Biometrika (2000), **87**, 1, *pp.* 99–112

Printed in Great Britain

Cholesky decomposition of a hyper inverse Wishart matrix

By ALBERTO ROVERATO

Dipartimento di Economia Politica, University of Modena and Reggio Emilia, Viale J. Berengario n. 51, 41100 Modena, Italy

roverato@unimo.it

Summary

The canonical parameter of a covariance selection model is the inverse covariance matrix Σ^{-1} whose zero pattern gives the conditional independence structure characterising the model. In this paper we consider the upper triangular matrix Φ obtained by the Cholesky decomposition $\Sigma^{-1}=\Phi^{\mathrm{T}}\Phi$. This provides an interesting alternative parameterisation of decomposable models since its upper triangle has the same zero structure as Σ^{-1} and its elements have an interpretation as parameters of certain conditional distributions. For a distribution for Σ, the strong hyper-Markov property is shown to be characterised by the mutual independence of the rows of Φ. This is further used to generalise to the hyper inverse Wishart distribution some well-known properties of the inverse Wishart distribution. In particular we show that a hyper inverse Wishart matrix can be decomposed into independent normal and chi-squared random variables, and we describe a family of transformations under which the family of hyper inverse Wishart distributions is closed.

Some key words: Cholesky decomposition; Conjugate distribution; Covariance selection model; Decomposable graph; Hyper-Markov distribution; Strong hyper-Markov property.

1. Introduction

In a Bayesian analysis of the multivariate normal distribution with mean vector equal to zero and covariance matrix Σ, the inverse Wishart family is the standard conjugate for the canonical parameter Σ^{-1} (Press, 1972, p. 185).

A covariance selection model (Dempster, 1972) is specified by assuming that some off-diagonal elements of Σ^{-1} are zero. In this way the pairwise conditional independence structure of the variables is stated (Wermuth, 1976). Speed & Kiiveri (1986) associated an undirected graph $\mathcal{G}=(V, E)$ with the inverse covariance matrix and showed that this model belongs to the family of graphical models introduced by Darroch, Lauritzen & Speed (1980). For the Bayesian analysis of decomposable covariance selection models, Dawid & Lauritzen (1993) proposed a generalisation of the inverse Wishart distribution. They called such a distribution hyper inverse Wishart and showed that it is conjugate for Σ^{-1} and strong hyper-Markov over $\mathcal{G}$.

In this paper, for a decomposable covariance selection model we consider the upper triangular matrix Φ obtained by the Cholesky decomposition $\Sigma^{-1}=\Phi^{\mathrm{T}}\Phi$. A first interesting feature of Φ is that, if the vertices V are enumerated according to a perfect vertex elimination scheme, then its upper triangle has the same zero pattern as Σ^{-1} (Wermuth, 1980; Paulsen, Power & Smith, 1989). Thus it provides an alternative parameterisation of decomposable covariance selection models. Its elements are variation independent and are shown to have an interpretation as parameters of the conditional distributions involved in the recursive factorisation of the density in directed models (Wermuth, 1980).

In the context of a probability distribution for Σ, we show that the strong hyper-Markov property is characterised by the mutual independence of the rows of Φ. This turns out to be useful in the analysis of hyper-Markov distributions for Σ, but can also be applied to construct alternative prior distributions for the model.

We finally assume the covariance matrix to have a hyper inverse Wishart distribution, $\Sigma \sim \text{HIW}_{\mathcal{G}}(\delta, B)$, and provide a generalisation of some properties of the inverse Wishart distribution. We see that in the standard case, where B is the identity matrix, the elements of Φ are mutually independent; the nonzero off-diagonal elements are standard normal variables and the diagonal elements are square roots of chi-squared variables. For a positive definite matrix D, we describe the procedure to transform Σ to obtain a $\text{HIW}_{\mathcal{G}}(\delta, D)$ matrix. The distribution of the determinant of Σ is also considered.

The general theory relating to covariance selection models, Cholesky decompositions and hyper inverse Wishart distributions is presented in § 2. In § 3 a statistical interpretation of the elements of Φ is provided. In § 4 we discuss the relationship between the strong hyper-Markov property and the independence structure of Φ. In § 5 we derive the distribution of the inverse of a hyper inverse Wishart matrix. Finally, in § 6 some properties of the hyper inverse Wishart distribution are presented.

2. BACKGROUND

2·1. *Covariance selection models and Cholesky decomposition*

In this section we introduce the notation related to the graph and covariance selection model theory required in this paper and present a result concerning the Cholesky decomposition of inverse covariance matrices. For a comprehensive account of covariance selection model theory and related graph theory see Lauritzen (1996).

Consider a decomposable graph $\mathcal{G} = (V, E)$ with vertex set $V = \{\alpha_1, \ldots, \alpha_p\}$ and set of edges E, and let $C_1, \ldots, C_k$ be a perfect sequence of its cliques. The histories and the residuals of the sequence are denoted by $H_j = C_1 \cup \ldots \cup C_j$ and $R_j = C_j \backslash H_{j-1}$ respectively, whereas $S_j = H_{j-1} \cap C_j$ are the clique separators. We denote the cardinality of cliques and separators by $c_j = |C_j|$ and $s_j = |S_j|$ respectively. The subgraph induced by $W \subseteq V$ is denoted by $\mathcal{G}_W$.

The numbering of the vertices V obtained by taking first the vertices in C_1, then those in R_2, R_3 and so on, is perfect (Lauritzen, 1996, Lemma 2.12) and has a perfect directed acyclic version $\mathcal{G}^<$ of $\mathcal{G}$ associated with it. We assume that the vertices V are ordered according to this perfect numbering taken in reverse order, so that α_1 is the last vertex in the perfect numbering and α_p is the first. In this way the vertices are enumerated according to a perfect vertex elimination scheme for $\mathcal{G}$.

Let $v_i = |\text{bd}(\alpha_i) \cap \text{pd}(\alpha_i)|$, where predecessors $\text{pd}(\alpha_i) = \{\alpha_{i+1}, \ldots, \alpha_p\}$ denote the number of vertices adjacent to α_i following α_i in the elimination scheme. With respect to $\mathcal{G}^<$, we can also write $v_i = |\text{pa}(\alpha_i)|$, the number of arrows pointing to α_i, so that the number of edges $|E|$ can be written as $v = \sum_{i=1}^{p} v_i$. It is straightforward to see that, for the vertices $\alpha_i \in C_1$, the values taken by v_i are $0, 1, \ldots, c_1 - 1$. Moreover, since for all $j = 2, \ldots, k$ (H_{j-1}, R_j, S_j) decomposes $\mathcal{G}_{H_j}$ (Lauritzen, 1996, Lemma 2.11), the values v_i associated with the sets R_j are $s_j, s_j + 1, \ldots, c_j - 1$.

