WISHART DISTRIBUTIONS FOR DECOMPOSABLE COVARIANCE GRAPH MODELS

BY KSHITIJ KHARE[1] AND BALA RAJARATNAM[2]

University of Florida and Stanford University

Gaussian covariance graph models encode marginal independence among the components of a multivariate random vector by means of a graph G. These models are distinctly different from the traditional concentration graph models (often also referred to as Gaussian graphical models or covariance selection models) since the zeros in the parameter are now reflected in the covariance matrix Σ, as compared to the concentration matrix $\Omega = \Sigma^{-1}$. The parameter space of interest for covariance graph models is the cone P_G of positive definite matrices with fixed zeros corresponding to the missing edges of G. As in Letac and Massam [*Ann. Statist.* **35** (2007) 1278–1323], we consider the case where G is decomposable. In this paper, we construct on the cone P_G a family of Wishart distributions which serve a similar purpose in the covariance graph setting as those constructed by Letac and Massam [*Ann. Statist.* **35** (2007) 1278–1323] and Dawid and Lauritzen [*Ann. Statist.* **21** (1993) 1272–1317] do in the concentration graph setting. We proceed to undertake a rigorous study of these "covariance" Wishart distributions and derive several deep and useful properties of this class. First, they form a rich conjugate family of priors with multiple shape parameters for covariance graph models. Second, we show how to sample from these distributions by using a block Gibbs sampling algorithm and prove convergence of this block Gibbs sampler. Development of this class of distributions enables Bayesian inference, which, in turn, allows for the estimation of Σ, even in the case when the sample size is less than the dimension of the data (i.e., when "$n < p$"), otherwise not generally possible in the maximum likelihood framework. Third, we prove that when G is a homogeneous graph, our covariance priors correspond to standard conjugate priors for appropriate directed acyclic graph (DAG) models. This correspondence enables closed form expressions for normalizing constants and expected values, and also establishes hyper-Markov properties for our class of priors. We also note that when G is homogeneous, the family IW_{Q_G} of Letac and Massam [*Ann. Statist.* **35** (2007) 1278–1323] is a special case of our covariance Wishart distributions. Fourth, and finally, we illustrate the use of our family of conjugate priors on real and simulated data.

1. Introduction. Due to recent advances in science and information technology, there has been a huge influx of high-dimensional data from various fields such as genomics, environmental sciences, finance and the social sciences. Making sense of all the many complex relationships and multivariate dependencies present in the data, formulating correct models and developing inferential procedures is one of the major challenges in modern day statistics. In parametric models, the covariance or correlation matrix (or its inverse) is the fundamental object that quantifies relationships between random variables. Estimating the covariance matrix in a sparse way is crucial in high-dimensional problems and enables the detection of the most important relationships. In this light, graphical models have served as tools to discover structure in high-dimensional data.

The primary aim of this paper is to develop a new family of conjugate prior distributions for covariance graph models (a subclass of graphical models) and study the properties of this family of distributions. It is shown in this paper that these properties are highly attractive for Bayesian inference in high-dimensional settings. In covariance graph models, specific entries of the covariance matrix are restricted to be zero, which implies marginal independence in the Gaussian case. Covariance graph models correspond to curved exponential families and are distinctly different from the well-studied concentration graph models, which, in turn, correspond to natural exponential families.

A rich framework for Bayesian inference for natural exponential families has been established in the last three decades, starting with the seminal and celebrated work of Diaconis and Ylvisaker [10] that laid the foundations for constructing conjugate prior distributions for natural exponential family models. The Diaconis–Ylvisaker (henceforth referred to as "DY") conjugate priors are characterized by posterior linearity of the mean. An analogous framework for curved exponential families is not available in the literature.

Concentration graph models (or covariance selection models) were one of the first graphical models to be formally introduced to the statistics community. These models reflect conditional independencies in multivariate probability distributions by means of a graph. In the Gaussian case, they induce sparsity or zeros in the inverse covariance matrix and correspond to natural exponential families. In their pioneering work, Dawid and Lauritzen [9] developed the DY prior for this class of models. In particular, they introduced the hyper-inverse Wishart as the DY conjugate prior for concentration graph models. In a recent major contribution to this field, a rich family of conjugate priors that subsumes the DY class has been developed by Letac and Massam [20]. Both the hyper-inverse Wishart priors and the "Letac–Massam" priors have attractive properties which enable Bayesian inference, with the latter allowing multiple shape parameters and hence being suitable in high-dimensional settings. Bayesian procedures corresponding to these Letac–Massam priors have been derived in a decision theoretic framework in the recent work of Rajaratnam, Massam and Carvalho [26].

Consider an undirected[3] graph G with a finite set of vertices V (of size p) and a finite set E of edges between these vertices, that is, $G = (V, E)$. The Gaussian covariance graph model corresponding to the graph G is the collection of p-variate Gaussian distributions with covariance matrix Σ such that $\Sigma_{ij} = 0$ whenever $(i, j) \notin E$. This class of models was first formally introduced by Cox and Wermuth [6, 7]. In the frequentist setting, maximum likelihood estimation in covariance graph models has been a topic of interest in recent years. Many iterative methods that obtain the maximum likelihood estimate have been proposed in the literature. The graphical modeling software MIM in Edwards [12] fits these models by using the "dual likelihood method" from Kauermann [17]. In Wermuth, Cox and Marchetti [31], the authors derive asymptotically efficient approximations to the maximum likelihood estimate in covariance graph models for exponential families. Chaudhuri, Drton and Richardson [4] propose an iterative conditional fitting algorithm for maximum likelihood estimation in this class of models. Covariance graph models have also been used in applications in Butte et al. [3], Grzebyk, Wild and Chouaniere [15], Mao, Kschischang and Frey [21] and others.

Although Gaussian covariance graph models are simple and intuitive to understand, no comprehensive theoretical framework for Bayesian inference for this class of models has been developed in the literature. In that sense, Bayesian inference for covariance graph models has been an open problem since the introduction of these models by Cox and Wermuth [6, 7] more than fifteen years ago. The main difficulty is that these models give rise to curved exponential families. The zero restrictions on the entries of the covariance matrix Σ translate into complicated restrictions on the corresponding entries of the natural parameter, $\Omega = \Sigma^{-1}$. Hence, the sparseness in Σ does not translate into sparseness in Σ^{-1} and thus a covariance graph model cannot be viewed as a concentration graph model. No general theory is available for Bayesian inference in curved exponential families for continuous random variables, akin to the Diaconis–Ylvisaker [10] or standard conjugate theory for natural exponential families.

There are several desirable properties that one might want when constructing a class of priors, but one of the foremost requirements is to be able to compute quantities such as the mean or mode of the posterior distribution, either in closed form or by sampling from the posterior distribution by a simple mechanism. This is especially important in high-dimensional situations, where computations are complex and can become infeasible very quickly. Another desirable and related feature is conjugacy, that is, the class of priors is such that the posterior distribution also belongs to this class. Among other things, this increases the prospects of obtaining closed form Bayes estimators and can also add to the interpretability of the hyper-parameters. The class of Wishart distributions developed by Letac and Massam [20] (and later used for flexible Bayesian inference for concentration graph

---

[3]We shall use dotted edges for our graphs, in keeping with the notation in the literature; bi-directed edges have also been used for representing covariance graphs.

models by Rajaratnam, Massam and Carvalho [26]), known as the IW_{P_G} family of distributions, are not appropriate for the covariance graph setting. There is the additional option of using the IW_{Q_G} class as priors for this situation. We, however, establish that the posterior distribution fails to belong to the same class and there are no known results for computing the posterior mean or mode, either in closed form or by sampling from the posterior distribution.

A principal objective of this paper is to develop a framework for Bayesian inference for Gaussian covariance graph models. We proceed to construct a rich and flexible class of conjugate Wishart distributions, with multiple shape parameters, on the space of positive definite matrices with fixed zeros, that corresponds to a decomposable graph G. This class of distributions is specified up to a normalizing constant, and conditions under which this normalizing constant can be evaluated in closed form are derived. We explore the distributional properties of our class of priors and, in particular, show that the parameter can be partitioned into blocks so that the conditional distribution of each block, given the others, is tractable. Based on this property, we propose a block Gibbs sampling algorithm to simulate from the posterior distribution. We proceed to formally prove the convergence of this block Gibbs sampler. Our priors yield proper inferential procedures, even in the case when the sample size n is less than the dimension p of the data, whereas maximum likelihood estimation is, in general, only possible when $n \geq p$ (in fact, in the homogeneous case, it can be shown that the condition $n \geq p$ is actually also necessary, thus highlighting the fact that results from the concentration graph setting do not carry over to the covariance model setting). We also show that our covariance Wishart distributions are, in the decomposable nonhomogeneous case, very different from the Letac–Massam priors W_{P_G} and IW_{Q_G}. However, when the underlying graph G is homogeneous, the Letac–Massam IW_{Q_G} priors are a special case of our distributions. We establish, in the homogeneous setting, a correspondence between the covariance priors in this paper and the natural conjugate priors for appropriate directed acyclic graph (DAG) models. This correspondence helps us to explicitly evaluate quantities like the normalizing constant and the posterior mean of the covariance matrix in closed form. In this scenario, we also show that our class of priors satisfies the strong directed hyper-Markov property (as introduced in Dawid and Lauritzen [9] for concentration graph models). It should be pointed out that these aforementioned results for homogeneous graphs can also be established directly, without exploiting the correspondence with the DAG models. The direct approach is self-contained, whereas the latter invokes an external result which states that for the restrictive class of homogeneous graphs, covariance graph models and DAGs are Markov equivalent.

We noted above that for concentration graph models or the traditional Gaussian graphical models, a rich theory has been established by Dawid and Lauritzen [9], who derive the single parameter DY conjugate prior for these models, and by Letac and Massam [20], who derive a larger flexible class with multiple shape parameters. In essence, this paper is the analog of the results in the two aforementioned

papers in the covariance graph model setting, with parallel results, all of which are contained in a single comprehensive piece. Hence, this work completes the powerful theory that has been developed in the mathematical statistics literature for decomposable models.

We also point out that a class of priors in the recent work [29] is a special case of our class of flexible covariance Wishart distributions.[4] Our family allows multiple shape parameters, as compared to a single shape parameter, and hence yields a richer class suitable to high-dimensional problems. Moreover, we show that their iterative algorithm to sample from the posterior is different from ours. Since the authors do not undertake a theoretical investigation of the convergence properties of their algorithm, it is not clear if it does indeed converge to the desired distribution. On the other hand, we proceed to formally prove that our algorithm converges to the desired distribution. The remaining sections of this paper are considerably different from [29] since we undertake a rigorous probabilistic analysis of our conjugate Wishart distributions for covariance graph models, whereas they give a useful and novel treatment of latent variables and mixed graph models in a machine learning context.

This paper is structured as follows. Section 2 introduces the required preliminaries and notation. In Section 3, the class of covariance Wishart distributions is formally constructed. Conjugacy to the class of covariance graph models and sufficient conditions for integrability are established. Comparison with the Letac–Massam IW_{Q_G} priors, which are not, in general, conjugate in the covariance graph setting, is also undertaken. In Section 4, a block Gibbs sampler which enables sampling from the posterior distribution is proposed and the corresponding conditional distributions are derived. Thereafter, a formal proof of convergence of this block Gibbs sampler is provided. In Section 5, we restrict ourselves to the case when G is a homogeneous graph. We examine the distributional properties of our class of priors in this section and prove that the covariance priors introduced in this paper correspond to natural conjugate priors for DAG models in the homogeneous setting. This correspondence helps in establishing closed form expressions for normalizing constants, expected values and hyper-Markov properties for our class of priors for G homogeneous. Finally, we illustrate the use of our family of conjugate priors and the methodology developed in this paper on a real example, as well as on simulated data. The Appendix contains the proofs of some of the results stated in the main text.

2. Preliminaries. In this section, we give the necessary notation, background and preliminaries that are needed in subsequent sections.

---

[4]This is in a similar spirit to the way in which the HIW prior of Dawid and Lauritzen [9] is a special case of the generalized family of Wishart distributions proposed by Letac and Massam [20] for the concentration graph setting.

2.1. *Modified Cholesky decomposition.* If Σ is a positive definite matrix, then there exists a unique decomposition

$$(2.1) \qquad \Sigma = LDL^T,$$

where L is a lower-triangular matrix with diagonal entries equal to 1 and D a diagonal matrix with positive diagonal entries. This decomposition of Σ is referred to as the *modified Cholesky decomposition* of Σ (see [25]). We now provide a formula that explicitly computes the inverse of a lower-triangular matrix with 1's on the diagonal, such as those that appear in (2.1).

PROPOSITION 1. *Let L be an $m \times m$ lower-triangular matrix with 1's on the diagonal. Let*

$$\mathcal{A} = \bigcup_{r=2}^{m} \{\boldsymbol{\tau} : \boldsymbol{\tau} \in \{1, 2, \ldots, m\}^r, \tau_i < \tau_{i-1} \ \forall 2 \leq i \leq r\}$$

and

$$L_{\boldsymbol{\tau}} = \prod_{i=2}^{\dim(\boldsymbol{\tau})} L_{\tau_{i-1}\tau_i} \qquad \forall \boldsymbol{\tau} \in \mathcal{A},$$

where $\dim(\tau)$ denotes the length of the vector τ. Then, $L^{-1} = N$, where N is lower-triangular matrix with 1's on the diagonal and, for $i > j$,

$$N_{ij} = \sum_{\boldsymbol{\tau} \in \mathcal{A}, \tau_1 = i, \tau_{\dim \tau} = j} (-1)^{\dim(\tau)-1} \prod_{i=2}^{\dim(\tau)} L_{\tau_{i-1}\tau_i}.$$

The proof is provided in the Appendix.

An undirected graph G is a pair (V, E), where V is a permutation[5] of the set $\{1, 2, \ldots, m\}$ denoting the set of vertices of G. The set $E \subseteq V \times V$ denotes the set of edges in the graph. If vertices u and v are such that $(u, v) \in E$, then we say that there is an edge between u and v. It is also understood that $(u, v) \in E$ implies that $(v, u) \in E$, that is, the edges are undirected. Although the dependence of $G = (V, E)$ on the particular ordering in V is often suppressed, the reader should bear in mind that unlike traditional graphs, the graphs defined above are not equivalent up to permutation of the vertices[6] modulo the edge structure. Below, we describe two classes of graphs which play a central role in this paper.

---

[5]The ordering in V is emphasized here since the elements of V will later correspond to rows or columns of matrices.

[6]This has been done for notational convenience, as will be seen later.

2.2. *Decomposable graphs.* An undirected graph G is said to be *decomposable* if any induced subgraph does not contain a cycle of length greater than or equal to four. The reader is referred to Lauritzen [19] for all of the common notions of graphical models (and, in particular, decomposable graphs) that we will use here. One such important notion is that of a perfect order of the cliques. Every decomposable graph admits a perfect order of its cliques. Let $(C_1, C_2, \ldots, C_k)$ be one such perfect order of the cliques of the graph G. The *history* for the graph is given by $H_1 = C_1$ and

$$H_j = C_1 \cup C_2 \cup \cdots \cup C_j, \qquad j = 2, 3, \ldots, k,$$

and the *minimal separators* of the graph are given by

$$S_j = H_{j-1} \cap C_j, \qquad j = 2, 3, \ldots, k.$$

Let

$$R_j = C_j \setminus H_{j-1} \qquad \text{for } j = 2, 3, \ldots, k.$$

Let $k' \leq k - 1$ denote the number of distinct separators and $\nu(S)$ denote the multiplicity of S, that is, the number of j such that $S_j = S$. Generally, we will denote by $\mathcal{C}$ the set of cliques of a graph and by $\mathcal{S}$ its set of separators.

Now, let Σ be an arbitrary positive definite matrix with zero restrictions according to $G = (V, E)$,[7] that is, $\Sigma_{ij} = 0$ whenever $(i, j) \notin E$. It is known that if G is decomposable, then there exists an ordering of the vertices such that if $\Sigma = LDL^T$ is the modified Cholesky decomposition corresponding to this ordering, then, for $i > j$,

(2.2) $L_{ij} = 0$ whenever $(i, j) \notin E$.

