WISHART DISTRIBUTIONS FOR COVARIANCE GRAPH MODELS

By

Kshitij Khare  
Bala Rajaratnam

Technical Report No. 2009-1  
July 2009

Department of Statistics  
STANFORD UNIVERSITY  
Stanford, California 94305-4065
WISHART DISTRIBUTIONS FOR COVARIANCE GRAPH MODELS

By

Kshitij Khare
Bala Rajaratnam

Department of Statistics
Stanford University

Technical Report No. 2009-1
July 2009

This research was supported in part by National Science Foundation grant DMS 0505303.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu
WISHART DISTRIBUTIONS FOR COVARIANCE GRAPH MODELS

By Kshitij Khare† and Bala Rajaratnam*

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), as the zeroes in the parameter are now reflected in the covariance matrix $\Sigma$, 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 zeroes corresponding to the missing edges of $G$. As in Letac and Massam [20] we consider the case when $G$ is decomposable. In this paper we construct on the cone $P_G$ a family of Wishart distributions, that serve a similar purpose in the covariance graph setting, as those constructed by Letac and Massam [20] and Dawid and Lauritzen [8] 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 $\Sigma$ even in the case when the sample size is less than the dimension of the data (i.e., when $n < p$), otherwise not possible in general in the maximum likelihood framework. Third, we show that when $G$ is a homogeneous graph, the normalizing constant can be evaluated in closed form. In fact, in this particular case, the family $IW_{Q_G}$ of [20] is a special case of our covariance Wishart distributions. Fourth, we also show that this family of priors satisfies an analogue of the strong directed hyper Markov property, when $G$ is homogeneous, and proceed to compute moments and Laplace transforms. We then introduce a new concept called ‘generalized posterior linearity’ for curved exponential families. We proceed to demonstrate that a subclass of our priors satisfies this notion of posterior linearity - thereby drawing parallels with the Diaconis-Ylvisaker priors for natural exponential families. Fifth 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 and 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 consequently study the

---

* Bala Rajaratnam was supported in part by NSF grant DMS 0505303
† Kshitij Khare was supported in part by the B.C. and E.J. Eaves Stanford graduate fellowship.
properties of this family of distributions. 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. 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 zeroes in the inverse covariance matrix, and correspond to natural exponential families. In their pioneering work, Dawid and Lauritzen [8] 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 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 et al. [25].

Consider an undirected\(^1\) graph \(G\) with a finite set of vertices \(V\) (of size \(p\)) and a finite set \(E\) of edges between these vertices, i.e., \(G = (V, E)\). The Gaussian covariance graph model corresponding to the graph \(G\) is the collection of \(p\)-variate Gaussian distributions with covariance matrix \(\Sigma\), such that \(\Sigma_{ij} = 0\) whenever \((i, j) \notin E\). This class of models was first formally introduced by Cox and Wermuth [4, 5]. 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 et al. [29], the authors derive asymptotically efficient approximations to the maximum likelihood estimate in covariance graph models for exponential families. Chaudhuri et al. [3] 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. [2], Grzebyk et al. [15], Mao et al. [21] among 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

---

\(^1\)We shall use dotted edges for our graphs keeping in line with notation in the literature, bi-directed edges have also been used for representing covariance graphs.
been an open problem since the introduction of these models by Cox and Wermuth [4, 5] more than 15 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 $\Sigma$ translate into complicated restrictions on the corresponding entries of the natural parameter, $\Omega = \Sigma^{-1}$. Hence, the sparseness in $\Sigma$ does not translate into sparseness in $\Sigma^{-1}$. No general theory 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, is available.

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, i.e., 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 classes of Wishart distributions developed by Letac and Massam [20], and used later on in the concentration graph setting by Rajaratnam et al. [25], seem promising as priors for this situation. We however establish that the posterior distribution fails to belong to the same class, and computing the posterior mean or mode either in closed form or by sampling from the posterior distribution is intractable.

A principal objective of this paper is to develop a framework for Bayesian inference for Gaussian covariance graph models. We proceed to define a rich class of Wishart distributions on the space of positive definite matrices with fixed zeroes, that correspond to a decomposable graph $G$. Our approach yields a flexible class of conjugate priors with multiple shape parameters for Gaussian covariance graph models. 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 are 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 $n \geq p$. We also show that our covariance Wishart distributions are in general very different from the Letac-Massam priors. However, when the underlying graph $G$ is homogeneous, the Letac-Massam $IW_{Q_G}$ priors are a special case of our distributions. We also show that when the underlying graph $G$ is homogeneous, not only can one explicitly evaluate the normalizing constant, quantities like the posterior mean of the covariance matrix are also available in closed form. In this scenario, we show that our class of priors satisfies the strong directed “covariance” hyper Markov property (parallel to the one introduced in Dawid and Lauritzen [8] for concentration graph models). We noted above that for concentration graph models or the traditional Gaussian graphical models, a rich theory has been established by Dawid and Lauritzen [8], 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 analogue of these two above mentioned papers in the covariance graph model setting, with parallel results all of which are contained in one 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 a recent unpublished manuscript [27] are a narrow special case of our flexible covariance Wishart distributions\(^2\). 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. In fact, it turns out that their algorithm is not a Gibbs sampler in the true sense. Since the authors do not undertake a theoretical investigation of the convergence properties of this algorithm, it is not clear if it does indeed converge to the desired distribution. On the other hand, our algorithm is indeed a Gibbs sampler and we proceed to formally prove that our algorithm converges to the desired distribution. Moreover, contrary to what is stated in [27], we show that it is possible to explicitly evaluate the normalizing constant for a non-trivial subclass of covariance graph models. This property allows for Bayesian model selection using closed form expressions. All the remaining sections of our paper are vastly different from [27], as we undertake a rigorous theoretical analysis of our conjugate Wishart distributions for covariance graph models, whereas they give a treatment of latent variables and mixed graph models in a machine learning context.

This paper is structured as follows. Section 2 introduces required preliminaries and notation. In Section 3, the class of covariance Wishart distributions are formally constructed. Conjugacy to the class of covariance graph models and sufficient conditions for integrability are established. Comparison with the Letac-Massam priors, which are not in general conjugate in the covariance graph setting, are 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 subsequent sections, we restrict ourselves to the case when \(G\) is a homogeneous graph. In Section 5, we show that our family of priors satisfies a version of the strong directed hyper Markov property. In Section 6, we compute moments and Laplace transforms for our class of distributions, and demonstrate that a subclass of our priors satisfies a generalized notion of posterior linearity - thereby drawing parallels to the DY approach for natural exponential families. 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 various 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.

2.1. Modified Cholesky decomposition. If \(\Sigma\) is a positive definite matrix, then there exists a unique decomposition

\[
\Sigma = LDL^T,
\]

\(^2\)In a similar spirit to the way in which the HIW prior of Dawid and Lauritzen [8] is a special case of the generalized family of Wishart distributions proposed by Letac and Massam [20] for the concentration graph setting.
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 $\Sigma$ is referred to as the modified Cholesky decomposition of $\Sigma$ (see [24]). 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

$$A = \cup_{r=2}^{m} \{ \tau : \tau \in \{1, 2, \cdots, m\}^r, \tau_i < \tau_{i-1} \ \forall \ 2 \leq i \leq r \}$$

and

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

where $\dim(\tau)$ denotes the length of the vector $\tau$. Then $L^{-1} = N$, where

$$N_{ij} = \begin{cases} 
0 & \text{if } i < j \\
1 & \text{if } i = j \\
\sum_{\tau \in A, \tau_1 = i, \tau_{\dim(\tau)} = j} (-1)^{\dim(\tau)-1} L_{\tau} & \text{if } i > j.
\end{cases}$$

The proof is provided in the appendix.

An undirected graph $G$ is a pair $(V,E)$, where $V$ is a permutation\(^3\) of the set $\{1, 2, \cdots, m\}$ 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 $(v, u) \in E$, i.e., the edges are undirected. Though 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\(^4\) modulo the edge structure. We describe below two classes of graphs which play a central role in this paper.

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 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 $(C_1, C_2, \cdots, C_k)$ 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

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

and the minimal separators of the graph are given by

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

\(^3\) The ordering in $V$ is emphasized here as the elements of $V$ will later correspond to rows or columns of matrices.

\(^4\) This has been done for notational convenience, as will be seen later.
Let \( R_j = C_j \setminus H_{j-1} \) for \( j = 2, 3, \ldots, k \).

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

Now, let \( \Sigma \) be an arbitrary positive definite matrix with zero restrictions according to \( G = (V, E) \) \(^5\), i.e., \( \Sigma_{ij} = 0 \) whenever \( (i, j) \notin E \). It is known that if \( G \) is decomposable, there exists an ordering of the vertices such that if \( \Sigma = LDL^T \) is the modified Cholesky decomposition corresponding to this ordering, then for \( i > j \),

\[
L_{ij} = 0 \quad \text{whenever} \quad (i, j) \notin E.
\]

Although the ordering is not unique in general, the existence of such an ordering characterizes decomposable graphs (see [23]). 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. Lauritzen [19] shows that this ordering provides a ‘perfect vertex elimination scheme’ for \( G \), and in [23] the authors show that such an ordering satisfies (2.2). Two related articles are [26, 28].

2.3. The spaces \( P_G, Q_G \) and \( L_G \). An \( m \)-dimensional Gaussian covariance graph model\(^6\) 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 \( (i, j) \notin E \), the \( i \)-th and \( j \)-th components of the multivariate random vector are marginally independent. Without loss of generality, we can assume that these models have mean zero and are characterized by the parameter set \( P_G \) of positive definite covariance matrices \( \Sigma \) such that \( \Sigma_{ij} = 0 \) whenever the edge \((i,j)\) is not in \( E \). Following the notation in [20, 25] for \( G \) decomposable, we define \( Q_G \) to be the space on which the free elements of the precision matrices (or inverse covariance matrices) \( \Omega \) live.

More formally let \( M \) denote the set of symmetric matrices of order \( m \), \( M^+ \subset M \) the cone of positive definite matrices (abbreviated \( > 0 \)), \( I_G \) the linear space of symmetric incomplete matrices \( x \) with missing entries \( x_{ij}, (i, j) \notin E \) and \( \kappa : M \rightarrow I_G \) 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 \( \Omega = \kappa(\Sigma^{-1}), \Sigma \in P_G \). Indeed it is easy to verify that the entries \( \Omega_{ij}, (i, j) \notin E \) are such that

\[
\Omega_{ij} = \Omega_{i, V \setminus \{i, j\}} \Omega_{V \setminus \{i, j\}V \setminus \{i, j\}}^{-1} \Omega_{V \setminus \{i, j\}, j},
\]

and are therefore not free parameters of the precision matrix for Gaussian covariance graph models. We are therefore led to consider the two cones

\[
P_G = \{ y \in M^+_m | y_{ij} = 0, (i, j) \notin E \}
\]

\[
Q_G = \{ x \in I_G | x_{Ci} > 0, i = 1, \ldots, k \}.
\]

\(^5\)It is emphasized here that the ordering of the vertices reflected in \( V \) plays a crucial role in the definitions and results that follow.

\(^6\)A brief overview of the literature in this area is provided in the introduction.
where $P_G \subset Z_G$ and $Q_G \subset I_G$, where $Z_G$ denotes the linear space of symmetric matrices with zero entries $y_{ij}, (i, j) \notin E$.

Gröne et al. [14] proved the following:

**Proposition 2.** When $G$ is decomposable, for any $x$ in $Q_G$ there exists a unique $\hat{x}$ in $M_m^+$ such that $\forall (i, j) \in E$ we have $x_{ij} = \hat{x}_{ij}$ and such that $\hat{x}^{-1}$ is in $P_G$.

This defines an isomorphism $^7$ between $P_G$ and $Q_G$:

$$\varphi : y = (\hat{x})^{-1} \in P_G \mapsto x = \varphi(y) = \kappa(y^{-1}) \in Q_G,$$

where $\kappa$ denotes the projection of $M$ into $I_G$. The matrix $\hat{x}$ is said to be the completion of $x$.

We now introduce new spaces that we shall need in our subsequent analysis $^8$. Let $L_G$ denote the space of all lower triangular matrices with diagonal entries equal to 1, such that the entries in the lower triangle have zero restrictions corresponding to $G$, i.e.,

$$L_G = \{L : L_{ij} = 0 \text{ whenever } i < j, \text{ or } (i, j) \notin E, \text{ and } L_{ii} = 1, \forall 1 \leq i, j \leq m\}.$$

Define $\Theta_G$ (the modified Cholesky space) by

$$\Theta_G = \{\theta = (L, D) : L \in L_G, D \text{ diagonal with } D_{ii} > 0 \\forall 1 \leq i \leq m\}.$$

Denote the mapping $\psi : \Theta_G \to M_m^+$ as defined by

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

The above mapping $\psi$ 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 $\bullet - \bullet - 3 - 4$, denoted by $A_4$, as an induced subgraph. Homogeneous graphs have an equivalent representation in terms of directed rooted 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

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

$^7$Furthermore, it also defines a diffeomorphism.

$^8$These spaces are not defined in [20, 25].
Denote by $R$ the equivalence relation on $V$ defined by

$$iRj \iff i \rightarrow j \text{ and } j \rightarrow i.$$  

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

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 proved 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 ($w(\tilde{i}) = |\tilde{i}|$), such that no vertex has exactly one child. Also when $iRj$, we say $i$ 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.

![Hasse diagram](image)

**Figure 1.** (a) An example of a homogeneous graph with 7 vertices (b) The corresponding Hasse diagram

We prove the following proposition for homogeneous graphs, which will play an important role in our analysis.

**Proposition 3.** If $G$ is a homogeneous graph, there exists an ordering of the vertices, such that for this ordering

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

The proof of this proposition can be found in the appendix.

