Bayesian Analysis (0000) **00**, Number 0, pp. 1

High dimensional Bayesian inference for Gaussian DAG models

SUPPLEMENTARY MATERIAL

Emanuel Ben-David*, Tianxi Li† Hélène Massam‡ and Bala Rajaratnam§

1 Graph theory, Markov properties and Gaussian DAGs

1.1 Graphs

A graph $\mathcal{G}$ is a pair of objects (V, E), where V and E are two disjoint finite sets representing, respectively, the vertices and the edges of $\mathcal{G}$. Each edge in E is either an ordered pair (i, j) or an unordered pair $\{i, j\}$, for some $i, j \in V$. An edge $(i, j) \in E$ is called directed where i is said to be a parent of j, and j is said to be a child of i, when $i \neq j$. We write this as $i \to j$. The set of parents of i is denoted by $\mathrm{pa}(i)$, and the set of children of i is denoted by $\mathrm{ch}(i)$. The family of i is $\mathrm{fa}(i) = \mathrm{pa}(i) \cup \{i\}$. An edge $\{i, j\} \in E$ is called undirected where i is said to be a neighbor of j, or j a neighbor of i, when $i \neq j$. We write this $i \sim_{\mathcal{G}} j$. The set of all neighbors of i is denoted by $\mathrm{ne}(i)$. We say i and j are adjacent if there exists either a directed or an undirected edge between them. A loop in $\mathcal{G}$ is an ordered pair (i, i), or an unordered pair $\{i, i\}$ in E. For ease of notation, in this paper we always shall assume that the edge set E contains all the loops, although we shall draw the respective graphs without the loops.

We say that the graph $\mathcal{G}' = (V', E')$ is a subgraph of $\mathcal{G} = (V, E)$, denoted by $\mathcal{G}' \subset \mathcal{G}$, if $V' \subset V$ and $E' \subset E$. In addition, if $\mathcal{G}' \subset \mathcal{G}$ and $E' = V' \times V' \cap E$, we say that $\mathcal{G}'$ is an induced subgraph of $\mathcal{G}$. We shall consider only induced subgraphs in what follows. For a subset $A \subset V$, the induced subgraph $\mathcal{G}_A = (A, A \times A \cap E)$ is said to be the graph induced by A. A graph $\mathcal{G}$ is called complete if every pair of vertices are adjacent. A clique of $\mathcal{G}$ is an induced complete subgraph of $\mathcal{G}$ that is not a subset of any other induced complete subgraphs of $\mathcal{G}$. More simply, a subset $A \subset V$ is called a clique if the induced subgraph $\mathcal{G}_A$ is a clique of $\mathcal{G}$. The set of the cliques of $\mathcal{G}$ is denoted by $\mathscr{C}_{\mathcal{G}}$.

A path in $\mathcal{G}$ of length $n \geq 1$ from a vertex i to a vertex j is a finite sequence of distinct vertices $i_0 = i, \ldots, i_n = j$ in V such that $(i_{\nu-1}, i_\nu)$ or $\{i_{\nu-1}, \nu_k\}$ are in E for each $\nu = 1, \ldots, n$. We say that the path is directed if at least one of the edges is directed. We say i leads to j, denoted by $i \to \cdots \to j$, if there is a directed path from i to j. A graph $\mathcal{G} = (V, E)$ is called connected if for any pair of distinct vertices $i, j \in V$ there exists a path between them. An n-cycle in $\mathcal{G}$ is a path of length n with the additional requirement that the end points are identical. A directed n-cycle is defined accordingly.