Although the ordering is not unique in general, the existence of such an ordering characterizes decomposable graphs (see [24]). A constructive way to obtain such an ordering is given as follows. Label the vertices in descending order, starting with vertices in $C_1, R_2, R_3, \ldots, R_k$, with vertices belonging to a particular set being ordered arbitrarily (see [19, 24, 30] for more details).

2.3. *The spaces P_G, Q_G and $\mathcal{L}_G$.* An m-dimensional Gaussian covariance graph model[8] can be represented by the class of multivariate normal distributions with fixed zeros in the covariance parameter (i.e., marginal independencies) described by a given graph $G = (V, E)$. That is, if $(i, j) \notin E$, then the ith and jth components of the multivariate random vector are marginally independent. Without loss of generality, we can assume that these models have mean zero and are

---

[7]It is emphasized here that the ordering of the vertices reflected in V plays a crucial role in the definitions and results that follow.

[8]A brief overview of the literature in this area is provided in the Introduction.

characterized by the parameter set P_G of positive definite covariance matrices Σ such that $\Sigma_{ij} = 0$ whenever the edge (i, j) is not in E. Following the notation in [20, 26] for G decomposable, we define Q_G to be the space on which the free elements of the precision matrices (or inverse covariance matrices) Ω live.

More formally, let M denote the set of symmetric matrices of order m, $M_m^+ \subset M$ the cone of positive definite matrices (abbreviated as "> 0"), I_G the linear space of symmetric incomplete matrices x with missing entries x_{ij}, $(i, j) \notin E$, and $\kappa : M \mapsto I_G$ the projection of M into I_G. The parameter set of the precision matrices of Gaussian covariance graph models can also be described as the set of incomplete matrices $\Omega = \kappa(\Sigma^{-1})$, $\Sigma \in P_G$. The entries Ω_{ij}, $(i, j) \notin E$, are not free parameters of the precision matrix for Gaussian covariance graph models (see [20, 26] for details). We are therefore led to consider the two cones

$$(2.3) \qquad P_G = \{y \in M_m^+ | y_{ij} = 0, (i, j) \notin E\},$$

$$(2.4) \qquad Q_G = \{x \in I_G | x_{C_i} > 0, i = 1, \ldots, k\},$$

where $P_G \subset Z_G$, $Q_G \subset I_G$ and Z_G denotes the linear space of symmetric matrices with zero entries y_{ij}, $(i, j) \notin E$. Furthermore Grone et al. [14] prove that for G decomposable, the spaces P_G and Q_G are isomorphic (once more, see [20, 26] for details).

We now introduce new spaces $\mathcal{L}_G$ and $\boldsymbol{\Theta}_G$ (the modified Cholesky space) that will be needed in our subsequent analysis[9]:

$$\mathcal{L}_G = \{L : L_{ij} = 0 \text{ whenever } i < j, \text{ or } (i, j) \notin E, \text{ and } L_{ii} = 1, \forall 1 \le i, j \le m\};$$

$$\boldsymbol{\Theta}_G = \{\boldsymbol{\theta} = (L, D) : L \in \mathcal{L}_G, D \text{ diagonal with } D_{ii} > 0 \, \forall 1 \le i \le m\}.$$

We define the mapping $\psi : \boldsymbol{\Theta}_G \to M_m^+$ as follows:

$$(2.5) \qquad \psi(L, D) = LDL^T.$$

This mapping ψ plays an important role in our analysis and shall be studied later.

2.4. *Homogeneous graphs.* A graph $G = (V, E)$ is defined to be homogeneous if, for all $(i, j) \in E$, either

$$\{u : u = j \text{ or } (u, j) \in E\} \subseteq \{u : u = i \text{ or } (u, i) \in E\}$$

or

$$\{u : u = i \text{ or } (u, i) \in E\} \subseteq \{u : u = j \text{ or } (u, j) \in E\}.$$

Equivalently, a graph G is said to be homogeneous if it is decomposable and does not contain the graph $\overset{1}{\bullet} - \overset{2}{\bullet} - \overset{3}{\bullet} - \overset{4}{\bullet}$, denoted by A_4, as an induced subgraph. Homogeneous graphs have an equivalent representation in terms of directed rooted

---

[9]These spaces are not defined in [20, 26].

trees, called *Hasse diagrams*. The reader is referred to [20] for a detailed account of the properties of homogeneous graphs. We write $i \to j$ whenever

$$\{u : u = j \text{ or } (u, j) \in E\} \subseteq \{u : u = i \text{ or } (u, i) \in E\}.$$

Denote by R the equivalence relation on V defined by

$$i R j \quad \Leftrightarrow \quad i \to j \text{ and } j \to i.$$

Let $\bar{i}$ denote the equivalence class in V/R containing i. The Hasse diagram of G is defined as a directed graph with vertex set $V_H = V/R = \{\bar{i} : i \in V\}$ and edge set E_H consisting of directed edges with $(\bar{i}, \bar{j}) \in E_H$ for $\bar{i} \neq \bar{j}$ if the following holds: $i \to j$ and $\nexists k$ such that $i \to k \to j$, $\bar{k} \neq \bar{i}$, $\bar{k} \neq \bar{j}$.

If G is a homogeneous graph, then the Hasse diagram described above is a directed rooted tree such that the number of children of a vertex is never equal to one. It was proven in [20] that there is a one-to-one correspondence between the set of homogeneous graphs and the set of directed rooted trees with vertices weighted by positive integers $[w(\bar{i}) = |\bar{i}|]$, such that no vertex has exactly one child. Also, when $i R j$, we say that i and j are *twins* in the Hasse diagram of G. Figure 1 provides an example of a homogeneous graph with seven vertices and the corresponding Hasse diagram.

The following proposition for homogeneous graphs plays an important role in our analysis.

PROPOSITION 2. *If G is a homogeneous graph, then there exists an ordering of the vertices, such that, for this ordering*:

1. $\Sigma \in P_G \Leftrightarrow L \in \mathcal{L}_G$, *where $\Sigma = LDL^T$ is the modified Cholesky decomposition of Σ*;
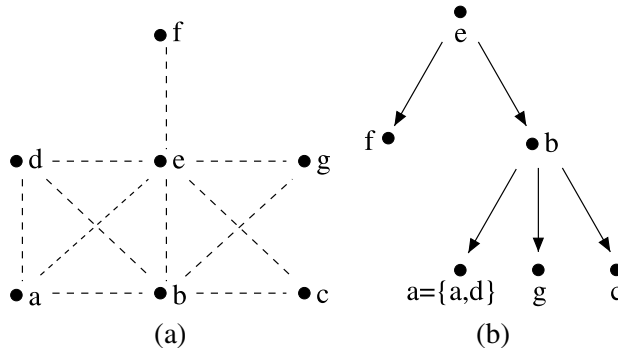2. $L \in \mathcal{L}_G \Leftrightarrow L^{-1} \in \mathcal{L}_G$.

FIG. 1. (a) *An example of a homogeneous graph with 7 vertices*; (b) *the corresponding Hasse diagram.*

The proof of this proposition is well known and so is omitted for the sake of brevity (see [1, 18, 27]). We now describe a procedure for ordering the vertices, under which Proposition 2 holds. Given a homogeneous graph G, we first construct the Hasse diagram for G. The vertices are labeled in descending order, starting from the root of the tree. If the equivalence class at any node has more than one element, then they are labeled in any order. Hereafter, we shall refer to this ordering scheme as the *Hasse perfect vertex elimination scheme*. For example, if we apply this ordering procedure to the graph in Figure 1, then the resulting labels are $\{a, b, c, d, e, f, g\} \to \{4, 5, 1, 3, 7, 6, 2\}$.

2.5. *Vertex ordering.* Let $G = (V, E)$ be an undirected decomposable graph with vertex set $V = \{1, 2, \ldots, m\}$ and edge set E. Let S_V denote the permutation group associated with V. For any $\sigma \in S_V$, let $G_\sigma := (\sigma(V), E_\sigma)$, where $(u, v) \in E_\sigma$ if and only if $(\sigma^{-1}(u), \sigma^{-1}(v)) \in E$. Let $S_D \subseteq S_V$ denote the subset of permutations σ of V such that, for any $\Sigma \in M_m^+$ with $\Sigma = LDL^T$, $L \in \mathcal{L}_{G_\sigma} \Leftrightarrow \Sigma \in P_{G_\sigma}$. Hence, for every $\sigma \in S_D$, the mapping $\psi_\sigma : \mathbf{\Theta}_{G_\sigma} \to M_m^+$ defined in (2.5) is a bijection from $\mathbf{\Theta}_{G_\sigma}$ to P_{G_σ}. In particular, the ordering corresponding to any perfect vertex elimination scheme lies in S_D (see Section 2.2). If G is homogeneous, let $S_H \subseteq S_D$ denote the subset of permutations σ of V such that $L \in \mathcal{L}_{G_\sigma} \Leftrightarrow L^{-1} \in \mathcal{L}_{G_\sigma}$. In particular, any ordering of the vertices corresponding to the Hasse perfect vertex elimination scheme lies in S_H (see Section 2.4). The above defines a nested triplet of permutations of V given by $S_H \subseteq S_D \subseteq S_V$.

3. Wishart distributions for covariance graphs. Let $G = (V, E)$ be an undirected decomposable graph with vertex set V and edge set E. We assume that the vertices in V are ordered so that $V \in S_D$. The covariance graph model associated with G is the family of distributions

$$\mathcal{G} = \{\mathcal{N}_m(\mathbf{0}, \Sigma) : \Sigma \in P_G\}$$
$$\cong \{\mathcal{N}_m(\mathbf{0}, LDL^T) : (L, D) \in \mathbf{\Theta}_G\}.$$

Consider the class of measures on $\mathbf{\Theta}_G$ with density [with respect to $\prod_{i > j, (i, j) \in E} dL_{ij} \prod_{i=1}^m dD_{ii}$]

(3.1) $\tilde{\pi}_{U, \boldsymbol{\alpha}}(L, D) = e^{-(\text{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})/2}, \qquad \theta = (L, D) \in \mathbf{\Theta}_G.$

These measures are parameterized by a positive definite matrix U and a vector $\boldsymbol{\alpha} \in \mathbb{R}^m$ with nonnegative entries. Let us first establish some notation:

- $\mathcal{N}(i) := \{j : (i, j) \in E\}$;
- $\mathcal{N}^{\prec}(i) := \{j : (i, j) \in E, i > j\}$;
- $U^{\prec i} := ((U_{kl}))_{k, l \in \mathcal{N}^{\prec}(i)}$;
- $U^{\preceq i} := ((U_{kl}))_{k, l \in \mathcal{N}^{\prec}(i) \cup \{i\}}$;
- $U_{\cdot i}^{\prec} := (U_{ki})_{k \in \mathcal{N}^{\prec}(i)}.$

Let

$$z_G(U, \boldsymbol{\alpha}) := \int e^{-(\text{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})/2} \, dL \, dD.$$

If $z_G(U, \boldsymbol{\alpha}) < \infty$, then $\widetilde{\pi}_{U,\boldsymbol{\alpha}}$ can be normalized to obtain a probability measure. A sufficient condition for the existence of a normalizing constant for $\widetilde{\pi}_{U,\boldsymbol{\alpha}}(L, D)$ is provided in the following proposition.

THEOREM 1. *Let* $dL := \prod_{(i,j) \in E, i > j} dL_{ij}$ *and* $dD := \prod_{i=1}^m dD_{ii}$. *Then,*

$$\int_{\boldsymbol{\Theta}_G} e^{-(\text{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})/2} \, dL \, dD < \infty$$

if

$$\alpha_i > |\mathcal{N}^{\prec}(i)| + 2 \qquad \forall i = 1, 2, \ldots, m.$$

As the proof of this proposition is rather long and technical, it is deferred to the Appendix. The normalizing constant $z_G(U, \boldsymbol{\alpha})$ is not generally available in closed form. Let us consider a simple example to illustrate the difficulty of computing the normalizing constant explicitly.

Let $G = A_4$, that is, the path on four vertices, or $\overset{1}{\bullet} - \overset{2}{\bullet} - \overset{3}{\bullet} - \overset{4}{\bullet}$. Note that this is a decomposable (but not homogeneous) graph. The restrictions on L are $L_{31} = L_{41} = L_{42} = 0$. Let $U \in P_G$ and $\boldsymbol{\alpha} = (4, 4, 4, 4)$. Then, after integrating out the elements $D_{ii}, 1 \le i \le 4$ (recognizing them as inverse-gamma integrals) and transforming the entries of L to the independent entries of L^{-1} (as in the proof of Proposition 1), the normalizing constant reduces to an integral of the form

$$\int_{\mathbb{R}^3} (U_{22} + 2U_{12}x_1 + U_{11}x_1^2)^{-1}$$
$$\times (U_{11}x_1^2 x_2^2 + U_{22}x_2^2 + U_{33} + 2U_{12}x_1 x_2^2 + 2U_{13}x_1 x_2 + 2U_{23}x_2)^{-1}$$
$$\times (U_{11}x_1^2 x_2^2 x_3^2 + U_{22}x_2^2 x_3^2 + U_{33}x_3^2 + U_{44} + 2U_{12}x_1 x_2^2 x_3^2$$
$$+ 2U_{13}x_1 x_2 x_3^2 + 2U_{14}x_1 x_2 x_3 + 2U_{23}x_2 x_3^2 + 2U_{24}x_2 x_3 + U_{34}x_3)^{-1} \, d\mathbf{x}.$$

The above integral does not seem to be computable by standard techniques for general U. Despite this inherent difficulty, we propose a novel method which allows sampling from this rich family of distributions (see Section 4).

We will show later that the condition in Theorem 1 is necessary and sufficient for the existence of a normalizing constant for homogeneous graphs. Moreover, in this case, the normalizing constant can be computed in closed form. We denote by $\pi_{U,\boldsymbol{\alpha}}$ the normalized version of $\widetilde{\pi}_{U,\boldsymbol{\alpha}}$ whenever $z_G(U, \boldsymbol{\alpha}) < \infty$. The following lemma shows that the family $\pi_{U,\boldsymbol{\alpha}}$ is a conjugate family for Gaussian covariance graph models.

LEMMA 1. *Let $G = (V, E)$ be a decomposable graph, where vertices in V are ordered so that $V \in S_D$. Let $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ be an i.i.d. sample from $\mathcal{N}_m(\mathbf{0}, LDL^T)$, where $(L, D) \in \mathbf{\Theta}_G$. Let $S = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i \mathbf{Y}_i^T$ denote the empirical covariance matrix. If the prior distribution on (L, D) is $\pi_{U,\alpha}$, then the posterior distribution of (L, D) is given by $\pi_{\widetilde{U}, \widetilde{\alpha}}$, where $\widetilde{U} = nS + U$ and $\widetilde{\alpha} = (n + \alpha_1, n + \alpha_2, \ldots, n + \alpha_m)$.*

PROOF. The likelihood of the data is given by

$$f(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \mid L, D) = \frac{1}{(\sqrt{2\pi})^{nm}} e^{-(\mathrm{tr}((LDL^T)^{-1}(nS)) + n \log |D|)/2}.$$

Using $\pi_{U,\alpha}$ as a prior for (L, D), the posterior distribution of (L, D) given the data $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$ is

$$\pi_{U,\alpha}(L, D \mid \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$$
$$\propto e^{-(\mathrm{tr}((LDL^T)^{-1}(nS+U)) + \sum_{i=1}^{m}(n+\alpha_i) \log D_{ii})/2}, \qquad \boldsymbol{\theta} \in \mathbf{\Theta}_G.$$

Hence, the posterior distribution belongs to the same family as the prior, that is,

$$\pi_{U,\alpha}(\cdot \mid \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n) = \pi_{\widetilde{U}, \widetilde{\alpha}}(\cdot),$$

where $\widetilde{U} = nS + U$ and $\widetilde{\alpha} = (n + \alpha_1, n + \alpha_2, \ldots, n + \alpha_m)$. $\square$

REMARK. If we assume that the observations have unknown mean μ, that is, $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ are i.i.d. $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} \in \mathbb{R}^m$, $\Sigma \in P_G$, then

$$\widetilde{S} := \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$$

is the minimal sufficient statistic for Σ. Here, $n\widetilde{S}$ has a Wishart distribution with parameter Σ and $n - 1$ degrees of freedom. Hence, if we assume a prior $\pi_{U,\alpha}$ for (L, D), then the posterior distribution is given by

$$\pi_{U,\alpha}(\cdot \mid \widetilde{S}) = \pi_{\widetilde{U}, \widetilde{\alpha}}(\cdot),$$

where $\widetilde{U} = n\widetilde{S} + U$ and $\boldsymbol{\alpha} = (n - 1 + \alpha_1, n - 1 + \alpha_2, \ldots, n - 1 + \alpha_m)$.

