomly chosen positive definite matrices*

	$\delta = 3$	$\delta = 5$	$\delta = 10$	$\delta = 20$	$\delta = 40$	$\delta = 60$
LA	−13.476	−27.448	−68.653	−176.70	−447.49	−759.98
I	−18.444	−29.530	−69.513	−177.10	−447.68	−760.11
D_1	−18.446	−29.529	−69.511	−177.10	−447.68	−760.11
D_2	−18.443	−29.531	−69.514	−177.10	−447.68	−760.11
D_3	−18.440	−29.533	−69.516	−177.10	−447.68	−760.11
D_4	−18.444	−29.535	−69.514	−177.10	−447.68	−760.11

Table 3. *Estimated* $\log\{h_G(\delta, D^{\mathscr{V}})\}$ *for the graph in Fig. 2 (b) (13 parameters). LA: Laplace approximation. I: identity matrix.* D_i for $i = 1, \ldots, 4$: *randomly chosen positive definite matrices*

	$\delta = 3$	$\delta = 5$	$\delta = 10$	$\delta = 20$	$\delta = 40$	$\delta = 60$
LA	−11.179	−21.559	−48.784	−117.53	−287.89	−483.84
I	−16.360	−23.887	−49.783	−118.00	−288.12	−483.99
D_1	−16.372	−23.905	−49.793	−118.00	−288.12	−483.99
D_2	−16.397	−23.897	−49.804	−118.01	−288.12	−483.99
D_3	−16.387	−23.915	−49.794	−118.01	−288.12	−483.99
D_4	−16.371	−23.892	−49.801	−118.01	−288.12	−483.99

where $\delta' = \delta - 2$. Note that (6) is not a function of $D^{\mathscr{V}}$ and depends on G only through $|V|$ and $|\mathscr{V}|$.

Tables 2 and 3 give the estimated $\log\{h_G(\delta, D^{\mathscr{V}})\}$ for different values of the hyperparameters δ and $D^{\mathscr{V}}$. As mentioned above, these tables provide an empirical evidence that, as well as for the decomposable case, also when G is non-decomposable the normalizing constant does not depend on $D^{\mathscr{V}}$.

Figure 3 describes the evolution of the estimators of $\log\{h_G(3, I)\}$ as the sample size increases. These examples show that our importance sampler works extremely well. For all the simulations the sample size was set to 15 000; nevertheless, for a satisfactory estimate of the constant a much smaller number of iterations seems to suffice. Furthermore, we observed that the convergence rate improves as δ increases.

All the computations were done on a Pentium II-266 using the interprete language *R*. It is remarkable that, despite the lack of efficiency of such a language, the CPU time required by the procedure was always acceptable.

6. Data distribution and application

We now apply the results of this paper to the analysis of the Fisher's iris data in Table 1.

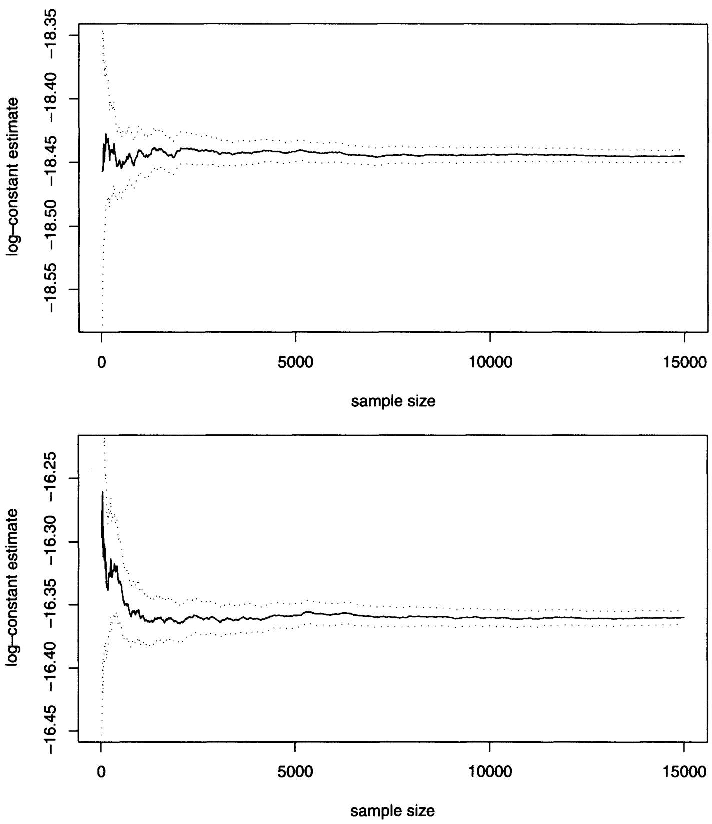

Fig. 3. Convergence of the estimator of $\log\{h_G(3,I)\}$ for the graphs in Fig. 2 (a) and (b). The envelope provides a 95% Monte Carlo confidence interval.