An undirected graph, which we denote by $\mathcal{G} = (V, \mathscr{E})$, is a graph with all of its edges undirected. The undirected graph $\mathcal{G}$ is said to be decomposable if it has no induced cycle of length greater than or equal to four, excluding the loops. A constructive definition in terms of the cliques and the separators of the graph $\mathcal{G}$ can also be specified (The reader is referred to Lauritzen (1996) for details.) A directed graph, denoted by $\mathcal{D} = (V, E)$, is now a graph with all of its edges directed. The directed graph $\mathcal{D}$ is said to be acyclic if it has no cycles, excluding the loops. The undirected version of $\mathcal{D} = (V, E)$, denoted by $\mathcal{D}^u = (V, E^u)$, is

* U.S. Census Bureau, Washington, DC 20233 e.h.bendavid@gmail.com
†Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107 tianxili@umich.edu
‡Department of Mathematics and Statistics,York University, Toronto, ON M3J 1P3 massamh@mathstat.yorku.ca
§Department of Statistics, Stanford University, Stanford, CA 94305 brajaratnam01@gmail.com

 DOI: 0000

the undirected graph obtained by replacing all the directed edges of $\mathcal{D}$ by undirected ones. An immorality in a directed graph $\mathcal{D}$ is an induced subgraph of the from $i \to k \leftarrow j$. Moralizing an immorality entails adding an undirected edge between the pair of parents that have the same children. Then the moral graph of $\mathcal{D}$, denoted by $\mathcal{D}^p = (V, E^p)$, is the undirected graph obtained by first moralizing each immorality of $\mathcal{D}$ and then making the undirected version of the resulting graph. A DAG $\mathcal{D}$ is said to be perfect if it has no immoralities; that is the parents of all vertices are adjacent, or equivalently if the set of parents of each vertex induces a complete subgraph of $\mathcal{D}$. Decomposable (undirected) graphs and (directed) perfect graphs have a deep connection. In particular, when $\mathcal{G}$ is decomposable there exists a DAG version of $\mathcal{G}$, that is a DAG $\mathcal{D}$ with the property that $\mathcal{D}^u = \mathcal{G}$, such that $\mathcal{D}$ is perfect (Lauritzen, 1996).

Given a DAG, the set of ancestors of a vertex j, denoted by $\mathrm{an}(j)$, is the set of those vertices i such that $i \to \cdots \to j$. Similarly, the set of descendants of a vertex i, denoted by $\mathrm{de}(i)$, is the set of those vertices j such that $i \to \cdots \to j$. The set of non-descendants of i is $\mathrm{nd}(i) = V \setminus (\mathrm{de}(i) \cup \{i\})$. A set $A \subseteq V$ is called ancestral when A contains the parents of its members. The smallest ancestral set containing the subset B of V is denoted by $\mathrm{An}(B)$.

1.2 Markov properties for DAG models

Let V be a finite set of indices and $X = (X_i)_{i \in V}$ a collection of random variables, where each X_i is a random variable on the probability space $\mathcal{X}_i$. Let the probability space $\mathcal{X}$ be defined as the product space $\mathcal{X} = \times_{i \in V} \mathcal{X}_i$. Now let $\mathcal{D} = (V, E)$ be a DAG. For simplicity, and without loss of generality, we always assume that the given $\mathcal{D}$ is connected and the edge set E contains all the loops $(i, i), i \in V$. We say that a probability distribution P on $\mathcal{X}$ has the recursive factorization property with respect to $\mathcal{D}$, denoted by DF (the directed factorization property), if there are σ-finite measures μ_i on $\mathcal{X}_i$ and non-negative functions $k^i(x_i, x_{pa(i)})$, referred to as kernels, defined on $\mathcal{X}_{fa(i)}$ such that, for each $i \in V$,

$$\int k^i(y_i, x_{pa(i)}) d\mu_i(y_i) = 1.$$

The density of P with respect to the product measure $\mu = \otimes_{i \in V} \mu_i$ is

$$p(x) = \prod_{i \in V} k^i(x_i, x_{pa(i)}).$$

In this case, each kernel $k^i(x_i, x_{\mathrm{pa}(i)})$ is in fact a version of $p(x_i \mid x_{\mathrm{pa}(i)})$, the conditional distribution of X_i given $X_{\mathrm{pa}(i)}$. An immediate consequence of this definition is the following lemma.