REMARK. Note that, as with the distributions in [20, 26], the functional form of the prior distribution depends on the ordering of the vertices specified—but this is not as restrictive as it first appears. In this sense, an ordering is essentially another "parameter" to be specified and thus can also be viewed as imposing extra information. We return to this point in the examples section where we investigate the impact of ordering on a real-world example (see Section 6). But, more importantly, given a perfect ordering of the vertices, any rearrangement of the vertices within the residuals $R_j = C_j \backslash H_{j-1}$ will still preserve the zeros between Σ and L,

and will thus be sufficient for our purposes. In this sense, the covariance Wishart distributions introduced in this paper do not actually depend on a full ordering of the vertices. In fact, for the class of decomposable graphs, any perfect ordering is sufficient, that is, any ordering that is used in [20] will also be relevant for the covariance Wishart distributions defined above. In this sense, these decomposable covariance Wishart distributions are very flexible, especially since we are working in the curved exponential family setting and are still able to use any ordering that is appropriate for the [20] distributions which address the natural exponential family (NEF) concentration graph situation. The technical reason why any perfect ordering will suffice is that any perfect ordering will preserve the zeros between Σ and the matrix L from its Cholesky decomposition [24, 27]. Moreover, from an applications perspective, since matrix operations are not invariant with respect to ordering of the nodes, an ordering that facilitates calculations is desirable. All that the ordering does is to relabel the vertices, but the edge structure is completely and fully retained. To further clarify what is meant, if one has a list of a genes/proteins called ABLIM1, BCL6, etc. and their names are replaced with the numbers 1, 2, 3, etc., the problem can first be analyzed with the integer labels and one can then go back to the original labels after the analysis is done. So, in many applications, the ordering is not a real restriction.

3.1. *Induced prior on P_G and Q_G.* The prior $\pi_{U,\alpha}$ on $\boldsymbol{\Theta}_G$ (the modified Cholesky space) induces a prior on P_G (the covariance matrix space) and Q_G. We provide an expression for the induced priors on these spaces in order to compare our Wishart distributions with other classes of distributions. Note that since the vertices have been ordered so that $V \in S_D$, the transformation

$$\psi : \boldsymbol{\Theta}_G \to M_m^+$$

defined by

$$\psi(L, D) = LDL^T =: \Sigma$$

is a bijection from $\boldsymbol{\Theta}_G$ to P_G. The lemma below provides the required Jacobians for deriving the induced priors on P_G and Q_G. The reader is referred to Section 2.2 for notation on decomposable graphs. Note that if x is a matrix, then $|x|$ denotes its determinant, while if C is a set, then $|C|$ denotes its cardinality.

LEMMA 2 (Jacobians of transformations).

1. *The Jacobian of the transformation $\psi : (L, D) \to \Sigma$ from $\boldsymbol{\Theta}_G$ to P_G is*

$$\prod_{i=1}^{m} D_{jj}(\Sigma)^{-n_j}.$$

Here, $D_{jj}(\Sigma)$ denotes that D_{jj} is a function of Σ, and $n_j := |\{i : (i, j) \in E, i > j\}|$ for $j = 1, 2, \ldots, m$.

2. *The absolute value of the Jacobian of the bijection* $\zeta : x \to \hat{x}^{-1}$ *from* Q_G *to* P_G *is*

$$\prod_{C \in \mathcal{C}} |x_C|^{-|C|-1} \prod_{S \in \mathcal{S}} |x_S|^{(|S|+1)\nu(S)}.$$

PROOF.   The first part is a direct consequence of a result in [28] and the proof of the second part can be found in [27].   □

These Jacobians allow us to compute the induced priors on P_G and Q_G. The induced prior corresponding to $\tilde{\pi}_{U,\alpha}$ on P_G is given by

$$(3.2) \qquad \tilde{\pi}_{U,\boldsymbol{\alpha}}^{P_G}(\Sigma) \propto e^{-(\operatorname{tr}(\Sigma^{-1}U) + \sum_{i=1}^{m}(2n_i+\alpha_i)\log D_{ii}(\Sigma))/2}, \qquad \Sigma \in P_G.$$

We first note that the traditional inverse Wishart distribution (see [23]) with parameters U and n is a special case of (3.2) when G is the complete graph and $\alpha_i = n - 2m + 2i, \forall 1 \le i \le m$. We also note that the $\mathcal{G}$-inverse Wishart priors introduced in [29] have a one-dimensional shape parameter δ and are a very special case of our richer class $\tilde{\pi}_{U,\boldsymbol{\alpha}}^{P_G}$. The single shape parameter δ is given by the relationship $\alpha_i + 2n_i = \delta + 2m, 1 \le i \le m$.[10]

We now proceed to derive the induced prior on Q_G. Let $x = \kappa(\Sigma^{-1})$ denote the image of Σ in Q_G and let $\hat{x}$ denote Σ^{-1} (see [20, 26] for more details). Using the second part of Lemma 2, the induced prior corresponding to $\tilde{\pi}_{U,\boldsymbol{\alpha}}$ on Q_G is given by

$$\tilde{\pi}_{U,\boldsymbol{\alpha}}^{Q_G}(x) \propto e^{-(\operatorname{tr}(\hat{x}U) + \sum_{i=1}^{m}(2n_i+\alpha_i)\log D_{ii}((\hat{x})^{-1}))/2}$$

$$\times \frac{\prod_{S \in \mathcal{S}} |x_S|^{(|S|+1)\nu(S)}}{\prod_{C \in \mathcal{C}} |x_C|^{|C|+1}}, \qquad x \in Q_G.$$

3.2. *Comparison with the Letac–Massam priors.*   We now carefully compare our class of priors to those proposed in Letac and Massam [20]. In [20], the authors construct two classes of distributions, named W_{P_G} and W_{Q_G}, on the spaces P_G and Q_G, respectively, for G decomposable (see [20], Section 3.1). These distributions are generalizations of the Wishart distribution on these convex cones and have been found to be very useful for high-dimensional Bayesian inference, as illustrated in [26]. These priors lead to corresponding classes of inverse Wishart distributions IW_{P_G} (on Q_G) and IW_{Q_G} (on P_G), that is, $U \sim IW_{P_G}$ whenever $\hat{U}^{-1} \sim W_{P_G}$, and

[10]There is an interesting parallel here that becomes apparent from our derivations above. In the concentration graph setting, the single shape parameter hyper-inverse Wishart (HIW) prior of Dawid and Lauritzen [9] is a special case of the multiple shape parameter class of priors introduced by Letac and Massam [20], in the sense that $\alpha_i = -\frac{1}{2}(\delta + c_i - 1)$ (see [26] for notation). In a similar spirit, we discover that the single shape parameter class of priors in [29] is a special case of the multiple shape parameter class of priors $\tilde{\pi}_{U,\boldsymbol{\alpha}}$ introduced in this paper, in the sense that $\alpha_i = \delta - 2n_i + 2m$.

$V \sim IW_{Q_G}$ whenever $\kappa(V^{-1}) \sim W_{Q_G}$. In [20], it is shown that the family of distributions IW_{P_G} yields a family of conjugate priors in the Gaussian concentration graph setting, that is, when $\Sigma \in Q_G$.

As the IW_{Q_G} priors of [20] are defined on the space P_G, in principle, they can potentially serve as priors[11] in the covariance graph setting since the parameter of interest Σ lives in P_G. Let us examine this class more carefully, first with a view to understanding their use in the covariance graph setting and second to compare them to our priors. Following the notation for decomposable graphs in Section 2.2 and in [20], the density of the IW_{Q_G} distribution is given by

$$IW_{Q_G}^{U,\boldsymbol{\alpha},\boldsymbol{\beta}}(\Sigma) \propto e^{\mathrm{tr}(\Sigma^{-1}U)/2} \frac{\prod_{C \in \mathcal{C}} |(\Sigma^{-1})_C|^{\alpha(C)+(c+1)/2}}{\prod_{S \in \mathcal{S}} |(\Sigma^{-1})_S|^{\nu(S)(\beta(S)+(s+1)/2)}}, \qquad \Sigma \in P_G,$$

where $U \in P_G$, and $\alpha(C)$, $C \in \mathcal{C}$ and $\beta(S)$, $S \in \mathcal{S}$ are real numbers. The posterior density of Σ under this prior is given by

$$\pi_{U,\boldsymbol{\alpha},\boldsymbol{\beta}}^{\mathrm{IW}}(\Sigma \mid \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n) \propto e^{-\mathrm{tr}(\Sigma^{-1}(U+nS))/2}$$

$$\times \frac{\prod_{C \in \mathcal{C}} |(\Sigma^{-1})_C|^{\alpha(C)+(c+1)/2+n/2}}{\prod_{S \in \mathcal{S}} |(\Sigma^{-1})_S|^{\nu(S)(\beta(S)+(s+1)/2)+n\nu(S)/2}}.$$

However, $U + nS$ may not, in general, be in P_G, which is a crucial assumption in the analysis in [20]. Hence, the conjugacy breaks down.

We now investigate similarities and differences between our class of priors and the IW_{Q_G} class. Since the $IW_{Q_G}^{U,\boldsymbol{\alpha},\boldsymbol{\beta}}$ density is defined only for $U \in P_G$, a pertinent question is whether our class of priors has the same functional form when $U \in P_G$. We discover that this is not the case and demonstrate this through an example. Consider the 4-chain A_4. One can easily verify that the terms $e^{-\mathrm{tr}(\Sigma^{-1}U)/2}$ are identical in both priors. We now show that the remaining terms are not identical. If $\Sigma = LDL^T$ is the modified Cholesky decomposition of Σ, then, for this particular graph with $C_1 = \{1, 2\}$, $C_2 = \{2, 3\}$, $C_3 = \{3, 4\}$ and $S_2 = \{3\}$, $S_3 = \{4\}$, the expression that is not in the exponential term for the IW_{Q_G} density is of the form

$$\frac{\prod_{i=1}^{3} |(\Sigma^{-1})_{C_i}|^{\alpha_i}}{\prod_{i=2}^{3} |(\Sigma^{-1})_{S_i}|^{\beta_i}} = \left(\frac{1}{D_{11}}\right)^{\alpha_1} \left(\frac{1}{D_{22}} + \frac{L_{32}^2}{D_{33}} + \frac{L_{32}^2 L_{43}^2}{D_{44}}\right)^{\alpha_1 - \beta_1}$$

$$\times \left(\frac{1}{D_{22}}\right)^{\alpha_2} \left(\frac{1}{D_{33}} + \frac{L_{43}^2}{D_{44}}\right)^{\alpha_2 - \beta_2} \left(\frac{1}{D_{33}D_{44}}\right)^{\alpha_3}.$$

This expression is clearly different from the term, other than the exponent $e^{-\mathrm{tr}(\Sigma^{-1}U)/2}$ in $\pi_{U,\boldsymbol{\alpha}}^{P_G}$, which is a product of different powers of D_{ii}, $i = 1, 2, 3, 4$.

---

[11]The use of this class of nonconjugate priors for Bayesian inference in covariance graph models was already explored in [22].

However, an interesting property emerges when G is homogeneous. Note that, in this case, for any clique C and any separator S,

$$|(\Sigma^{-1})_C| = \prod_{i \in C} \frac{1}{D_{ii}}, |(\Sigma^{-1})_S| = \prod_{i \in S} \frac{1}{D_{ii}}.$$

Hence, when G is homogeneous, the class IW_{Q_G} is contained in the class π^{P_G}. The containment is strict because U need not be in P_G for our class π^{P_G}. Also, in IW_{Q_G}, the exponent of D_{ii} and D_{jj} is the same if $i\,R\,j$, that is, the shape parameter, is shared for vertices in the same equivalence class, as defined by the relation R. We, however, note that the difference in the number of shape parameters is not a major difference, due to the result of Consonni and Veronese [5], together with fact that for the W_{Q_G} (and, correspondingly, for the IW_{Q_G}), each one of the blocks $x_{[i]}$ has a Wishart distribution (see Theorem 4.5 of [20]).

We therefore note that in the restrictive case when G is homogeneous and when $U \in P_G$, the two classes of distributions π^{P_G} and IW_{Q_G} have the same functional form. The fact that we do not restrict $U \in P_G$ is an important difference since, even in the homogeneous case, they yield a larger class of distributions on the homogeneous cone P_G compared to those in Andersson and Wojnar [2], resulting in nonsuperficial consequences for inference in covariance graph models.[12]

4. Sampling from the posterior distribution. In this section, we study the properties of our family of distributions and thereby provide a method that allows us to generate samples from the posterior distribution corresponding to the priors defined in Section 3. In particular, we prove that $\theta = (L, D) \in \Theta_G$ can be partitioned into blocks so that the conditional distribution of each block given the others is a standard distribution in statistics and hence easy to sample from. We can therefore generate samples from the posterior distribution by using the block Gibbs sampling algorithm.

4.1. *Distributional properties and the block Gibbs sampler.* Let us introduce some notation before deriving the required conditional distributions. Let $G = (V, E)$ be a decomposable graph such that $V \in S_D$. For a lower-triangular matrix L with diagonal entries equal to 1,

$$L_{u\cdot} := u\text{th row of } L, \qquad u = 1, 2, \ldots, m,$$
$$L_{\cdot v} := v\text{th column of } L, \qquad v = 1, 2, \ldots, m,$$
$$L_{\cdot v}^G := (L_{uv})_{u>v,(u,v)\in E}, \qquad v = 1, 2, \ldots, m-1.$$

---

[12]We note, however, that the distributions in [2] are quite general since the authors consider other homogeneous cones and not just P_G.

So, $L^G_{.v}$ is the vth column of L without the components which are specified to be zero under the model $\mathcal{G}$ (and without the vth diagonal entry, which is 1). In terms of this notation, the parameter space can be represented as

$$\boldsymbol{\Theta}_G = \{(L^G_{.1}, L^G_{.2}, L^G_{.3}, \ldots, L^G_{.m-1}, D):$$

(4.1)

$$L_{ij} \in \mathbb{R}, \forall 1 \le j < i \le m, (i, j) \in E, D_{ii} > 0, \forall 1 \le i \le m\}.$$

Suppose that $\boldsymbol{\theta} \sim \pi_{U,\alpha}$ for some positive definite U and $\boldsymbol{\alpha} \in \mathbb{R}^m$ with nonnegative entries. The posterior distribution is then $\pi_{\widetilde{U},\widetilde{\alpha}}$, where $\widetilde{U} = nS + U, \widetilde{\boldsymbol{\alpha}} = (n + \alpha_1, n + \alpha_2, \ldots, n + \alpha_m)$. In the following proposition, we derive the distributional properties which provide the essential ingredients for our block Gibbs sampling procedure.