The hyper inverse Wishart distribution is conjugate for the model and, for $G=(V,\mathscr{V})$, if the prior distribution of $\Sigma^{\mathscr{V}}$ is $\mathrm{HIW}_G(\delta, D^{\mathscr{V}})$ and a random sample X of size n has been observed then the posterior is $\mathrm{HIW}_G\{\delta+n,(D+S)^{\mathscr{V}}\}$, where $S=n\hat{\Sigma}$. For G decomposable, the density of the marginal distribution of the data was given by Dawid & Lauritzen (1993, eq. (45)) and, by using (4) this can be generalized to the non-decomposable case as

$$u(x|G)=(2\pi)^{-np/2}\frac{h_G(\delta,D^{\mathscr{V}})|\mathrm{Iss}(D)_{\mathscr{V}\mathscr{V}}|^{1/2}|D|^{(\delta-2)/2}}{h_G\{\delta+n,(D+S)^{\mathscr{V}}\}|\mathrm{Iss}(D+S)_{\mathscr{V}\mathscr{V}}|^{1/2}|(D+S)|^{(\delta+n-2)/2}} \tag{7}$$

where D and $(D+S)$ are the PD-completions of $D^{\mathscr{V}}$ and $(D+S)^{\mathscr{V}}$ respectively.

Table 1 involves four variables. The set of all models for four variables is made up of 64 elements $\{G_1,\ldots,G_{64}\}$, three of which are non-decomposable. Here we use the marginal likelihood $\tilde{L}(G_i)=u(x|G_i)$ to derive the posterior distribution $\Pr(G_i|x)$.

For our analysis we assumed the prior distribution for the models to be constant, $\Pr(G_i)=64^{-1}$ for $i=1,\ldots,64$, so that $\Pr(G_i|x)\propto\tilde{L}(G_i)$. As hyperparameters of the prior distribution for $\Sigma^{\mathscr{V}}$ we set $D=I$ and $\delta=3$.

Figure 4 gives the 16 most probable models; they represent the 25% of all possible models and the sum of their probabilities is 0.987. The model with highest posterior probability is non-decomposable but several models have rather close posterior probabilities. Note that also model number 11 is non-decomposable.

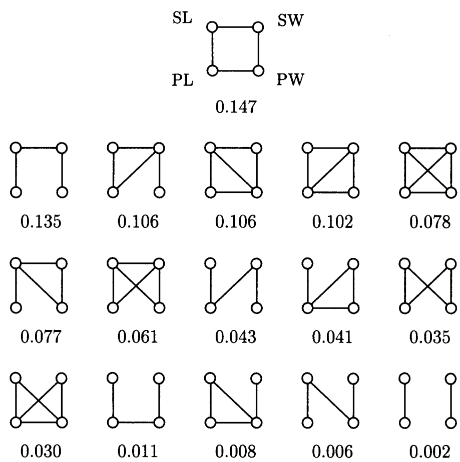

Fig. 4. Fisher's iris data: sixteen most probable models, together with the associated probabilities.

7. Discussion

The main contribution of this paper concerns two distributions: the generalization of the hyper inverse Wishart and the importance sampler. Nevertheless, it is worth pointing out the importance of the theory of T-completion for the development of all the material here presented. In fact, we deem that it might provide a tool with more general applications in the theory of non-decomposable Gaussian graphical models.

We conclude by noting that the assumption, here empirically proved, that $h_G(\delta, D^{\mathcal{V}})$ is not a function of $D^{\mathcal{V}}$ substantially reduces the number of constants to be numerically evaluated in a model search procedure. For example, in the application in section 6 the three non-decomposable models on four variables are all chordless cycles so that, given δ, $h_G(\cdot)$ has the same value for all of them. An analysis of all models for five variables would just require the numerical evaluation of the constant for five graphs: the chordless four-cycle plus four prime graphs on five vertices.

Acknowledgements

The author wishes to thank Professor Steffen Lauritzen and two reviewers for useful suggestions. We have used the free software *R* for calculations. This research was supported by the Italian National Research Council grant number CNRC008704–004.

References

Asmussen, S. & Edwards, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70**, 567–578.

Bjerg, A. M. & Nielsen, T. H. (1993). Modelselektion i kovariansselektionsmodeller, MSc Thesis, Department of Mathematics and Computer Science, Aalborg University, Denmark.

Consonni, G. & Veronese, P. (2001). Conditionally reducible natural exponential families and enriched conjugate priors. *Scand. J. Statist.* **28**, 377–406.

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

Cox, D. R. & Wermuth, N. (2000). On the generation of the chordless four-cycle. *Biometrika* **87**, 206–212.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68**, 265–274.