Lemma 1.1. *(Lauritzen, 1996) If P admits a recursive factorization with respect to the directed graph $\mathcal{D}$, then it also admits a factorization with respect to the undirected graph $\mathcal{D}^{\mathrm{m}}$, and, consequently, obeys the global Markov property with respect to $\mathcal{D}^{\mathrm{m}}$.*

Proof. For each vertex $i \in V$ the set $fa(i)$ is a complete subset of $\mathcal{D}^{\mathrm{m}}$. Thus if we define $\psi_{fa(i)}(x_{fa(i)}) = k^i(x_i, x_{\mathrm{pa}(i)})$, then $p(x) = \prod_{i \in V} p(x_i \mid x_{\mathrm{pa}(i)}) = \prod_{i \in V} k^i(x_i, x_{\mathrm{pa}(i)}) = \prod_{i \in V} \psi_{fa(i)}(x_{fa(i)})$. Therefore, P admits a factorization with respect to $\mathcal{D}^{\mathrm{m}}$ and by proposition 3.8 in Lauritzen (1996) it also obeys the global Markov property with respect to $\mathcal{D}^{\mathrm{m}}$. □

Another direct implication of the DF property is that if P admits a recursive factorization with respect to $\mathcal{D}$, then, for each ancestral set A, the marginal distribution P_A admits a recursive factorization with respect to the induced graph $\mathcal{D}_A$. Combining this result with Lemma 1.1 we obtain the following: P admits

a recursive factorization with respect to $\mathcal{D}$ then $A \perp\!\!\!\perp B \mid S\ [P]$, whenever A and B are separated by S in $(\mathcal{D}_{An(A\cup B\cup S)})^{\mathrm{m}}$. We call this property the directed global Markov property, DG, and any distribution that satisfies this property is said to be a directed Markov field over $\mathcal{D}$. For DAGs the directed Markov property plays the same role as the global Markov property does for undirected graphs, in the sense that it provides an optimal rule for recovering the conditional independence relations encoded by the directed graph.

We now introduce below another Markov property for DAGs. A distribution P on $\mathcal{X}$ is said to obey the directed local Markov property (DL) with respect to $\mathcal{D}$ if for each $i \in V$

$$i \perp\!\!\!\perp nd(i) \mid \mathrm{pa}(i).$$

Now for a given DAG $\mathcal{D}$ consider the parent-order graph defined as a DAG obtained by relabeling the vertices in V with $1, 2, \ldots, |V|$, in such a way that $\mathrm{pa}(i) \subseteq \{i+1, \ldots, |V|\}$ for each vertex $i \in V$. When $\mathcal{D}$ is parent ordered P obeys the parent-order-Markov property (PO) with respect to $\mathcal{D}$ if for every vertex i,

$$i \perp\!\!\!\perp \{i+1, \ldots, |V|\} \setminus \mathrm{pa}(i) \mid \mathrm{pa}(i).$$

Suppose P has a density with respect to μ. Then P obeys one of the directed Markov properties DF, DG, DL, PO if and only if it obeys all of them, that is the four Markov properties for DAGs are equivalent under a mild condition (Lauritzen, 1996).

1.3 Linear recursive properties of Gaussian DAGs

Let $X = (X_1, \ldots, X_p)^\top$ be a random vector in $\mathbb{R}^p$ with the multivariate distribution $\mathrm{N}_p(0, \Sigma)$. Consider the system of linear recursive regression equations:

$$\begin{array}{rcl} X_1 & = & -\beta_{12}X_2 - \beta_{13}X_3 - \cdots - \beta_{1p}X_p + \epsilon_1 \\ X_2 & = & -\beta_{23}X_3 - \cdots - \beta_{2p}X_p + \epsilon_2 \\ & \vdots & \\ X_p & = & \epsilon_p, \end{array}$$