THEOREM 2.   *Using the notation above, the conditional distributions of each component of $\boldsymbol{\theta}$ [as in (4.1)] given the other components and the data* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ *are as follows*:

1.

$$L^G_{.v} \mid (L \backslash L^G_{.v}, D, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n) \sim \mathcal{N}(\boldsymbol{\mu}^{v,G}, M^{v,G}) \qquad \forall v = 1, 2, \ldots, m-1,$$

where

$$\mu^{v,G}_u := \mu^v_u + \sum_{u' > v:(u',v) \in E} \sum_{\substack{w > v:(w,v) \notin E \\ or\ w < v, L^{-1}_{vw} = 0}} M^{v,G}_{uu'} (L^{-1} \widetilde{U} (L^T)^{-1})_{vv}$$

$$\times (LDL^T)^{-1}_{u'w} \mu^v_w$$

$$\forall u > v, (u, v) \in E,$$

$$\mu^v_u := \frac{(L^{-1}\widetilde{U})_{vu}}{(L^{-1}\widetilde{U}(L^T)^{-1})_{vv}} \qquad \forall u \text{ such that } L^{-1}_{vu} = 0,$$

$$(M^{v,G})^{-1}_{uu'} := (L^{-1}\widetilde{U}(L^T)^{-1})_{vv}(LDL^T)^{-1}_{uu'} \qquad \forall u, u' > v, (u, v), (u', v) \in E;$$

2.

$$D_{ii} \mid L, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m \sim \text{IG}\left(\frac{\widetilde{\alpha}_i}{2} - 1, \frac{(L^{-1}\widetilde{U}(L^T)^{-1})_{ii}}{2}\right),$$

independently for $i = 1, 2, \ldots, m$, where "IG" represents the inverse-gamma distribution.

REMARK.   The notation $w : L^{-1}_{vw} = 0$ in the definition of $\boldsymbol{\mu}^{v,G}$ above means indices w for which L^{-1}_{vw} is 0 as a function of entries of L.

Deriving the required conditional distributions in Theorem 2 entails careful analysis. We first state two lemmas which are essential for deriving these distributions.

LEMMA 3.   *Let $u > v$, $(u, v) \in E$. Then,*

$$\frac{\partial L_{ij}^{-1}}{\partial L_{uv}} = -L_{iu}^{-1}L_{vj}^{-1} \qquad \forall 1 \le j < i \le m.$$

PROOF.   The proof is straightforward and is therefore omitted for brevity.   □

Recall from Proposition 1 that L_{ij}^{-1} functionally depends on L_{uv} only if $i \ge u > v \ge j$. We use this observation repeatedly in our arguments. For a given v, to prove conditional multivariate normality of the conditional distribution of $L_{\cdot v}^G$ given the others, we shall demonstrate that if we treat D and the other columns of L as constants, then $\text{tr}((LDL^T)^{-1}\widetilde{U})$ is a quadratic form in the entries of $L_{\cdot v}^G$.

LEMMA 4.   *Let $u, u' > v$, $(u, v), (u', v) \in E$. Then,*

$$\frac{\partial^2}{\partial L_{uv}\,\partial L_{u'v}} \text{tr}((LDL^T)^{-1}\widetilde{U}) = 2(L^{-1}\widetilde{U}(L^T)^{-1})_{vv}(LDL^T)_{uu'}^{-1},$$

which is functionally independent of the elements of $L_{\cdot v}^G$.

PROOF.   First, note that,

$$\frac{\partial}{\partial L_{uv}} \text{tr}((LDL^T)^{-1}\widetilde{U})$$

$$= \frac{\partial}{\partial L_{uv}} \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \frac{L_{ki}^{-1}L_{kj}^{-1}}{D_{kk}} \widetilde{U}_{ij} \right)$$

$$= -\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \left(\frac{L_{ku}^{-1}L_{vi}^{-1}L_{kj}^{-1} + L_{ki}^{-1}L_{ku}^{-1}L_{vj}^{-1}}{D_{kk}} \right)\widetilde{U}_{ij} \qquad \text{(by Lemma 3)}$$

$$= -2\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \frac{L_{ku}^{-1}L_{vi}^{-1}L_{kj}^{-1}}{D_{kk}} \widetilde{U}_{ij}.$$

Note that L^{-1} is a lower-triangular matrix. Hence,

$$\frac{\partial^2}{\partial L_{uv}\,\partial L_{u'v}} \text{tr}((LDL^T)^{-1}\widetilde{U})$$

$$= -2\frac{\partial}{\partial L_{u'v}} \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \frac{L_{ku}^{-1}L_{vi}^{-1}L_{kj}^{-1}}{D_{kk}} \widetilde{U}_{ij} \right)$$

$$= 2\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \frac{L_{ku}^{-1}L_{vi}^{-1}L_{ku'}^{-1}L_{vj}^{-1}}{D_{kk}} \widetilde{U}_{ij}$$

$$= 2\left(\sum_{i=1}^{m}\sum_{j=1}^{m}L_{vi}^{-1}\widetilde{U}_{ij}L_{vj}^{-1}\right)\left(\sum_{k=1}^{m}\frac{L_{ku}^{-1}L_{ku'}^{-1}}{D_{kk}}\right)$$

$$= 2(L^{-1}\widetilde{U}(L^{T})^{-1})_{vv}(LDL^{T})_{uu'}^{-1}.$$

The second equality above follows by noting that by Proposition 1, L_{vi}^{-1} is functionally independent of $L_{\cdot v}^{G}$ for all $1 \leq i \leq m$ and L_{ku}^{-1} is functionally independent of $L_{\cdot v}^{G}$ for all $1 \leq k \leq m$ and $u > v$, and then applying Lemma 3. Using this functional independence argument above once more, we thereby conclude that $2(L^{-1}\widetilde{U}(L^{T})^{-1})_{vv}(LDL^{T})_{uu'}^{-1}$ is independent of $L_{\cdot v}^{G}$. $\square$

PROOF OF THEOREM 2. An immediate consequence of Lemma 4 and the preceding remark is that we can write $\text{tr}((LDL^{T})^{-1}\widetilde{U})$ as follows:

$$\text{tr}((LDL^{T})^{-1}\widetilde{U})$$

$$= \sum_{u>v,(u,v)\in E}\sum_{u'>v,(u',v)\in E}((L^{-1}\widetilde{U}(L^{T})^{-1})_{vv}(LDL^{T})_{uu'}^{-1})$$

$$\times (L_{uv} - b_{u})(L_{u'v} - b_{u'}) + C,$$

where $\mathbf{b} = (b_{u})_{u>v,(u,v)\in E}$ and C are independent of $L_{\cdot v}^{G}$. In order to evaluate $(b_{u})_{u>v,(u,v)\in E}$, note that the term in $\frac{\partial}{\partial L_{uv}}\text{tr}((LDL^{T})^{-1}\widetilde{U})$ which is independent of $L_{\cdot v}^{G}$ is given by

$$(4.2) \qquad -2\sum_{u'>v,(u',v)\in E}((L^{-1}\widetilde{U}(L^{T})^{-1})_{vv}(LDL^{T})_{uu'}^{-1})b_{u'}$$

for every $u > v$, $(u, v) \in E$. However, from the proof of Lemma 4, we alternatively know that

$$\frac{\partial}{\partial L_{uv}}\text{tr}((LDL^{T})^{-1}\widetilde{U}) = -2\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\frac{L_{ku}^{-1}L_{vi}^{-1}L_{kj}^{-1}}{D_{kk}}\widetilde{U}_{ij}.$$

Note that by Lemma 3, $L_{ku}^{-1}L_{kj}^{-1}$ is functionally dependent on $L_{\cdot v}^{G}$ if and only if $L_{ku}^{-1} \neq 0$ and $L_{vj}^{-1} \neq 0$ (as a function of L). Hence, the term in $\frac{\partial}{\partial L_{uv}}\text{tr}((LDL^{T})^{-1}\widetilde{U})$ which is independent of $L_{\cdot v}^{G}$ is given by

$$-2\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\frac{L_{ku}^{-1}L_{vi}^{-1}L_{kj}^{-1}}{D_{kk}}\widetilde{U}_{ij}\mathbf{1}_{\{L_{vj}^{-1}=0 \text{ or } L_{ku}^{-1}=0\}}$$

$$(4.3) \qquad = -2\sum_{j:L_{vj}^{-1}=0}\left(\sum_{i=1}^{m}L_{vi}^{-1}\widetilde{U}_{ij}\right)\left(\sum_{k=1}^{m}\frac{L_{ku}^{-1}L_{kj}^{-1}}{D_{kk}}\right)$$

$$= -2 \sum_{j:L_{vj}^{-1}=0} (L^{-1}\widetilde{U})_{vj}(LDL^T)_{uj}^{-1}$$

$$= -2 \sum_{j:L_{vj}^{-1}=0} \frac{(L^{-1}\widetilde{U})_{vj}}{(L^{-1}\widetilde{U}(L^T)^{-1})_{vv}}(L^{-1}\widetilde{U}(L^T)^{-1})_{vv}(LDL^T)_{uj}^{-1}.$$

Now, observe the following facts:

1. the expressions in (4.2) and (4.3) should be the same for every $u > v$, $(u, v) \in E$;

2. if $A = \begin{pmatrix} A_1 & A_2 \\ A_2^T & A_3 \end{pmatrix}$, $\boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$ and $\boldsymbol{\eta}$ are such that

$$A_1\boldsymbol{\xi}_1 + A_2\boldsymbol{\xi}_2 = A_1\boldsymbol{\eta},$$

then

$$\boldsymbol{\eta} = \boldsymbol{\xi}_1 + A_1^{-1}A_2\boldsymbol{\xi}_2.$$

If we choose A, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ as

$$A_{uu'} := (L^{-1}\widetilde{U}(L^T)^{-1})_{vv}(LDL^T)_{uu'}^{-1} \qquad \forall u, u' \text{ such that } L_{vu}^{-1}, L_{vu'}^{-1} = 0,$$

$$\xi_u := \frac{(L^{-1}\widetilde{U})_{vu}}{(L^{-1}\widetilde{U}(L^T)^{-1})_{vv}} \qquad \forall u \text{ such that } L_{vu}^{-1} = 0,$$

$$\eta_u := b_u \qquad \forall u > v, (u, v) \in E,$$

then combining the observations above with (4.2) and (4.3), we obtain

$$\mathrm{tr}((LDL^T)^{-1}\widetilde{U})$$

$$= \sum_{u>v,(u,v)\in E} \sum_{u'>v,(u',v)\in E} ((L^{-1}\widetilde{U}(L^T)^{-1})_{vv}(LDL^T)_{uu'}^{-1})$$

$$\times (L_{uv} - \mu_u^{v,G})(L_{u'v} - \mu_{u'}^{v,G}) + C.$$

As defined earlier,

$$\mu_u^v = \frac{(L^{-1}\widetilde{U})_{vu}}{(L^{-1}\widetilde{U}(L^T)^{-1})_{vv}} \qquad \forall u \text{ such that } L_{vu}^{-1} = 0,$$

$$\mu_u^{v,G} = \mu_u^v + \sum_{u'>v,(u',v)\in E} \sum_{\substack{w>v,(w,v)\notin E \\ \text{or } w<v, L_{vw}^{-1}=0}} M_{uu'}^{v,G}(L^{-1}\widetilde{U}(L^T)^{-1})_{vv}(LDL^T)_{u'w}^{-1}\mu_w^v$$

$$\forall u > v, (u, v) \in E$$

and C is independent of $L_{\cdot v}^G$. It follows that under $\pi_{\widetilde{U},\widetilde{\alpha}}$, the conditional distribution of $L_{\cdot v}^G$ given the other parameters and the data $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ is $\mathcal{N}(\boldsymbol{\mu}^{v,G}, M^{v,G})$.

For deriving the conditional distribution of the entries of D, we note that

$$e^{-(\mathrm{tr}((LDL^T)^{-1}\tilde{U})+\sum_{i=1}^m \tilde{\alpha}_i \log D_{ii})/2} = \prod_{j=1}^m \frac{1}{D_{jj}^{\tilde{\alpha}_j/2}} e^{-(L^{-1}\tilde{U}(L^T)^{-1})_{jj}/(2D_{jj})}.$$

The above leads us to conclude that the conditional distribution of D_{jj} given the other parameters and the data $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ is $\mathrm{IG}(\frac{\tilde{\alpha}_j}{2} - 1, \frac{(L^{-1}\tilde{U}(L^T)^{-1})_{jj}}{2})$, independently for every $j = 1, 2, \ldots, m$. $\quad\square$

4.2. *Convergence of block Gibbs sampler.* The block Gibbs sampling procedure, based on the conditional distributions derived above, gives rise to a Markov chain. It is natural to ask whether this Markov chain converges to the desired distribution $\pi_{\tilde{U}, \tilde{\alpha}}$. Convergence properties are sometimes overlooked due to the theoretical demands in establishing them. However, they yield theoretical safeguards that the block Gibbs sampling algorithm can be used for sampling from the posterior distribution.

We now prove that sufficient conditions for convergence of a Gibbs sampling Markov chain to its stationary distribution (see [2], Theorem 6) are satisfied by the Markov chain corresponding to our block Gibbs sampler. Let $\phi(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density evaluated at $\mathbf{x}$. Let $f_{\mathrm{IG}}(d \mid \alpha, \lambda)$ denote the $\mathrm{IG}(\alpha, \lambda)$ density evaluated at d. Let us fix $\psi, d_1, d_2 > 0$ arbitrarily. Let

$$\boldsymbol{\Theta}_{\psi, d_1, d_2} := \{\boldsymbol{\theta} = (L, D) \in \boldsymbol{\Theta}_G : |L_{ij}| \leq \psi, d_1 \leq D_{ii} \leq d_2 \,\forall i > j, (i, j) \in E\}.$$

We now formally prove the conditions which are sufficient for establishing convergence.

PROPOSITION 3. *There exists some $\delta > 0$ such that, uniformly for all $\boldsymbol{\theta} = (L, D) \in \boldsymbol{\Theta}_{\psi, d_1, d_2}$,*

$$\phi(L_{\cdot,v}^G \mid \boldsymbol{\mu}^{v,G}, M^{v,G}) > \delta \qquad \forall v = 1, 2, \ldots, m - 1,$$

$$f_{\mathrm{IG}}\left(D_{ii} \mid \frac{\tilde{\alpha}_i}{2} - 1, \frac{(L^{-1}\tilde{U}(L^T)^{-1})_{ii}}{2}\right) > \delta \qquad \forall i = 1, 2, \ldots, m.$$

PROOF. First, by Proposition 1, all entries of L^{-1} are polynomials in the entries of L. Since $\boldsymbol{\Theta}_{\psi, d_1, d_2}$ is bounded and closed, there exists $\psi_1 > 0$ such that

$$(L, D) \in \boldsymbol{\Theta}_{\psi, d_1, d_2} \Rightarrow |L_{uv}^{-1}| \leq \psi_1 \qquad \forall u > v, (u, v) \in E.$$

Using the above, there exists a constant $\psi_2 > 0$ such that if $(L, D) \in \boldsymbol{\Theta}_{\psi, d_1, d_2}$, then

$$(4.4) \quad |(L^{-1}\tilde{U})_{vu}| \leq \psi_2, \qquad |(LDL^T)_{uu'}^{-1}| \leq \psi_2, \qquad |(L^{-1}\tilde{U}(L^T)^{-1})_{vv}| \leq \psi_2$$

for every $1 \leq v, u, u' \leq m$. Second, since $L_{vv}^{-1} = 1$ for all $1 \leq v \leq m$ and $\widetilde{U}$ is positive definite, it follows that there exists a constant $\psi_3 > 0$ such that if $(L, D) \in \Theta_{\psi, d_1, d_2}$, then

$$(4.5) \qquad \qquad \psi_3 \leq (L^{-1}\widetilde{U}(L^T)^{-1})_{vv}$$

for every $1 \leq v \leq m$.