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

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

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

Geiger, D. & Heckerman, D. (2000). Parameters priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* to appear.

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

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

Grone, R., Johnson, C. R., Sà, E. M. & Wolkowice, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.* **58**, 109–124.

Gutiérrez-Peña, E. & Smith, A. F. M. (1997). Exponential and Bayesian conjugate families: review and extensions. *Test* **6**, 1–90.

Isserlis, L. (1918). On a formula for the product-moment correlation of any order of a normal frequency distribution in any number of variables. *Biometrika* **12**, 134–139.

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

Leimer, H. G. (1993). Optimal decomposition by clique separators. *Discrete Math.* **113**, 99–123.

Massam, H. & Neher, E. (1997). On transformation and determinants of Wishart variables on symmetric cones. *J. Theoret. Probab.* **10**, 867–902.

Massam, H. & Neher, E. (1998). Estimation and testing for lattice conditional independence models on euclidean Jordan algebras. *Ann. Statist.* **26**, 1051–1082.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory.* Wiley, New York.

Paulsen, V. I., Power, S. C. & Smith, R. R. (1989). Schur products and matrix completions. *J. Funct. Anal.* **85**, 151–178.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika* **86**, 677–690.

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

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

Speed, T. P. & Kiiveri, H. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 138–150.

Tanner, M. (1996). *Tools for statistical inference*; 2nd edn. Springer-Verlag, New York.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 82–86.

Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95–108.

Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *J. Amer. Statist. Assoc.* **75**, 963–972.

Wermuth, N. (1989). Modelling effects in multivariate normal distributions. *Methodika* **III**, 74–93.

Wermuth, N. & Cox, D. R. (1998). On association models defined over independence graphs. *Bernoulli* **3**, 477–495.

Wermuth, N. & Cox, D. R. (2000). A sweep operator for triangular matrices and its statistical applications. Research report, ZUMA 00–04.

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

Received June 2000, in final form July 2001

Alberto Roverato, Dipartimento di Economia Politica, University of Modena and Reggio Emilia, Viale J. Berengario, 51, 41100 Modena, Italy.
E-mail: roverato@unimo.it

Appendix

A. Cholesky decomposition of a concentration matrix

In stochastic simulation, when a positive definite matrix K has to be generated the Cholesky decomposition $K = \Phi^{\mathrm{T}}\Phi$ is widely used because the elements of Φ are variation independent.

However, the variation independence property is not retained when, for an undirected graph $G = (V, \mathcal{V})$, K is restricted to belong to $M^+(G)$. Here we consider $K \in M^+(G)$ and partition Φ into $(\Phi^{\mathcal{V}}, \Phi^{\bar{\mathcal{V}}})$ so that Φ is the T-completion of $\Phi^{\mathcal{V}}$, defined is section 2.1. In our approach, we assume that the elements of $\Phi^{\mathcal{V}}$, hereafter referred to as the free entries of Φ, are variation independent, $\Phi^{\mathcal{V}} \in M_*^{\triangleleft}(G)$, and show that the fixed entries $\Phi^{\bar{\mathcal{V}}}$ of Φ are uniquely identified by the constraint $\Phi^{\mathrm{T}}\Phi \in M^+(G)$. As a consequence, the transformation $\Phi^{\mathcal{V}} \to K$, i.e. the transformation between the free entries of Φ and the non-zero entries of $K = \Phi^{\mathrm{T}}\Phi$, defines a bijective mapping between $M_*^{\triangleleft}(G)$ and $M^+(G)$. The material presented in this appendix is concerned with this transformation and is mainly technical. We show how the T-completion of $\Phi^{\mathcal{V}}$ can be used to deal efficiently with the parameter space of an arbitrary Gaussian graphical model, describe the local computation property of T-completion operations and finally, provide the Jacobian of the inverse transformation $K \to \Phi^{\mathcal{V}}$.

We first describe the procedure for computing the T-completion Φ from the incomplete matrix $\Phi^{\mathcal{V}} \in M^+(G)$, where $G = (V, \mathcal{V})$ is an arbitrary undirected graph. In this way the existence and uniqueness of Φ are also shown.

Assume first $\Phi \in M^{\triangleleft}$ and let $K = \Phi^{\mathrm{T}}\Phi$. Because of the triangular form of Φ, for any $(r, s) \in \mathcal{W}$

$$\kappa_{rs} = \sum_{i=1}^{r} \phi_{ir}\phi_{is}, \tag{8}$$

so that κ_{rs} is not a function of any ϕ_{ij} with either $i > r$ or $j > s$ and we can write

$$\kappa_{rs} = \kappa_{rs}(\Phi_{\{r\}\{s\}}), \tag{9}$$

where we recall that $\Phi_{\{r\}\{s\}} = \Phi_{\{1,\dots,r\}\{1,\dots,s\}}$.