For an undirected graph $\mathcal{G} = (V, E)$ we denote by $M(\mathcal{G})$ the set of $|V| \times |V|$ matrices $A = \{\alpha_{rs}\}$ satisfying $a_{rs} = a_{sr} = 0$ for all pairs (r, s) such that $(\alpha_r, \alpha_s) \notin E$, and by $M^+(\mathcal{G})$

those elements of $M(\mathcal{G})$ that are positive definite, and therefore symmetric. Recall that, if $\mathcal{G}$ is decomposable and $B^{-1}=A$ belongs to $M^+(\mathcal{G})$, then the following properties hold (Lauritzen, 1996, p. 145):

$$B^{-1}=\sum_{i=1}^{k}[B_{C_iC_i}^{-1}]^0-\sum_{i=2}^{k}[B_{S_iS_i}^{-1}]^0, \tag{1}$$

$$|B|=\frac{\prod_{i=1}^{k}|B_{C_iC_i}|}{\prod_{i=2}^{k}|B_{S_iS_i}|}, \tag{2}$$

where $[B_{CC}]^0$ denotes the $|V|\times|V|$ matrix obtained by padding zero entries around B_{CC} to obtain full dimension and $|B|$ the determinant of B.

Let $X\equiv X_V$ be a $|V|$-variate normal random vector with mean equal to zero and covariance matrix Σ. The covariance selection model, or Gaussian graphical model, for X with graph $\mathcal{G}=(V,E)$ is specified by assuming that $K=\Sigma^{-1}$ belongs to $M^+(\mathcal{G})$ (Dempster, 1972; Wermuth, 1976). A model whose graph is decomposable is itself called decomposable. We remark that the covariance selection model is a regular exponential family with canonical parameter K (Lauritzen, 1996, p. 132).

For a subset $W\subset V$, the submatrix Σ_{WW} is the parameter of the marginal distribution of X_W and, for $Z=V\backslash W$,

$$\Sigma_{Z|W}=\Sigma_{ZZ}-\Sigma_{ZW}\Sigma_{WW}^{-1}\Sigma_{WZ},\quad \Gamma_{Z|W}=\Sigma_{ZW}\Sigma_{WW}^{-1}$$

are the parameters related to the conditional distribution of X_Z given X_W or, more compactly, of $X_Z\,|\,X_W$.

The following theorem provides a characterisation of decomposable graphs based on the preservation of the zero pattern in the Cholesky decomposition of matrices $M^+(\mathcal{G})$.

THEOREM 1 (*Paulsen et al.*, 1989). *The following are equivalent for an undirected graph* $\mathcal{G}=(V,E)$:

(i) *the graph* $\mathcal{G}$ *is decomposable*;

(ii) *there exists a permutation of the vertices* $\{\alpha_1,\ldots,\alpha_{|V|}\}$ *such that with respect to this renumbering every* $A\in M^+(\mathcal{G})$ *factors as* $A=T^{\mathrm{T}}T$ *with* $T\in M(\mathcal{G})$ *and* T *upper triangular.*

Although we do not provide a proof of this theorem, we point out that Paulsen et al. (1989) showed that, if $\mathcal{G}=(V,E)$ is decomposable, any permutation of the vertices V following a perfect vertex elimination scheme for $\mathcal{G}$ satisfies point (ii); see also Wermuth (1980). The matrix T is unique up to multiplication by a unitary diagonal matrix, and thus we can identify it by assuming that it has positive diagonal.

Throughout this paper, for a decomposable graph $\mathcal{G}=(V,E)$, we always consider the vertices V enumerated according to a perfect vertex elimination scheme. Furthermore, for a matrix $A\in M^+(\mathcal{G})$ we assume that the row and column numbers follow the vertex ordering, and write $A=T^{\mathrm{T}}T$ to define the Cholesky decomposition of A, where T is upper triangular with positive diagonal and belongs to $M(\mathcal{G})$.

We close this section by noting that Theorem 1 provides an alternative parameterisation of decomposable covariance selection models by means of a triangular matrix T whose elements are variation independent.

2·2. *The hyper inverse Wishart distribution*

Dawid & Lauritzen (1993) defined the weak hyper-Markov property and used it to construct a family of probability distributions for the conjugate Bayesian analysis of decomposable covariance selection models. They called such a family hyper inverse Wishart since it can be regarded as a generalisation of the inverse Wishart family.

The inverse Wishart distribution arises in the conjugate analysis of the saturated model where $\mathscr{G}$ is the complete graph (Press, 1972, p. 187). In this case the moment parameter Σ is constrained only to be positive definite and, using the parameterisation of Dawid (1981), we write $\Sigma \sim \text{IW}(\delta, B)$ with density

$$f_V(\Sigma \mid \delta, B) = h(\delta, V) \frac{|\Sigma|^{-(\delta+2|V|)/2}}{|B|^{-(\delta+|V|-1)/2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}B)\right\},$$

where

$$h(\delta, V) = \frac{2^{-|V|(\delta+|V|-1)/2}}{\Gamma_{|V|}\{\frac{1}{2}(\delta+|V|-1)\}} \tag{3}$$

is the normalising constant (Muirhead, 1982, p. 113). The hyperparameters are $\delta > 0$ and a positive definite matrix B. We recall that the multivariate gamma function has the form $\Gamma_{|V|}(y) = \pi^{|V|(|V|-1)/4} \prod_{i=1}^{|V|} \Gamma\{y-(i-1)/2\}$.

For a model with arbitrary decomposable graph $\mathscr{G} = (V, E)$, the moment parameter is the collection of clique-marginal covariance matrices $\{\Sigma_{C_jC_j};\, j = 1, \ldots, k\}$. By using (1) we can combine these to obtain a full matrix Σ such that $\Sigma^{-1} \in M^+(\mathscr{G})$. The hyper inverse Wishart is defined as the unique hyper-Markov distribution for Σ with marginals $\Sigma_{C_jC_j} \sim \text{IW}(\delta, B_{C_jC_j})$, for all $j = 1, \ldots, k$. The hyperparameter $\{B_{C_jC_j};\, j = 1, \ldots, k\}$ is a collection of submatrices of a $|V| \times |V|$ positive definite matrix B which, without loss of generality, we assume constructed by a formula such as (1). Hence, following Dawid & Lauritzen (1993), we write $\Sigma \sim \text{HIW}_{\mathscr{G}}(\delta, B)$. Thus the hyper inverse Wishart distribution is constructed so as to satisfy the weak hyper-Markov property with respect to $\mathscr{G}$ and its density can be written as a function of the clique-marginal densities,

$$f_{\mathscr{G}}(\Sigma \mid \delta, B) = \frac{\prod_{j=1}^{k} f_{C_j}(\Sigma_{C_jC_j} \mid \delta, B_{C_jC_j})}{\prod_{j=2}^{k} f_{S_j}(\Sigma_{S_jS_j} \mid \delta, B_{S_jS_j})}. \tag{4}$$

We point out that the mathematically independent arguments of (4) are the entries σ_{ij} of Σ with $(\alpha_i, \alpha_j) \in E$, which, under the graphical conditional independence constraints, fully determine Σ.

Dawid & Lauritzen (1993) showed that the hyper inverse Wishart distribution also satisfies the strong hyper-Markov property, which implies stronger independence properties than those resulting from the constructing procedure. The definitions of hyper-Markov properties and the theory of hyper-Markov distributions required for this paper can be found in Dawid & Lauritzen (1993). Here we use the notation $X \perp\!\!\!\perp Y$ to write that the random quantities X and Y are independent and $\perp\!\!\!\perp\{X, Y, Z\}$ to denote the mutual independence of X, Y and Z.

We now turn to the saturated model. In this case the hyper inverse Wishart prior simplifies to an inverse Wishart distribution, whose properties are well known. We recall some of them.