Let $(L, D) \in \Theta_{\psi, d_1, d_2}$. Note that

$$(L_{\cdot v}^G - \boldsymbol{\mu}^{v,G})^T (M^{v,G})^{-1} (L_{\cdot v}^G - \boldsymbol{\mu}^{v,G})$$
$$= (L_{\cdot v}^G)^T (M^{v,G})^{-1} L_{\cdot v}^G - 2(L_{\cdot v}^G)^T (M^{v,G})^{-1} \boldsymbol{\mu}^{v,G} + (\boldsymbol{\mu}^{v,G})^T (M^{v,G})^{-1} \boldsymbol{\mu}^{v,G}.$$

Observe that if $\boldsymbol{\zeta} = \left(\begin{smallmatrix} \zeta_1 \\ \zeta_2 \end{smallmatrix}\right) \in \mathbb{R}^m$ and $\Sigma = \left(\begin{smallmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{smallmatrix}\right)$ is a positive definite matrix, then

$$(\boldsymbol{\zeta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{\zeta}_2)^T \Sigma_{11} (\boldsymbol{\zeta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{\zeta}_2) \leq \zeta^T \Sigma \zeta \qquad \forall \zeta \in \mathbb{R}^m.$$

If we choose $\boldsymbol{\zeta}$ and Σ as

$$\Sigma_{uu'} := (L^{-1}\widetilde{U}(L^T)^{-1})_{vv}(LDL^T)_{uu'}^{-1} \qquad \forall u, u' \text{ such that } L_{vu}^{-1}, L_{vu'}^{-1} = 0,$$

$$\zeta_u = \mu_u^v \qquad \forall u \text{ such that } L_{vu}^{-1} = 0,$$

then combining the observation above and the definition of $\boldsymbol{\mu}^{v,G}$, we get that

$$(\boldsymbol{\mu}^{v,G})^T (M^{v,G})^{-1} \boldsymbol{\mu}^{v,G} \leq \sum_{u: L_{vu}^{-1}=0} \sum_{u': L_{vu'}^{-1}=0} \mu_u^v (L^{-1}\widetilde{U}(L^T)^{-1})_{vv}(LDL^T)_{uu'}^{-1}\mu_{u'}^v.$$

From the definitions in Theorem 2, we also have

$$((M^{v,G})^{-1}\boldsymbol{\mu}^{v,G})_u = \sum_{j: L_{vj}^{-1}=0} (L^{-1}\widetilde{U})_{vj}(LDL^T)_{uj}^{-1} \qquad \forall u > v, (u, v) \in E.$$

It follows by (4.4) that for $(L, D) \in \Theta_{\psi, d_1, d_2}$, there exists $\psi_4 > 0$ such that

$$(4.6) \qquad (L_{\cdot v}^G - \boldsymbol{\mu}^{v,G})^T (M^{v,G})^{-1} (L_{\cdot v}^G - \boldsymbol{\mu}^{v,G}) \leq \psi_4$$

for every $v = 1, 2, \ldots, m - 1$. Also, by the definition of $M^{v,G}$, it follows that for $(L, D) \in \Theta_{\psi, d_1, d_2}, 0 < |M^{v,G}| < \infty$ and $|M^{v,G}|$ is a continuous function of (L, D). Recall that for a matrix $A, |A|$ denotes the determinant of A. Since Θ_{ψ, d_1, d_2} is a bounded and closed set, both the maximum and minimum of the function $|M^{v,G}|$ are attained in Θ_{ψ, d_1, d_2}. It follows that for $(L, D) \in \Theta_{\psi, d_1, d_2}$, there exist $0 < \kappa_1 < \kappa_2$ such that

$$(4.7) \qquad \qquad \kappa_1 < |M^{v,G}| < \kappa_2$$

for every $v = 1, 2, \ldots, m - 1$. It follows by (4.4), (4.5), (4.6) and (4.7) that for $(L, D) \in \Theta_{\psi, d_1, d_2}$, there exists $\delta_1 > 0$ such that

$$\phi(L_{\cdot v}^G \mid \boldsymbol{\mu}^{v,G}, M^{v,G}) > \delta_1 \qquad \forall v = 1, 2, \ldots, m - 1.$$

Note, furthermore, that if $(L, D) \in \mathbf{\Theta}_{\psi, d_1, d_2}$, then, from (4.4) and (4.5),

$$\psi_3 \le (L^{-1}\widetilde{U}(L^T)^{-1})_{ii} \le \psi_2$$

for every $1 \le i \le m$. Hence, there exists $\delta_2 > 0$ such that

$$f_{\text{IG}}\left(D_{ii}\Big|\frac{\widetilde{\alpha}_i}{2} - 1, \frac{(L^{-1}\widetilde{U}(L^T)^{-1})_{ii}}{2}\right) > \delta_2 \qquad \forall i = 1, 2, \ldots, m.$$

Let $\delta = \min(\delta_1, \delta_2)$. Hence, for $\boldsymbol{\theta} = (L, D) \in \mathbf{\Theta}_{\psi, d_1, d_2}$,

$$\phi(L^G_{\cdot v} \mid \boldsymbol{\mu}^{v,G}, M^{v,G}) > \delta \qquad \forall v = 1, 2, \ldots, m-1,$$

$$f_{\text{IG}}\left(D_{ii}\Big|\frac{\widetilde{\alpha}_i}{2} - 1, \frac{(L^{-1}\widetilde{U}(L^T)^{-1})_{ii}}{2}\right) > \delta \qquad \forall i = 1, 2, \ldots, m. \qquad \square$$

Recall that $n_v = |\{u : u > v, (u, v) \in E\}|$. Note that the measures corresponding to $\mathcal{N}(\boldsymbol{\mu}^{v,G}, M^{v,G})$ and $\mathcal{N}(\mathbf{0}, I_{n_v})$ are mutually absolutely continuous and the corresponding densities with respect to Lebesgue measure are positive everywhere on $\mathbb{R}^{n_v}$ for all $v = 1, 2, \ldots, m-1$. In addition, the measures corresponding to $\text{IG}(\frac{\widetilde{\alpha}_i}{2} - 1, \frac{(L^{-1}\widetilde{U}(L^T)^{-1})_{ii}}{2})$ and $\text{IG}(\frac{\widetilde{\alpha}_i}{2} - 1, 1)$ are mutually absolutely continuous and the corresponding densities with respect to Lebesgue measure are positive everywhere on $(0, \infty)$ for all $i = 1, 2, \ldots, m$. Also, since $\mathbf{\Theta}_{\psi, d_1, d_2}$ is bounded and closed, $\prod_{v=1}^{m-1} \phi(L^G_{\cdot v} \mid \mathbf{0}, I_{n_v}) \prod_{i=1}^{m} f_{\text{IG}}(\frac{\widetilde{\alpha}_i}{2} - 1, 1)$ is bounded on $\mathbf{\Theta}_{\psi, d_1, d_2}$. Combining this with Proposition 3, all required conditions in [2], Theorem 6 are satisfied. Hence, the block Gibbs sampling Markov chain, based on the derived conditional distributions, converges to the desired stationary distribution $\pi_{U, \boldsymbol{\alpha}}$.

We note that in [29], page 18, the authors introduce a procedure to sample from the $\mathcal{G}$-inverse Wishart distributions (these are a narrow subclass of our priors $\pi^{P_G}_{U, \boldsymbol{\alpha}}$). Essentially, at every iteration, they cycle through all of the rows of Σ. At the ith step in an iteration, they sample the vector $\Sigma^G_{i \cdot} := (\Sigma_{ij})_{j \in \mathcal{N}(i)}$ from its conditional distribution (Gaussian) given the other entries of Σ and then sample $\gamma_i := \frac{1}{\Sigma^{-1}_{ii}}$ from its conditional distribution (inverse-gamma) given the other entries of Σ. Since Σ is a symmetric matrix, for $(i, j) \in E$, the variable Σ_{ij} appears in $\Sigma^G_{i \cdot}$ as well as $\Sigma^G_{j \cdot}$. Hence, $((\Sigma^G_{1 \cdot}, \gamma_1), (\Sigma^G_{2 \cdot}, \gamma_2), \ldots, (\Sigma^G_{m \cdot}, \gamma_m))$ is not a disjoint partition of the variable space. Therefore, their procedure is not strictly a Gibbs sampling procedure and its convergence properties are not clear. On the other hand, in our procedure, we cycle through $(L^G_{\cdot 1}, L^G_{\cdot 2}, \ldots, L^G_{\cdot m}, D)$, which is a disjoint partition of the variable space. Hence, our procedure is a Gibbs sampler in the true sense. There are also other differences between the two procedures, such as the fact that $\gamma_i \ne D_{ii}$ unless $i = m$.

REMARK.   It is useful to compare our covariance priors to the conditionally conjugate priors introduced by [8] in the complete case. Upon closer investigation we discover that the priors of [8] are quite different from $\widetilde{\pi}_{U, \boldsymbol{\alpha}}(L, D)$. First,

they do not consider structural zeros. More importantly, under their posterior, the distribution of L^{-1} conditional on D is jointly multivariate normal. In the general decomposable covariance graph setting however, the zeros of L do not carry over to L^{-1}, and so it is not possible in our framework for the distribution of the (constrained or unconstrained) elements of L^{-1}, conditional on D, to be jointly multivariate normal.

5. The special case of homogeneous graphs: Closed form expressions. Note that the covariance graph model, that is, the family of distributions

$$\mathcal{G} = \{\mathcal{N}_m(\mathbf{0}, \Sigma) : \Sigma \in P_G\}$$

(supported on $\mathbb{R}^m$) is a curved exponential family for any connected noncomplete graph G. As discussed earlier, the fact that the family is curved renders the Diaconis–Ylvisaker framework no longer applicable in this setting. Hence, a rich and flexible class of distributions was introduced in order to serve as priors for the class of covariance graph models. A natural question to ask is whether the class of priors itself belongs to a curved exponential family.[13] Indeed, this class of priors is interesting in its own right and warrants an independent investigation. Such analysis has the potential to place the class of priors in a known framework and thus exploit this property.

Let us therefore now turn our attention to the class of priors $\{\widetilde{\pi}_{U,\boldsymbol{\alpha}}^{P_G}\}_{U \in M_m^+}$ as a family of distributions supported on P_G, with U as a parameter. We now state a lemma which formally establishes that the class of priors can be framed in the context of natural exponential families.

LEMMA 5.   *For arbitrarily fixed $\boldsymbol{\alpha}$, the family of distributions $\{\widetilde{\pi}_{U,\boldsymbol{\alpha}}^{P_G}\}_{U \in M_m^+}$ is a general exponential family, that is, it can be transformed into a natural exponential family. The natural parameter is $U = ((U_{ij}))_{1 \le i \le j \le m}$, the corresponding set of sufficient statistics is $\Sigma^{-1} = ((\Sigma_{ij}^{-1}))_{1 \le i \le j \le m}$ and the cumulant generating function is $\log z_G(U, \boldsymbol{\alpha})$.*

PROOF.   The proof is straightforward and is therefore omitted.   □

Placing the class of covariance priors in a natural exponential family framework yields insights into the structure and functional form of this class of distributions. As noted earlier, $z_G(U, \boldsymbol{\alpha})$ is not generally available in closed form. A question that naturally arises is whether there are any conditions under which $z_G(U, \boldsymbol{\alpha})$ can be evaluated in closed form. In this section, we establish that when G is homogeneous, $z_G(U, \boldsymbol{\alpha})$ and $\mathbf{E}_{U,\boldsymbol{\alpha}}$ can be evaluated in closed form. It is known that when

---

[13]Note: not the class of distributions associated with the covariance graph probability model but rather the class of priors that is introduced in this paper.

G is a homogeneous graph, the covariance graph model is Markov equivalent to an appropriate DAG (see [11]). It is, however, important to clarify that the Markov equivalence of covariance graph models and DAGs does not immediately imply that Bayesian inference for covariance graph models using our priors automatically follows. We also need to establish a correspondence between our priors and known priors for DAG models. We now prove that in the special case when G is homogeneous, our priors correspond to the standard conjugate priors for an appropriate DAG. This yields yet another property of our class of priors. The following theorem is the main result of this section and helps us establish the aforementioned correspondence.

THEOREM 3.   *Let $G = (V, E)$ be homogeneous, with vertices ordered according to the Hasse perfect vertex elimination scheme specified in Section* 2.4, *that is, $V \in S_H$. If $\Sigma \sim \pi_{U,\alpha}^{P_G}$ and $\Sigma = LDL^T$ is its modified Cholesky decomposition, then*

$$\{(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec})\}_{1 \leq i \leq m}$$

are mutually independent. Furthermore, the distributions of these quantities are specified as follows:

$$(\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec} \mid D_{ii} \sim \mathcal{N}((U^{\prec i})^{-1}U_{\cdot i}^{\prec}, D_{ii}(U^{\prec i})^{-1});$$

$$D_{ii} \sim \mathrm{IG}\left(\frac{\alpha_i}{2} - \frac{|\mathcal{N}^{\prec}(i)|}{2} - 1, \frac{U_{ii} - (U_{\cdot i}^{\prec})^T(U^{\prec i})^{-1}U_{\cdot i}^{\prec}}{2}\right)$$

$$\forall i = 1, 2, \ldots, m.$$

REMARK.   The above result decomposes Σ into mutually independent coordinates. Note that for any i such that $\bar{i}$ is a leaf of the Hasse tree and i has the minimal label in its equivalence class $\bar{i}$, we have

$$\mathcal{N}^{\prec}(i) = \phi.$$

In this case, it is understood that $\Sigma^{\prec i}$ and $\Sigma_{\cdot i}^{\prec}$ are vacuous parameters and that $D_{ii} = \Sigma_{ii}$.

PROOF OF THEOREM 3.   Let G be a homogeneous graph with m vertices, with the vertices ordered according to the Hasse perfect elimination scheme specified in Section 2.4. Recall that the vertices of the Hasse diagram of G are equivalence classes formed by the relation R defined in Section 2.4. The vertex labeled m clearly lies in the equivalence class of vertices at the root of the corresponding Hasse diagram. Let us remove the vertex labeled m from the graph G and let G' denote the induced graph on the remaining $m - 1$ vertices. The graph G' can be of the following two types.

- Case I: If the equivalence class of m contains more than one element, then G' is a homogeneous graph with the Hasse diagram having the same depth as the

Hasse diagram of G, but with one less vertex in the equivalence class at the root. Recall that the depth of a tree is the length of the longest path from its root to any leaf.

- Case II: If the equivalence class of m contains only one element, then G' is a disconnected graph, with the connected components being homogeneous graphs with the Hasse diagram having depth one less than the depth of the Hasse diagram of G.

Note that for every $1 \le i \le m$ such that $\mathcal{N}^{\prec}(i) \ne \phi$, $\Sigma^{\preceq i}$ can be partitioned as

$$\Sigma^{\preceq i} = \begin{bmatrix} \Sigma^{\prec i} & \Sigma^{\prec}_{\cdot i} \\ (\Sigma^{\prec}_{\cdot i})^T & \Sigma_{ii} \end{bmatrix}.$$

Also, note that if $Z \sim \mathcal{N}(\mathbf{0}, \Sigma)$, then D_{ii} is the conditional variance of Z_i given $Z_1, Z_2, \ldots, Z_{i-1}$ (see [16]). Note that $\Sigma_{kl} = 0$ for all $1 \le k, l \le i, k \in \mathcal{N}^{\prec}(i), l \notin \mathcal{N}^{\prec}(i)$. It follows that $D_{ii} = \Sigma_{ii} - (\Sigma^{\prec}_{\cdot i})^T (\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i}$. Hence, by the formula for the inverse of a partitioned matrix, it follows that

$$(\Sigma^{\preceq i})^{-1} = \begin{bmatrix} (\Sigma^{\prec i})^{-1} + \dfrac{((\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i})((\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i})^T}{D_{ii}} & -\dfrac{(\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i}}{D_{ii}} \\ -\dfrac{((\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i})^T}{D_{ii}} & \dfrac{1}{D_{ii}} \end{bmatrix}.$$

Hence,

$$
\begin{aligned}
\operatorname{tr}((\Sigma^{\preceq i})^{-1} U^{\preceq i}) \\
&= \operatorname{tr}((\Sigma^{\prec i})^{-1} U^{\prec i}) \\
&\quad + \frac{1}{D_{ii}} ((\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i} - (U^{\prec i})^{-1} U^{\prec}_{\cdot i})^T U^{\prec i} ((\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i} - (U^{\prec i})^{-1} U^{\prec}_{\cdot i}) \\
&\quad + \frac{1}{D_{ii}} (U_{ii} - (U^{\prec}_{\cdot i})^T (U^{\prec i})^{-1} U^{\prec}_{\cdot i}).
\end{aligned}
$$
(5.1)

We again note that from our argument at the beginning of the proof, $\Sigma^{\prec i} = \Sigma^{\preceq(i-1)}$ or $\Sigma^{\prec i}$ has a block diagonal structure (after an appropriate permutation of the rows and columns) with blocks $\Sigma^{\preceq i_1}, \Sigma^{\preceq i_2}, \ldots, \Sigma^{\preceq i_k}$ for some $k > 1, 1 \le i_1, i_2, \ldots, i_k < i$.[14] It follows that

$$\operatorname{tr}((\Sigma^{\prec i})^{-1} U^{\prec i}) = \sum_{j=1}^{k} \operatorname{tr}((\Sigma^{\preceq i_j})^{-1} U^{\preceq i_j}).$$

---

[14]If the equivalence class of i has k children in the Hasse diagram of G and V_j is the set of vertices in V belonging to the subtree rooted at the jth child, then V_j, for $1 \le j \le k$, are disjoint subsets. In fact, if $i_j = \max\{i' : i' \in V_j\}$, then it follows by the construction of the Hasse diagram that $V_j = \mathcal{N}^{\preceq}(i_j)$ for $1 \le j \le k$.