Consider now $\Phi^{\mathcal{V}} \in M_*^{\triangleleft}(G)$ and let $(r, s) \in \bar{\mathcal{V}}$, so that ϕ_{rs} is one of the specified entries of $\Phi^{\bar{\mathcal{V}}}$; i.e. one of the fixed entries of Φ to be identified under the constraint $\Phi^{\mathrm{T}}\Phi = K \in M^+(G)$. Since in this case $\kappa_{rs} = 0$, by (8) we have

$$\phi_{rs} = \begin{cases} 0 & \text{for } r = 1 \\ -\frac{1}{\phi_{rr}} \sum\limits_{i=1}^{r-1} \phi_{ir}\phi_{is} & \text{for } r > 1. \end{cases} \tag{10}$$

In (10), ϕ_{rs} is written as a function of both the free elements $\Phi^{\mathcal{V}}_{\{r\}\{s\}}$ and the fixed elements $\Phi^{\bar{\mathcal{V}}}_{\{r-1\}\{s\}}$. Hence, if we order the specified entries of $\Phi^{\bar{\mathcal{V}}}$ according to the rows of $\Phi^{\bar{\mathcal{V}}}$, then ϕ_{rs} is not a function of any of the other fixed elements following it in this sequence. Consequently, by following this ordering, (10) can be used to recursively compute the T-completion of $\Phi^{\mathcal{V}}$. More precisely, the fixed elements ϕ_{rs} with $r = 1$ are all zero and occupy the very first positions in the sequence. Thus, the first element of the sequence with $r > 1$, say ϕ_{rs}, is a function of $\Phi^{\mathcal{V}}_{\{r\}\{s\}}$. Such a function, $\phi_{rs} = \phi_{rs}(\Phi^{\mathcal{V}}_{\{r\}\{s\}})$, is well defined because $\phi_{rr} > 0$ for all $\Phi^{\mathcal{V}} \in M_*^{\triangleleft}(G)$. Similarly, the next element of the sequence, say ϕ_{ij}, is a well defined function possibly involving $\phi_{rs}(\Phi^{\mathcal{V}}_{\{r\}\{s\}})$, of $\Phi^{\mathcal{V}}_{\{i\}\{j\}}$, and so on. We can conclude that every fixed element ϕ_{rs} is uniquely identified as a function of $\Phi^{\mathcal{V}}_{\{r\}\{s\}}$ so that

$$\Phi_{\{r\}\{s\}} = \Phi_{\{r\}\{s\}}\left(\Phi^{\mathcal{V}}_{\{r\}\{s\}}\right) \tag{11}$$

is a one-to-one function available in closed form and, for $r = s = p$

$$\Phi = \Phi(\Phi^{\mathcal{V}}) \tag{12}$$

is well defined for all $\Phi^{\mathcal{V}} \in M_*^{\triangleleft}(G)$ and identifies the unique T-completion of $\Phi^{\mathcal{V}}$.

Example 5. For the graph G in Fig. 1, the T-completion of $\Phi^{\mathscr{V}} \in M_*^{\triangleleft}(G)$ is

$$\Phi(\Phi^{\mathscr{V}}) = \begin{pmatrix} \phi_{11} & \phi_{12} & 0 & \phi_{14} \\ 0 & \phi_{22} & \phi_{23} & -\frac{\phi_{12}\phi_{14}}{\phi_{22}} \\ 0 & 0 & \phi_{33} & \phi_{34} \\ 0 & 0 & 0 & \phi_{44} \end{pmatrix}.$$

The procedure $\Phi^{\mathscr{V}} \to \Phi(\Phi^{\mathscr{V}}) \to \Phi^{\mathrm{T}}\Phi$ provides an efficient way for generating elements from the set $M^+(G)$: an incomplete matrix $\Phi^{\mathscr{V}}$ can be easily constructed since its elements are variation independent and the recursive computation of $\Phi(\Phi^{\mathscr{V}})$ is efficient because the elements in the same row of $\Phi^{\bar{\mathscr{V}}}$ can be computed in parallel. It should also be noticed that the determinant of K can be directly computed from $\Phi^{\mathscr{V}}$, $|K| = \prod_{r=1}^{p} \phi_{rr}^2$, and the trace from Φ, $\mathrm{tr}(K) = \sum_{r=1}^{p}\sum_{s=r}^{p}\phi_{rs}^2$.