where $-\beta_{ij}$ is the partial regression coefficient of X_j $(j > i)$ in the regression of X_i on its predecessors $X_{i+1}, \ldots, X_j, \ldots, X_p$. Now β_{ij} is zero if and only if $i \perp\!\!\!\perp \{i+1, \ldots, |V|\} \setminus \mathrm{pa}(i) \mid \mathrm{pa}(i)$. Hence the partial regression coefficient β_{ij} is zero if there does not exist an arrow from j to i, that is $j \notin \mathrm{pa}(i)$, $j > i$. In addition, the residuals ϵ_i are normally distributed and mutually independent with mean zero and variance $\sigma^2_{ii|\mathrm{pa}(i)}$. We can rewrite the first system of equations in the form of a linear system $BX = \epsilon$, where B is the upper triangular matrix

$$B = \begin{pmatrix} 1 & \beta_{12} & \ldots & \beta_{1p} \\ 0 & 1 & \ldots & \beta_{2p} \\ & & & \\ 0 & \ldots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \quad \text{and} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}.$$

From this we obtain

$$Var(BX) = Var(\epsilon)$$

$$\begin{aligned}
&\Rightarrow B\Sigma B^\top = Diag(\sigma^2_{1|\mathrm{pa}(1)}, \ldots, \sigma^2_{p-1|\mathrm{pa}(p-1)}, \sigma^2_{pp}) =: D \\
&\Rightarrow \Sigma = B^{-1} D (B^\top)^{-1} \\
&\Rightarrow \Sigma^{-1} = B^\top D^{-1} B. \qquad (1)
\end{aligned}$$

Let $L = B^\top$. Then $\Sigma^{-1} = LD^{-1}L^\top$ is the modified Cholesky decomposition of Σ^{-1}, in terms of the lower triangular matrix L and the diagonal matrix D^{-1}. Let $\mathcal{D} = (V, E)$ denote a DAG. The Gaussian distribution $\mathrm{N}_p(0, \Sigma)$ obeys the directed Markov property with respect to $\mathcal{D}$ if and only if $L_{ij} = 0$ whenever there is no arrow between i and j (Wermuth, 1980). Thus (1) gives a convenient description of $\mathscr{N}(\mathcal{D})$. We explore this model in more detail below.

2 The DAG-Wishart distributions

2.1 The normalizing constant of the DAG-Wishart distribution

Theorem 2.1. *Let* $dL = \prod_{(i,j)\in E, i>j} dL_{ij}$ *and* $dD = \prod_{i=1}^p dD_{ii}$ *denote, respectively, the canonical Lebesgue measures on* $\mathcal{L}_\mathcal{D}$ *and* $\mathbb{R}^p_+$ *and let* $\mathrm{pa}_i = |\mathrm{pa}(i)|$. *Then,*

$$z_\mathcal{D}(U, \alpha) = \int_{\Theta_\mathcal{D}} \exp\left\{-\frac{1}{2}\mathrm{tr}(LD^{-1}L^\top U)\right\} \prod_{i=1}^p D_{ii}^{-\alpha_i/2} dL dD < \infty$$

if and only if $\alpha_i > \mathrm{pa}_i + 2$ *for each* $i = 1, \ldots, p$. *Furthermore, in this case*

$$z_\mathcal{D}(U, \alpha) = \prod_{i=1}^p \frac{\Gamma\left(\frac{\alpha_i}{2} - \frac{\mathrm{pa}_i}{2} - 1\right) 2^{\alpha_i/2-1} (\sqrt{\pi})^{\mathrm{pa}_i} \det(U_{\mathrm{pa}(i)})^{\alpha_i/2-\mathrm{pa}_i/2-3/2}}{\det(U_{\mathrm{fa}(i)})^{\alpha_i/2-\mathrm{pa}_i/2-1}}. \qquad (2)$$

Proof. First we integrate out the terms involving D_{ii}'s.