The distribution induced on $K = \Sigma^{-1}$ by $\Sigma \sim \text{IW}(\delta, B)$ is Wishart, $K \sim W(\delta + |V| - 1, A)$, and its density function is

$$q_V(K \mid \delta, A) = h(\delta, V) \frac{|K|^{(\delta-2)/2}}{|A|^{(\delta+|V|-1)/2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(KA^{-1})\right\}, \tag{5}$$

where $A = B^{-1}$. It can be checked that this is the standard conjugate distribution for the canonical parameter K (Press, 1972, p. 185). As a consequence it follows that the inverse Wishart is a Diaconis–Ylvisaker-conjugate family (Diaconis & Ylvisaker, 1979) and has the related properties; for a definition of standard and Diaconis–Ylvisaker-conjugate families and a review of their properties see Gutiérrez-Peña & Smith (1997).

We now mention a result, useful in stochastic simulation, concerning the Cholesky decomposition of a standard Wishart matrix. Odell & Feiveson (1966) showed that, if $\Psi^{\mathrm{T}}\Psi = K \sim W(\delta + |V| - 1, I)$, then the main diagonal elements ψ_{ii} of Ψ are square roots of $\chi^2_{\delta+|V|-i}$ quantities, the off-diagonal elements ψ_{rs} of Ψ are $N(0, 1)$ quantities and these random variables are mutually independent. A Wishart matrix with parameter $A = T^{\mathrm{T}}T$ may successively be obtained by the transformation $(\Psi T)^{\mathrm{T}}(\Psi T) = T^{\mathrm{T}}KT \sim W(\delta + |V| - 1, A)$.

The following section provides a statistical interpretation of the Cholesky decomposition of an inverse covariance matrix.

3. Cholesky decomposition of Σ^{-1}

Covariance selection models are usually parameterised by either the covariance matrix Σ or its inverse, the concentration matrix K. The latter represents the natural parameterisation of undirected models since its zero pattern gives the conditional independence structure characterising the model and its nonzero off-diagonal elements are unnormalised partial correlations. Nevertheless the covariance matrix Σ is often preferred since it allows a straightforward implementation of the procedures locally computable on marginal models.

As we pointed out in § 2·1, when the model is decomposable an alternative parameterisation is given by the Cholesky decomposition $\Phi^{\mathrm{T}}\Phi$ of K. Provided that the vertices of $\mathcal{G} = (V, E)$ are enumerated according to a perfect vertex elimination scheme, the upper triangle of Φ has the same zero pattern as K. Furthermore the elements of Φ are variation independent and have an interesting interpretation as parameters of the conditional distributions involved in the recursive factorisation of the density function of X according to $\mathcal{G}^{<}$ (Wermuth, 1980).

To see this we first assume $\mathcal{G}$ to be the complete graph so that K is only constrained to be positive definite. If we partition the vertex set into the subsets $A \subset V$ and $B = V \backslash A$ and order the vertices V accordingly, the Cholesky decomposition of K can be written as

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

Applying the rules for the inverse of a partitioned matrix we obtain

$$\Phi^{\mathrm{T}}_{AA}\Phi_{AA} = (\Sigma_{A|B})^{-1}, \quad -\Phi^{-1}_{AA}\Phi_{AB} = \Gamma_{A|B}, \quad \Phi^{\mathrm{T}}_{BB}\Phi_{BB} = \Sigma^{-1}_{BB}. \tag{6}$$

Hence $(\Phi_{AA}, \Phi_{AB}) = \phi_A(\Gamma_{A|B}, \Sigma_{A|B})$ is a one-to-one transformation of $(\Gamma_{A|B}, \Sigma_{A|B})$, and $\Phi_{BB} = \phi_B(\Sigma_{BB})$ is a one-to-one transformation of Σ_{BB}. Thus the upper row-block (Φ_{AA}, Φ_{AB}) of Φ provides an alternative parameterisation of the conditional distribution of $X_A | X_B$, whereas the lower row-block Φ_{BB} can be obtained from the Cholesky decomposition of Σ^{-1}_{BB}, the parameter of the marginal distribution of X_B. The same procedure can then be applied recursively with respect to Φ_{BB} and a given partition of the vertex subset B.

For an arbitrary decomposable covariance selection model with graph $\mathcal{G} = (V, E)$ the inverse covariance matrix K belongs to $M^{+}(\mathcal{G})$. Since the vertices V are ordered according to a perfect vertex elimination scheme, in the upper triangle of Φ the ith row, $(\Phi_{\{\alpha_i\}\{\alpha_i\}}, \Phi_{\{\alpha_i\}\mathrm{pd}(\alpha_i)})$, has exactly $v_i + 1$ nonzero elements which, with respect to $\mathscr{G}^<$, can be written as $(\Phi_{\{\alpha_i\}\{\alpha_i\}}, \Phi_{\{\alpha_i\}\mathrm{pa}(\alpha_i)})$. Applying the recursive procedure above one vertex at a time, we obtain that the ith row of Φ, for $i = 1, \ldots, |V|$, provides an alternative parameterisation of the conditional distribution of $X_{\{\alpha_i\}} \mid X_{\mathrm{pd}(\alpha_i)}$, or equivalently of $X_{\{\alpha_i\}} \mid X_{\mathrm{pa}(\alpha_i)}$. Note that we allow $\mathrm{pd}(\alpha_i)$ to be the empty set.

For decomposable covariance selection models there exists an equivalence between the factorisation property on $\mathscr{G}$ and the recursive factorisation property on $\mathscr{G}^<$ of the distribution of X; see Lauritzen (1996, Proposition 3.28). The factorisation of the distribution of X on $\mathscr{G}$ involves the clique-marginal distributions of X, parameterised by submatrices of Σ, and is a consequence of the zero structure of K. In a similar way we can say that the recursive factorisation on $\mathscr{G}^<$ involves the conditional distributions parameterised by the rows of Φ and is a consequence of the zero structure of Φ itself. We can conclude that K and Φ play a similar role in the parameterisation of covariance selection models. The former is more suitable for undirected models whereas the use of the latter seems more appropriate where directed models are concerned. The equation $K = \Phi^{\mathrm{T}}\Phi$ is the link between the two parameterisations.

4. Independence structure of Φ and the strong hyper-Markov property

In Bayesian analysis of covariance selection models the parameter Σ has a probability distribution associated with it. Dawid & Lauritzen (1993) showed that a desirable feature of a distribution for Σ is the strong hyper-Markov property. This allows considerable simplification of the inference process both in the implementation of the procedures and in the interpretation of the results. In this section we investigate the implications of such a property for the Cholesky decomposition of Σ^{-1}. In particular we see that for covariance selection models the strong hyper-Markov property can be defined in terms of the independence structure of Φ.

For a probability distribution for the parameter of the saturated model we have the following result.