We now describe a procedure for ordering the vertices, under which Proposition 3 holds. Given a homogeneous graph $G$, we first construct the Hasse diagram for $G$. The vertices are labelled in descending order starting from the root of the tree. If the equivalence class at any node has more than one element, they are labelled in any order. We start from the left at each level in the tree and exhaust all nodes in that level before going down to the next level. The process is terminated after reaching the rightmost leaf of the tree. Hereafter, we shall refer to this ordering scheme as the **Hasse perfect vertex elimination scheme**. For example, if we apply this ordering procedure for the graph in Figure 1, then the resulting labels are $\{a, b, c, d, e, f, g\} \rightarrow \{4, 5, 1, 3, 7, 6, 2\}$. 
2.5. Vertex ordering. Let $G = (V, E)$ be an undirected decomposable graph with vertex set $V = \{1, 2, \cdots, m\}$ 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 $(\sigma^{-1}(u), \sigma^{-1}(v)) \in E$. Let $S_D \subset S_V$ denote the subset of permutations $\sigma$ of $V$, such that for any $\Sigma \in M_m^+$ with $\Sigma = LDL^T$, $L \in L_{G_\sigma} \iff \Sigma \in P_{G_\sigma}$. Hence for every $\sigma \in S_D$, the mapping $\psi_\sigma : \Theta_{G_\sigma} \rightarrow M_m^+$ defined in (2.7), is a bijection from $\Theta_{G_\sigma}$ to $P_{G_\sigma}$. In particular, the ordering corresponding to any perfect vertex elimination scheme lies in $S_D$ (see Section 2.2). If $G$ is homogeneous, let $S_H \subset S_D$ denote the subset of permutations $\sigma$ of $V$, such that $L \in L_{G_\sigma} \iff L^{-1} \in L_{G_\sigma}$. 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 $S_H \subset S_D \subset S_V$.

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

\[
G = \{N_m(0, \Sigma) : \Sigma \in P_G\} \\
\cong \{N_m(0, LDL^T) : (L, D) \in \Theta_G\}.
\]

Consider the class of measures on $\Theta_G$ with density (w.r.t. $\prod_{i>j, (i,j) \in E} dL_{ij} \prod_{i=1}^m dD_{ii}$)

\[
\tilde{\pi}_{U,\alpha}(L, D) = e^{-\left(\text{tr}(LDDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii}\right)}, \quad \theta = (L, D) \in \Theta_G.
\]

(3.1)

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

- $N(i) := \{j : (i,j) \in E\}$
- $N^>(i) := \{j : (i,j) \in E, i > j\}$
- $U^{-i} := ((U_{kl}))_{k,l \in N^>(i)}$
- $U^{\leq i} := ((U_{kl}))_{k,l \in N^<\cup(i)}$
- $U^{\prec i} := (U_{kl})_{k \in N^<\cup(i)}$

Let

\[
z_G(U, \alpha) := \int e^{-\left(\text{tr}(LDDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii}\right)} dLdD.
\]

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

**Theorem 1.** Let $dL := \prod_{i>j, (i,j) \in E} dL_{ij}$ and $dD := \prod_{i=1}^m dD_{ii}$. Then, \[
\int_{\Theta_G} e^{-\left(\text{tr}(LDDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii}\right)} dLdD < \infty
\]

if \[
\alpha_i > |N^>(i)| + 2 \quad \forall \; i = 1, 2, \cdots, m.
\]
As the proof of this proposition is rather long and technical, it is given in the Appendix. The normalizing constant $z_G(U, \alpha)$ is not available in closed form in general. Let us consider a simple example to illustrate the difficulty in computing the normalizing constant explicitly.

Let $G = A_4$, i.e., the path on 4 vertices or $\bullet - \bullet - \bullet - \bullet$. Note that this is a decomposable (but not homogeneous) graph. The restrictions on $L$ are $L_{31} = L_{41} = L_{42} = 0$. Let $\tilde{U} \in P_G$, and $\alpha = (4, 4, 4, 4)$. Then after integrating out the elements $D_{ij}$, $1 \leq i \leq 4$ (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

$$
\int_{k^3} \frac{1}{(U_{22} + 2U_{12}x_1 + U_{11}x_1^2)(U_{11}x_1^2 + U_{22}x_2^2 + 2U_{12}x_1x_2 + 2U_{13}x_1x_3 + 2U_{23}x_2x_3)} dx.
$$

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 (see Section 3.3). Moreover, in this case the normalizing constant can be computed in closed form. We denote by $\pi_{U, \alpha}$, the normalized version of $\tilde{\pi}_{U, \alpha}$ whenever $z_G(U, \alpha) < \infty$. The following lemma shows that the family $\pi_{U, \alpha}$ 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 $Y_1, Y_2, \cdots, Y_n$ be an i.i.d. sample from $N_m(0, LDL^T)$, where $(L, D) \in \Theta_G$. Let $S = \frac{1}{n} \sum_{i=1}^{n} Y_iY_i^T$ 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 $\pi_{\tilde{U}, \tilde{\alpha}}$, where $\tilde{U} = nS + U$ and $\tilde{\alpha} = (n + \alpha_1, n + \alpha_2, \cdots, n + \alpha_m)$.

**Proof.** The likelihood of the data is given by

$$
f(Y_1, Y_2, \cdots, Y_n \mid L, D) = \frac{1}{(\sqrt{2\pi})^{mn}} e^{-\frac{1}{2} \left( \text{tr}((LDL^T)^{-1}(nS)) + n \log |D| \right)}.
$$

Using $\pi_{U, \alpha}$ as a prior for $(L, D)$, the posterior distribution of $(L, D)$ given the data $(Y_1, Y_2, \cdots, Y_n)$ is

$$
\pi_{U, \alpha}(L, D \mid Y_1, Y_2, \cdots, Y_n) \propto e^{-\frac{1}{2} \left( \text{tr}((LDL^T)^{-1}(nS + U)) + \sum_{i=1}^{m} (n + \alpha_i) \log D_{ii} \right)}, \quad \theta \in \Theta_G.
$$

Hence the posterior distribution belongs to the same family as the prior, i.e.,

$$
\pi_{U, \alpha}(\cdot \mid Y_1, Y_2, \cdots, Y_n) = \pi_{\tilde{U}, \tilde{\alpha}}(\cdot),
$$

where $\tilde{U} = nS + U$ and $\tilde{\alpha} = (n + \alpha_1, n + \alpha_2, \cdots, n + \alpha_m)$. \hfill \Box

**Remark.** If we assume that the observations do not have mean zero, i.e., $Y_1, Y_2, \cdots, Y_n$ are i.i.d. $N(\mu, \Sigma)$, with $\mu \in \mathbb{R}^m$, $\Sigma \in P_G$, then

$$
\tilde{S} := \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})^T.
$$
is the minimal sufficient statistic for $\Sigma$. Here, $n\tilde{S}$ has a Wishart distribution with parameter $\Sigma$ 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

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

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

### 3.1. Induced prior on $P_G$ and $Q_G$.

The prior $\pi_{U,\alpha}$ on $\Theta_G$ (the modified Cholesky space) induces a prior on $P_G$ (the covariance matrix space) and $Q_G$. Recall from Section 2.3 that $P_G$ is the space of positive definite matrices with zero restrictions according to $G$, and $Q_G$ is the space of incomplete matrices isomorphic to $P_G$. 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

$$\psi : \Theta_G \to M^+_m$$

defined by

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

is a bijection from $\Theta_G$ to $P_G$. The lemma below provides the required Jacobians for deriving the induced priors on $P_G$ and $Q_G$. We prove the first part, and the proof of the second part can be found in [26]. The reader is referred to Section 2.2 for notation on decomposable graphs.

Note that if $x$ is a matrix, $|x|$ 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 $\Theta_G$ 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 $\Sigma$, 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

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

**Proof (Part 1).** Let $\Sigma \in P_G$, and $(L, D) = (\psi|_{P_G})^{-1}(\Sigma)$. Note that for $(i, j) \in E, i > j$,

$$\Sigma_{ij} = (LDL^T)_{ij} = \sum_{k=1}^{m} L_{ik} L_{jk} D_{kk} = \sum_{k=1}^{j} L_{ik} L_{jk} D_{kk},$$

since $L$ is lower triangular. Now note that,

$$\frac{\partial}{\partial L_{ij}} (LDL^T)_{ij} = D_{jj}.$$
For $1 \leq i \leq m$, 
\[ \frac{\partial}{\partial D_{ii}} (LDLT)_{ii} = 1. \]

Arrange the entries of $\theta = (L, D) \in \Theta_G$ as $D_{11}, \{L_{2k} : (2, k) \in E, 1 \leq k < 2\}, D_{22}, \{L_{3k} : (3, k) \in E, 1 \leq k < 3\}, \ldots, D_{m-1,m-1}, \{L_{mk} : (m, k) \in E, 1 \leq k < m\}, D_{mm}$, and the entries of $\Sigma \in \Sigma_G$ as $\Sigma_{11}, \{\Sigma_{2k} : (2, k) \in E, 1 \leq k < 2\}, \Sigma_{22}, \{\Sigma_{3k} : (3, k) \in E, 1 \leq k < 3\}, \ldots, \Sigma_{m-1,m-1}, \{\Sigma_{mk} : (m, k) \in E, 1 \leq k < m\}, \Sigma_{mm}$. Since, $\Sigma_{ij}$ depends on $\{L_{jk} : (j, k) \in E, 1 \leq k < j\}, \{L_{ik} : (i, k) \in E, 1 \leq k < j\}$ and $\{D_{kk}, 1 \leq k \leq j\}$, it is clear that $\Sigma_{ij}$ is functionally independent of elements of $\Theta_G$ that follow it in the arrangement described above. Hence the gradient matrix of the transformation $\psi$ (with this arrangement) is a lower triangular matrix, and the Jacobian of the transformation $\psi$ is therefore given as

\[ \prod_{i=1}^{m} \prod_{(i,j) \in E, i > j} \frac{1}{D_{jj}(\Sigma)}. \]

It follows from the expression above that the Jacobian of the transformation $\psi$ is

\[ \prod_{j=1}^{m} D_{jj}(\Sigma)^{-n_j}. \]

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

\[ \bar{\pi}_{U,\alpha}^G(\Sigma) \propto e^{-(tr(\Sigma^{-1}U) + \sum_{i=1}^{m} (2n_i + \alpha_i) \log D_{ii}(\Sigma))/2}, \quad \Sigma \in \Sigma_G. \tag{3.2} \]

We first note that the traditional Inverse-Wishart distribution (see [22]) with parameters $U$ and $n$ is a special case of (3.2) when $G$ is the complete graph, and $\alpha_i = n - 2m + 2i, \forall 1 \leq i \leq m$. We also note that the $G$-inverse Wishart priors introduced in [27] have a one-dimensional shape parameter $\delta$, and are a very special case of our richer class $\bar{\pi}_{U,\alpha}^G$. The single shape parameter $\delta$ is given by the relationship $\alpha_i + 2n_i = \delta + 2m, \ 1 \leq i \leq m$.  

We now proceed to derive the induced prior on $Q_G$. Let $x = \varphi(\Sigma) = \kappa(\Sigma^{-1})$ denote the image of $\Sigma$ in $Q_G$. Using the second part of Lemma 2, the induced prior corresponding to $\bar{\pi}_{U,\alpha}$ on $Q_G$ is given by

\[ \bar{\pi}_{U,\alpha}^G(x) \propto e^{-\frac{(tr(\delta U) + \sum_{i=1}^{m} (2n_i + \alpha_i) \log D_{ii}(\delta^{-1}))}{2}} \times \prod_{S \in \mathcal{S}} \frac{|x_S|^{(|S|+1)\nu(S)}}{\prod_{C \in \mathcal{G}} |x_C|^{|C|+1}}, \quad x \in Q_G. \]

\[ \text{There is an interesting parallel here that becomes apparent from our derivations above. In the concentration graph setting, the single shape parameter hyper inverse Wishart (HIW) prior of Dawid and Lauritzen [8] is a special case of the multiple shape parameter class of priors introduced by Letac and Massam [20], in the sense $\alpha_i = -\frac{1}{2}(\delta + c_i - 1)$ (see [25] for notation). In a similar spirit, we discover that the single shape parameter class of priors in [27] is a special case of the multiple shape parameter class of priors $\bar{\pi}_{U,\alpha}$ introduced in this paper, in the sense $\alpha_i = \delta - 2n_i + 2m$.} \]
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 [25]. These priors lead to corresponding classes of inverse Wishart distributions $IW_{P_G}$ (on $Q_G$) and $IW_{Q_G}$ (on $P_G$), i.e., $U \sim IW_{P_G}$ whenever $U^{-1} \sim W_{P_G}$ and $V \sim IW_{Q_G}$ 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, i.e., 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 in the covariance graph setting, since the parameter of interest $\Sigma$ 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

$$IW_{Q_G}^{U,\alpha,\beta}(\Sigma) \propto e^{-\frac{1}{2}tr(\Sigma^{-1}U)} \frac{\prod_{C \in C} |\Sigma_C^{-1}|^{\alpha(C)} \prod_{S \in S} |\Sigma_S^{-1}|^{\beta(S)\nu(S)}}{\prod_{S \in S} |\Sigma_S^{-1}|^{(\beta(S)\nu(S) + \frac{n_S(S)}{2})}},$$

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

$$\pi_{U,\alpha,\beta}^{IW}(\Sigma | Y_1, Y_2, \ldots, Y_n) \propto e^{-\frac{1}{2}tr(\Sigma^{-1}(U + nS))} \frac{\prod_{C \in C} |\Sigma_C^{-1}|^{\alpha(C) + \frac{n}{2}}}{\prod_{S \in S} |\Sigma_S^{-1}|^{(\beta(S)\nu(S) + \frac{n_S(S)}{2})}}.$$

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 $IW_{Q_G}^{U,\alpha,\beta}$ 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-path $A_4$. One can easily verify, that the terms $e^{-\frac{1}{2}tr(\Sigma^{-1}U)}$ 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 $\Sigma$, then for this particular graph with $C_1 = \{1, 2\}, C_2 = \{2, 3\}, C_3 = \{3, 4\}$ and $S_1 = \{2\}, S_2 = \{3\}, S_3 = \{4\}$, the expression that is not in the exponential term for the $IW_{Q_G}$ density is given as follows:

$$\prod_{i=1}^{D_1} |\Sigma_{C_i}^{-1}|^{\alpha_i} \prod_{i=1}^{D_3} |\Sigma_{S_i}^{-1}|^{\beta_i} = \left(\frac{1}{D_{11}}\right)^{\alpha_1} \left(\frac{1}{D_{22}} + \frac{L_{23}^2}{D_{33}} + \frac{L_{23}^2 L_{43}^2}{D_{44}}\right)^{\alpha_1 - \beta_1} \times \left(\frac{1}{D_{22}}\right)^{\alpha_2} \left(\frac{1}{D_{33}} + \frac{L_{34}^2}{D_{44}}\right)^{\alpha_2 - \beta_2} \left(\frac{1}{D_{33}D_{44}}\right)^{\alpha_3}.$$

This expression is clearly different from the term other than the exponent $e^{-\frac{1}{2}tr(\Sigma^{-1}U)}$ in $\pi_{U,\alpha,\beta}^{IW}$, which is a product of different powers of $D_{ii}$, $i = 1, 2, 3, 4$. Yet another way to see that our
class of priors is not the same in general as the $\text{IW}_{QG}$ is to note that the number of shape parameters in $\text{IW}_{QG}$ is equal to the sum of the number of cliques plus one, whereas the number of shape parameters in $\pi_{PG}$ is equal to the number of vertices in the graph.