Given $K \in M^+(G)$ the matrix Φ defined by $K = \Phi^{\mathrm{T}}\Phi$ is uniquely identified. Nevertheless, it depends on the permutation of the indices V; that is on the permutation of the rows and columns of K. This has also an influence on the functional form of (12) and a different vertex ordering may bring an efficiency improvement in its computation. Wermuth (1980) showed that when G is decomposable, taking a vertex ordering that follows a perfect vertex elimination scheme for G implies $\Phi^{\bar{\mathscr{V}}} = 0$ for all $K \in M^+(G)$. She also showed that this cannot be generalized to non-decomposable graphs (see also Paulsen *et al.*, 1989; Roverato, 2000). Nevertheless, Wermuth's (1980) result can be exploited to show the following.

Proposition 3

For $G = (V, \mathscr{V})$ let Φ be the T-completion of $\Phi^{\mathscr{V}} \in M_*^{\triangleleft}(G)$. If $G^* = (V, \mathscr{V}^*)$ is a decomposable graph with $\mathscr{V} \subseteq \mathscr{V}^*$ and such that the vertices V are enumerated along a perfect sequence of prime components of G^*, then $\Phi^{\bar{\mathscr{V}}^*} = 0$.

Proof. Wermuth (1980) showed that $\Phi^{\bar{\mathscr{V}}^*} = 0$ for all Φ such that $\Phi^{\mathrm{T}}\Phi \in M^+(G^*)$ and the result follows because $M^+(G) \subseteq M^+(G^*)$.

An important property for the analysis of graphical models is that of model collapsibility (Asmussen & Edwards, 1983; Wermuth, 1989). It is strictly related with the collapsibility of G and allows to compute several statistical quantities locally on marginal models. We show now that the T-completion of $\Phi^{\mathscr{V}}$ only requires to consider the incomplete prime components of G and that these can be computed locally and in parallel.

We first show that collapsibility of G implies that T-completion can be performed locally on different row-blocks of the incomplete matrix.

Proposition 4

For $G = (V, \mathscr{V})$ let Φ be the T-completion of $\Phi^{\mathscr{V}} \in M_*^{\triangleleft}(G)$. If G is collapsible onto $B = \{b, b+1, \ldots, p\}$ and $A = V \backslash B$, then

(i) *Φ_{AV} is a one-to-one function of $\Phi^{\mathscr{V}}_{AV}$;*
(ii) *Φ_{BB} is the T-completion of $\Phi^{\mathscr{V}}_{BB} = \Phi^{\mathscr{B}}_{BB}$ with respect to G_B.*

Proof. (i) is always true. In fact, by putting $a = b - 1$ we can write $\Phi_{AV} = \Phi_{\{a\}\{p\}}$ so that by (11) $\Phi_{\{a\}\{p\}} = \Phi_{\{a\}\{p\}}(\Phi^{\mathscr{V}}_{\{a\}\{p\}})$ is a one-to-one function. We now show (ii). It is always true that $\Phi^{\mathscr{V}}_{BB} = \Phi^{\mathscr{B}}_{BB} \in M_*^{\triangleleft}(G_B)$. By the results on parametric collapsibility of Gaussian graphical models (Wermuth, 1989) we can deduce that, if $K = \Phi^{\mathrm{T}}\Phi$, then $K_{BB|A} \in M^+(G_B)$ and the result follows because $K_{BB|A} = \Phi^{\mathrm{T}}_{BB}\Phi_{BB}$; see also (17) below.

Consider now $\Phi^{\mathcal{V}} \in M_*^{\lhd}(G)$ where $G = (V, \mathcal{V})$ is an arbitrary undirected graph such that the vertices V are ordered along a perfect sequence $(P_1, \dots, P_k)$ of prime components of G. In this case the decomposable fill-in augmented graph $G^* = (V, \mathcal{V}^*)$ obtained by completing the prime components of G, that is $(r,s) \in \mathcal{V}^*$ if either $(r,s) \in \mathcal{V}$ or $\{r,s\} \subseteq P_j$ for some $j = 1, \dots, k$, satisfies the conditions of proposition 3. Hence, the T-completion Φ of $\Phi^{\mathcal{V}}$ can be partitioned into $(\Phi^{\mathcal{V}^*}, \Phi^{\bar{\mathcal{V}}^*})$ with $\Phi^{\bar{\mathcal{V}}^*} = 0$ whereas $\Phi^{\mathcal{V}^*}$ is made up of the row-blocks

$$\Phi_{P_1P_1}, \quad (\Phi_{R_2R_2}\Phi_{R_2S_2}), \dots, (\Phi_{R_kR_k}\Phi_{R_kS_k}). \tag{13}$$