Note that $\Sigma = \Sigma^{\preceq m}$. Using (5.1) recursively, we get

$$\text{tr}(\Sigma^{-1}U)$$

$$(5.2) \quad = \sum_{i=1}^{m} \frac{1}{D_{ii}} \big((\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec} - (U^{\prec i})^{-1}U_{\cdot i}^{\prec}\big)^{T} U^{\prec i} \big((\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec} - (U^{\prec i})^{-1}U_{\cdot i}^{\prec}\big)$$

$$+ \frac{1}{D_{ii}} \big(U_{ii} - (U_{\cdot i}^{\prec})^{T}(U^{\prec i})^{-1}U_{\cdot i}^{\prec}\big).$$

Let us now evaluate the Jacobian of the transformation

$$\Sigma \to \{(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec})\}_{1 \le i \le m}.$$

It follows by simple matrix manipulations that the Jacobian of the transformation

$$\Sigma^{\preceq i} \to (\Sigma^{\prec i}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec}, D_{ii})$$

is given by $|\Sigma^{\prec i}|$.

Once more, note that $\Sigma = \Sigma^{\preceq m}$ and, as mentioned earlier, $\Sigma^{\prec i} = \Sigma^{\preceq (i-1)}$ or $\Sigma^{\prec i}$ (after an appropriate permutation of the rows and columns) has a block diagonal structure with blocks $\Sigma^{\preceq i_1}, \Sigma^{\preceq i_2}, \dots, \Sigma^{\preceq i_k}$ for some $k > 1$, $1 \le i_1, i_2, \dots,$ $i_k < i$. Hence, by regarding the transformation

$$\Sigma \to \{(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec})\}_{1 \le i \le m}$$

as a series of transformations of the type $\Sigma^{\preceq i} \to (\Sigma^{\prec i}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec}, D_{ii})$, it follows that the determinant of the Jacobian is given by

$$(5.3) \qquad \prod_{i=1}^{m} |\Sigma^{\prec i}| = \prod_{i=1}^{m} \prod_{j \in \mathcal{N}^{\prec}(i)} D_{jj} = \prod_{j=1}^{m} D_{jj}^{n_j}.$$

Here, as in Section 3.1 Lemma 2,

$$n_j = |\{(i > j : (i, j) \in E\}| \qquad \forall j = 1, 2, \dots, m.$$

Also, from Section 3.1,

$$\pi_{U,\boldsymbol{\alpha}}^{P_G}(\Sigma) = \frac{1}{z_G(U, \boldsymbol{\alpha})} e^{-(\text{tr}(\Sigma^{-1}U) + \sum_{j=1}^{m}(2n_j + \alpha_j)\log D_{jj})/2}, \qquad \Sigma \in P_G.$$

Let

$$\boldsymbol{\Gamma} = \{(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec})\}_{1 \le i \le m}.$$

It follows from the decomposition of $\text{tr}(\Sigma^{-1}U)$ from (5.2) and the computation of the determinant of the Jacobian (5.3) that

$$\pi_{U,\boldsymbol{\alpha}}^{\boldsymbol{\Gamma}}(\{(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec})_{1 \le i \le m}\})$$

$$(5.4) \quad = \frac{1}{z_G(U, \boldsymbol{\alpha})} \prod_{i=1}^{m} e^{-1/(2D_{ii})((\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec} - (U^{\prec i})^{-1}U_{\cdot i}^{\prec})^{T} U^{\prec i}((\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec} - (U^{\prec i})^{-1}U_{\cdot i}^{\prec})}$$

$$\times \prod_{i=1}^{m} e^{-1/(2D_{ii})(U_{ii} - (U_{\cdot i}^{\prec})^{T}(U^{\prec i})^{-1}U_{\cdot i}^{\prec})} D_{ii}^{-\alpha_i/2}.$$

The above proves the mutual independence of $\{(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma^{\prec}_{\cdot i})\}_{1 \le i \le m}$. From the joint density of $(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma^{\prec}_{\cdot i})$, it is clear that

$$(\Sigma^{\prec i})^{-1}\Sigma^{\prec}_{\cdot i} \mid D_{ii} \sim \mathcal{N}((U^{\prec i})^{-1}U^{\prec}_{\cdot i}, D_{ii}(U^{\prec i})^{-1}).$$

To evaluate the marginal density of D_{ii}, we integrate out $(\Sigma^{\prec i})^{-1}\Sigma^{\prec}_{\cdot i}$ from the joint density of $(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma^{\prec}_{\cdot i})$. Note that

$$\int_{\mathbb{R}^{|\mathcal{N}^{\prec}(i)|}} e^{-1/(2D_{ii})((\Sigma^{\prec i})^{-1}\Sigma^{\prec}_{\cdot i}-(U^{\prec i})^{-1}U^{\prec}_{\cdot i})^T U^{\prec i}((\Sigma^{\prec i})^{-1}\Sigma^{\prec}_{\cdot i}-(U^{\prec i})^{-1}U^{\prec}_{\cdot i})} d((\Sigma^{\prec i})^{-1}\Sigma^{\prec}_{\cdot i})$$

$$= C D_{ii}^{-|\mathcal{N}^{\prec}(i)|/2},$$

where C is a constant, since the above integral is essentially an unnormalized multivariate normal integral. Hence, the marginal density of D_{ii} is given by

$$\pi^{D_{ii}}_{U,\boldsymbol{\alpha}}(d) \propto e^{-(U_{ii}-(U^{\prec}_{\cdot i})^T(U^{\prec i})^{-1}U^{\prec}_{\cdot i})/(2d)} d^{-(\alpha_i/2+|\mathcal{N}^{\prec}(i)|/2)}.$$

We can therefore conclude that

$$D_{ii} \sim \mathrm{IG}\left(\frac{\alpha_i}{2} - \frac{|\mathcal{N}^{\prec}(i)|}{2} - 1, \frac{U_{ii} - (U^{\prec}_{\cdot i})^T(U^{\prec i})^{-1}U^{\prec}_{\cdot i}}{2}\right). \qquad \square$$

REMARK. At first glance, it seems as if the only part of U that appears in Theorem 3 is $(U_{ij})_{(i,j)\in E}$, that is, the projection of U onto I_G. Hence, one could incorrectly conclude that up to the number of shape parameters in each equivalence class, in the homogeneous case, the priors introduced in this paper and the IW_{Q_G} are identical. However, a careful inspection shows that this is not the case. Note that the conditional covariance of $(\Sigma^{\prec i})^{-1}\Sigma^{\prec}_{\cdot i}$ is $D_{ii}(U^{\prec i})^{-1}$, and $U^{\prec i}$ can contain entries of the form U_{kl} such that $(k,l) \notin E$. For example, suppose that $G = \overset{1}{\bullet} - \overset{3}{\bullet} - \overset{2}{\bullet}$. Then,

$$U^{\prec 3} = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix},$$

but $(1,2) \notin E$. Hence, in the homogeneous setting, $\pi^{P_G}_{U,\boldsymbol{\alpha}}$ is truly a larger class than the IW_{Q_G} family of distributions.

We now establish the correspondence between $\pi^{\Gamma}_{U,\boldsymbol{\alpha}}$ in (5.4) and the conjugate prior for an appropriate Gaussian DAG model. Let $G = (V, E)$ be a homogeneous graph with $V \in S_H$, that is, the vertices have been ordered according to the perfect vertex elimination scheme for homogeneous graphs outlined in Section 2.4. Let us construct a DAG as follows:

1. Consider the Hasse diagram of G (for simplicity and clarity of exposition, assume that the equivalence class at each vertex has just one element).
2. Assign a directed edge from u to v if u is a descendant of v in the Hasse tree, that is, reverse the directions of all the arrows, including those that do not appear in the Hasse tree, but which are implied by transitivity.
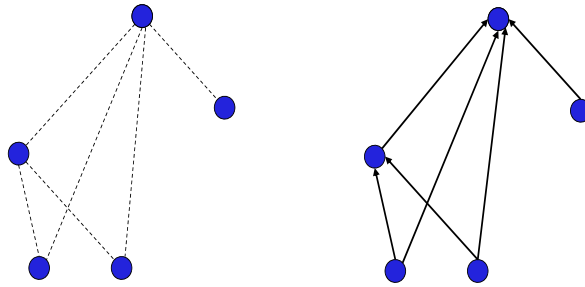
FIG. 2.   *Example of homogeneous covariance graph* (*left*); *associated DAG model* (*right*).

An example of a DAG constructed in this manner is given in Figure 2.

Now, let pa(i) denote the set of parents of i according to the direction specified above. If $\mathbf{Y} \sim \mathcal{N}_m(\mathbf{0}, \Sigma)$ has a distribution which is Markov with respect to the above DAG, then the density of $\mathbf{Y}$ factorizes as

$$f(\mathbf{y}) = \prod_{i=1}^{m} f\big(y_i \mid \mathbf{y}_{\text{pa}(i)}\big)$$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi} D_{ii}^{1/2}} e^{-(y_i - ((\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i})^T \mathbf{y}_{\text{pa}(i)})^T D_{ii}^{-1} (y_i - ((\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i})^T \mathbf{y}_{\text{pa}(i)})},$$

where $D_{ii} := \Sigma_{ii} - (\Sigma^{\prec}_{\cdot i})^T (\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i} = \Sigma_{i|\text{pa}(i)}$.

The standard conjugate prior for each factor of the product above can be obtained as follows. Given an arbitrary positive definite matrix U and $\alpha_i' > 0$ for $i = 1, 2, \ldots, m$, let

$$(\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i} \mid D_{ii} \sim \mathcal{N}((U^{\prec i})^{-1} U^{\prec}_{\cdot i}, D_{ii}(U^{\prec i})^{-1}),$$

$$D_{ii} \sim \text{IG}\bigg(\alpha_i', \frac{U_{ii} - (U^{\prec}_{\cdot i})^T (U^{\prec i})^{-1} U^{\prec}_{\cdot i}}{2}\bigg)$$

for $i = 1, 2, \ldots, m$, where $\{(D_{ii}, (\Sigma^{\prec i})^{-1} \Sigma^{\prec}_{\cdot i})\}_{1 \leq i \leq m}$ are mutually independent. This corresponds precisely to the $\pi_{U,\boldsymbol{\alpha}}^{\Gamma}$ density in (5.4).

We now proceed to state, without proof, results for homogeneous covariance graph models by exploiting the correspondence of our priors to the standard conjugate priors for DAGs. In particular, hyper-Markov properties, the normalizing constant and expected values for covariance graph models, for G homogeneous, are formally stated below.

Let $G = (V, E)$ be a homogeneous graph with $V \in S_H$, that is, the vertices have been ordered according to the perfect vertex elimination scheme for homogeneous graphs outlined in Section 2.4. Let D be the directed graph obtained from G by directing all edges in G from the vertex with the smallest number to the vertex with the highest number. Let pa(i) denote the set of parents of i according to the direction specified in D. It follows that $\text{pa}(i) = \mathcal{N}^{\prec}(i)$. As in [9, 20],

let $\mathrm{pr}(i) = \{1, 2, \ldots, i - 1\}$ denote the set of predecessors of i according to the direction specified in D. We now proceed to define the hyper-Markov property.

DEFINITION 1.   A family of priors $\mathcal{F}$ on P_G satisfies the strong hyper-Markov property with respect to the direction D if, whenever $\pi \in \mathcal{F}$ and $\Sigma \sim \pi$,

$$\Sigma_{i|\mathrm{pa}(i)} \perp \Sigma_{\mathrm{pr}(i)} \qquad \forall 1 \le i \le m,$$

where $\Sigma_{i|\mathrm{pa}(i)} := \Sigma_{ii} - (\Sigma_{\cdot i}^{\prec})^T (\Sigma^{\prec i})^{-1} \Sigma_{\cdot i}^{\prec} = D_{ii}$.

In the following corollary, we state, without proof, that the family of priors $\pi_{U,\boldsymbol{\alpha}}^{P_G}$ satisfies the strong hyper-Markov property with respect to the direction D.

COROLLARY 1.   *Let $G = (V, E)$ be homogeneous with $V \in S_H$. If $\Sigma \sim \pi_{U,\boldsymbol{\alpha}}^{P_G}$, then*

$$D_{ii} \perp \Sigma_{\{1,2,\ldots,i-1\}} \qquad \forall 1 \le i \le m.$$

REMARK.   Recall that

$$\Sigma^{\prec i} = ((\Sigma_{uv}))_{u,v \in \mathcal{N}^{\prec}(i)}$$

is different from

$$\Sigma_{\{1,2,\ldots,i-1\}} = ((\Sigma_{uv}))_{1 \le u, v \le i-1}.$$

We demonstrated in Section 3.2 that the family IW_{Q_G} of Letac and Massam [20] is a subfamily of our class of priors $\pi_{U,\boldsymbol{\alpha}}^{P_G}$ when G is homogeneous. Consequently, we can now prove hyper-Markov properties for the IW_{Q_G} family.

COROLLARY 2.   *Let $G = (V, E)$ be homogeneous with $V \in S_H$. Let D be the directed graph obtained from G by directing all edges in G from the vertex with the smallest number to the vertex with the highest number. The family IW_{Q_G} is then strong hyper-Markov with respect to the direction G_H.*

Hyper-Markov properties for the IW_{Q_G} family were not established in [20]. Hence, we note that the corollary above is a new result for this family.

We now proceed to state, without proof, the functional form of the normalizing constant for homogeneous graphs, once again exploiting the correspondence between our covariance priors and the conjugate priors for DAGs. In particular, below, we state necessary and sufficient conditions for existence of the normalizing constant and give an explicit expression for it in such cases.