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

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

Hence when $G$ is homogeneous, the class $\text{IW}_{QG}$ is contained in the class $\pi_{PG}$ because in $\text{IW}_{QG}$, the exponent of $D_{ii}$ and $D_{jj}$ is the same if $i \sim j$, i.e., the shape parameter is shared for vertices in the same equivalence class as defined by the relation $R$. Hence the containment is strict, notwithstanding the fact that $U$ need not be in $P_G$ for our class $\pi_{PG}$. In the restrictive case first when $G$ is homogeneous, second when no two vertices belong to the same equivalence class as defined by the relation $R$, and third when $U \in P_G$, the two classes of distributions $\pi_{PG}$ and $\text{IW}_{QG}$ have the same functional form.

3.3. Normalizing constant for homogeneous graphs. Suppose $G = (V, E)$ is a homogeneous graph, and the vertices are ordered such that $V \in S_H$. In this case the normalizing constant can be computed explicitly, thus enabling closed form Bayesian model selection for this class of graphs. Recall that

$$z_G(U, \alpha) := \int e^{-\frac{1}{2} \text{tr}((LDLT)^{-1}U) + \sum_{i=1}^{m} \alpha_i \log D_{ii}} dLdD.$$ 

We provide below necessary and sufficient conditions for existence of the normalizing constant, and an explicit expression for it in such cases.

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

$$z_G(U, \alpha) = \prod_{i=1}^{m} \Gamma \left( \frac{\alpha_i}{2} - \frac{|\mathcal{N}_\prec(i)|}{2} - 1 \right) \left( \frac{2}{\sqrt{\pi}} \right)^{-1} \left( \frac{\alpha_i}{2} - \frac{|\mathcal{N}_\prec(i)|}{2} - \frac{3}{2} \right)^{-1}.$$

The proof is provided in the Appendix. Furthermore, it can be shown that the result is maximal in the sense that in general, if the graph $G$ is not homogeneous, the normalizing constant cannot be evaluated in closed form (see [18] for more details).

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 are standard distributions 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,

$$L_u \ := \ u^{th} \text{ row of } L, \ u = 1, 2, \cdots, m,$$

$$L_v \ := \ v^{th} \text{ column of } L, \ v = 1, 2, \cdots, m,$$

$$L^G_{uv} \ := \ (L_{uv})_{u>v, (u,v) \in E}, \ v = 1, 2, \cdots, m - 1.$$

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

$$(4.1) \quad \Theta_G = \left\{ (L^G_1, L^G_2, L^G_3, \ldots, L^G_{m-1}, D) : L_{ij} \in \mathbb{R}, \ \forall \ 1 \leq j < i \leq m, (i, j) \in E, D_{ii} > 0, \ \forall \ 1 \leq i \leq m \right\}.$$

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

**Theorem 3.** Using the notation above, the conditional distributions of each component of $\theta$ (as in (4.1)) given the other components and the data $Y_1, Y_2, \ldots, Y_n$ are as follows.

1. $L^G_v \mid (L \setminus L^G_v, D, Y_1, Y_2, \ldots, Y_n) \sim \mathcal{N}(\mu^{v,G}, M^{v,G}) \ \forall \ v = 1, 2, \cdots, m - 1,$

where

$$\mu^{v,G}_u \ := \ \mu^{v}_u + \sum_{w>v: (u,w) \in E} \sum_{w>v, (w,v) \notin E \text{ or } w<v, L_{uw}=0} M^{v,G}_{uw} L^{-1}(U^T)^{-1} \left( L^{-1} U (L^T)^{-1} \right)^{-1}_{uv} (L^{v,G})^{-1}_{uw} \mu^{v}_w \ \forall u > v, (u, v) \in E,$$

$$\mu^{v}_u \ := \ \frac{(L^{-1} \tilde{U})_{vu}}{(L^{-1} \tilde{U})_{uu}} \ \forall u \text{ such that } L^{-1}_{uu} = 0,$$

$$(M^{v,G})^{-1}_{uv} \ := \ L^{-1} \tilde{U} (L^T)^{-1} \left( L^{-1} U (L^T)^{-1} \right)^{-1}_{uv} \ \forall u, u' > v, (u, v), (u', v) \in E.$$

2. $D_{ii} \mid L, Y_1, Y_2, \ldots, Y_n \sim IG \left( \frac{\alpha_{ii}}{2} - 1, \frac{(L^{-1} \tilde{U} (L^T)^{-1})_{ii}}{2} \right)$

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

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

Deriving the required conditional distributions in Theorem 3 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,

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

Proof. Let $u > v$. If $i < u$ or $j > v$, $L_{ij}^{-1}$ does not depend on $L_{uv}$ from Proposition 1, hence,

$$\frac{\partial}{\partial L_{uv}} L_{ij}^{-1} = 0 = -L_{iu}^{-1}L_{vj}^{-1},$$

with the second equality above being a consequence of $L^{-1}$ being lower triangular. If $i \geq u > v \geq j$, then using the identity $L^{-1}LL^{-1} = L^{-1}$,

$$L_{ij}^{-1} = \sum_{i'=1}^{m} \sum_{j'=1}^{m} L_{i'i'}^{-1}L_{ij'}^{-1}L_{j'j}^{-1} = 2L_{ij}^{-1} + L_{iu}^{-1}L_{uv}L_{vj}^{-1} + C,$$

where $C$ is functionally independent of $L_{uv}$. Since both $L_{iu}^{-1}$ and $L_{vj}^{-1}$ are functionally independent of $L_{uv}$, we get that,

$$\frac{\partial L_{ij}^{-1}}{\partial L_{uv}} = 2\frac{\partial L_{ij}^{-1}}{\partial L_{uv}} + L_{iu}^{-1}L_{vj}^{-1},$$

and hence,

$$\frac{\partial L_{ij}^{-1}}{\partial L_{uv}} = -L_{iu}^{-1}L_{vj}^{-1}.$$

□

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

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

$$\frac{\partial^2}{\partial L_{uv}\partial L_{u'v}} \text{tr}((LDLT)_{v}^{-1}\tilde{U}) = 2 \left(L^{-1}\tilde{U}(LT)^{-1}\right)_{vv} (LDLT)_{u'u},$$

which is functionally independent of the elements of $L_{G}v$. 
WISHART DISTRIBUTIONS

Proof. First note that,
\[
\frac{\partial}{\partial L_{uv}} \text{tr}((LDLT)^{-1} \bar{U})
= \frac{\partial}{\partial L_{uv}} \left( \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{L_{ku}^{-1} L_{vi}^{-1} L_{kj}^{-1}}{D_{kk}} \tilde{U}_{ij} \right)
\]
\[
= -2 \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{L_{ku}^{-1} L_{vi}^{-1} L_{kj}^{-1}}{D_{kk}} \tilde{U}_{ij}
\]
\[
\text{(by Lemma 3)}
\]
\[
= -2 \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{L_{ku}^{-1} L_{vi}^{-1} L_{kj}^{-1}}{D_{kk}} \tilde{U}_{ij}.
\]

Note that \(L^{-1}\) is a lower triangular matrix. Hence,
\[
\frac{\partial^2}{\partial L_{uv} \partial L_{u'v'}} \text{tr}((LDLT)^{-1} \bar{U})
= -2 \frac{\partial}{\partial L_{u'v'}} \left( \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{L_{ku}^{-1} L_{vi}^{-1} L_{kj}^{-1}}{D_{kk}} \tilde{U}_{ij} \right)
\]
\[
= 2 \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{L_{ku}^{-1} L_{vi}^{-1} L_{kj}^{-1}}{D_{kk}} \tilde{U}_{ij}
\]
\[
\text{(by Lemma 3)}
\]
\[
= 2 \left( \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{L_{vi}^{-1} \tilde{U}_{ij} L_{kj}^{-1}}{D_{kk}} \right) \left( \sum_{k=1}^{m} \frac{L_{ku}^{-1} L_{kv}^{-1}}{D_{kk}} \right)
\]
\[
= 2 \left( L^{-1} \tilde{U} (LT)^{-1} \right)_{uv} \left( LDLT \right)_{u'v'}^{-1}.
\]

Note that by Proposition 1, \(L_{vi}^{-1}\) is functionally independent of \(L_{vi}^{G}\) for all \(1 \leq i \leq m\), and \(L_{ku}^{-1}\) is functionally independent of \(L_{kv}^{G}\) for all \(1 \leq k \leq m\) and \(u > v\). We thereby conclude that
\[
2 \left( L^{-1} \tilde{U} (LT)^{-1} \right)_{uv} \left( LDLT \right)_{u'v'}^{-1}
\]
is independent of \(L_{vi}^{G}\). \(\square\)

Proof of Theorem 3. An immediate consequence of Lemma 4 is that we can write \(\text{tr}((LDLT)^{-1} \bar{U})\) as follows:
\[
\text{tr}((LDLT)^{-1} \bar{U})
= \sum_{u>v, (u,v) \in E} \sum_{u'v, (u',v) \in E} \left( L^{-1} \tilde{U} (LT)^{-1} \right)_{uv} \left( LDLT \right)_{u'v'}^{-1} (L_{uv} - b_u)(L_{u'v} - b_{u'}) + C,
\]
where \(b = (b_u)_{u>v, (u,v) \in E}\) and \(C\) are independent of \(L_{vi}^{G}\). In order to evaluate \((b_u)_{u>v, (u,v) \in E}\), note that the term in \(\frac{\partial}{\partial L_{uv}} \text{tr}((LDLT)^{-1} \bar{U})\) which is independent of \(L_{vi}^{G}\) is given by
\[
(4.2)
-2 \sum_{u'>v, (u',v) \in E} \left( L^{-1} \tilde{U} (LT)^{-1} \right)_{uv} \left( LDLT \right)_{u'v'}^{-1} b_{u'},
\]
for every $u > v$, $(u, v) \in E$. However, from the proof of Lemma 4 we alternatively know that
\[
\frac{\partial}{\partial L_{uv}} \text{tr}((LDL^T)^{-1} \tilde{U}) = -2 \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{L_{ku}^{-1} L_{vi}^{-1} L_{kj}^{-1}}{D_{kk}} \tilde{U}_{ij} \{ L_{ij}^{-1} = 0 \text{ or } L_{ku}^{-1} = 0 \}
\]

Note that by Lemma 3, $L_{ku}^{-1} L_{kj}^{-1}$ is functionally dependent on $L_{ij}^{-1}$ if and only if $L_{ku}^{-1} \neq 0$ and $L_{ij}^{-1} \neq 0$ (as a function of $L$). And hence the term in $\frac{\partial}{\partial L_{uv}} \text{tr}((LDL^T)^{-1} \tilde{U})$ which is independent of $L_{ij}^{-1}$ is given by
\[
-2 \sum_{j:L_{v}^{-1}=0} \left( \sum_{i=1}^{m} L^{-1}_{vi} \tilde{U}_{ij} \right) \left( \sum_{k=1}^{m} \frac{L_{ku}^{-1} L_{kj}^{-1}}{D_{kk}} \right)
\]
\[
= -2 \sum_{j:L_{v}^{-1}=0} \left( L^{-1} \tilde{U}_{v} \right)_{uj} \left( (LDL^T)^{-1} \right)_{uj}
\]
(4.3)
\[
= -2 \sum_{j:L_{v}^{-1}=0} \left( L^{-1} \tilde{U}_{v} \right)_{uj} \left( L^{-1} \tilde{U}(L^T)^{-1} \right)_{uu} \left( (LDL^T)^{-1} \right)_{uj}.
\]

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

\[
A_1 \xi_1 + A_2 \xi_2 = A_1 \eta,
\]
then
\[
\eta = \xi_1 + A_1^{-1} A_2 \xi_2.
\]

If we choose $A, \xi$ and $\eta$ as
\[
A_{uu'} := \left( L^{-1} \tilde{U}(L^T)^{-1} \right)_{vv} \left( (LDL^T)^{-1} \right)_{uu'} \forall u, u' \text{ such that } L_{uu}^{-1}, L_{uu'}^{-1} = 0,
\]
\[
\xi_u := \frac{(L^{-1} \tilde{U})_{uu}}{(L^{-1} \tilde{U}(L^T)^{-1})_{vv}} \forall u \text{ such that } L_{uu}^{-1} = 0,
\]
\[
\eta_u := b_u \forall u > v, (u, v) \in E,
\]
then combining the observations above with (4.2) and (4.3), we obtain
\[
\text{tr}((LDL^T)^{-1} \tilde{U}) = \sum_{u > v, (u, v) \in E} \sum_{u' > v, (u', v) \in E} \left( (L^{-1} \tilde{U}(L^T)^{-1})_{vv} \left( (LDL^T)^{-1} \right)_{uu'} \left( L_{uv} - \mu_u^v G \right) \left( L_{u'v} - \mu_{u'}^v G \right) + C.\right)
\]
As defined earlier,

\[
\mu^v_u = \frac{(L^{-1} U^T)_vu}{(L^{-1} U (LT)^{-1})_{vu}} \quad \forall u \text{ such that } L^{-1}_{vu} = 0,
\]

\[
\mu^v_u \sum_{u' > v, (u',v) \in E} \sum_{w > v, (w,v) \notin E} M^v_{uw'} (L^{-1} U (LT)^{-1})_{vw} (LDLT)^{-1}_{uw} \mu^v_w \quad \forall u > v, (u,v) \in E,
\]

and \( C \) is independent of \( L^G_u \). It follows that under \( \pi_{U,G} \), the conditional distribution of \( L^G_u \) given the other parameters and the data \( Y_1, Y_2, \ldots, Y_n \) is \( N(\mu^v_u, M^v_{uG}) \).