Proposition 1. *For a $|V| \times |V|$ positive definite random matrix Σ the following are equivalent*:

(i) *for every pair of subsets A and $B = V \backslash A$ of V it holds that $(\Gamma_{A|B}, \Sigma_{A|B}) \perp\!\!\!\perp \Sigma_{BB}$*;

(ii) *for every permutation of the vertices V the rows of the matrix Φ defined by $\Sigma^{-1} = \Phi^{\mathrm{T}}\Phi$ are mutually independent.*

Proof. For two subsets A and $B = V \backslash A$ of V consider the ordering of the vertices V obtained by taking first the vertices in A. In this case, by (6), $(\Gamma_{A|B}, \Sigma_{A|B})$ is a one-to-one transformation of (Φ_{AA}, Φ_{AB}) and Σ_{BB} is a one-to-one transformation of Φ_{BB}, so that $(\Gamma_{A|B}, \Sigma_{A|B}) \perp\!\!\!\perp \Sigma_{BB}$ is equivalent to

$$(\Phi_{AA}, \Phi_{AB}) \perp\!\!\!\perp \Phi_{BB}. \tag{7}$$

Since the submatrices (Φ_{AA}, Φ_{AB}) and Φ_{BB} are distinct row-blocks of Φ, it is clear that (ii) $\Rightarrow$ (i).

We now show that (i) $\Rightarrow$ (ii). Consider a permutation $\{\alpha_1, \ldots, \alpha_{|V|}\}$ of V. For $i = 1$ we set $A = \{\alpha_1\}$ and $B = \{\alpha_2, \ldots, \alpha_{|V|}\}$. In this case (i) implies (7), which states that the first row of Φ is independent of the following $|V| - 1$ rows Φ_{BB}. Similarly, if for $i = 2$

we set $A = \{\alpha_1, \alpha_2\}$, then by (7) the first two rows of Φ are independent of the following $|V| - 2$ rows. The desired result follows by iterating this procedure for all $i = 1, \ldots, |V| - 1$. □

We now obtain a characterisation of the strong hyper-Markov property for covariance selection models based on Φ.

THEOREM 2. *For the parameter Σ of a decomposable covariance selection model with graph $\mathcal{G} = (V, E)$ the following are equivalent*:

(i) *the distribution of Σ is strong hyper-Markov over $\mathcal{G}$*;
(ii) *for every enumeration of the vertices V following a perfect vertex elimination scheme for $\mathcal{G}$ the rows of the matrix Φ defined by $\Sigma^{-1} = \Phi^{\mathrm{T}}\Phi$ are mutually independent.*

Proof. Dawid & Lauritzen (1993, Proposition 3.15) showed that the distribution of Σ is strong hyper-Markov over $\mathcal{G}$ if and only if it is directed strong hyper-Markov over $\mathcal{G}^<$ for every perfect directed version $\mathcal{G}^<$ of $\mathcal{G}$; that is if and only if, for every enumeration of the vertices V following a perfect vertex elimination scheme for $\mathcal{G}$, and for all $\alpha_i \in V$, we have

$$(\Gamma_{\{\alpha_i\}|\mathrm{pd}(\alpha_i)}, \Sigma_{\{\alpha_i\}|\mathrm{pd}(\alpha_i)}) \perp\!\!\!\perp \Sigma_{\mathrm{pd}(\alpha_i)\mathrm{pd}(\alpha_i)}, \tag{8}$$

where $\mathrm{pd}(\alpha_i) = \{\alpha_{i+1}, \alpha_{i+2}, \ldots, \alpha_{|V|}\}$. As shown in the proof of Proposition 1, (8) is equivalent to

$$(\Phi_{\{\alpha_i\}\{\alpha_i\}}, \Phi_{\{\alpha_i\}\mathrm{pd}(\alpha_i)}) \perp\!\!\!\perp \Phi_{\mathrm{pd}(\alpha_i)\mathrm{pd}(\alpha_i)}, \tag{9}$$

so that, since the submatrices $(\Phi_{\{\alpha_i\}\{\alpha_i\}}, \Phi_{\{\alpha_i)\mathrm{pd}(\alpha_i)})$ and $\Phi_{\mathrm{pd}(\alpha_i)\mathrm{pd}(\alpha_i)}$ are distinct row-blocks of Φ, we have that (ii)$\Rightarrow$(i).

We now show that (i)$\Rightarrow$(ii). Assume the vertices of V are enumerated according to a perfect vertex elimination scheme for $\mathcal{G}$. Then (i) implies (9) for all $i = 1, \ldots, |V| - 1$. The desired result follows because (9) states that the ith row of Φ is independent of the following $|V| - i$ rows. □

For the complete graph every permutation of the vertices is a perfect vertex elimination scheme, so that Proposition 1 can be regarded as a special case of Theorem 2.

When the variables X_V have a recursive response structure such that, for $i = 1, \ldots, p$, the variable X_i is considered as a response to variables $X_{i+1}, \ldots, X_{|V|}$ and explanatory to variables $X_1, \ldots, X_{i-1}$, the covariance selection model is called directed and has a directed acyclic graph associated with it. In this case the vertex set has a natural ordering which follows the variable structure. If the directed graph is perfect then its undirected version obtained by substituting directed edges for undirected ones is decomposable, and we have the following result.

COROLLARY 1. *Let $\mathcal{G}^<$ be a perfect directed acyclic graph and $\mathcal{G} = (V, E)$ its undirected version. Then for the parameter Σ of a decomposable covariance selection model with graph $\mathcal{G}$ the following are equivalent*:

(i) *the distribution of Σ is strong directed hyper-Markov over $\mathcal{G}^<$*;
(ii) *the rows of the matrix Φ defined by $\Sigma^{-1} = \Phi^{\mathrm{T}}\Phi$ are mutually independent.*

Note that, from Proposition 3.8 of Dawid & Lauritzen (1993), point (i) of Corollary 1 also implies that the distribution of Σ is weak hyper-Markov over $\mathcal{G}$.

The nonzero elements of Φ can be compactly written as collection of row-block submatrices of Φ, $\{\Phi_{C_1C_1}, (\Phi_{R_2R_2}\, \Phi_{R_2S_2}), \ldots, (\Phi_{R_kR_k}\, \Phi_{R_kS_k})\}$, and a direct consequence of Theorem 2 is as follows.

COROLLARY 2. *If the distribution of* Σ *is strong hyper-Markov over* $\mathscr{G}$ *then*

$$\perp\!\!\!\perp \{\Phi_{C_1C_1}, (\Phi_{R_2R_2}\, \Phi_{R_2S_2}), \ldots, (\Phi_{R_kR_k}\, \Phi_{R_kS_k})\}. \tag{10}$$

An immediate application of the independence properties of Φ is to the hyper inverse Wishart distribution, which is strong hyper-Markov. In particular, point (i) of Proposition 1 is a property of the inverse Wishart distribution; see for example Dawid & Lauritzen (1993, Lemma 7.4). However, we also note that Corollary 1 provides a procedure for constructing alternative, either weak or strong directed, hyper-Markov distributions for covariance selection models.

5. INVERSE OF A HYPER INVERSE WISHART MATRIX

In the remainder of this paper we assume that Σ has a hyper inverse Wishart distribution. Here we derive the distribution induced by Σ on $K = \Sigma^{-1}$ and discuss some of its properties.