COROLLARY 3. *Let $G = (V, E)$ be a homogeneous graph with vertices ordered such that $V \in S_H$. Then, $z_G(U, \boldsymbol{\alpha}) < \infty$ if and only if $\boldsymbol{\alpha}$ satisfies the conditions in Proposition 1, that is, $\alpha_i > |\mathcal{N}^{\prec}(i)| + 2 \; \forall i = 1, 2, \ldots, m$. In this case,*

$$z_G(U, \boldsymbol{\alpha}) = \prod_{i=1}^{m} \Gamma\left(\frac{\alpha_i}{2} - \frac{|\mathcal{N}^{\prec}(i)|}{2} - 1\right) 2^{\alpha_i/2-1} (\sqrt{\pi})^{|\mathcal{N}^{\prec}(i)|}$$

(5.5)

$$\times |U^{\prec i}|^{\alpha_i/2 - |\mathcal{N}^{\prec}(i)|/2 - 3/2} \big/ |U^{\preceq i}|^{\alpha_i/2 - |\mathcal{N}^{\prec}(i)|/2 - 1}.$$

We now proceed to state, without proof, expected values related to our class of priors $\pi_{U,\boldsymbol{\alpha}}^{P_G}$ when G is homogeneous, again by exploiting the correspondence between our covariance priors and conjugate priors for DAGs. In particular, we now provide a recursive method that gives closed form expressions for the expected value of the covariance matrix when $\Sigma \sim \pi_{U,\boldsymbol{\alpha}}^{P_G}$. Since $\Sigma_{uv} = 0 \; \forall (u, v) \notin E$, we only need to evaluate the expectation of Σ_{ii} and $\Sigma_{\cdot i}^{\prec}$ for every $1 \le i \le m$. Let

$$A_1 := \{i \in V : \mathcal{N}^{\prec}(i) = \phi\}.$$

Clearly, if $i \in A_1$, then $\Sigma^{\prec i}$ and $\Sigma_{\cdot i}$ are vacuous parameters and $D_{ii} = \Sigma_{ii}$. It follows from Theorem 3 that for $i \in A_1$,

$$\mathbf{E}_{U,\boldsymbol{\alpha}}[\Sigma_{ii}] = \mathbf{E}_{U,\boldsymbol{\alpha}}[D_{ii}] = \frac{U_{ii} - (U_{\cdot i}^{\prec})^T (U^{\prec i})^{-1} U_{\cdot i}^{\prec}}{\alpha_i - 4},$$

assuming that $\alpha_i > 4$, since $X \sim \mathrm{IG}(\lambda, \gamma)$ implies that $\mathbf{E}[X] = \frac{\lambda}{\gamma - 1}$.

For $k = 2, 3, 4, \ldots$, define

$$A_k = \left\{i \in V : \mathcal{N}^{\prec}(i) \subseteq \bigcup_{l=1}^{k-1} A_l\right\} \setminus \left(\bigcup_{l=1}^{k-1} A_l\right).$$

Since there are finitely many vertices in V, there exists some k^* such that $A_k = \phi$ for $k > k^*$. The sets $\{A_k\}_{1 \le k \le k^*}$ essentially provide a way of computing $\mathbf{E}_{U,\boldsymbol{\alpha}}[\Sigma]$, by starting at the bottom of the Hasse diagram of G and then moving up sequentially.

COROLLARY 4. *Let G be a homogeneous graph. Given the expectations of $\Sigma_{\cdot j}^{\prec}$ and Σ_{jj} for $j \in \bigcup_{l=1}^{k-1} A_l$, the expectations of $\Sigma_{\cdot i}^{\prec}$ and Σ_{ii} for $i \in A_k$ are given, respectively, by the expressions*

$$\mathbf{E}_{U,\boldsymbol{\alpha}}[\Sigma_{\cdot i}^{\prec}] = \mathbf{E}_{U,\boldsymbol{\alpha}}[\Sigma^{\prec i}] (U^{\prec i})^{-1} U_{\cdot i}^{\prec};$$

$$\mathbf{E}_{U,\boldsymbol{\alpha}}[\Sigma_{ii}] = \frac{U_{ii} - (U_{\cdot i}^{\prec})^T (U^{\prec i})^{-1} U_{\cdot i}^{\prec}}{\alpha_i - |\mathcal{N}^{\prec}(i)| - 4}$$

$$+ \mathrm{tr}\left(\mathbf{E}_{U,\boldsymbol{\alpha}}[\Sigma^{\prec i}] \left(\frac{(U^{\prec i})^{-1} (U_{ii} - (U_{\cdot i}^{\prec})^T (U^{\prec i})^{-1} U_{\cdot i}^{\prec})}{\alpha_i - |\mathcal{N}^{\prec}(i)| - 4}\right.\right.$$

$$\left.\left. + (U^{\prec i})^{-1} U_{\cdot i}^{\prec} (U_{\cdot i}^{\prec})^T (U^{\prec i})^{-1}\right)\right),$$

provided that $\alpha_i > |\mathcal{N}^{\prec}(i)| + 4$.

The corollary is not formally proved since it follows directly from the correspondence between our covariance priors and the natural conjugate priors for DAGs. We note once more that the expressions above yield a recursive but closed form method to calculate $\mathbf{E}[\Sigma]$ when $\Sigma \sim \pi_{U,\boldsymbol{\alpha}}^{P_G}$.

REMARK. There is an intriguing parallel between the expressions for the normalizing constant and the expected values for the $\pi_{U,\boldsymbol{\alpha}}^{P_G}$ distribution and the IW_{P_G} distribution (as derived in [20, 26]) when G is homogeneous. This automatically leads one to wonder if the $\pi_{U,\boldsymbol{\alpha}}^{P_G}$ and the IW_{P_G} distributions are the same. We now show that this is not the case.

If one compares the density of $(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec})_{1 \leq i \leq m}$ in (5.4) and the $IW_{P_G}^*$ density in (3.16) of [20], they initially appear to have the same functional form. We now proceed to show that they are supported on different spaces. This difference is illustrated by the following example. Let $G = \overset{1}{\bullet} - \overset{3}{\bullet} - \overset{2}{\bullet}$. In this case, each equivalence class in the Hasse tree of G has exactly one vertex. Note that vertex 3 has two descendants, and vertices 1 and 2 do not have any descendants in the Hasse tree of G. Hence, it follows that the density of $(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec})_{1 \leq i \leq 3}$ is supported on $(\mathbb{R} \times NULL, \mathbb{R} \times NULL, \mathbb{R} \times \mathbb{R}^2)$. On the other hand, vertices 1 and 2 have one ancestor, and vertex 3 has no ancestors in the Hasse tree of G. Hence, it follows that the $IW_{P_G}^*$ density is supported on $(\mathbb{R} \times \mathbb{R}, \mathbb{R} \times \mathbb{R}, \mathbb{R} \times NULL)$.

So, at first glance, it looks as if $(D_{ii}, (\Sigma^{\prec i})^{-1}\Sigma_{\cdot i}^{\prec})_{1 \leq i \leq m}$ has the same form as $IW_{P_G}^*$, but, upon further examination, we see that even for the simplest homogeneous graph, they are structurally different. In fact, $(\mathbb{R} \times NULL, \mathbb{R} \times NULL, \mathbb{R} \times \mathbb{R}^2)$ does not support the $IW_{P_G}^*$ distribution for any G that is homogeneous.

6. Examples. The main purpose of this paper is to undertake a theoretical investigation of our class of distributions and their efficacy for use in Bayesian estimation in covariance graph models. We nevertheless provide two examples (one real and one simulated) to demonstrate how the methodology developed in this paper can be implemented.

6.1. *Genomics example.* We provide an illustration of our methods on a data set consisting of gene expression data from microarray experiments with yeast strands from Gasch et al. [13]. This data set has also been analyzed in [4, 11]. As in [4, 11], we consider a subset of eight genes involved in galactose utilization. There are $n = 134$ experiments and the empirical covariance matrix for these measurements is provided in Table 1. Note that the sample covariance matrix is obtained after centering since the mean is not assumed to be zero.

We consider the covariance graph model specified by the graph G in Figure 3 with the overall aim of estimating Σ under this covariance graph model.

TABLE 1
Empirical covariance matrix for yeast data

	GAL11	GAL4	GAL80	GAL3	GAL7	GAL10	GAL1	GAL2
GAL11	0.152							
GAL4	0.034	0.130						
GAL80	0.015	0.039	0.221					
GAL3	−0.055	0.034	0.073	0.608				
GAL7	−0.051	−0.053	0.183	0.722	3.423			
GAL10	−0.048	−0.039	−0.188	0.553	2.503	2.372		
GAL1	−0.066	−0.061	0.224	0.517	2.768	2.409	2.890	
GAL2	−0.119	−0.018	0.208	0.583	2.547	2.278	2.514	2.890

The maximum likelihood estimate for $\Sigma \in P_G$, provided by the iterative conditional fitting algorithm described in [4], yields a deviance of 4.694 over 7 degrees of freedom, thus indicating a good model fit. The maximum likelihood estimate is provided in Table 1. We use the following ordering for our analysis: {GAL11, GAL4, GAL80, GAL3, GAL7, GAL10, GAL1, GAL2}.

Our goal is to obtain the posterior mean for Σ under our new class of priors and then to provide Bayes estimators for Σ. We use two diffuse priors to illustrate our methodology. The first prior is denoted as $\widetilde{\pi}_{U_1, \boldsymbol{\alpha}^1}$, where $U_1 = \frac{\mathrm{tr}(S)}{8} I_8$, $\alpha_i^1 = 5 + |\mathcal{N}^{\prec}(i)|$, $i = 1, 2, \ldots, 8$, that is, $\boldsymbol{\alpha}_1 = (5, 6, 6, 8, 7, 8, 9, 12)$. The second prior
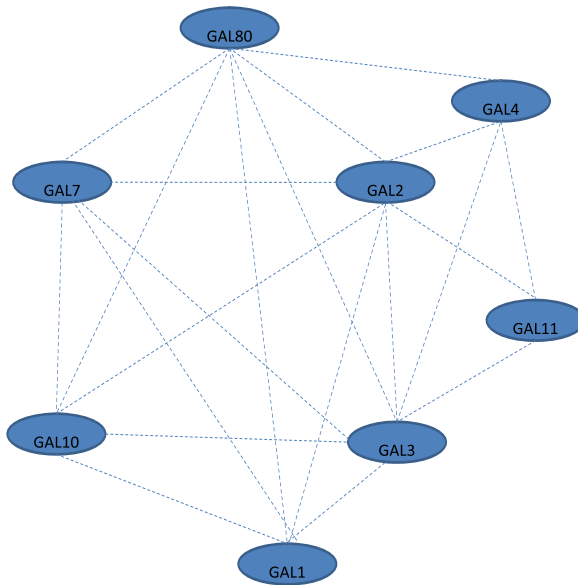

FIG. 3.   *Covariance graph for yeast data.*

used is $\tilde{\pi}_{U_2,\boldsymbol{\alpha}^2}$, where $U_2 = 0$, $\alpha_i^2 = 2$, $i = 1, 2, \ldots, 8$. Note that we could have used any ordering in S_D for our analysis. As an example, we select an alternate ordering, {GAL11, GAL4, GAL80, GAL10, GAL2, GAL3, GAL1, GAL7}, and also consider the two priors mentioned above under this alternative ordering. The block Gibbs sampling procedure was run for the four priors as specified in Section 4. The burn-in period was chosen to be 1,000 iterations and the subsequent 1,000 iterations were used to compute the posterior mean. Increasing the burn-in period to more than 1,000 iterations results in insignificant changes to our estimates, thus indicating that the burn-in period chosen is sufficient. The posterior mean estimates for both the priors, together with the MLE estimates, are provided in Table 2. The running time for the Gibbs sampling procedure for each prior is approximately 26 seconds on a Pentium M 1.6 GHz processor. We find that the Bayesian approach using our priors and the corresponding block Gibbs sampler gives stable estimates and thus yields a useful alternative methodology for inference in covariance graph models. We also note that the two different vertex orderings yield very similar results.

6.2. *Simulation example.* A proof of convergence of the block Gibbs sampling algorithm proposed in Section 4.1 was provided in Section 4.2. The speed at which convergence occurs is also a very important concern for implementation of the algorithm. The number of steps that are required before one can generate a reasonable approximate sample from the posterior distribution is reflective of the rate of convergence. Understanding this is important for the accuracy of Bayes estimates such as the posterior mean. We proceed to investigate the performance of the block Gibbs sampling algorithm in a situation where the posterior mean is known exactly and hence allows a direct comparison. Consider a homogeneous graph G with 50 vertices, with the corresponding Hasse tree given by Figure 4. Let $\Sigma \in P_G$, where the vertices have been ordered according to the Hasse perfect vertex elimination scheme of Section 2.4, the diagonal entries are 50 and all other nonzero entries are 1. We simulate 100 observation vectors $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \ldots, \mathbf{Y}_{99}, \mathbf{Y}_{100}$ from $\mathcal{N}_{50}(0, \Sigma)$. For illustration purposes, we choose a diffuse prior $\pi_{U,\boldsymbol{\alpha}}^{\Sigma}$ with $U = 0$ and $\alpha_i = 2|\mathcal{N}^<(i)| + 5, i = 1, 2, \ldots, 50$.

Since the graph G is homogeneous, we can compute the posterior mean $\Sigma_{\text{mean}} := \mathbf{E}_{U,\boldsymbol{\alpha}}[\Sigma \mid \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{100}]$ explicitly. We can therefore assess the ability of the block Gibbs sampling algorithm to estimate the posterior mean by comparing it to the true value of the mean. We run the block Gibbs sampling algorithm to sample from the posterior distribution and subsequently check its performance in estimating Σ_{mean}. We use an initial burn-in period of B iterations and then average over the next I iterations to get the estimate $\widehat{\Sigma}$. The times needed for computation (using the R software) and the relative errors $\frac{\|\widehat{\Sigma}-\Sigma_{\text{mean}}\|_2}{\|\Sigma_{\text{mean}}\|_2}$ corresponding to various choices of B and I are provided in Table 3. The diagnostics in Table 3 indicate that the block Gibbs sampling algorithm performs exceptionally well, yielding

TABLE 2
ICF: *Maximum likelihood estimate from iterative conditional fitting.* BY1: *Bayesian posterior mean estimate for prior* $\pi_{U_1, \boldsymbol{\alpha}^1}$. BY2: *Bayesian posterior mean estimate for prior* $\pi_{U_2, \boldsymbol{\alpha}^2}$. $\widetilde{BY}1$: *Bayesian posterior mean estimate for prior* $\pi_{U_1, \boldsymbol{\alpha}^1}$ *with a different ordering.* $\widetilde{BY}2$: *Bayesian posterior mean estimate for prior* $\pi_{U_2, \boldsymbol{\alpha}^2}$ *with a different ordering*

|	GAL11	GAL4	GAL80	GAL3	GAL7	GAL10	GAL1	GAL2	Method
GAL11	0.152	0.030	0	−0.052	0	0	0	−0.068	ICF
	0.164	0.030	0	−0.050	0	0	0	−0.068	BY1
	0.156	0.030	0	−0.052	0	0	0	−0.068	BY2
	0.152	0.030	0	−0.051	0	0	0	−0.069	$\widetilde{BY}1$
	0.155	0.030	0	−0.052	0	0	0	−0.070	$\widetilde{BY}2$
GAL4		0.128	0.040	0.042	0	0	0	0.030	ICF
		0.142	0.040	0.041	0	0	0	0.027	BY1
		0.133	0.041	0.042	0	0	0	0.028	BY2
		0.128	0.039	0.040	0	0	0	0.026	$\widetilde{BY}1$
		0.132	0.040	0.042	0	0	0	0.0278	$\widetilde{BY}2$
GAL80			0.223	0.082	0.197	0.198	0.239	0.227	ICF
			0.237	0.072	0.193	0.194	0.235	0.216	BY1
			0.232	0.076	0.199	0.2	0.243	0.223	BY2
			0.224	0.076	0.197	0.197	0.240	0.218	$\widetilde{BY}1$
			0.232	0.076	0.202	0.203	0.245	0.227	$\widetilde{BY}2$
GAL3				0.612	0.723	0.549	0.515	0.582	ICF
				0.626	0.713	0.544	0.509	0.575	BY1
				0.643	0.747	0.568	0.532	0.599	BY2
				0.628	0.719	0.549	0.517	0.582	$\widetilde{BY}1$
				0.667	0.749	0.574	0.531	0.605	$\widetilde{BY}2$
GAL7					3.422	2.593	2.768	2.540	ICF
					3.462	2.584	2.756	2.533	BY1
					3.588	2.682	2.866	2.636	BY2
					3.541	2.588	2.761	2.532	$\widetilde{BY}1$
					3.708	2.681	2.865	2.627	$\widetilde{BY}2$
GAL10						2.372	2.409	2.267	ICF
						2.373	2.400	2.266	BY1
						2.453	2.497	2.358	BY2
						2.389	2.407	2.277	$\widetilde{BY}1$
						2.473	2.489	2.356	$\widetilde{BY}2$
GAL1							2.890	2.502	ICF
							2.961	2.501	BY1
							3.086	2.604	BY2
							2.969	2.496	$\widetilde{BY}1$
							3.087	2.582	$\widetilde{BY}2$
GAL2								2.870	ICF
								3.003	BY1
								3.153	BY2
								2.892	$\widetilde{BY}1$
								3.005	$\widetilde{BY}2$
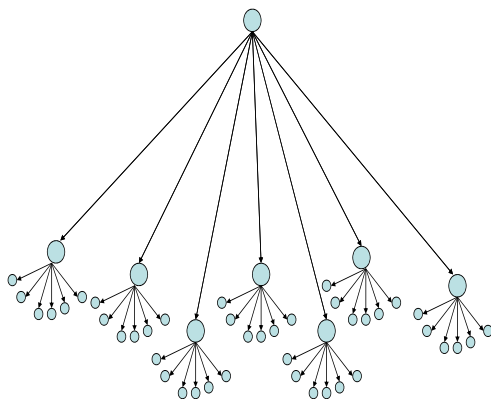
FIG. 4. *Hasse diagram for a homogeneous graph with* 50 *vertices.*

estimates that approach the true mean in only a few thousand steps. The time taken for running the algorithm is also provided in Table 3.