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 \( \pi_{U,G} \). We now formally prove the conditions which are sufficient for establishing convergence.

We now prove that sufficient conditions for convergence of a Gibbs sampling Markov chain to its stationary distribution (see [1, Theorem 6]) are satisfied by the Markov chain corresponding to our block Gibbs sampler. Let \( \phi(x | \mu, \Sigma) \) denote the \( N(\mu, \Sigma) \) density evaluated at \( x \). Let \( f_{IG}(d | \alpha, \lambda) \) denote the \( IG(\alpha, \lambda) \) density evaluated at \( d \). Let us fix \( \psi, d_1, d_2 > 0 \) arbitrarily. Let

\[
\Theta_{\psi,d_1,d_2} := \{ \theta = (L, D) \in \Theta_G : |L_{ij}| \leq \psi, d_1 \leq D_{ii} \leq d_2 \forall i > j, (i,j) \in E \}.
\]

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

**Proposition 4.** \( \exists \delta > 0 \) such that for all \( \theta = (L, D) \in \Theta_{\psi,d_1,d_2}, \)

\[
\phi(L^G_.,v | \mu^v_u, M^v_{uG}) > \delta \forall v = 1, 2, \ldots, m - 1,
\]

\[
f_{IG}(D_{ii} \bar{\alpha}_i/2 - 1, \frac{(L^{-1} U (LT)^{-1})_{ii}}{2}) > \delta \forall i = 1, 2, \ldots, m.
\]
Proof Firstly, by Proposition 1, all entries of $L^{-1}$ are polynomials in the entries of $L$. Since $\Theta_{\psi,d_1,d_2}$ is bounded and closed, there exists $\psi_1 > 0$ such that

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

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

$$(4.4) \quad \left| \left( L^{-1}U \right)_{vu} \right| \leq \psi_2, \quad \left| (LDLT)^{-1}_{uu'} \right| \leq \psi_2, \quad \left| \left( L^{-1}U(L^T)^{-1} \right)_{vv} \right| \leq \psi_2,$$

for every $1 \leq v, u, u' \leq m$. Secondly, since $L^{-1}_{vu} = 1 \forall 1 \leq v \leq m$, and $\tilde{U}$ 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

$$(4.5) \quad \psi_3 \leq \left( L^{-1}U(L^T)^{-1} \right)_{vv}.$$  

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

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

$$
\begin{align*}
\left( L^G_v - \mu^vG \right)^T \left( M^vG \right)^{-1} \left( L^G_v - \mu^vG \right) & = \left( L^G_v \right)^T \left( M^vG \right)^{-1} L^G_v - 2 \left( L^G_v \right)^T \left( M^vG \right)^{-1} \mu^vG + \left( \mu^vG \right)^T \left( M^vG \right)^{-1} \mu^vG. \\
\end{align*}
$$

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

$$
\left( \zeta_1 + \Sigma^{-1}_{11} \Sigma_{12} \zeta_2 \right)^T \Sigma_{11} \left( \zeta_1 + \Sigma^{-1}_{11} \Sigma_{12} \zeta_2 \right) \leq \zeta^T \Sigma \zeta, \quad \forall \zeta \in \mathbb{R}^m.
$$

If we choose $\zeta$ and $\Sigma$ as

$$
\Sigma_{uv} := \left( L^{-1}U(L^T)^{-1} \right)_{vv} (LDLT)^{-1}_{uv} \forall u, u' \text{ such that } L^{-1}_{vu}, L^{-1}_{uv'} = 0,
$$

$$
\zeta_u = \mu^v_u \forall u \text{ such that } L^{-1}_{vu} = 0,
$$

then combining the observation above and the definition of $\mu^vG$, we get that,

$$
\left( \mu^vG \right)^T \left( M^vG \right)^{-1} \mu^vG \leq \sum_{u:L^{-1}_{vu}=0} \sum_{u':L^{-1}_{vu'}=0} \mu^v_u \left( L^{-1}U(L^T)^{-1} \right)_{vv} (LDLT)^{-1}_{uu} \mu^v_{u'}.
$$

From the definitions in Theorem 3 we also have,

$$
\left( \left( M^vG \right)^{-1} \mu^vG \right)_u = \sum_{j:L^{-1}_{vj}=0} \left( L^{-1}U \right)_{vj} (LDLT)^{-1}_{uj} \forall v, (u, v) \in E.
$$

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

$$
(4.6) \quad \left( L^G_v - \mu^vG \right)^T \left( M^vG \right)^{-1} \left( L^G_v - \mu^vG \right) \leq \psi_4,
$$
for every \( v = 1, 2, \cdots, m - 1 \). 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 \( \Theta_{\psi,d_1,d_2} \) is a bounded and closed set, both the maximum and minimum of the function \(|M^{v,G}|\) are attained in \( \Theta_{\psi,d_1,d_2} \). It follows that for \((L, D) \in \Theta_{\psi,d_1,d_2}\), there exist \( 0 < \kappa_1 < \kappa_2 \) such that

\[
\kappa_1 < |M^{v,G}| < \kappa_2 \tag{4.7}
\]

for every \( v = 1, 2, \cdots, m - 1 \). 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

\[
\phi \left( L^G_v | \mu^{v,G}, M^{v,G} \right) > \delta_1 \quad \forall \ v = 1, 2, \cdots, m - 1.
\]

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

\[
\psi_3 \leq \left( L^{-1} \bar{U}(L^T)^{-1} \right)_{ii} \leq \psi_2
\]

for every \( 1 \leq i \leq m \). Hence, there exists \( \delta_2 > 0 \) such that

\[
f_{IG} \left( D_{ii} | \bar{\alpha}_i \right) - 1, \frac{\left( L^{-1} \bar{U}(L^T)^{-1} \right)_{ii}}{2} \right) > \delta_2 \quad \forall \ i = 1, 2, \cdots, m.
\]

Let \( \delta = \min(\delta_1, \delta_2) \). Hence, for \( \theta = (L, D) \in \Theta_{\psi,d_1,d_2}\),

\[
\phi \left( L^G_v | \mu^{v,G}, M^{v,G} \right) > \delta \quad \forall \ v = 1, 2, \cdots, m - 1,
\]

\[
f_{IG} \left( D_{ii} | \bar{\alpha}_i \right) - 1, \frac{\left( L^{-1} \bar{U}(L^T)^{-1} \right)_{ii}}{2} \right) > \delta \quad \forall \ i = 1, 2, \cdots, m.
\]

Recall that \( n_v = |\{u : u > v, (u,v) \in E\}| \). Note that the measures corresponding to \( \mathcal{N} \left( \mu^{v,G}, M^{v,G} \right) \) and \( \mathcal{N} \left( 0, I_{n_v} \right) \) are mutually absolutely continuous, and the corresponding densities w.r.t. Lebesgue measure are positive everywhere on \( \mathbb{R}^{n_v}, \forall \ v = 1, 2, \cdots, m - 1 \). In addition, the measures corresponding to \( IG \left( \frac{\bar{\alpha}_v}{2} - 1, \frac{\left( L^{-1} \bar{U}(L^T)^{-1} \right)_{ii}}{2} \right) \) and the \( IG \left( \frac{\bar{\alpha}_v}{2} - 1, \frac{\left( L^{-1} \bar{U}(L^T)^{-1} \right)_{ii}}{2} \right) \) are mutually absolutely continuous, and the corresponding densities w.r.t. Lebesgue measure are positive everywhere on \( (0, \infty), \forall \ i = 1, 2, \cdots, m \). Also, since \( \Theta_{\psi,d_1,d_2} \) is bounded and closed, \( \prod_{i=1}^{m-1} \phi \left( L^G_i | 0, I_{n_v} \right) \prod_{i=1}^{m} f_{IG} \left( \frac{\bar{\alpha}_v}{2} - 1, 1 \right) \) is bounded on \( \Theta_{\psi,d_1,d_2} \). Combining this with Proposition 4, all required conditions in [1, Theorem 6] are satisfied. Hence, the block Gibbs sampling Markov chain, based on the derived conditional distributions converges to the desired stationary distribution \( \pi_{U,\alpha} \).

We note that in [27, Page 18], the authors introduce a procedure to sample from the \( G \)-inverse Wishart distributions (these are a narrow subclass of our priors \( \pi_{U,\alpha}^{G,G} \)). Essentially, at
every iteration, they cycle through all the rows of $\Sigma$. At the $i^{th}$ step in an iteration, they sample the vector $\Sigma_i^{G} := (\Sigma_{ij})_{j \in N(i)}$ from its conditional distribution (Gaussian) given the other entries of $\Sigma$, and then sample $\gamma_i := \frac{1}{\Sigma_{ii}}$ from its conditional distribution (Inverse-Gamma) given the other entries of $\Sigma$. Since $\Sigma$ is a symmetric matrix, for $(i, j) \in E$, the variable $\Sigma_{ij}$ appears in $\Sigma_i^{G}$ as well as $\Sigma_j^{G}$. Hence, $\left((\Sigma_1^{G}, \gamma_1), (\Sigma_2^{G}, \gamma_2), \ldots, (\Sigma_m^{G}, \gamma_m)\right)$ 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 $\left(L_1^{G}, L_2^{G}, \ldots, L_m^{G}, D\right)$, 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 $\gamma_i \neq D_{ii}$ unless $i = m$.

5. Covariance hyper Markov properties. In this section we study the “covariance” hyper Markov properties of the class of distributions introduced in this paper. The conceptual foundations of hyper Markov properties in the context of concentration graph models were laid in Dawid and Lauritzen [8]. The reader is referred to Letac and Massam [20] for a brief overview. We now prove a property of our class of priors that will be used to establish properties that are analogous to hyper Markov properties in the covariance graph setting. This property is also crucial in evaluating $E[\Sigma]$ when $\Sigma \sim \pi_{U,\alpha}^{PD}$.

Theorem 4. Let $G = (V, E)$ be homogeneous, with vertices ordered according to the Hasse perfect vertex elimination scheme specified in Section 2.4, i.e., $V \in S_H$. If $\Sigma \sim \pi_{U,\alpha}^{PD}$, and $\Sigma = LDL^T$ is its modified Cholesky decomposition, then

$$\left\{\left(D_{ii}, (\Sigma^{-i})^{-1}\Sigma_i^{-1}\right)\right\}_{1 \leq i \leq m}$$

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

$$\left((\Sigma^{-i})^{-1}\Sigma_i^{-1}\right) | D_{ii} \sim \mathcal{N}\left((U^{-i})^{-1}U_i^{-1}, D_{ii}(U^{-i})^{-1}\right), \text{ and}$$

$$D_{ii} \sim IG\left(\frac{\alpha_i}{2} - \frac{|N^{-i}(i)|}{2} - 1, \frac{U_{ii} - (U^{-i})^T(U^{-i})^{-1}U^{-i}}{2}\right) \forall i = 1, 2, \ldots, m.$$

Remark. The above result decomposes $\Sigma$ into mutually independent coordinates. Note that for any $i$ such that $i$ is a leaf of the Hasse tree and $i$ has the minimal label in its equivalence class $i$, $N^{-i}(i) = \phi$. In this case, it is understood that $\Sigma^{-i}$ and $\Sigma_i^{-1}$ are vacuous parameters and that $D_{ii} = \Sigma_{ii}$.