We first write the density (4) explicitly and give a compact formulation of the normalising constant. For a decomposable graph $\mathscr{G} = (V, E)$ let Σ be $\textsc{hiw}_{\mathscr{G}}(\delta, B)$. By applying (1) and (2) to (4) we can write the density of Σ as

$$f_{\mathscr{G}}(\Sigma \mid \delta, B) = h_{\mathscr{G}}(\delta, V) \left(\frac{\prod_{j=1}^{k} |\Sigma_{C_jC_j}|^{|C_j|+1}}{\prod_{j=2}^{k} |\Sigma_{S_jS_j}|^{|S_j|+1}} \right)^{-1} \left(\frac{\prod_{j=1}^{k} |B_{C_jC_j}|^{|C_j|+1}}{\prod_{j=2}^{k} |B_{S_jS_j}|^{|S_j|+1}} \right)^{\frac{1}{2}}$$

$$\times \frac{|\Sigma|^{-(\delta-2)/2}}{|B|^{-(\delta-2)/2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}(\Sigma^{-1}B) \right\}. \tag{11}$$

It is interesting to point out that the quantities in round brackets can be interpreted as determinants of Isserlis matrices for Σ and B respectively (Roverato & Whittaker, 1998).

By (3) and (4), the normalising constant is

$$h_{\mathscr{G}}(\delta, V) = \frac{\prod_{j=1}^{k} h(\delta, C_j)}{\prod_{j=2}^{k} h(\delta, S_j)}$$

$$= \frac{\prod_{j=2}^{k} \Gamma_{s_j}\{(\delta + s_j - 1)/2\}}{\prod_{j=1}^{k} \Gamma_{c_j}\{(\delta + c_j - 1)/2\}} \times \frac{\prod_{j=1}^{k} 2^{-c_j(\delta + c_j - 1)/2}}{\prod_{j=2}^{k} 2^{-s_j(\delta + s_j - 1)/2}}. \tag{12}$$

When evaluation of $f_{\mathscr{G}}(\Sigma \mid \delta, B)$ is required, equation (12) turns out to provide an inefficient formulation of $h_{\mathscr{G}}(\delta, V)$. By making use of indexes v_i in the following proposition we give a more efficient formulation of the normalising constant.

PROPOSITION 2. *The normalising constant* $h_{\mathscr{G}}(\delta, V)$, *where* $\delta > 0$ *is an integer, can be written in the form*

$$h_{\mathscr{G}}(\delta, V) = (2\pi)^{-\nu/2} \prod_{i=1}^{p} \frac{2^{-(\delta + v_i)/2}}{\Gamma\{(\delta + v_i)/2\}}.$$

For the proof see the Appendix.

Proposition 2 is applied in the proof of Theorem 3 in § 6. Nevertheless, it is worth presenting it here since the quantity $h_{\mathscr{G}}(\delta, V)$ may be required for the computation of the densities considered here as well as of other quantities relating to covariance selection models, such as the Bayes factor.

Consider now the change of variables to $K = \Sigma^{-1}$; that is, the transformation between the collection of entries σ_{ij} of Σ with $(\alpha_i, \alpha_j) \in E$ and the nonzero entries of K. The Jacobian matrix J of this transformation was computed by Roverato & Whittaker (1998), who showed that

$$|J| = \frac{\prod_{i=1}^{k} |\Sigma_{C_iC_i}|^{|C_i|+1}}{\prod_{i=2}^{k} |\Sigma_{S_iS_i}|^{|S_i|+1}}.$$

The density function of K is then obtained by replacing Σ by K^{-1} in (11) and multiplying it by $|J|$, which cancels out the first term in round brackets. In order to emphasise the similarities with the density in (5) we put $A = B^{-1}$ and write the density function of K in the form

$$q_{\mathscr{G}}(K \mid \delta, A) \propto |K|^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(KA^{-1})\right\}, \quad K \in M^{+}(\mathscr{G}). \tag{13}$$

Thus we can see that if $\mathscr{G}$ is the complete graph then K has a Wishart distribution. Otherwise (13) is proportional to the density of a Wishart matrix conditioned on the event $\{K \in M^{+}(\mathscr{G})\}$. Therefore, we call such a distribution $\mathscr{G}$-conditional Wishart and write $K \sim W_{\mathscr{G}}(\delta + |V| - 1, A)$.

We close this section with a remark about the use of the hyper inverse Wishart distribution as a conjugate prior for covariance selection models. Equation (13) makes it clear that the density function of K is proportional to the likelihood function for the canonical parameter of the covariance selection model with graph $\mathscr{G}$; see for example Lauritzen (1996, eqn (5.14)). Consequently, the $\mathscr{G}$-conditional Wishart is the standard conjugate family for K and, as for the inverse Wishart family, the hyper inverse Wishart is a Diaconis–Ylvisaker-family.

6. Decomposition of the hyper inverse Wishart matrix

The discussion of § 4 suggests that the Cholesky decomposition $\Phi^{\mathrm{T}}\Phi$ of Σ^{-1} can be used to compare the properties of the inverse Wishart and the hyper inverse Wishart distributions. For example, Proposition 1 and Theorem 2 show that the independence properties of the two distributions can be expressed in the same way when described in terms of Φ. Here we carry out an analysis of the probability structure of Φ and use it to extend some properties of the inverse Wishart to the hyper inverse Wishart distribution. In particular we show how, given a positive definite matrix D, $\Sigma \sim \mathrm{HIW}_{\mathscr{G}}(\delta, B)$ can be transformed to obtain a $\mathrm{HIW}_{\mathscr{G}}(\delta, D)$ matrix. This is further applied to derive the distribution of the determinant of Σ.

We first assume Σ to have a standard hyper inverse Wishart distribution, $\Sigma \sim \mathrm{HIW}_{\mathscr{G}}(\delta, I)$, where $I \equiv I_{|V|}$ is the $|V| \times |V|$ identity matrix, and give an extension of Odell & Feiveson's (1966) result; see also Muirhead (1982, Theorem 3.2.14).

Theorem 3. *Suppose $\Sigma \sim \mathrm{HIW}_{\mathscr{G}}(\delta, I)$, where $\mathscr{G} = (V, E)$ is a decomposable graph and $\delta > 0$ is an integer, and put $\Sigma^{-1} = \Psi^{\mathrm{T}}\Psi$. Then the main diagonal elements ψ_{ii} of Ψ are square roots of $\chi^2_{\delta+\nu_i}$, the off-diagonal elements ψ_{rs} of Ψ, with $r < s$ and $(\alpha_r, \alpha_s) \in E$, are $N(0, 1)$, and these random variables are mutually independent.*

For the proof see the Appendix.

We turn now to a hyper inverse Wishart distribution with matrix parameter B, that is $\Sigma \sim \mathrm{HIW}_{\mathscr{G}}(\delta, B)$. The distribution of Σ is strong hyper-Markov so that the independence structure of Φ is given in Theorem 2. The joint distribution of the elements of Φ can be obtained by considering the row-block submatrices in Corollary 2 independently, and by noting that, as a consequence of the zero structure of matrices $\Sigma^{-1}_{H_jH_j} = \Phi^{\mathrm{T}}_{H_jH_j}\Phi_{H_jH_j}$ or, more generally, of parametric collapsibility properties of graphical models with graphs $\mathscr{G}_{H_j}$ (Whittaker, 1990, p. 397), we can write

$$\Phi^{\mathrm{T}}_{C_1C_1}\Phi_{C_1C_1} = \Sigma^{-1}_{C_1C_1} \tag{14}$$

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

$$\begin{pmatrix} \Phi^{\mathrm{T}}_{R_jR_j} & 0 \\ \Phi_{S_jR_j} & * \end{pmatrix}\begin{pmatrix} \Phi_{R_jR_j} & \Phi_{R_jS_j} \\ 0 & * \end{pmatrix} = \Sigma^{-1}_{C_jC_j}, \tag{15}$$

where asterisks denote submatrices which need not be specified explicitly. By the definition of the hyper inverse Wishart distribution we have $\Sigma^{-1}_{C_jC_j} \sim W(\delta + c_j - 1, B^{-1}_{C_jC_j})$, so that the independent row-block submatrices in (10) can be obtained from the Cholesky decomposition of suitably specified independent Wishart matrices.