It follows that, in order to obtain the T-completion of $\Phi^{\mathcal{V}}$ it is sufficient to derive the unspecified entries of the submatrices

$$\Phi_{P_1P_1}^{\mathcal{P}_1} \quad \text{and, for } j = 2, \dots, k, \quad \Phi_{P_jP_j}^{\mathcal{P}_j} = \begin{pmatrix} \Phi_{R_jR_j} & \Phi_{R_jS_j} \\ 0 & \Phi_{S_jS_j} \end{pmatrix}^{\mathcal{P}_j} \tag{14}$$

which are incomplete only if the corresponding prime components are incomplete. It can also be checked that , as a consequence of the zero structure of Φ (see Roverato, 2000, eq. (14) and (15))

$$\Phi_{P_1P_1}^{\mathrm{T}}\Phi_{P_1P_1} = \Sigma_{P_1P_1}^{-1} \tag{15}$$

and, for $j = 2, \dots, k$,

$$\begin{pmatrix} \Phi_{R_jR_j}^{\mathrm{T}} & 0 \\ \Phi_{S_jR_j} & \cdot \end{pmatrix}\begin{pmatrix} \Phi_{R_jR_j} & \Phi_{R_jS_j} \\ 0 & \cdot \end{pmatrix} = \Sigma_{P_jP_j}^{-1} \tag{16}$$

where dots denote submatrices which need not to be specified explicitly. We can now give the main result of this appendix.

Theorem 1

For $G = (V, \mathcal{V})$ let Φ be the T-completion of $\Phi^{\mathcal{V}} \in M_*^{\lhd}(G)$. If the vertices V are enumerated along a perfect sequence of prime components $(P_1, \dots, P_k)$ of G, then for every incomplete prime component P_j of G, the submatrix $\Phi_{P_jP_j}$ of Φ is the T-completion of $\Phi_{P_jP_j}^{\mathcal{P}_j}$ with respect to G_{P_j}. All the remaining unspecified elements of $\Phi^{\mathcal{V}}$ are zero.

Proof. We have shown above that under the conditions of the theorem, the T-completion of $\Phi^{\mathcal{V}}$ only requires consideration of the unspecified entries of the submatrices $\Phi_{P_jP_j}^{\mathcal{P}_j}$ where P_j is an incomplete prime component of G. Thus, we have to show that the submatrix $\Phi_{P_jP_j}$ of Φ is the T-completion of $\Phi_{P_jP_j}^{\mathcal{P}_j} \in M_*^{\lhd}(G_{P_j})$ for $j = 1, \dots, k$.

Consider first $j = k$. Since $G_{H_k} = G$ is collapsible onto $G_{H_{k-1}}$, we can apply proposition 4 with $A = R_k$ and $B = H_{k-1}$ to obtain that $(\Phi_{R_kR_k}\Phi_{R_kS_k})$ is a one-to-one function of $(\Phi_{R_kR_k}\Phi_{R_kS_k})^{\mathcal{P}_k}$. If P_k is complete such a function is trivial, otherwise it is given by (11) and, since S_k is complete, by (14) is the T-completion of $\Phi_{P_kP_k}^{\mathcal{P}_k}$. By proposition 4, $\Phi_{H_{k-1}H_{k-1}}$ is the T-completion of $\Phi_{H_{k-1}H_{k-1}}^{\mathcal{H}_{k-1}} \in M(G_{H_{k-1}})$ and the result for $j = 1, \dots, k-1$ follows by recursive application of the the same procedure.

Throughout this paper $K = \Sigma^{-1}$ is the canonical parameter of $X_V \sim N(0, \Sigma)$. The following statistical interpretation of Φ was given by Wermuth (1980); see also Massam & Neher (1997) and Roverato (2000).

Assume $B = \{b, b+1, \dots, p\}$ and let $A = V \backslash B$ so that the pair (A, B) is a partition of V. The triangular matrix Φ can be partitioned accordingly

$$K = \begin{pmatrix} \Phi_{AA}^{\mathrm{T}} & 0 \\ \Phi_{BA} & \Phi_{BB}^{\mathrm{T}} \end{pmatrix}\begin{pmatrix} \Phi_{AA} & \Phi_{AB} \\ 0 & \Phi_{BB} \end{pmatrix}$$

and, by applying the rules for the inverse of a partitioned matrix it is straightforward to check that the two row-blocks $\Phi_{AV} = (\Phi_{AA}, \Phi_{AB})$ and Φ_{BB} are one-to-one functions of the parameters $(\Gamma_{A|B}, \Sigma_{AA|B})$ and Σ_{BB} of $X_A|X_B$ and X_B respectively

$$\begin{aligned} \Phi_{AA}^{\mathrm{T}}\Phi_{AA} &= (\Sigma_{AA|B})^{-1} \\ -\Phi_{AA}^{-1}\Phi_{AB} &= \Gamma_{A|B} \\ \Phi_{BB}^{\mathrm{T}}\Phi_{BB} &= \Sigma_{BB}^{-1}. \end{aligned} \tag{17}$$

Hence a consequence of proposition 4 is as follows.