Proof. 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 labelled $m$ clearly lies in the equivalence class of vertices at the root of the corresponding Hasse diagram. Let us remove the vertex labelled $m$ 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 \( 1 \leq i \leq m \) such that \( N^< (i) \neq \phi \), \( \Sigma^\leq i \) can be partitioned as

\[
\Sigma^\leq i = \begin{bmatrix}
\Sigma^\leq i_1 & \Sigma^\leq i_2 \\
(\Sigma^\leq i_1)^T & \Sigma^\leq i_i
\end{bmatrix}.
\]

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

\[
(S^\leq i)^{-1} = \begin{bmatrix}
(S^i)^{-1} + \frac{((S^i)^{-1}\Sigma_i^1)(S^i)^{-1}\Sigma_i^1)^T}{D_{ii}} & -(S^i)^{-1}\Sigma_i^1 \\
-(S^i)^{-1}\Sigma_i^1 & \frac{1}{D_{ii}}
\end{bmatrix}.
\]

Hence,

\[
\text{tr} \left( (S^\leq i)^{-1}U^\leq i \right) = \text{tr} \left( (S^\leq i)^{-1}U^i \right) + \frac{1}{D_{ii}} \left( (S^i)^{-1}\Sigma_i^1 - (U^i)^{-1}U^\leq i \right)^TU^\leq i \left( (S^i)^{-1}\Sigma_i^1 - (U^i)^{-1}U^\leq i \right) + \frac{1}{D_{ii}} U_{ii} \left( U^\leq i - (U^i)^TU^\leq i \right) \left( U^\leq i - (U^i)^TU^\leq i \right)^T U^\leq i \right).
\]

(5.1)

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

\[
\text{tr} \left( (S^\leq i)^{-1}U^\leq i \right) = \sum_{j=1}^{k} \text{tr} \left( (S^\leq i_j)^{-1}U^\leq i_j \right).
\]

Note that \( \Sigma = \sum_{i=1}^{m} \). Using (5.1) recursively, we get,

\[
\text{tr}(\Sigma^{-1}U) = \sum_{i=1}^{m} \frac{1}{D_{ii}} \left( (S^i)^{-1}\Sigma_i^1 - (U^i)^{-1}U^\leq i \right)^TU^\leq i \left( (S^i)^{-1}\Sigma_i^1 - (U^i)^{-1}U^\leq i \right) + \frac{1}{D_{ii}} U_{ii} \left( U^\leq i - (U^i)^TU^\leq i \right) \left( U^\leq i - (U^i)^TU^\leq i \right)^T U^\leq i \right).
\]

(5.2)

\footnote{If the equivalence class of \( i \) has \( k \) children in the Hasse diagram of \( G \), and \( V_j \) the set of vertices in \( V \) belonging to the subtree rooted at the \( j \)-th child, then \( V_j \), for \( 1 \leq j \leq k \), are disjoint subsets. Infact, if \( i_{ij} = \max \{ i' : i' \in V_j \} \), then it follows by the construction of the Hasse diagram that \( V_j = N^< (i_{ij}) \) for \( 1 \leq j \leq k \).}
Let us now evaluate the Jacobian of the transformation

$$\Sigma \rightarrow \left\{ \left( D_{ii}, (\Sigma^{<i})^{-1}\Sigma_{ii}^{<i} \right) \right\}_{1 \leq i \leq m}.$$ 

Suppose \( M = \begin{bmatrix} R & u \\ u^T & w \end{bmatrix} \) is a \( p \times p \) matrix, and consider the transformation

$$M = (R, u, w) \rightarrow (R, R^{-1}u, w - u^TR^{-1}u).$$

Since \( R^{-1}u \) depends only on \( (R, u) \) and \( w - u^TR^{-1}u \) depends on \( (R, u, w) \), it follows that the gradient matrix

$$\left( \frac{\partial R}{\partial R} \right) \cdot \left( \frac{\partial (R^{-1}u)}{\partial u} \right) \left( \frac{\partial (w - u^TR^{-1}u)}{\partial w} \right)^{-1},$$

which simplifies to \(|R|\). In particular, the Jacobian of the transformation

$$\Sigma^{<i} \rightarrow \left( \Sigma^{<i}, (\Sigma^{<i})^{-1}\Sigma_{ii}^{<i}, D_{ii} \right)$$

is given by \(|\Sigma^{<i}|\).

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

$$\Sigma \rightarrow \left\{ \left( D_{ii}, (\Sigma^{<i})^{-1}\Sigma_{ii}^{<i} \right) \right\}_{1 \leq i \leq m}$$

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

$$\prod_{i=1}^{m} |\Sigma^{<i}| = \prod_{i=1}^{m} \prod_{j \in N^{<i}(i)} D_{jj}$$

$$= \prod_{j=1}^{m} \prod_{i:j \in N^{<i}(i)} D_{jj}$$

$$= \prod_{j=1}^{m} D_{jj}^{n_j}.$$ 

(5.3)

Here as in Section 3.1 Lemma 2,

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

Also from Section 3.1,

$$\pi_{U,\alpha}^{PG}(\Sigma) = \frac{1}{z_G(U, \alpha)} e^{-\frac{1}{2} \left( \text{tr}(\Sigma^{-1}U) + \sum_{j=1}^{m} (2n_j + \alpha_j) \log D_{jj} \right)} \cdot \Sigma \in PG.$$
Let
\[ \Gamma = \left\{ \left( D_{ii}, (\Sigma^{-1})_{ii} \right) \right\}_{1 \leq i \leq m}. \]

It follows from the decomposition of \( tr (\Sigma^{-1} U) \) from (5.2) and the computation of the determinant of the Jacobian (5.3) that,
\[
\begin{align*}
\pi_{U, \alpha}^\Gamma \left( \left\{ \left( D_{ii}, (\Sigma^{-1})_{ii} \right) \right\}_{1 \leq i \leq m} \right) &= \frac{1}{z_G(U, \alpha)} \prod_{i=1}^{m} e^{- \frac{1}{2 D_{ii}} \left( (\Sigma^{-1})_{ii} - (U^{-1})_{ii}^{-1} \right)^T U \left( (\Sigma^{-1})_{ii} - (U^{-1})_{ii}^{-1} \right)} \\
&\quad \times \prod_{i=1}^{m} e^{- \frac{1}{2 D_{ii}} \left( U_{ii} - (U^{-1})_{ii} \right)^T (U^{-1})_{ii}} D_{ii}^{- \frac{\alpha_i}{2}}.
\end{align*}
\]

The above proves the mutual independence of \( \left\{ \left( D_{ii}, (\Sigma^{-1})_{ii} \right) \right\}_{1 \leq i \leq m} \). From the joint density of \( (D_{ii}, (\Sigma^{-1})_{ii}) \), it is clear that
\[
(\Sigma^{-1})_{ii} \mid D_{ii} \sim \mathcal{N} \left( (U^{-1})_{ii}, D_{ii}(U^{-1})_{ii} \right).
\]

To evaluate the marginal density of \( D_{ii} \), we integrate out \( (\Sigma^{-1})_{ii} \) from the joint density of \( (D_{ii}, (\Sigma^{-1})_{ii}) \). Note that
\[
\int_{\mathbb{R}^{[N^\prec(i)]}} e^{- \frac{1}{2 D_{ii}} \left( (\Sigma^{-1})_{ii} - (U^{-1})_{ii}^{-1} \right)^T U \left( (\Sigma^{-1})_{ii} - (U^{-1})_{ii}^{-1} \right)} d((\Sigma^{-1})_{ii}) = CD_{ii}^{- \frac{|N^\prec(i)|}{2}},
\]
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
\[
\pi_{D_{ii}, \alpha}^U (d) \propto e^{- \frac{1}{2d} \left( U_{ii} - (U^{-1})_{ii} \right)^T (U^{-1})_{ii}} d \left( \frac{\alpha_i + |N^\prec(i)|}{2} \right).
\]

We can therefore conclude that,
\[
D_{ii} \sim IG \left( \frac{\alpha_i}{2} - \frac{|N^\prec(i)|}{2} - 1, \frac{U_{ii} - (U^{-1})_{ii}}{2} \right).
\]

The mutual independence of the components of \( \Gamma \) provides the ingredients to establish the strong directed “covariance” hyper Markov property of our class of priors. Let \( G = (V, E) \) be a decomposable graph with \( V \in S_D \), i.e., the vertices have been ordered according to the perfect vertex elimination scheme for decomposable graphs outlined in Section 2.2. 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 \( pa(i) = N^\prec(i) \). As in [8, 20], let \( pr(i) = \{1, 2, \cdots, i-1\} \) denote the set of predecessors of \( i \) according to the direction specified in \( D \). Analogous to the strong hyper Markov property in the context of concetration graph models defined in [8, 20], we now define the “covariance” hyper Markov property in the context of covariance graph models.
Definition 1. A family of priors $\mathcal{F}$ on $P_G$ satisfies the strong “covariance” hyper Markov property with respect to the direction $D$, if whenever $\pi \in \mathcal{F}$ and $\Sigma \sim \pi$,

$$\Sigma_{i \text{pa}(i)} \perp \Sigma_{\text{pr}(i)} \forall 1 \leq i \leq m,$$

where $\Sigma_{i \text{pa}(i)} := \Sigma_{ii} - (\Sigma_i^T \Sigma_i)^{-1} \Sigma_i = D_{ii}$.

Let $G = (V, E)$ be a homogeneous graph with $V \in S_H$, i.e., the vertices have been ordered according to the perfect vertex elimination scheme for homogeneous graphs outlined in Section 2.4. Recall that a homogeneous graph is decomposable and $S_H \subseteq S_D$. 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 Hasse diagram of $G$ is directed in the reverse way). We show in the following proposition that in this situation, the family of priors $\pi_{U,\alpha}^{P_G}$ satisfies the strong “covariance” hyper Markov property with respect to the direction $D$.

**Proposition 5.** Let $G = (V, E)$ be homogeneous, with $V \in S_H$. If $\Sigma \sim \pi_{U,\alpha}^{P_G}$, then

$$D_{ii} \perp \Sigma_{(1,2,\ldots,i-1)} \forall 1 \leq i \leq m.$$ 

Remark. Recall that

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

is different from

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

Proof Since the vertices have been labelled according to the Hasse perfect vertex elimination scheme, $\Sigma_{(1,2,\ldots,i-1)}$ (after an appropriate permutation of the rows and columns) has a block diagonal structure with diagonal blocks $\Sigma_{<i}$ and $\Sigma_{\leq j_1}, \Sigma_{\leq j_2}, \ldots, \Sigma_{\leq j_l}$ for some $l \geq 0$, $1 \leq j_1, j_2, \ldots, j_l < i$ with $(i,j_1) \notin E \forall 1 \leq q \leq l$. This means that $\Sigma_{(1,2,\ldots,i-1)}$ is a function of $\{(D_{kk}, (\Sigma_{<k})^{-1}\Sigma_{k})\}_{1 \leq k < i}$ because $\Sigma_{\leq j} \rightarrow (\Sigma_{<j}, (\Sigma_{<j})^{-1}\Sigma_{<j}, D_{jj})$ is a one-to-one transformation for every $1 \leq j \leq m$. By Proposition 4, $(D_{ii}, (\Sigma_{<i})^{-1}\Sigma_{<i})$ is independent of $(D_{kk}, (\Sigma_{<k})^{-1}\Sigma_{k})_{1 \leq k < i}$, therefore we can conclude that

$$D_{ii}(\Sigma) \perp \Sigma_{(1,2,\ldots,i-1)}.$$ 

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 $\pi_{U,\alpha}^{P_G}$ when $G$ is homogeneous. Consequently, we can now prove “covariance” hyper Markov properties for the $IW_{Q_G}$ family.

Corollary 1. 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. Then, the family $IW_{Q_G}$ is 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.
6. Laplace transforms, expected values and generalized posterior linearity. We now consider the task of computing the Laplace transform and expected values related to our class of priors $\pi_{U,\alpha}^{PG}$. In particular, we provide closed form expressions for these quantities, when $G$ is homogeneous. Let us first establish notation that is needed in this section. Suppose $A$ is a matrix of dimension $m$ and $M$ is a subset of $\{1, 2, \ldots, m\}$. We use the following notation.

$$A_M := ((A_{ij}))_{i,j \in M},$$

$$A^0_M := \begin{cases} A_{ij} & i, j \in M, \\ 0 & \text{otherwise}. \end{cases}$$

We now provide an expression for the expected value of the inverse covariance matrix in the following proposition.

**Proposition 6.** Suppose $\alpha$ satisfies the condition in Proposition 1. Then, for $\Sigma \sim \pi_{U,\alpha}^{PG}$,

$$E_{U,\alpha}[\Sigma^{-1}] = \sum_{i=1}^{m} (\alpha_i - |N\prec(i)| - 2) \left( (U^{-\prec i})^{-1} \right)^0 - \sum_{i=1}^{m} (\alpha_i - |N\prec(i)| - 3) \left( (U^{-\prec i})^{-1} \right)^0.$$

The proposition is proved in the Appendix. We can also obtain the Laplace transform of $\Sigma^{-1}$ when $\Sigma \sim \pi_{U,\alpha}^{PG}$.

**Lemma 5.** For positive definite matrices $U, B$, and $\alpha$ satisfying the condition in Proposition 1, the Laplace transform of the distribution of $\Sigma^{-1}$ defined as $L_{U,\alpha}(B)$ is given as follows:

$$L_{U,\alpha}(B) = E_{U,\alpha} \left[ e^{-\frac{\text{tr}(\Sigma^{-1} B)}{2}} \right] = \frac{z_G(U + B, \alpha)}{z_G(U, \alpha)}.$$

This lemma follows immediately from the definition of $\pi_{U,\alpha}^{PG}$ and $z_G(U, \alpha)$.