The probability structure of Φ can be exploited to derive some useful properties of the hyper inverse Wishart distribution.

Consider first $\Sigma \sim \mathrm{IW}(\delta, B)$ and let $B^{-1} = Q^{\mathrm{T}}Q$. If for a positive definite matrix D we put $D^{-1} = P^{\mathrm{T}}P$ and $O = Q^{-1}P$, then we have

$$\begin{array}{ll} (\Phi^{\mathrm{T}}\Phi)^{-1} & \sim \mathrm{IW}(\delta, B) \\ \Downarrow & \\ \Phi^{\mathrm{T}}\Phi & \sim W(\delta + |V| - 1, B^{-1}) \\ \Downarrow & \\ (\Phi O)^{\mathrm{T}}(\Phi O) & \sim W(\delta + |V| - 1, D^{-1}) \\ \Downarrow & \\ \{(\Phi O)^{\mathrm{T}}(\Phi O)\}^{-1} & \sim \mathrm{IW}(\delta, D). \end{array} \tag{16}$$

This procedure can be applied recursively to the quantities (14) and (15) to establish the following theorem.

THEOREM 4. *Suppose* $\Sigma \sim \mathrm{HIW}_{\mathscr{G}}(\delta, B)$, *where* $\mathscr{G} = (V, E)$ *is a decomposable graph and* $\delta > 0$ *is an integer, and let* D *be a* $|V| \times |V|$ *positive definite matrix. Put* $\Sigma^{-1} = \Phi^{\mathrm{T}}\Phi$ *and, for* $j = 1, \ldots, k$, *let* $B^{-1}_{C_jC_j} = Q^{j\mathrm{T}}Q^j$, $D^{-1}_{C_jC_j} = P^{j\mathrm{T}}P^j$ *and* $O^j = (Q^j)^{-1}P^j$. *Then the upper triangular matrix* $\Upsilon = \mathscr{T}(\Phi; O^1, \ldots, O^k)$ *defined by* $\Upsilon \in M(\mathscr{G})$,

$$\Upsilon_{C_1C_1} = \Phi_{C_1C_1}O^1,$$

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

$$\Upsilon_{R_jR_j} = \Phi_{R_jR_j}O^j_{R_jR_j}, \quad \Upsilon_{R_jS_j} = \Phi_{R_jR_j}O^j_{R_jS_j} + \Phi_{R_jS_j}O^j_{S_jS_j},$$

is such that $(\Upsilon^{\mathrm{T}}\Upsilon)^{-1}$ *has distribution* $\mathrm{HIW}_{\mathscr{G}}(\delta, D)$.

For the proof see the Appendix.

Note that the procedure (16), involving the Cholesky decomposition of both B and D, is not efficient, and it is usually more effectively presented with respect to an arbitrary matrix $O^{\mathrm{T}}O$, where O is not necessarily triangular; see for example Muirhead (1982, Theorem 3.2.11). However, in Theorem 4 this is not possible since, like $\{(Q^{1\mathrm{T}}Q^1)^{-1}, \ldots, (Q^{k\mathrm{T}}Q^k)^{-1}\}$ and $\{(P^{1\mathrm{T}}P^1)^{-1}, \ldots, (P^{k\mathrm{T}}P^k)^{-1}\}$, the collection $\{(O^{1\mathrm{T}}O^1)^{-1}, \ldots, (O^{k\mathrm{T}}O^k)^{-1}\}$ is in general not made up of submatrices of the same matrix. Moreover, the upper triangular form of matrices O^j is a fundamental requirement for the proof of the theorem.

For $D=I$, Theorem 4 gives a procedure for standardising a hyper inverse Wishart matrix, whereas for $B=I$ it can be applied together with Theorem 3 to generate hyper inverse Wishart random variates efficiently.

Although we use a positive definite matrix B as parameter of the hyper inverse Wishart distribution, only the collection of submatrices $\{B_{C_1C_1}, \ldots, B_{C_kC_k}\}$ enters into the specification of the distribution. Therefore, in Theorem 4 also only the corresponding submatrices of D need to be specified. In § 2·2 we identified B uniquely by requiring that $B^{-1} \in M^+(\mathscr{G})$. The following corollary motivates such a choice.

COROLLARY 3. *For a hyper inverse Wishart matrix $\Sigma \sim \text{HIW}_{\mathscr{G}}(\delta, B)$ let $\Xi \sim \text{HIW}_{\mathscr{G}}(\delta, I)$ be its standardisation obtained by setting $D=I$ in Theorem* 4. *Then*

$$|\Sigma| = \frac{\prod_{i=1}^{k} |B_{C_iC_i}|}{\prod_{i=2}^{k} |B_{S_iS_i}|} |\Xi|,$$

so that if $B^{-1} \in M^+(\mathscr{G})$ we have $|\Sigma| = |B|\,|\Xi|$.

Proof. For $B^{-1} \in M^+(\mathscr{G})$ put $B^{-1} = Q^{\mathrm{T}}Q$. Using the notation of Theorem 4, by (2) we have to show that $|(\Phi^{\mathrm{T}}\Phi)^{-1}| = |(\Upsilon^{\mathrm{T}}\Upsilon)^{-1}|\,|(Q^{\mathrm{T}}Q)^{-1}|$, which is equivalent to $|\Phi| = |\Upsilon|\,|Q|$. The product of two upper triangular matrices Υ and Q is an upper triangular matrix with main diagonal given by the product, element by element, of the diagonals of Υ and Q. Consequently, in order to have $|\Phi| = |\Upsilon|\,|Q|$ it is sufficient that Φ and ΥQ have the same diagonal.

For $D=I$, the matrix Υ is defined by $\Upsilon = \mathscr{T}\{\Phi; (Q^1)^{-1}, \ldots, (Q^k)^{-1}\}$ so that $\Phi = \mathscr{T}\{\Upsilon; Q^1, \ldots, Q^k\}$. Thus, by construction, the main diagonal of Φ is the product, element by element, of the main diagonals of $\Upsilon_{C_1C_1}$ and Q^1, $\Upsilon_{R_2R_2}$ and $Q^2_{R_2R_2}$, and so on. The desired result follows because, by a procedure similar to that used to show (14) and (15), the block-diagonal elements of Q can be shown to be $(Q^1, Q^2_{R_2R_2}, \ldots, Q^k_{R_kR_k})$. □

Since $|\Sigma^{-1}| = |\Sigma|^{-1}$, Theorem 3 and Corollary 3 allow the following immediate generalisations of Theorems 3.2.15 and 3.2.16 of Muirhead (1982).