The diagnostics in Table 3 indicate that the block Gibbs sampling algorithm performs exceptionally well, yielding estimates that approach the true mean in only a few thousand steps. The time taken for running the algorithm is also provided in Table 3.

7. Closing remarks. In this paper, we have proposed a theoretical framework for Bayesian inference in covariance graph models. The main challenge was the unexplored terrain of working with curved exponential families in the continuous setting. A rich class of conjugate priors has been developed in this paper for covariance graph models where the underlying graph is decomposable.

We have been able to exploit the structure of the conjugate priors to develop a block Gibbs sampler to effectively sample from the posterior distribution. A rigorous proof of convergence is also given. Comparison with other classes of priors is also undertaken. We are able to compute the normalizing constant for homoge-

TABLE 3
Performance assessment of the Gibbs sampler in simulation example

Burn-in (B)	Average (I)	Time (seconds)	Relative error
1000	1000	139.77	0.01748220
2000	1000	209.72	0.01240595
3000	1000	279.52	0.01300910
4000	1000	349.44	0.01142864
4000	3000	489.19	0.01246141
4000	5000	631.21	0.01081264
4000	7000	769.70	0.009244206

neous graphs, thereby making Bayesian model selection possible in a tractable way for this class of models. The Bayesian approach yields additional dividends, in the sense that we can now carry out inference in covariance graph models, even when the sample size n is less than the dimension p of the data, something which is otherwise not generally possible in the maximum likelihood framework. Furthermore, we thoroughly explore the theoretical properties of our class of conjugate priors. In particular, in the homogeneous case, hyper-Markov properties and closed form expressions for the expected value of the covariance matrix are established. Furthermore, the usefulness of the methodology that is developed is illustrated through examples. A couple of open problems are worth mentioning:

- What are the necessary conditions for the existence of the normalizing constant for decomposable graphs?
- Does the hyper-Markov property for the class of priors developed in this paper hold for decomposable graphs?

We conclude by noting that the use of the class of Wishart distributions introduced in this paper for Bayesian inference, along with a detailed study of Bayes estimators in this context, is clearly an important topic and is the focus of current research.

APPENDIX

PROOF OF PROPOSITION 1. From the definition of N and L, it is easy to verify that

$$(LN)_{ii} = 1 \qquad \forall 1 \leq i \leq m,$$

$$(LN)_{ij} = 0 \qquad \forall 1 \leq i < j \leq m.$$

Now, let $i > j$. It follows, by the definition of N, that

$$(LN)_{ij} = \sum_{k=j}^{i} L_{ik} N_{kj}$$

$$= N_{ij} + \sum_{k=j+1}^{i-1} L_{ik} \sum_{\tau \in A, \tau_1 = k, \tau_{\dim(\tau)} = j} (-1)^{\dim(\tau)-1} L_{\tau} + L_{ij}$$

$$= N_{ij} - \sum_{k=j+1}^{i-1} \sum_{\tau \in A, \tau_1 = k, \tau_{\dim(\tau)} = j} (-1)^{\dim(\tau)} L_{ik} L_{\tau} + L_{ij}.$$

Note that any $\tau' \in \mathcal{A}$ with $\tau'_1 = i$, $\tau'_{\dim(\tau')} = j$, $\dim(\tau') > 2$ can be uniquely expressed as $\tau' = (i, \tau)$, where $j + 1 \leq \tau_1 \leq i - 1$, $\tau_{\dim(\tau)} = j$. Recall that, by definition, $L_{\tau'} = L_{i\tau_1} L_{\tau}$. Also, if $\tau' \in \mathcal{A}$ with $\tau'_1 = i$, $\tau'_{\dim(\tau')} = j$, $\dim(\tau') = 2$, then

$\tau' = (i, j)$ and $L_{\tau'} = L_{ij}$. Hence,

$$(LN)_{ij} = N_{ij} - \sum_{\tau' \in A, \tau'_1 = i, \tau'_{\dim(\tau')} = j} (-1)^{\dim(\tau')-1} L_{\tau'}$$

$$= N_{ij} - N_{ij}$$

$$= 0.$$

Hence, $LN = I$ and thus $L^{-1} = N$. □

PROOF OF THEOREM 1. Let us simplify the integral by integrating out the terms D_{ii}, $1 \leq i \leq m$:

$$\int e^{-(\mathrm{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})/2} \, dL \, dD$$

$$= \int e^{-(\mathrm{tr}(D^{-1}(L^{-1}U(L^T)^{-1})) + \sum_{i=1}^m \alpha_i \log D_{ii})/2} \, dL \, dD$$

$$= \int \prod_{i=1}^m e^{-(L^{-1}U(L^T)^{-1})_{ii}/(2D_{ii})} D_{ii}^{-\alpha_i/2} \, dD \, dL$$

$$= \int \prod_{i=1}^m \frac{\Gamma(\alpha_i/2 - 1)2^{\alpha_i/2-1}}{((L^{-1}U(L^T)^{-1})_{ii})^{\alpha_i/2-1}} \, dL$$

$$\text{(assuming } \alpha_i > 2 \, \forall i = 1, 2, \ldots, m)$$

$$= \int \prod_{i=1}^m \frac{\Gamma(\alpha_i/2 - 1)2^{\alpha_i/2-1}}{((L^{-1})_{i.} U((L^{-1})_{i.})^T)^{\alpha_i/2-1}} \, dL. \qquad (**)$$

In order to simplify this integral, we perform a change of measure by transforming the nonzero elements of L to the corresponding elements of L^{-1}. For convenience and brevity, the notation L_{ij}^{-1} is used in place of $(L^{-1})_{ij}$. Now, note the following facts.

1. Let $L \in \mathcal{L}_G$. From Proposition 1, for $(i, j) \in E$, $i > j$,

$$(\text{A.1}) \qquad L_{ij}^{-1} = -L_{ij} + f\big((L_{uv})_{(u,v) \in E, j \leq u < i, j \leq v < u \text{ or } u=i, j < v < i}\big),$$

that is, $L_{ij}^{-1} + L_{ij}$ is a function (f) of L_{uv}, $(u, v) \in E$, $j \leq u < i$, $j \leq v < u$ or $u = i$, $j < v < i$, such that f is zero when all its arguments are zero. We use the above to show that L is a function of $\{L_{uv}^{-1}\}_{u>v, (u,v) \in E}$. Let $i^* = \min\{i : L_{ij} \neq 0 \text{ for some } j < i\}$. Let $j^* = \max\{j : L_{i^*j} \neq 0\}$. By (A.1) and the definition of i^* and j^*, we have $L_{i^*j^*} = -L_{i^*j^*}^{-1}$. We proceed by induction. Let $i > j$, $(i, j) \in E$ and suppose that the hypothesis is true for all $(u, v) \in E$, $1 \leq u < i$, $1 \leq v < u$ or $u = i$, $j < v < i$. Then,

$$L_{ij} = -L_{ij}^{-1} + f\big((L_{uv})_{(u,v) \in E, j \leq u < i, j \leq v < u \text{ or } u=i, j < v < i}\big)$$

and, by the induction hypothesis, the right-hand side of the above equation is a function of $\{L_{uv}^{-1}\}_{u>v,(u,v)\in E}$. Hence, the matrix L is a function of $\{L_{uv}^{-1}\}_{u>v,(u,v)\in E}$.

It follows that the transformation

$$\{L_{ij}\}_{(i,j)\in E,i>j} \to \{L_{ij}^{-1}\}_{(i,j)\in E,i>j}$$

is a bijection and the absolute value of the Jacobian of this transformation is 1 since it is the determinant of a lower-triangular matrix with diagonal entries 1.

2. If $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$ and $U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}$ is a positive definite matrix, then

(A.2) $\mathbf{x}^T U \mathbf{x} = z^T z + \mathbf{x}_2^T (U_{22} - U_{21} U_{11}^{-1} U_{12}) \mathbf{x}_2 \geq \mathbf{x}_2^T (U_{22} - U_{21} U_{11}^{-1} U_{12}) \mathbf{x}_2$,

where $z = U_{11}^{1/2} \mathbf{x}_1 + U_{11}^{-1/2} U_{12} \mathbf{x}_2$.

Hence, after transforming the nonzero entries of L to the corresponding entries of L^{-1} and using (A.2) to eliminate the dependent entries of L^{-1} from the integrand, we get

$$\int e^{-(\operatorname{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})/2} \, dL \, dD$$

$$= \int \prod_{i=1}^m \frac{\Gamma(\alpha_i/2 - 1) 2^{\alpha_i/2 - 1}}{((L^{-1})_{i.} U ((L^{-1})_{i.})^T)^{\alpha_i/2 - 1}} \, dL$$

$$\leq K \prod_{i=2}^m \int_{\mathbb{R}^{|\mathcal{N}^{\prec}(i)|}} \frac{1}{\left(\begin{pmatrix} \mathbf{a}_i^T & 1 \end{pmatrix} U_i^* \begin{pmatrix} \mathbf{a}_i \\ 1 \end{pmatrix} \right)^{\alpha_i/2 - 1}} \, d\mathbf{a}_i.$$

Here, K is a constant, U_i^* is an appropriate positive definite matrix and $\mathbf{a}_i$ represents the independent entries in the ith row of L^{-1}. By a suitable linear transformation $\mathbf{b}_i$ of each of the $\mathbf{a}_i$, $i = 2, 3, \ldots, m$, we get

$$\int e^{-(\operatorname{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})/2} \, dL \, dD$$

$$\leq K^* \prod_{i=2}^m \int_{\mathbb{R}^{|\mathcal{N}^{\prec}(i)|}} \frac{1}{(\mathbf{b}^T \mathbf{b} + u_i^{**})^{\alpha_i/2 - 1}} \, d\mathbf{b}_i.$$

Here, K^* and u_i^{**}, $i = 2, 3, \ldots, m$, are constants. Using the standard fact that

$$\int_{\mathbb{R}^k} \frac{1}{(\mathbf{x}^T \mathbf{x} + 1)^\gamma} \, d\mathbf{x} < \infty \qquad \text{if } \gamma > \frac{k}{2},$$

we conclude that

$$\int e^{-(\operatorname{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})/2} \, dL \, dD < \infty$$

if $\alpha_i > |\mathcal{N}^{\prec}(i)| + 2$ for all $i = 1, 2, \ldots, m$. $\square$

REFERENCES

[1] ANDERSSON, S. A. and WOJNAR, G. G. (2004). Wishart distributions on homogeneous cones. *J. Theoret. Probab.* **17** 781–818. MR2105736

[2] ATHREYA, K. B., DOSS, H. and SETHURAMAN, J. (1996). On the convergence of the Markov chain simulaton method. *Ann. Statist.* **24** 69–100. MR1389881

[3] BUTTE, A. J., TAMAYO, P., SLONIM, D., GOLUB, T. R. and KOHANE, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci.* **97** 12182–12186.

[4] CHAUDHURI, S., DRTON, M. and RICHARDSON, T. S. (2007). Estimation of a covariance matrix with zeroes. *Biometrika* **94** 199–216. MR2307904

[5] CONSONNI, G. and VERONESE, P. (2003). Enriched conjugate and reference priors for the Wishart family on the symmetric cones. *Ann. Statist.* **31** 1491–1516. MR2012823

[6] COX, D. R. and WERMUTH, M. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.* **8** 204–218, 247–277. MR1243593

[7] COX, D. R. and WERMUTH, M. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. Chapman & Hall, London. MR1456990

[8] DANIELS, M. J. and POURAHMADI, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89** 553–566. MR1929162

[9] DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267

[10] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238

[11] DRTON, M. and RICHARDSON, T. S. (2008). Graphical methods for efficient likelihood inference in Gaussian covariance models. *J. Mach. Learn. Res.* **9** 893–914. MR2417257

[12] EDWARDS, D. M. (2000). *Introduction to Graphical Modelling*, 2nd ed. Springer, New York. MR1880319

[13] GASCH, A. P., SPELLMAN, P. T., KAO, C. M., CARMEL-HAREL, O., EISEN, M. B., STORZ, G., BOTSTEIN, D. and BROWN, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11** 4241–4257.

[14] GRONE, R., JOHNSON, C. R., SÁ, E. M. and WOLKOWICZ, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.* **58** 109–124. MR0739282

[15] GRZEBYK, M., WILD, P. and CHOUANIERE, D. (2004). On identification of multifactor models with correlated residuals. *Biometrika* **91** 141–151. MR2050465

[16] HUANG, J., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 8598. MR2277742

[17] KAUERMANN, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist.* **23** 105–116. MR1380485

[18] KHARE, K. and RAJARATNAM, B. (2008). Wishart distributions for covariance graph models. Technical report 2008-11, Dept. Statistics, Stanford Univ.

[19] LAURITZEN, S. L. (1996). *Graphical Models*. Oxford Univ. Press, New York. MR1419991

[20] LETAC, G. and MASSAM, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35** 1278–1323. MR2341706

[21] MAO, Y., KSCHISCHANG, F. R. and FREY, B. J. (2004). Convolutional factor graphs as probabilistic models. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (M. Chickering and J. Halperin, eds.) 374–381. AUAI Press, Arlington, MA.

[22] MASSAM, H. (2007). The IW_{Q_G} as a prior for the variance parameter of covariance graph models. Working paper.

[23] MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York. MR0652932

[24] PAULSEN, V. I., POWER, S. C. and SMITH, R. R. (1989). Schur products and matrix completions. *J. Funct. Anal.* **85** 151–178. MR1005860

[25] POURAHMADI, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance–correlation parameters. *Biometrika* **94** 1006–1013. MR2376812

[26] RAJARATNAM, B., MASSAM, H. and CARVALHO, C. (2008). Flexible covariance estimation in graphical models. *Ann. Statist.* **36** 2818–2849. MR2485014

[27] ROVERATO, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika* **87** 99–112. MR1766831

[28] ROVERATO, A. (2002). Hyper inverse Wishart distribution for non decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.* **29** 391–411. MR1925566

[29] SILVA, R. and GHAHRAMANI, Z. (2009). The hidden life of latent variables: Bayesian learning with mixed graph models. *J. Mach. Learn. Res.* **10** 1187–1238. MR2520804

[30] WERMUTH, N. (1980). Linear recursive equations, covariance selection and path analysis, *J. Amer. Statist. Assoc.* **75** 963–972. MR0600984

[31] WERMUTH, M., COX, D. R. and MARCHETTI, G. M. (2006). Covariance chains. *Bernoulli* **12** 841–862. MR2265345

DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
102 GRIFFIN-FLOYD HALL
GAINESVILLE, FLORIDA 32606
USA
E-MAIL: kdkhare@stat.ufl.edu

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
390 SERRA MALL
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: brajarat@stanford.edu