We now provide a recursive method that gives closed form expressions for the expected value of the covariance matrix, when $\Sigma \sim \pi_{U,\alpha}^{PG}$. The basis for the method is the block mutual independence result given in Theorem 4. Since $\Sigma_{uv} = 0 \forall (u, v) \notin E$, we only need to evaluate the expectation of $\Sigma_{ii}$ and $\Sigma_{i}^\prec$ for every $1 \leq i \leq m$. The procedure is given as follows. Let

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

Recall from Section 2.4 that since $G$ is homogeneous, there exists an associated Hasse diagram on the equivalence classes of the vertices generated by the relation $R$. All vertices whose equivalence classes are leaves of the Hasse diagram, and which have the smallest label in their equivalence classes, are members of the set $A_1$. Clearly, if $i \in A_1$, then $\Sigma_{i}^\prec$ and $\Sigma_{i}^\succ$ are vacuous parameters and $D_{ii} = \Sigma_{ii}$. It follows from Theorem 4 that for $i \in A_1$,

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

assuming that $\alpha_i > 4$, since $X \sim IG(\lambda, \gamma)$ implies that $E[X] = \frac{\lambda}{\gamma - 1}$.
For \( k = 2, 3, 4, \cdots \), define
\[
A_k = \left\{ i \in V : N^{\prec}(i) \subseteq \bigcup_{l=1}^{k-1} A_l \right\} \setminus \left( \bigcup_{l=1}^{k-1} A_l \right).
\]
Since there are finitely many vertices in \( V \), \( \exists k^* \) such that \( A_k = \phi \) for \( k > k^* \). The sets \( \{A_k\}_{1 \leq k \leq k^*} \) essentially provide a way of computing \( \mathbf{E}_{U,\alpha}[\Sigma] \) by starting at the bottom of the Hasse diagram of \( G \) and then moving up sequentially.

**Proposition 7.** Given the expectations of \( \Sigma_{jj} \) and \( \Sigma_{j \prec j} \) for \( j \in \bigcup_{l=1}^{k-1} A_l \), the expectations of \( \Sigma_{ii} \) and \( \Sigma_{i \prec i} \) for \( i \in A_k \) are given by the following expressions.

\[
\mathbf{E}_{U,\alpha}[\Sigma_{i \prec i}] = \mathbf{E}_{U,\alpha}[\Sigma_{i \prec i}] (U^{\prec i})^{-1} U_{i}^{\prec i},
\]
\[
\mathbf{E}_{U,\alpha}[\Sigma_{ii}] = \frac{U_{ii} - (U_{i}^{\prec i})^T (U^{\prec i})^{-1} U_{i}^{\prec i}}{\alpha_i - |N^{\prec}(i)| - 4} + \frac{tr \left( \mathbf{E}_{U,\alpha}[\Sigma_{i \prec i}] \right)}{\alpha_i - |N^{\prec}(i)| - 4}
\]
\[
- \frac{(U^{\prec i})^{-1} (U_{ii} - (U_{i}^{\prec i})^T (U^{\prec i})^{-1} U_{i}^{\prec i})}{\alpha_i - |N^{\prec}(i)| - 4} + (U^{\prec i})^{-1} U_{i}^{\prec i} (U_{i}^{\prec i})^T (U^{\prec i})^{-1}
\]
provided \( \alpha_i > |N^{\prec}(i)| + 4 \).

The proposition is proved in the Appendix. The expressions above yield a recursive but closed form method to calculate \( \mathbf{E}[\Sigma] \) when \( \Sigma \sim \pi_{U,\alpha}^{P_G} \).

6.1. **Posterior linearity.** In their seminal paper, Diaconis and Ylvisaker [10] laid out a general framework for the construction of families of conjugate priors for natural exponential families. These conjugate priors have a specific form. More specifically, if \( X \) is a random vector distributed according to a natural exponential family with canonical parameter \( \eta \in \mathbb{R}^d \), their conjugate priors are characterized through the property of linear posterior expectation of the mean parameter of \( X \): \( \mathbf{E}_{\pi(\eta|X)}[\mathbf{E}_{P_{\eta}}[X | \eta]] = ax + b, \ a \in \mathbb{R}, b \in \mathbb{R}^d \). As covariance graph models give rise to curved exponential families, it is natural to ask if our class of priors has any posterior linearity properties. Note that for this class of models, the canonical parameter \( \Omega = \Sigma^{-1} \) lies in a lower dimensional space \( (\kappa(\Omega) \in Q_G) \), and

\[
\mathbf{E}[S | \Omega] = \Sigma.
\]

A concrete example of posterior linearity is available in the concentration graph setting. The hyper inverse Wishart (HIW) prior of Dawid and Lauritzen [8] and is a special case of the larger flexible family of conjugate priors by Letac and Massam [20] and satisfies posterior linearity, i.e., the HIW is a DY prior whereas the entire family of Letac-Massam priors does not satisfy this property in general.

In the same spirit, a relevant question is whether any member of our flexible class of priors has this posterior linearity property, i.e., if it is a DY conjugate prior. More precisely, does there exist \( U \) and \( \alpha \) such that

\[
(6.1) \quad \mathbf{E}_{U,\alpha}[\Sigma | S] = \mathbf{E}_{U,\alpha}[\Sigma] = aS + bU,
\]
for some $a, b \in \mathbb{R}$ and $\tilde{U} = nS + U$ and $\tilde{\alpha} = (\alpha_1 + n, \alpha_2 + n, \cdots, \alpha_m + n)$. It is immediately apparent that it is not possible to obtain such a relation because the left hand side of (6.1) lies in $P_G$, whereas the minimal sufficient statistic $S$ almost surely does not lie in $P_G$ but lies in the higher dimensional space $\mathbb{P}^+$ (although if the sample size $n$ is large enough, the entries of $S$ corresponding to edges not in $G$ will be close to zero, i.e., $S_{ij} \to 0$ a.s. if $(i, j) \notin E$).

Nevertheless, in the case of curved exponential families, we introduce the notion of “pseudo posterior linearity” or “generalized posterior linearity”, which is a natural extension of the concept of posterior linearity in regular exponential family settings. We show that this pseudo/generalized posterior linearity is satisfied by certain members of our class of priors, and therefore draw parallels to the natural exponential family setting. We shall first introduce some new concepts that allow us to formalize this property.

**Definition 2.** Consider a curved exponential family with density (w.r.t. some measure $\mu$) given by

$$e^{-\left(\sum_{i=1}^{d} \eta_i T_i - \phi(\eta)\right)}, \eta \in \Xi,$$

where $\eta \in \mathbb{R}^d$ is the canonical parameter and $\Xi$ a lower-dimensional curve in $\mathbb{R}^d$. This curved exponential family is called “separable” if $\exists d' < d$ such that on $\Xi$, $\eta_1, \eta_2, \cdots, \eta_{d'}$ are functionally independent variables, with the remaining variables $\eta_{d'+1}, \eta_{d'+2}, \cdots, \eta_d$ being functions of $\eta_1, \eta_2, \cdots, \eta_{d'}$.

For these curved exponential families, we can “separate” out the functionally independent components of the canonical parameter from the dependent components. Since the minimal sufficient statistic and the canonical parameter in a regular exponential family are of the same dimension, the concept of posterior linearity is well defined for regular exponential family models. In order to obtain posterior linearity in a curved setting akin to the regular exponential family setup, a natural approach is to modify the likelihood so that it resembles a regular exponential family and then pose the question of posterior linearity for this modified likelihood. We make this notion more precise in the following definition.

**Definition 3.** The pseudo-likelihood of a distribution that arises from a “separable” curved exponential family (6.2) is defined to be

$$e^{-\left(\sum_{i=1}^{d'} \eta_i T_i - \phi(\eta)\right)}, \eta \in \Xi.$$

The “pseudo likelihood” is obtained by considering only the functionally independent variables in the first term in the exponent in (6.2). We now apply this notion to covariance graph models. Consider the class of distributions $\mathcal{G}$ as in Section 3, with a homogeneous graph $G = (V, E)$. We assume that the vertices have been ordered according to the Hasse perfect vertex elimination scheme specified in Section 2.4. The family $\mathcal{G}$ is indeed a separable exponential family with canonical parameter $\Omega = \Sigma^{-1}$ and $\{\Omega_{ij} : i > j, (i, j) \in E\}$ being the functionally independent components. Hence, the “pseudo likelihood” for this class of models is given by

$$e^{-\frac{n tr (\Omega S^*) + \log |\Sigma|}{2}}.$$
where

\[ S^*_ij = \begin{cases} S_{ij} & \text{if } (i, j) \in E, \\ 0 & \text{otherwise}. \end{cases} \]

If \( \Sigma \) has a prior distribution given by \( \pi_{U, \alpha}^\Sigma \), then the posterior distribution based on the pseudo likelihood is given by \( \pi_{U^*, \alpha^*}^\Sigma \), where \( U^* = nS^* + U, \alpha^* = (n + \alpha_1, n + \alpha_2, \ldots, n + \alpha_m) \). We now prove the following proposition.

**Proposition 8.** Let \( U \in P_G \) and \( \alpha_i = 2|N^<(i)| + 4 + c \) for some \( c > 0 \) and \( i = 1, 2, \ldots, m \).

If \( \Sigma \sim \pi_{U, \alpha}^P \), then,

\[ E_{U, \alpha}[\Sigma] = \frac{1}{c} U. \]

The proposition is proved in the Appendix.

Now note that if \( U \in P_G \), then \( U^* \in P_G \) (since \( S^* \in P_G \)). Hence, if \( U \) is chosen so that \( U^* \) is positive definite \(^{11}\), then under the conditions of Proposition 8, the expectation of \( \Sigma \) under the distribution \( \pi_{U^*, \alpha^*}^\Sigma \) (the posterior distribution based on the pseudo likelihood) is given by

\[ E_{U^*, \alpha^*}[\Sigma] = \frac{nS^* + U}{n + c}. \]

which is a linear function of the data. Hence we conclude that a subclass of our family of distributions satisfies the pseudo posterior linearity property.

**7. Examples.** The main purpose of the 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.

**7.1. Genomics Example.** We provide an illustration of our methods on a dataset consisting of gene expression data from microarray experiments with yeast strands from Gasch et al. [13]. This dataset has also been analyzed in [3, 11]. As in [3, 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, as the mean is not assumed to be zero.

We consider the covariance graph model specified by the graph \( G \) in Figure 2 with the overall aim of estimating \( \Sigma \) under this covariance graph model. The maximum likelihood estimate for \( \Sigma \in P_G \), provided by the iterative conditional fitting algorithm provided in [3], 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 ordering \( \{GAL11, GAL4, GAL80, GAL3, GAL7, GAL10, GAL1, GAL2\} \) for our analysis.

Our goal is to obtain the posterior mean for \( \Sigma \) under our new class of priors and then provide Bayes estimators for \( \Sigma \). We use two diffuse priors to illustrate our methodology. The first prior is denoted as \( \tilde{\pi}_{U_1, \alpha^1} \), where \( U_1 = \frac{n(S)}{8} I_8, \alpha^1_i = 5 + |N^<(i)|; i = 1, 2, \ldots, 8 \), i.e., \( \alpha_1 = (5, 6, 6, 8, 7, 8, 9, 12) \). The second prior used is \( \tilde{\pi}_{U_2, \alpha^2} \), where \( U_2 = 0, \alpha^2_i = 2 i = 1, 2, \ldots, 8 \).

\(^{11}\)For example, \( U \) can be chosen so that \( U^* \) is diagonally dominant
<table>
<thead>
<tr>
<th></th>
<th>GAL11</th>
<th>GAL4</th>
<th>GAL80</th>
<th>GAL3</th>
<th>GAL7</th>
<th>GAL10</th>
<th>GAL1</th>
<th>GAL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAL11</td>
<td>0.152</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAL4</td>
<td>0.034</td>
<td>0.130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAL80</td>
<td>0.015</td>
<td>0.039</td>
<td>0.221</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAL3</td>
<td>-0.055</td>
<td>0.034</td>
<td>0.073</td>
<td>0.608</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAL7</td>
<td>-0.051</td>
<td>-0.053</td>
<td>0.183</td>
<td>0.722</td>
<td>3.423</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAL10</td>
<td>-0.048</td>
<td>-0.039</td>
<td>-0.188</td>
<td>0.553</td>
<td>2.503</td>
<td>2.372</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAL1</td>
<td>-0.066</td>
<td>-0.061</td>
<td>0.224</td>
<td>0.517</td>
<td>2.768</td>
<td>2.409</td>
<td>2.890</td>
<td></td>
</tr>
<tr>
<td>GAL2</td>
<td>-0.119</td>
<td>-0.018</td>
<td>0.208</td>
<td>0.583</td>
<td>2.547</td>
<td>2.278</td>
<td>2.514</td>
<td>2.890</td>
</tr>
</tbody>
</table>

Table 1: Empirical covariance matrix for yeast data

Fig 2. Covariance graph for yeast data
The block Gibbs sampling procedure was run for both priors as specified in Section 4. The burn-in period was chosen to be 1000 iterations, and the subsequent 1000 iterations were used to compute the posterior mean. Increasing the burn-in period to more than 1000 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.

7.2. 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, like 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 $G$
Fig 3. Hasse diagram for a homogeneous graph with 50 vertices

with 50 vertices, with the corresponding Hasse tree given by Figure 3. Let $\Sigma \in P_G$, where the vertices have been ordered according to the Hasse perfect vertex elimination scheme in Section 2.4, the diagonal entries are 50 and all other non-zero entries are 1. We simulate 100 observation vectors $Y_1, Y_2, Y_3, \ldots, Y_{99}, Y_{100}$ from $N_{50}(0, \Sigma)$. For illustration purposes, we choose a diffuse prior $\pi_{U, \alpha}^\Sigma$ with $U = 0$ and $\alpha_i = 2|N^<(i)| + 5$, $i = 1, 2, \ldots, 50$.

Since the graph $G$ is homogeneous, we can compute the posterior mean $\Sigma_{\text{mean}} \equiv E_{U, \alpha}[\Sigma \mid Y_1, Y_2, \ldots, Y_{100}]$ explicitly. We can therefore assess the ability of the block Gibbs sampling algorithm in estimating 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 $\Sigma_{\text{mean}}$. We use an initial burn-in period of $B$ iterations and then average over the next $I$ iterations to get the estimate $\hat{\Sigma}$. The times needed for computation (using the R software) and the relative errors $\frac{\|\hat{\Sigma} - \Sigma_{\text{mean}}\|_2}{\|\Sigma_{\text{mean}}\|_2}$ 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, and yields 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, and yields 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.