COROLLARY 4. *If $\Sigma \sim \text{HIW}_{\mathscr{G}}(\delta, B)$, where $\delta > 0$ is an integer and $B^{-1} \in M^+(\mathscr{G})$, then $(|\Sigma|/|B|)^{-1}$ has the same distribution as $\prod_{i=1}^{|V|} \chi^2_{\delta+\nu_i}$, where $\chi^2_{\delta+\nu_i}$ denote mutually independent chi-squared random variables.*

COROLLARY 5. *If $\Sigma \sim \text{HIW}_{\mathscr{G}}(\delta, \delta B)$, where $\delta > 0$ is an integer and $B^{-1} \in M^+(\mathscr{G})$, then the asymptotic distribution as $\delta \to \infty$ of*

$$\left(\frac{\delta}{2|V|}\right)^{\frac{1}{2}} \log\left(\frac{|\Sigma|}{|B|}\right)^{-1}$$

is $N(0, 1)$.

We point out that in this section the main emphasis could have been given to the Wishart and the $\mathscr{G}$-conditional Wishart distributions, thereby allowing a more straightforward comparison of the material presented here with the existing literature. However, for Bayesian analysis of covariance selection models it is more useful to make clear how the given procedures are to be implemented for the hyper inverse Wishart distribution.

Appendix

Proofs

Proof of Proposition 2. For any positive function $f(.)$ we have

$$\prod_{i=1}^{c_j} f\{(\delta+c_j-i)/2\} = \prod_{i=0}^{c_j-1} f\{(\delta+i)/2\}$$

so that

$$\frac{\prod_{i=1}^{c_j} f\{(\delta+c_j-i)/2\}}{\prod_{i=1}^{s_j} f\{(\delta+s_j-i)/2\}} = \prod_{i=s_j}^{c_j-1} f\{(\delta+i)/2\} = \prod_{\alpha_i \in R_j} f\{(\delta+v_i)/2\},$$

$$\frac{\prod_{j=1}^{k}\prod_{i=1}^{c_j} f\{(\delta+c_j-i)/2\}}{\prod_{j=2}^{k}\prod_{i=1}^{s_j} f\{(\delta+s_j-i)/2\}} = \prod_{i=1}^{p} f\{(\delta+v_i)/2\}. \tag{A1}$$

We also remark that

$$v = \sum_{j=1}^{k} c_j(c_j-1)/2 - \sum_{j=2}^{k} s_j(s_j-1)/2. \tag{A2}$$

Consider the first term of (12). Using the definition of the multivariate gamma function along with (A1) and (A2) we have

$$\frac{\prod_{j=1}^{k} \Gamma_{c_j}\{(\delta+c_j-1)/2\}}{\prod_{j=2}^{k} \Gamma_{s_j}\{(\delta+s_j-1)/2\}} = \frac{\prod_{j=1}^{k} \pi^{c_j(c_j-1)/4} \prod_{i=1}^{c_j} \Gamma\{(\delta+c_j-i)/2\}}{\prod_{j=2}^{k} \pi^{s_j(s_j-1)/4} \prod_{i=1}^{s_j} \Gamma\{(\delta+s_j-i)/2\}}$$

$$= \pi^{v/2} \prod_{i=1}^{p} \Gamma\{(\delta+v_i)/2\}. \tag{A3}$$

Consider now the second term of (12). It can be checked that for any integer $m > c_j - 1$ the following identity holds:

$$\frac{2^{c_j(c_j-1)/4}}{2^{c_j m/2}} = 2^{-c_j(2m-c_j+1)/4} = \prod_{i=1}^{c_j} 2^{-\{m-(i-1)\}/2}.$$

Thus if we set $m = \delta + c_j - 1$, for any integer $\delta > 0$ the above equation takes the form

$$\frac{2^{c_j(c_j-1)/4}}{2^{c_j\{\delta+c_j-1\}/2}} = \prod_{i=1}^{c_j} 2^{-(\delta+c_j-i)/2}. \tag{A4}$$

Using (A1), (A2) and (A4) we obtain

$$\frac{\prod_{j=1}^{k} 2^{-c_j(\delta+c_j-1)/2}}{\prod_{j=2}^{k} 2^{-s_j(\delta+s_j-1)/2}} = 2^{-v/2} \frac{\prod_{j=1}^{k} 2^{c_j(c_j-1)/4} 2^{-c_j(\delta+c_j-1)/2}}{\prod_{j=2}^{k} 2^{s_j(s_j-1)/4} 2^{-s_j(\delta+s_j-1)/2}}$$

$$= 2^{-v/2} \frac{\prod_{j=1}^{k} (2^{c_j(c_j-1)/4}/2^{c_j(\delta+c_j-1)/2})}{\prod_{j=2}^{k} (2^{s_j(s_j-1)/4}/2^{s_j(\delta+s_j-1)/2})}$$

$$= 2^{-v/2} \prod_{i=1}^{p} 2^{-(\delta+v_i)/2}. \tag{A5}$$

The result follows by substituting (A3) and (A5) in (12). □

Proof of Theorem 3. As shown in § 5, $K = \Sigma^{-1} \sim W_{\mathscr{G}}(\delta + |V| - 1, I)$, with density

$$q_{\mathscr{G}}(K \mid \delta, I) = h_{\mathscr{G}}(\delta, V)|K|^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(K)\right\}. \tag{A6}$$

Consider the change of variables to $K = \Psi^{\mathrm{T}}\Psi$. In order to compute the Jacobian of this transformation we take the distinct nonzero elements of $K = \{\kappa_{rs}\}$, ordered according to the columns of K, and similarly for the elements of Ψ. In this way the Jacobian matrix J is triangular and its determinant is the product of the diagonal elements:

$$|J| = \left(\prod_{r=1}^{p} \frac{d}{d\psi_{rr}} \kappa_{rr}\right) \times \left(\prod_{(\alpha_r,\alpha_s)\in E} \frac{d}{d\psi_{rs}} \kappa_{rs}\right), \tag{A7}$$

where $p = |V|$. The p elements of the first term of (A7) are

$$\frac{d}{d\psi_{rr}} \kappa_{rr} = \frac{d}{d\psi_{rr}} \sum_{i=1}^{r} \psi_{ir}^2 = 2\psi_{rr},$$

and the v elements of the second term are

$$\frac{d}{d\psi_{rs}} \kappa_{rs} = \frac{d}{d\psi_{rs}} \sum_{i=1}^{r} \psi_{ir}\psi_{is} = \psi_{rr}.$$

Since the derivatives above only depend on the row position of κ_{rs}, and the ith row of Ψ has exactly $v_i + 1$ nonzero elements, it follows that

$$|J| = 2^p \prod_{i=1}^{p} \psi_{ii}^{v_i+1}. \tag{A8}$$

If we substitute in (A6) the expressions

$$\operatorname{tr}(K) = \sum_{i=1}^{p} \psi_{ii}^2 + \sum_{(\alpha_r,\alpha_s)\in E} \psi_{rs}^2, \quad |K| = \prod_{i=1}^{p} \psi_{ii}^2,$$

multiply it by $|J|$ from (A8) and write $h_{\mathscr{G}}(\delta, V)$ in the form given in Proposition 1, the joint density of the elements ψ_{rs}, with $r < s$ and $(\alpha_r, \alpha_s) \in E$, and ψ_{ii}, for $i = 1, \ldots, p$, can be written as

$$\left(\prod_{(\alpha_r,\alpha_s)\in E} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\psi_{rs}^2}\right) \times \left[\prod_{i=1}^{p} \frac{2^{-(\delta+v_i)/2}}{\Gamma\{(\delta+v_i)/2\}} (\psi_{ii}^2)^{(\delta+v_i)/2-1} e^{-\frac{1}{2}\psi_{ii}^2} \frac{d\psi_{ii}^2}{d\psi_{ii}}\right],$$

which is the product of the marginal density functions for the elements of Ψ stated in the theorem.