Corollary 3
If $G = (V, \mathcal{V})$ is collapsible onto $B = \{b, b+1, \ldots, p\}$ and $\Phi^{\top}\Phi = \Sigma^{-1} \in M^{+}(G)$, then $(\Gamma_{A|B}, \Sigma_{AA|B})$ is a one-to-one function of $\Phi_{AV}^{\mathcal{V}}$ and Σ_{BB} is a one-to-one function of $\Phi_{BB}^{\mathcal{B}}$.

We close this appendix by computing the Jacobian of the transformation $K \to \Phi^{\mathcal{V}}$ which is used both in the proof of theorem 1 and in section 5.

Proposition 5
For $\Phi^{\mathcal{V}} \in M_{*}^{\lhd}(G)$, where $G = (V, \mathcal{V})$, let $K = \Phi(\Phi^{\mathcal{V}})^{\mathrm{T}}\Phi(\Phi^{\mathcal{V}})$. The Jacobian of the inverse transformation $K \to \Phi^{\mathcal{V}}$ is

$$J(K \to \Phi^{\mathcal{V}}) = 2^{p} \prod_{i=1}^{p} \phi_{ii}^{v_i+1}.$$

Proof. By applying (11) to (9) we obtain $\kappa_{rs} = \kappa_{rs}(\Phi_{\{r\}\{s\}}^{\mathcal{V}})$, so that whenever either $i > r$ or $j > s$ we have $d/(d\phi_{ij})\kappa_{rs} = 0$. Thus, if we take the distinct non-zero elements of K ordered according to the rows of K, and similarly for the elements of $\Phi^{\mathcal{V}}$, the Jacobian matrix $J = \partial/(\partial\Phi^{\mathcal{V}})K$ is triangular and its determinant is the product of the diagonal elements

$$|J| = \prod_{(r,s)\in\mathcal{V}} \frac{d}{d\phi_{rs}} \kappa_{rs}. \tag{18}$$

By (8) for all $(r,s) \in \mathcal{V}$ we can write $\kappa_{rs} = \sum_{i=1}^{r-1} \phi_{ir}\phi_{is} + \phi_{rr}\phi_{rs}$ where, when $r = 1$ the expression involving $\sum_{i=1}^{r-1}$ must be considered zero. By (11), none of the terms of the sum $\sum_{i=1}^{r-1} \phi_{ir}\phi_{is}$ is a function of ϕ_{rs} so that $d/(d\phi_{rs})\kappa_{rs} = d/(d\phi_{rs})\phi_{rr}\phi_{rs}$. For $r = s$ this derivative is $d/(d\phi_{rr})\kappa_{rr} = 2\phi_{rr}$ while for $r < s$ with $(r,s) \in \mathcal{V}$ it is $d/(d\phi_{rs})\kappa_{rs} = \phi_{rr}$. By noticing that such derivatives only depend on the row position of κ_{rs}, and the ith row of $\Phi^{\mathcal{V}}$ has exactly $v_i + 1$ specified elements, it follows that $|J| = 2^{p}\prod_{i=1}^{p} \phi_{ii}^{v_i+1}$ and the proof is complete.

B. Proof of lemma 1

(ii) and (iii) are equivalent because, by (17), $K_{BB|A} = \Phi_{BB}^{\mathrm{T}}\Phi_{BB}$, $K_{AB} = \Phi_{AA}^{\mathrm{T}}\Phi_{AB}$ and $K_{AA} = \Phi_{AA}^{\mathrm{T}}\Phi_{AA}$. Hence, only (i) and (iii) need to be proved.

We first consider (iii). By proposition 4, $(\Phi_{AA}, \Phi_{AB}) = \Phi_{AV}$ is independent of Φ_{BB} if and only if $\Phi_{AV}^{\mathcal{V}} \perp\!\!\!\perp \Phi_{BB}^{\mathcal{B}}$. To prove the independence of Φ_{AV} and Φ_{BB} it is therefore sufficient to show that the density function of the distribution induced on $\Phi^{\mathcal{V}}$ by K factorizes as

$$f(\Phi^{\mathcal{V}}|\delta, T) = f_{AV}(\Phi_{AV}^{\mathcal{V}}|\delta, T) \times f_{BB}(\Phi_{BB}^{\mathcal{B}}|\delta, T_{BB}) \tag{19}$$

where T is defined by $Q = T^{\mathrm{T}}T$.

The density function (19) can be derived by replacing K by $\Phi^{\mathrm{T}}\Phi$ in (1) and multiplying it by the Jacobian in proposition 5 and has the form

$$f(\Phi^{\mathscr{V}}|\delta, T) \propto \left(\prod_{r=1}^{p}\phi_{rr}^{\delta-2}\right)\exp\left[-\frac{1}{2}\mathrm{tr}\left\{\Phi^{\mathrm{T}}\Phi(T^{\mathrm{T}}T)^{-1}\right\}\right]\prod_{r=1}^{p}\phi_{rr}^{\nu_r+1}, \tag{20}$$

where T is defined by $Q = T^{\mathrm{T}}T$. Recall that $|K| = \prod_{r=1}^{p}\phi_{rr}^{2}$.