8. 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 parallel theory analogous to the standard conjugate theory for natural exponential families has been developed for this class of models. In particular, a rich class of conjugate priors has been developed for covariance
graph models, when 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 homogeneous 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 do inference in covariance graph models even when the sample size $n$ is less than the dimension $p$ of the data, otherwise not possible in general in the maximum likelihood framework. Furthermore, we thoroughly explore the theoretical properties of our class of conjugate priors. In particular, we establish “covariance” hyper-Markov properties, expectation of the covariance matrix, Laplace transforms and investigate posterior linearity. Furthermore the usefulness of the methodology that is developed is illustrated through examples. A few open problems are worthy of mentioning.

- What are the necessary conditions for existence of the normalizing constant for decomposable graphs?
- Does the “covariance” 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, and a detailed study of Bayes estimators in this context, is clearly an important topic, and is the focus of current research.

Acknowledgments. The authors gratefully acknowledge the faculty at the Stanford Statistics department for their feedback and tremendous enthusiasm for this work.

References.


**Appendix.** Proof of Proposition 1. From the definition of $N$ and $L$, it is easy to verify that

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

$$(LN)_{ij} = 0, \forall 1 \leq i < j \leq m.$$
Now let $i > j$. It follows by the definition of $N$ that,

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

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

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

Note that any $\tau' \in A$ with $\tau_{\dim(\tau')} = j, \dim(\tau') > 2$, 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 $L_{\tau'} = L_{\tau_1\tau}L_{\tau}$. Also, if $\tau' \in A$ with $\tau_{\dim(\tau')} = j, \dim(\tau') = 2$, then $\tau' = (i, j)$ and $L_{\tau'} = L_{ij}$. Hence,

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

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

$$= 0.$$  

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

\[\square\]

**Proof of Proposition 3.** Assume that the vertices have been ordered according to the Hasse perfect vertex elimination scheme in Section 2.4. Suppose $\Sigma \in P_G$. We wish to prove that $L \in \mathcal{L}_G$.

Let $i^* = \min_{1 \leq i \leq m} \{\exists j < i \text{ such that } (i, j) \notin E\}$. We first claim that $\Sigma_{i^*1} = 0$. In order to prove this result, assume to the contrary that $\Sigma_{i^*1} \neq 0$, which in turn implies that $(i^*, 1)^{} \in E$. By the definition of $i^*$, it follows that $(j, 1) \in E$ for every $j < i^*$. This implies that 1 is a descendant or twin of both $i^*$ and $j$ in the Hasse tree, which further implies that $j$ is a descendant or twin of $i^*$ for every $j < i^*$. In other words $(i^*, j) \in E \forall j < i^*$. This leads to a contradiction with the definition of $i^*$. Hence, $\Sigma_{i^*1} = 0$.

Since $\Sigma_{11} = L_{11} D_{11}$ for every $1 \leq i \leq m$, it follows that $L_{i^*1} = \frac{\Sigma_{i^*1}}{L_{11}} = 0$. We now proceed by induction. Let $u > v$ be such that $(u, v) \notin E$, i.e., $\Sigma_{uv} = 0$. Assume that $L_{ij} = 0 \forall (i, j) \notin E$ with $j < i, 1 \leq i \leq u - 1$ or $i = u, j < v$. Note that

$$\Sigma_{uv} = L_{uv}D_{uv} + \sum_{i=1}^{v-1} L_{ui}L_{vi}D_{ii}.$$  

Suppose $\exists i, 1 \leq i \leq v - 1$ such that $(u, i) \in E, (v, i) \in E$. This means that $i$ is a descendant or twin of both $u$ and $v$ in the Hasse diagram, which then implies that $v$ is a descendant or twin of $u$, which is not possible, because $(u, v) \notin E$. Hence, using the induction hypothesis, $L_{ui}L_{vi} = 0$ for every $1 \leq i \leq v - 1$. Hence $L_{uv} = 0$. We can therefore inductively prove that $L_{ij} = 0$ for all $i > j, (i, j) \notin E$.

Suppose $L \in \mathcal{L}_G$. Let $u > v$ be such that $(u, v) \notin E$. Then $L_{uv} = 0$. Suppose $\exists i, 1 \leq i \leq v - 1$ such that $(u, i) \in E, (v, i) \in E$. This means that $i$ is a descendant or twin of both $u$ and $v$
in the Hasse diagram, which then implies that \( v \) is a descendant or twin of \( u \), which is not possible, because \((u, v) \notin E\). Since \( L \in \mathcal{L}_G \), this implies \( L_{ui}L_{vi} = 0 \) for every \( 1 \leq i \leq v - 1 \). Hence, \( \Sigma_{uv} = 0 \). Hence, the first part of the proposition is proved.

For the second part, by Proposition 1, it follows that for \( u > v \), \((L^{-1})_{uv} = 0 \) \( \forall L \in \mathcal{L}_G \) if and only if there is no path from \( u \) to \( v \) in the graph \( G \), such that the vertices in the path are strictly decreasing. This happens if and only if \( v \) is not a descendant of \( u \) in the rooted Hasse tree of \( G \), i.e., \((u, v) \notin E\), which implies \( L_{uv} = 0 \). Hence the second part of the proposition is proved. \( \square \)

**Proof of Theorem 1.** Let us simplify the integral by integrating out the terms \( D_{ii}, 1 \leq i \leq m \).

\[
\int e^{-\frac{1}{2} \left( \text{tr}(LDDL^T)^{-1}L) + \sum_{i=1}^{m} \alpha_i \log D_{ii} \right)} dLdD \]

\[
= \int e^{-\frac{1}{2} \left( \text{tr}(D^{-1}(L^{-1}U(L^T)^{-1})) + \sum_{i=1}^{m} \alpha_i \log D_{ii} \right)} dLdD 
\]

\[
= \prod_{i=1}^{m} e^{-\frac{1}{2} \left( \frac{1}{D_{ii}} \right)} D_{ii}^{\alpha_i} dDdL 
\]

\[
= \prod_{i=1}^{m} \frac{\Gamma \left( \frac{\alpha_i}{2} - 1 \right) \frac{2^{\frac{\alpha_i}{2} - 1}}{(\frac{L^{-1}U(L^T)^{-1})_{ii}})^{\frac{\alpha_i}{2} - 1}} \right) dL 
\]

\[
= \prod_{i=1}^{m} \frac{\Gamma \left( \frac{\alpha_i}{2} - 1 \right) \frac{2^{\frac{\alpha_i}{2} - 1}}{(\frac{L^{-1}U((L^{-1})_i)^T)^{\frac{\alpha_i}{2} - 1}} \right) dL 
\]

In order to simplify this integral we perform a change of measure by transforming the non-zero elements of \( L \) to the corresponding elements of \( L^{-1} \). Note the following facts.

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

   \[
   L^{-1}_{ij} = -L_{ij} + f \left( (L_{uv})_{(u, v) \in E, j \leq u < i, j \leq u < v \text{ or } u=i,j<v<i} \right),
   \]

   i.e., \( L^{-1}_{ij} + L_{ij} \) 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 \( \{L^{-1}_{uv}\}_{u>v, (u, v) \in E}\). Let \( i^* = \min \{ i : L_{ij} \neq 0 \text{ for some } j < i \} \). Let \( j^* = \max \{ j : L_{i^*j} \neq 0 \} \). By (8.1) and the definition of \( i^* \) and \( j^* \), we have \( L_{i^*j^*} = -L^{-1}_{i^*j^*} \). We proceed by induction. Let \( i > j, (i, j) \in E \) and suppose 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

   \[
   L_{ij} = -L^{-1}_{ij} + f \left( (L_{uv})_{(u, v) \in E, j \leq u < i, j \leq v < u \text{ or } u=i,j<v<i} \right),
   \]

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

   It follows that the transformation

   \[
   \{L_{ij}\}_{(i, j) \in E, i > j} \rightarrow \{L^{-1}_{ij}\}_{(i, j) \in E, i > j}
   \]

   is a bijection and the absolute value of the Jacobian of this transformation is 1, as it is the determinant of a lower triangular matrix with diagonal entries 1.
2. If \( x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \) and \( U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \) is a positive definite matrix, then

\[
(8.2) \quad x^T U x = z^T z + x_2^T \left( U_{22} - U_{21} U_{11}^{-1} U_{12} \right) x_2 \geq x_2^T \left( U_{22} - U_{21} U_{11}^{-1} U_{12} \right) x_2,
\]

where \( z = U_{11}^{\frac{1}{2}} x_1 + U_{11}^{-\frac{1}{2}} U_{12} x_2. \)

Hence, after transforming the non-zero entries of \( L \) to the corresponding entries of \( L^{-1} \), and using (8.2) to eliminate the dependent entries of \( L^{-1} \) from the integrand, we get,

\[
\int e^{-\left( \text{tr}(L D L^T)^{-1} U + \sum_{i=1}^{m} \alpha_i \log D_{ii} \right)} dL dD \leq K \prod_{i=2}^{m} \int_{\mathbb{R}^{N \setminus \{i\}}} \frac{1}{(\alpha_i^T 1) U_i \left( \alpha_i 1 \right)} d\alpha_i.
\]

Here \( K \) is a constant, \( U_i \) is an appropriate positive definite matrix, and \( \alpha_i \) represents the independent entries in the \( i \)th row of \( L^{-1} \). By a suitable linear transformation \( b_i \) of each of the \( \alpha_i, i = 2, 3, \ldots, m \), we get,

\[
\int e^{-\left( \text{tr}(L D L^T)^{-1} U + \sum_{i=1}^{m} \alpha_i \log D_{ii} \right)} dL dD \leq K^* \prod_{i=2}^{m} \int_{\mathbb{R}^{N \setminus \{i\}}} \frac{1}{(b_i^T b + u_i^{**})} d\alpha_i.
\]

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

\[
\int_{\mathbb{R}^k} \frac{1}{(x^T x + 1)^\gamma} dx < \infty \quad \text{if} \quad \gamma > \frac{k}{2},
\]

we conclude that,

\[
\int e^{-\left( \text{tr}(L D L^T)^{-1} U + \sum_{i=1}^{m} \alpha_i \log D_{ii} \right)} dL dD < \infty
\]

if \( \alpha_i > |N \setminus \{i\}| + 2 \quad \forall i = 1, 2, \ldots, m. \)

**Proof of Theorem 2.** We first note that by the formula provided on [9, page 16] that,

\[
\int_{\mathbb{R}} \frac{1}{(1 + x^2)^\gamma} dx = \begin{cases} \frac{\sqrt{\pi} \Gamma\left(\gamma - \frac{1}{2}\right)}{\Gamma(\gamma)} & \gamma > \frac{1}{2}, \\ \infty & \text{otherwise}. \end{cases}
\]

By repeated application, we can generalize the above formula to,

\[
\int_{\mathbb{R}^d} \frac{1}{(x^T x + 1)^\gamma} dx = \begin{cases} \frac{((\sqrt{\pi})^d \Gamma\left(\gamma - \frac{d}{2}\right)}{\Gamma(\gamma)} & \gamma > \frac{d}{2}, \\ \infty & \text{otherwise}. \end{cases}
\]
Let us now consider an integral of a similar form given by

$$\int_{\mathbb{R}^d} \frac{1}{(x^T 1)U(x)}^\gamma dx,$$

where $U = \begin{pmatrix} U_0 & u \\ u^T & w \end{pmatrix}$ is a positive definite matrix. Consider the linear transformation $y = U_0^{-\frac{1}{2}}x + U_0^{-\frac{1}{2}}u$. It follows that for $\gamma > \frac{d}{2}$,

$$\int_{\mathbb{R}^d} \frac{1}{(x^T 1)U(x)}^\gamma dx = \frac{1}{|U_0|^\frac{d}{2}} \int_{\mathbb{R}^d} \frac{1}{(y^T y + w - u^T U_0^{-1} u)^\gamma} dy.$$

Using arguments as in the proof of Theorem 1 (see (**)),

$$z_G(U, \alpha) = \int \prod_{i=1}^m \frac{\Gamma \left( \frac{\alpha_i}{2} - 1 \right) 2^{\frac{\alpha_i}{2} - 1}}{(U_1^{\mathcal{N}^c(i)})^{\frac{\alpha_i}{2} - 1}} \frac{\Gamma \left( \frac{\alpha_i}{2} - 1 \right) 2^{\frac{\alpha_i}{2} - 1}}{(D^{-1})^{\frac{\alpha_i}{2} - 1}} dL.$$

Let us use the transformation

$$\{L_{ij}\}_{(i,j) \in E, i \geq j} \rightarrow \{L_{ij}^{-1}\}_{(i,j) \in E, i \geq j}.$$

As demonstrated in the proof of Theorem 1, this transformation is a bijection and the absolute value Jacobian of this transformation is 1. Also, Proposition 3 implies that

$$L \in \mathcal{L}_G \Leftrightarrow L^{-1} \in \mathcal{L}_G.$$

Hence, $L_{kl}^{-1} = 0$ for $k > l, (k, l) \notin E$. It follows that,

$$z_G(U, \alpha) = \Gamma \left( \frac{\alpha_i}{2} - 1 \right) 2^{\frac{\alpha_i}{2} - 1} \prod_{i=2}^m \int_{\mathbb{R}^{\mathcal{N}^c(i)}} \frac{\Gamma \left( \frac{\alpha_i}{2} - 1 \right) 2^{\frac{\alpha_i}{2} - 1}}{(x^T 1)U^{\leq i}(x)}^\gamma dx.$$

This implies that $z_G(U, \alpha) < \infty$ if and only if $\frac{\alpha_i}{2} - 1 > \frac{|\mathcal{N}^c(i)|}{2} \ orall \ 1 \leq i \leq m$. Assuming these are satisfied, we obtain using (8.3) that,

$$z_G(U, \alpha) = \frac{\Gamma \left( \frac{\alpha_i}{2} - 1 \right) 2^{\frac{\alpha_i}{2} - 1}}{(U_1^{\mathcal{N}^c(i)})^{\frac{\alpha_i}{2} - 1}} \prod_{i=2}^m \frac{\Gamma \left( \frac{\alpha_i}{2} - 1 \right) 2^{\frac{\alpha_i}{2} - 1}}{(U_1^{\mathcal{N}^c(i)})^{\frac{\alpha_i}{2} - 1}}.$$
It follows after a little simplification that,
\[
z_G(U, \alpha) = \prod_{i=1}^{m} \frac{\Gamma \left( \frac{\alpha_k}{2} - \frac{\kappa}{2} \right)}{\Gamma \left( \frac{\alpha_k}{2} \right)} \left( \sum_{i,j} \frac{\kappa}{2} \right)^{-1} \left( \sum_{j=1}^{m} \frac{\kappa}{2} \right)^{-1} \frac{1}{\sqrt{2}} \sqrt{\pi} |U_<| \prod_{i=1}^{m} \left( \frac{\kappa}{2} - \frac{\kappa}{2} \right)^{-1} \frac{1}{\sqrt{2}} \sqrt{\pi} |U_<| \prod_{i=1}^{m} \left( \frac{\kappa}{2} - \frac{\kappa}{2} \right)^{-1}
\]
where it is implicitly assumed that $|U_<| = 1$ whenever $\kappa < 0$.