$$\begin{aligned}
& \int \exp\left[-\frac{1}{2}\mathrm{tr}\left\{\left(LD^{-1}L^\top\right)U\right\}\right] \prod_{i=1}^p D_{ii}^{-\alpha_i/2} dL dD \\
= & \int \exp\left[-\frac{1}{2}\mathrm{tr}\left\{D^{-1}\left(L^\top U L\right)\right\}\right] \prod_{i=1}^p D_{ii}^{-\alpha_i/2} dL dD \\
= & \int \exp\left\{-\frac{1}{2}\sum_{i=1}^p D_{ii}^{-1}(L^\top U L)_{ii}\right\} \prod_{i=1}^p D_{ii}^{-\alpha_i/2} dD dL \\
= & \int \left[\prod_{i=1}^p \int \exp\left\{-\frac{1}{2} D_{ii}^{-1}(L^\top U L)_{ii}\right\} D_{ii}^{-\alpha_i/2} dD_{ii}\right] dL \\
= & \int \prod_{i=1}^p \frac{\Gamma\left(\frac{\alpha_i}{2} - 1\right) 2^{\alpha_i/2-1}}{\left((L^\top U L)_{ii}\right)^{\alpha_i/2-1}} dL \qquad (\text{ if and only if } \alpha_i > 2 \text{ for each } i = 1, 2, \ldots, p) \\
= & \int \prod_{i=1}^p \frac{\Gamma\left(\frac{\alpha_i}{2} - 1\right) 2^{\alpha_i/2-1}}{\left((L_{\cdot i})^\top U L_{\cdot i}\right)^{\alpha_i/2-1}} dL
\end{aligned}$$

$$= \int \prod_{i=1}^{p} \frac{\Gamma\left(\frac{\alpha_i}{2}-1\right) 2^{\alpha_i/2-1}}{\left\{\begin{pmatrix}1 & L_{\mathrm{pa}(i),i}^\top\end{pmatrix}\begin{pmatrix}U_{ii} & U_{i,\mathrm{pa}(i)} \\ U_{\mathrm{pa}(i),i} & U_{\mathrm{pa}(i)}\end{pmatrix}\begin{pmatrix}1 \\ L_{\mathrm{pa}(i),i}\end{pmatrix}\right\}^{\alpha_i/2-1}} dL$$

$$= \prod_{i=1}^{p} \int_{\mathbb{R}^{\mathrm{pa}_i}} \frac{\Gamma\left(\frac{\alpha_i}{2}-1\right) 2^{\alpha_i/2-1}}{\left\{\begin{pmatrix}1 & L_{\mathrm{pa}(i),i}^\top\end{pmatrix}\begin{pmatrix}U_{ii} & U_{i,\mathrm{pa}(i)} \\ U_{\mathrm{pa}(i),i} & U_{\mathrm{pa}(i)}\end{pmatrix}\begin{pmatrix}1 \\ L_{\mathrm{pa}(i),i}\end{pmatrix}\right\}^{\alpha_i/2-1}} dL_{\mathrm{pa}(i),i}. \qquad (\star)$$

We evaluate this integral by considering a more general form:

$$\int_{\mathbb{R}^d} \frac{dx}{\left\{\begin{pmatrix}1 & x^\top\end{pmatrix}\begin{pmatrix}\lambda & b^\top \\ b & A\end{pmatrix}\begin{pmatrix}1 \\ x\end{pmatrix}\right\}^{\gamma}}$$

where the block partitioned matrices, formed by $\lambda \in \mathbb{R}$, $b \in \mathbb{R}^d$ and the $(d-1)\times(d-1)$ matrix A, is positive definite. To simplify the integral above we proceed in two steps.