Let SS(M) denote the sum of the squares of the entries of a matrix M so that if M is upper triangular SS(M) $=$ tr($M^{\mathrm{T}}M$). By noticing that both T^{-1} and ΦT^{-1} are upper triangular and that $(\Phi T^{-1})_{AV} = \Phi_{AV}T^{-1}$ and $(\Phi T^{-1})_{BB} = \Phi_{BB}T_{BB}^{-1}$, we can write

$$\begin{aligned}
\mathrm{tr}\left\{\Phi^{\top}\Phi(T^{\top}T)^{-1}\right\} &= \mathrm{tr}(\Phi^{\mathrm{T}}\Phi T^{-1}T^{-\mathrm{T}}) \\
&= \mathrm{tr}\left\{(\Phi T^{-1})^{\mathrm{T}}(\Phi T^{-1})\right\} \\
&= \mathrm{SS}(\Phi T^{-1}) \\
&= \mathrm{SS}\{(\Phi T^{-1})_{AV}\} + \mathrm{SS}\{(\Phi T^{-1})_{BB}\} \\
&= \mathrm{SS}(\Phi_{AV}T^{-1}) + \mathrm{SS}(\Phi_{BB}T_{BB}^{-1}) \\
&= \mathrm{SS}(\Phi_{AV}T^{-1}) + \mathrm{tr}\{(\Phi_{BB}T_{BB}^{-1})^{\mathrm{T}}(\Phi_{BB}T_{BB}^{-1})\} \\
&= \mathrm{SS}(\Phi_{AV}T^{-1}) + \mathrm{tr}\left\{\Phi_{BB}^{\mathrm{T}}\Phi_{BB}(T_{BB}^{\mathrm{T}}T_{BB})^{-1}\right\}.
\end{aligned} \tag{21}$$

By using (21) in (20), $f(\Phi^{\mathscr{V}}|\delta, T)$, can be factorized as in (19) with

$$f_{AV}(\Phi_{AV}^{\mathscr{V}}|\delta, T) \propto \left(\prod_{r=1}^{b-1}\phi_{rr}^{\delta-2}\right)\exp\left\{-\frac{1}{2}\mathrm{SS}(\Phi_{AV}T^{-1})\right\}\prod_{r=1}^{b-1}\phi_{rr}^{\nu_r+1}$$

and

$$f_{BB}(\Phi_{BB}^{\mathscr{B}}|\delta, T_{BB}) \propto \left(\prod_{r=b}^{p}\phi_{rr}^{\delta-2}\right)\exp\left[-\frac{1}{2}\mathrm{tr}\left\{\Phi_{BB}^{\mathrm{T}}\Phi_{BB}(T_{BB}^{\mathrm{T}}T_{BB})^{-1}\right\}\right]\prod_{r=b}^{p}\phi_{rr}^{\nu_r+1} \tag{22}$$

where $\Phi_{AV} = \Phi_{AV}(\Phi_{AV}^{\mathscr{V}})$ and $\Phi_{BB} = \Phi_{BB}(\Phi_{BB}^{\mathscr{B}})$ by proposition 4. We have thus shown the independence Φ_{AV} and Φ_{BB}.

Consider now (i). Because of the zero structure of K, the support of the distribution of $K_{BB|A}$ is $M^{+}(G_B)$ (Whittaker, 1990, p. 397) whereas its density function can be derived by considering the transformation $K_{BB|A} = \Phi_{BB}(\Phi_{BB}^{\mathscr{B}})^{\mathrm{T}}\Phi_{BB}(\Phi_{BB}^{\mathscr{B}})$ where the distribution of $\Phi_{BB}^{\mathscr{B}}$ has density (22). Since $\Phi_{BB}^{\mathrm{T}}\Phi_{BB} = K_{BB|A}$ and $T_{BB}^{\mathrm{T}}T_{BB} = Q_{BB|A}$, we have that $\prod_{r=b}^{p}\phi_{rr}^{\delta-2} = |K_{BB|A}|^{(\delta-2)/2}$ and $\prod_{r=b}^{p}\phi_{rr}^{\nu_r+1} = J(K_{BB|A} \to \Phi_{BB}^{\mathscr{B}})$ by proposition 5. Consequently the distribution of $K_{BB|A}$ has density

$$q_{G_B}(K_{BB|A}|\delta, Q_{BB|A}) \propto |K_{BB|A}|^{(\delta-2)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left\{K_{BB|A}(Q_{BB|A})^{-1}\right\}\right] \tag{23}$$

and this establishes (i).