□

Proof of Theorem 4. We first show that if Ψ has the structure given in Theorem 3, so that $(\Psi^{\mathrm{T}}\Psi)^{-1} \sim \mathrm{HIW}_{\mathscr{G}}(\delta, I)$, the upper triangular matrix $\Phi = \mathscr{T}(\Psi; Q^1, \ldots, Q^k)$, with nonzero elements

$$\begin{aligned} \Phi_{C_1C_1} &= \Psi_{C_1C_1} Q^1, \\ \Phi_{R_jR_j} &= \Psi_{R_jR_j} Q^j_{R_jR_j} \quad (j = 2, \ldots, k), \\ \Phi_{R_jS_j} &= \Psi_{R_jR_j} Q^j_{R_jS_j} + \Psi_{R_jS_j} Q^j_{S_jS_j} \quad (j = 2, \ldots, k), \end{aligned} \tag{A9}$$

is such that $(\Phi^{\mathrm{T}}\Phi)^{-1} \sim \mathrm{HIW}_{\mathscr{G}}(\delta, B)$.

The submatrices $\Phi_{C_1C_1}$, $(\Phi_{R_2R_2}, \Phi_{R_2S_2}), \ldots, (\Phi_{R_kR_k}, \Phi_{R_kS_k})$ in (A9) are constructed as functions of independent submatrices of Ψ, and are themselves independent as required in Corollary 2. Consequently, to prove that $(\Phi^{\mathrm{T}}\Phi)^{-1} \sim \mathrm{HIW}_{\mathscr{G}}(\delta, B)$, it is sufficient to show that such submatrices can be obtained by Cholesky decomposition of $W(\delta + c_j - 1, B^{-1}_{C_jC_j})$ matrices as in (14) and (15).

For $j = 1$ we have $\Psi^{\mathrm{T}}_{C_1C_1}\Psi_{C_1C_1} \sim W(\delta + c_1 - 1, I_{c_1})$ so that $\Phi^{\mathrm{T}}_{C_1C_1}\Phi_{C_1C_1}$ is $W(\delta + c_1 - 1, B^{-1}_{C_1C_1})$ as required.

For all $j = 2, \ldots, k$, we can always construct an $s_j \times s_j$ upper triangular matrix $\Theta_{S_jS_j}$ independent of $(\Psi_{R_jR_j}, \Psi_{R_jS_j})$ and such that $\Theta^{\mathrm{T}}_{S_jS_j}\Theta_{S_jS_j} \sim W(\delta + s_j - 1, I_{s_j})$. Hence, if we put

$$\Theta_{C_jC_j} = \begin{pmatrix} \Psi_{R_jR_j} & \Psi_{R_jS_j} \\ 0 & \Theta_{S_jS_j} \end{pmatrix}$$

we have $\Theta^{\mathrm{T}}_{C_jC_j}\Theta_{C_jC_j} \sim W(\delta + c_j - 1, I_{c_j})$ so that

$$(\Theta_{C_jC_j} Q^j)^{\mathrm{T}} (\Theta_{C_jC_j} Q^j) \sim W(\delta + c_j - 1, B^{-1}_{C_jC_j}),$$

where

$$\Theta_{C_jC_j}Q^j = \begin{pmatrix} \Psi_{R_jR_j}Q^j_{R_jR_j} & \Psi_{R_jR_j}Q^j_{R_jS_j} + \Psi_{R_jS_j}Q^j_{S_jS_j} \\ 0 & \Theta_{S_jS_j}Q^j_{S_jS_j} \end{pmatrix} = \begin{pmatrix} \Phi_{R_jR_j} & \Phi_{R_jS_j} \\ 0 & * \end{pmatrix},$$

and the first part of the proof is complete.

Thus, we have shown that if $\Sigma = (\Phi^{\mathrm{T}}\Phi)^{-1} \sim \mathrm{HIW}_{\mathscr{G}}(\delta, B)$ we have

$$\Phi = \mathscr{T}(\Psi; Q^1, \ldots, Q^k),$$

in distribution. Consequently

$$\begin{aligned} \Upsilon &= \mathscr{T}(\Phi; O^1, \ldots, O^k) = \mathscr{T}\{\mathscr{T}(\Psi; Q^1, \ldots, Q^k); O^1, \ldots, O^k\} \\ &= \mathscr{T}\{\mathscr{T}(\Psi; Q^1, \ldots, Q^k); (Q^1)^{-1}P^1, \ldots, (Q^k)^{-1}P^k\}, \end{aligned} \tag{A10}$$

in distribution. If we recall that

$$(Q^j)^{-1}P^j = \begin{pmatrix} (Q^j_{R_jR_j})^{-1}P^j_{R_jR_j} & (Q^j_{R_jR_j})^{-1}P^j_{R_jS_j} - (Q^j_{R_jR_j})^{-1}Q^j_{R_jS_j}(Q^j_{S_jS_j})^{-1}P^j_{S_jS_j} \\ 0 & (Q^j_{S_jS_j})^{-1}P^j_{S_jS_j} \end{pmatrix},$$

it is not hard to check that (A10) simplifies to $\Upsilon = \mathscr{T}(\Psi; P^1, \ldots, P^k)$, in distribution. Replacing Q by P in (A9) we obtain that $(\Upsilon^{\mathrm{T}}\Upsilon)^{-1}$ is $\mathrm{HIW}_{\mathscr{G}}(\delta, D)$, and the proof is complete. □

REFERENCES

DARROCH, J. N., LAURITZEN, S. L. & SPEED, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8**, 522–39.
DAWID, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68**, 265–74.
DAWID, A. P. & LAURITZEN, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–317.
DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–75.
DIACONIS P. & YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269–81.
GUTIÉRREZ-PEÑA, E. & SMITH, A. F. M. (1997). Exponential and Bayesian conjugate families: review and extensions. *Test* **6**, 1–90.
LAURITZEN, S. L. (1996). *Graphical Models*. Oxford: Oxford University Press.
MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.
ODELL, P. L. & FEIVESON, A. G. (1966). A numerical procedure to generate a sample covariance matrix. *J. Am. Statist. Assoc.* **61**, 199–203.
PAULSEN, V. I., POWER, S. C. & SMITH, R. R. (1989). Schur products and matrix completions. *J. Funct. Anal.* **85**, 151–78.
PRESS, S. J. (1972). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. New York: Holt, Rineholt and Winston.
ROVERATO, A. & WHITTAKER, J. (1998). The Isserlis matrix and its application to non-decomposable graphical Gaussian models. *Biometrika* **85**, 711–25.
SPEED, T. P. & KIIVERI, H. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 138–50.
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. Am. Statist. Assoc.* **75**, 963–72.
WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

[*Received January* 1999. *Revised June* 1999]