1) When $x \in \mathbb{R}$, by the formula provided on (Diaconis et al., 2008, page 16) we have the one dimensional integral

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

The d-dimensional version of this integral by repeated application of the right-hand-side formula can be computed as

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

2) Consider the general integral

$$\int_{\mathbb{R}^d} \frac{dx}{\left\{\begin{pmatrix}1 & x^\top\end{pmatrix}\begin{pmatrix}\lambda & b^\top \\ b & A\end{pmatrix}\begin{pmatrix}1 \\ x\end{pmatrix}\right\}^{\gamma}}.$$

Under the linear transformation $y = A^{1/2}x + A^{-1/2}b$, for $\gamma > \frac{d}{2}$ we have

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{dx}{\left\{\begin{pmatrix}1 & x^\top\end{pmatrix}\begin{pmatrix}\lambda & b^\top \\ b & A\end{pmatrix}\begin{pmatrix}1 \\ x\end{pmatrix}\right\}^{\gamma}} &= \frac{1}{\det(A)^{1/2}} \int_{\mathbb{R}^d} \frac{1}{(y^\top y + a - b^\top A^{-1} b)^\gamma} dy \\ &= \frac{(\sqrt{\pi})^d \Gamma\left(\gamma - \frac{d}{2}\right)}{\Gamma(\gamma)\det(A)^{1/2}\left(a - b^\top A^{-1} b\right)^{\gamma - d/2}}. \end{aligned} \qquad (3)$$

By applying (3) to the integral in $(\star)$ we obtain

$$z_{\mathcal{D}}(U,\alpha) = \prod_{i=1}^{p} \int_{\mathbb{R}^{\mathrm{pa}_i}} \frac{\Gamma\left(\frac{\alpha_i}{2}-1\right) 2^{\alpha_i/2-1}}{\left\{\begin{pmatrix}1 & L_{\mathrm{pa}(i),i}^\top\end{pmatrix}\begin{pmatrix}U_{ii}U_{i,\mathrm{pa}(i)} & \\ U_{\mathrm{pa}(i),i} & U_{\mathrm{pa}(i)}\end{pmatrix}\begin{pmatrix}1 \\ L_{\mathrm{pa}(i),i}\end{pmatrix}\right\}^{\alpha_i/2-1}} dL_{\mathrm{pa}(i),i}$$

$$= \prod_{i=1}^{p} \frac{\Gamma\left(\frac{\alpha_i}{2} - \frac{\mathrm{pa}_i}{2} - 1\right) 2^{\alpha_i/2-1} (\sqrt{\pi})^{\mathrm{pa}_i} \det(U_{\mathrm{pa}(i)})^{\alpha_i/2-\mathrm{pa}_i/2-3/2}}{\det(U_{\mathrm{fa}(i)})^{\alpha_i/2-\mathrm{pa}_i/2-1}},$$

where $\det(U_{\mathrm{pa}(i)}) = 1$ whenever $pa(i) = \emptyset$. Thus $z_{\mathcal{D}}(U, \alpha)$ is finite if and only if $\alpha_i > \mathrm{pa}_i + 2$ for each $i = 1, \ldots, p$. □

2.2 Hyper Markov properties

Theorem 2.2. *Let $(D, L) \sim \pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$. Then $\{(D_{ii}, L_{\mathrm{pa}(i),i}) : i = 1, \ldots, p\}$ are mutually independent. Moreover,*

$$D_{ii} \sim IG\left(\frac{\alpha_i}{2} - \frac{\mathrm{pa}_i}{2} - 1, \frac{1}{2} U_{ii|\mathrm{pa}(i)}\right), \text{ and} \tag{4}$$

$$L_{\mathrm{pa}(i),i} \mid D_{ii} \sim \mathrm{N}_{\mathrm{pa}_i}\left(-U_{\mathrm{pa}(i)}^{-1} U_{\mathrm{pa}(i),i}, D_{ii} U_{\mathrm{pa}(i)}^{-1}\right). \tag{5}$$

Proof. First consider the bijective mapping from the Cholesky parametrization to the D-parametrization:

$$\phi \equiv \left((D, L) \mapsto \times_{i \in V}(D_{ii}, L_{\mathrm{pa}(i),i})\right) : \Theta_{\mathcal{D}} \to \Xi_{\mathcal{D}}, \tag{6}$$

with the inverse mapping $\left(\times_{i \in V}(\lambda_i, \beta_{\mathrm{pa}(i)}) \mapsto (D, L)\right) : \Xi_{\mathcal{D}} \to \Theta_{\mathcal{D}}$, where $D = \mathrm{Diag}(\lambda_1 \ldots, \lambda_p)$ and

$$L_{ij} = \begin{cases} 1 & i = j \\ L_{ij} = \beta_{ij} & i \in pa(j) \\ 0 & \text{otherwise} \end{cases}$$