**Proof of Proposition 6.** It follows from the proof of Theorem 2 that for $(k, i), (k, j) \in E, i, j < k$,
\[
E_{U, \alpha} \left[ L_{k,i}^{-1} L_{k,j}^{-1} D_{kk}^{-1} \right] = K \int_{\mathbb{R}^{k \kappa}} \frac{|x_i x_j|}{\left( x^T 1 \right) U_<^k \left( x \right)} dx.
\]
Here $K$ is a constant. Hence,
\[
E_{U, \alpha} \left[ L_{k,i}^{-1} L_{k,j}^{-1} D_{kk}^{-1} \right] < \infty \text{ if } \frac{\alpha_k}{2} > \frac{|\kappa|}{2} < 1.
\]
We observe that $L_{kk}^{-1} = 1 \forall 1 \leq k \leq m$. Therefore,
\[
E_{U, \alpha} \left[ \left( L D L^T \right)_{ij}^{-1} \right] < \infty \text{ if } \alpha_k > |\kappa| + 2 \forall 1 \leq k \leq m.
\]
Hence,
\[
E_{U, \alpha} \left[ \Sigma^{-1} \right] < \infty \text{ if } \alpha_k > |\kappa| + 2 \forall 1 \leq k \leq m.
\]
Note that given any positive definite matrix $U_0$, $\exists \epsilon > 0$ such that if we perturb any entry of $U_0$ by at most $\epsilon$, the resulting matrix is still positive definite. Also, $\pi_{U, \alpha}(L, D)$ is a continuous function of $U$ in this bounded neighborhood of $U_0$. Since by definition,
\[
z_G(U, \alpha) = \int e^{-\left( \text{tr} \left( \Sigma^{-1} U \right) + \sum_{i=1}^{m} \alpha_i \log D \right)} dLdD,
\]
it follows by combining the facts above to differentiate inside the integral w.r.t. $U_{ij}$ (with $i > j$), that,
\[
\frac{\partial z_G(U, \alpha)}{\partial U_{ij}} = - \int \Sigma_{ij}^{-1} e^{-\left( \text{tr} \left( \Sigma^{-1} U \right) + \sum_{i=1}^{m} \alpha_i \log D \right)} dLdD.
\]
This gives,
\[
E_{U, \alpha} \left[ \Sigma_{ij}^{-1} \right]
\]
\[
= - \frac{\partial}{\partial U_{ij}} \left[ \text{const} + \sum_{k=1}^{m} \left( \frac{\alpha_k}{2} - \frac{|\kappa|}{2} \right) \log |U_<| + \sum_{k=1}^{m} \left( \frac{\alpha_k}{2} - \frac{|\kappa|}{2} - 1 \right) \log |U_<| \right]
\]
\[
= \sum_{k=1}^{m} \left( \frac{\alpha_k}{2} - \frac{|\kappa|}{2} - 1 \right) \frac{\partial}{\partial U_{ij}} \log |U_<| - \sum_{i=1}^{m} \left( \frac{\alpha_k}{2} - \frac{|\kappa|}{2} - 3 \right) \frac{\partial}{\partial U_{ij}} \log |U_<|.
\]
Note that, for a matrix $A$ and $i > j$,
\[ \frac{\partial}{\partial A_{ij}} \log |A| = 2A_{ij}^{-1}. \]

Note that, $\log |U^{\preceq k}|$ depends on $U_{ij}$ if and only if $i, j \in \mathcal{N}^{\preceq}(k) \cup \{k\}$ and $\log |U^{\preceq k}|$ depends on $U_{ij}$ if and only if $i, j \in \mathcal{N}^{\preceq}(k)$. It follows that for $i > j$,
\[ E_{U,\alpha} \left[ \Sigma^{-1}_{ij} \right] = \sum_{k=1}^{m} (\alpha_k - |\mathcal{N}^{\preceq}(k)| - 2) \left( (U^{\preceq k})^{-1} \right)_{ij}^{0} - \sum_{k=1}^{m} (\alpha_k - |\mathcal{N}^{\preceq}(k)| - 3) \left( (U^{\prec k})^{-1} \right)_{ij}^{0}. \]

Recall that for a matrix $A$ and a subset $M$ of $\{1, 2, \cdots, m\}$,
\[ A^0_M = \begin{cases} A_{ij} & \text{if } i, j \in M, \\ 0 & \text{otherwise.} \end{cases} \]

Again by differentiating inside the integral w.r.t. $U_{ii}$, we get that,
\[ E_{U,\alpha} \left[ \frac{\Sigma^{-1}_{ii}}{2} \right] = -\frac{\partial}{\partial U_{ii}} \log z_G(U, \alpha). \]

We use the identity
\[ \frac{\partial}{\partial A_{ii}} \log |A| = A_{ii}^{-1}, \]
and similarly conclude that,
\[ E_{U,\alpha} \left[ \Sigma^{-1}_{ii} \right] = \sum_{k=1}^{m} (\alpha_k - |\mathcal{N}^{\preceq}(k)| - 2) \left( (U^{\preceq k})^{-1} \right)_{ii}^{0} - \sum_{k=1}^{m} (\alpha_k - |\mathcal{N}^{\preceq}(k)| - 3) \left( (U^{\prec k})^{-1} \right)_{ii}^{0}. \]

Hence it follows that,
\[ E_{U,\alpha} \left[ \Sigma^{-1} \right] = \sum_{k=1}^{m} (\alpha_k - |\mathcal{N}^{\preceq}(k)| - 2) \left( (U^{\preceq k})^{-1} \right) - \sum_{k=1}^{m} (\alpha_k - |\mathcal{N}^{\preceq}(k)| - 3) \left( (U^{\prec k})^{-1} \right). \]

**Proof of Proposition 7.** Clearly, the expectation of $\Sigma_{ii}$ for $i \in A_1$ can be computed explicitly as shown in Section 6. Assume that the expectations of $\Sigma_{ii}$ and $\Sigma^{-1}_{ii}$ for $i \in \cup_{l=1}^{k-1} A_l$ are given. Now note that for $i \in A_k$,
\[ E_{U,\alpha} [\Sigma^{-1}_{ij}] = E_{U,\alpha} \left[ \Sigma^{-\preceq i}(\Sigma^{\preceq i})^{-1}\Sigma^{-\preceq i}_{ij} \right] = E_{U,\alpha} \left[ \Sigma^{-\preceq i} \right] E_{U,\alpha} \left[ (\Sigma^{\preceq i})^{-1}\Sigma^{-\preceq i}_{ij} \right] = E_{U,\alpha} \left[ \Sigma^{-\preceq i} \right] (U^{\preceq i})^{-1}U^{\preceq i}_{ij}. \]
The last two equalities follow from Theorem 4. Note that if \( j \in \mathcal{N}^{-i}(i) \), then \( j \in \bigcup_{i=1}^{k-1} A_i \), and hence \( \mathbf{E}_{U,\alpha} [\Sigma^{-i}] \) has already been computed by this stage. Similarly,

\[
\mathbf{E}_{U,\alpha}[\Sigma_{ii}] = \mathbf{E}_{U,\alpha} \left[ D_{ii} + (\Sigma^{-i}_i)T(\Sigma^{-i})^{-1}\Sigma_{ii}^{-1} \right] = \frac{U_{ii} - (U_i^{-1})^T(U^{-i})^{-1}U_i^{-1}}{\alpha_i - |\mathcal{N}^{-i}(i)| - 4} + \mathbf{E}_{U,\alpha} \left[ (\Sigma^{-i}_i)T(\Sigma^{-i})^{-1}\Sigma_{ii}^{-1} \right],
\]

where \( \mathbf{E}[D_{ii}] \) is once more calculated by using the result from Theorem 4 that \( D_{ii} \) has an inverse Gamma distribution. Now note that,

\[
\mathbf{E}_{U,\alpha} \left[ (\Sigma_{ii}^{-1})^{-1} \right] = \text{tr} \left( \mathbf{E}_{U,\alpha} \left[ (\Sigma_{ii}^{-1})^{-1} \right] \right) = \text{tr} \left( \mathbf{E}_{U,\alpha} \left[ (\Sigma_{ii}^{-1}) \right] \right)
\]

\[
= \text{tr} \left( \mathbf{E}_{U,\alpha} \left[ (\Sigma_{ii}^{-1}) \right] (\Sigma_{ii}^{-1}) \right) = \text{tr} \left( \mathbf{E}_{U,\alpha} \left[ (\Sigma_{ii}^{-1})^{-1} \right] \right)
\]

\[
= \text{tr} \left( \mathbf{E}_{U,\alpha} [\Sigma^{-i}] \mathbf{E}_{U,\alpha} \left[ (\Sigma^{-i})^{-1} | D_{ii} \right] \right)
\]

\[
= \text{tr} \left( \mathbf{E}_{U,\alpha} [\Sigma^{-i}] \mathbf{E}_{U,\alpha} \left[ (U^{-i})^{-1} + (U^{-i})^{-1}U_i^{-1}U_i^{-1}T(U^{-i})^{-1} \right] \right)
\]

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

Note that we have once more used the joint distribution of \( (D_{ii}, (\Sigma^{-i})^{-1}\Sigma_{ii}^{-1}) \) as specified in Theorem 4. \( \square \)

**Proof of Proposition 8** We proceed by induction. Clearly if \( i = 1 \), then \( \Sigma^{-1} \) and \( \Sigma_i^{-1} \) are vacuous parameters and

\[
\mathbf{E}[\Sigma_{11}] = \mathbf{E}[D_{11}] = \frac{U_{11}}{\alpha_1 - 4} = \frac{1}{c} U_{11}.
\]

Assume now that \( \mathbf{E} [\Sigma^{-k}] = \frac{1}{c} U^{-k} \) \( \forall 1 \leq k < i \). Note that from the proof of Theorem 4, \( \Sigma^{-i} = \Sigma^{-i(1)} \) or \( \Sigma^{-i} \) has a block diagonal structure (after an appropriate permutation of the rows and columns) with blocks \( \Sigma_i^{\leq i_1}, \Sigma_i^{\leq i_2}, \cdots, \Sigma_i^{\leq i_k} \) for some \( k > 1, 1 \leq i_1, i_2, \cdots, i_k < i \) and \( \mathcal{N}^{-i}(i) \subseteq \{ 1, 2, \cdots, i - 1 \} \), and since \( U \in P_G \), \( U^{-i} \) has a similar structure. Hence, it follows that,

\[
\mathbf{E}_{U,\alpha} \left[ \Sigma^{-i} \right] = \frac{1}{c} U^{-i}.
\]

By Proposition 7,

\[
\mathbf{E}_{U,\alpha} \left[ \Sigma_{ii}^{-1} \right] = \mathbf{E}_{U,\alpha} \left[ \Sigma^{-i} \right] (U^{-i})^{-1} U_i^{-1} = \frac{1}{c} U_i^{-1},
\]
and

\[
\mathbb{E}_{U, \alpha}[\Sigma_{ii}] = \left( U_{ii} - (U_{\prec i}^\gamma)^T (U_{\prec i}^\gamma)^{-1} U_{\prec i}^\gamma \right) \\
\quad + \frac{1}{c} \text{tr} \left( \frac{U_{ii} - (U_{\prec i}^\gamma)^T (U_{\prec i}^\gamma)^{-1} U_{\prec i}^\gamma}{\alpha_i - |N_{\prec i}| + 4} I_{|N_{\prec i}|} + U_{\prec i}^\gamma (U_{\prec i}^\gamma)^T (U_{\prec i}^\gamma)^{-1} \right) \\
\quad = \frac{U_{ii}}{|N_{\prec i}| + c} \left( 1 + \frac{|N_{\prec i}|}{c} \right) + (U_{\prec i}^\gamma)^T (U_{\prec i}^\gamma)^{-1} U_{\prec i}^\gamma \left( \frac{1}{c} - \frac{1}{|N_{\prec i}| + c} - \frac{|N_{\prec i}|}{c(|N_{\prec i}| + c)} \right) \\
\quad = \frac{1}{c} U_{ii}.
\]

Hence \( \mathbb{E}_{U, \alpha}[\Sigma_{\prec i}] = \frac{1}{c} U_{\prec i}^\gamma \). This implies the proposition. \(\square\)