Since $\beta_{\mathrm{pa}(j)} = (\beta_{ij} : i \in pa(j))$ is in $\mathbb{R}^{\mathrm{pa}(j)}$, the mapping $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ induces a prior on $\Xi_{\mathcal{D}}$ that we denote by $\pi_{U,\alpha}^{\Xi_{\mathcal{D}}}$. The Cholesky parametrization and D-parametrization of $\mathscr{N}(\mathcal{D})$ are essentially the same, because, as the mapping ϕ shows, each element in one is an arrangement of the partial regression coefficients and partial variances given by the system of recursive regression equations. Thus the Jacobian of ϕ is equal to 1. To derive the density of $\pi_{U,\alpha}^{\Xi_{\mathcal{D}}}$ it suffices to find an expression for $\mathrm{tr}\left\{(LD^{-1}L^{\top})U\right\}$ in terms of $\prod_{i \in V}(D_{ii}, L_{\mathrm{pa}(i),i})$. To this end, we proceed as follows.

$$\begin{aligned}
&\mathrm{tr}\left\{(LD^{-1}L^{\top})U)\right\} = \mathrm{tr}\left\{(D^{-1}L^{\top})UL\right\} = \sum_{i \in V} D_{ii}^{-1}(L^{\top}UL)_{ii} \\
&= \sum_{i \in V} D_{ii}^{-1} \left(\sum_{k,l \in V} L_{ki} U_{kl} L_{li}\right) \\
&= \sum_{i \in V} D_{ii}^{-1} \begin{pmatrix} 1 \\ L_{\mathrm{pa}(i),i} \end{pmatrix}^{\top} \begin{pmatrix} U_{ii} & U_{i,\mathrm{pa}(i)} \\ U_{\mathrm{pa}(i),i} & U_{\mathrm{pa}(i)} \end{pmatrix} \begin{pmatrix} 1 \\ L_{\mathrm{pa}(i),i} \end{pmatrix} \\
&= \sum_{i \in V} D_{ii}^{-1} \left(U_{ii} + L_{\mathrm{pa}(i),i}^{\top} U_{\mathrm{pa}(i),i} + U_{i,\mathrm{pa}(i)} L_{\mathrm{pa}(i),i} + L_{\mathrm{pa}(i),i}^{\top} U_{\mathrm{pa}(i)} L_{\mathrm{pa}(i),i}\right) \\
&= \sum_{i \in V} \left\{ D_{ii}^{-1} \left(L_{\mathrm{pa}(i),i} + U_{\mathrm{pa}(i)}^{-1} U_{\mathrm{pa}(i),i}\right)^{\top} U_{\mathrm{pa}(i)} \left(L_{\mathrm{pa}(i),i} + U_{\mathrm{pa}(i)}^{-1} U_{\mathrm{pa}(i),i}\right) + D_{ii}^{-1} U_{ii|\mathrm{pa}(i)} \right\}.
\end{aligned}$$

Therefore, the density of $\pi_{U,\alpha}^{\Xi_{\mathcal{D}}}$ with respect to Lebesgue measure $\prod_{i\in V} d\lambda_i d\beta_{\mathrm{pa}(i)}$ on $\times_{i\in V}(\mathbb{R}_+, \mathbb{R}^{\mathrm{pa}(i),i})$ is

$$z_{\mathcal{D}}(\alpha,U)^{-1}\prod_{i\in V}\exp\left[-\frac{1}{2}\left\{\lambda_i^{-1}(\beta_{\mathrm{pa}(i)}+U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i})^\top U_{\mathrm{pa}(i)}(\beta_{\mathrm{pa}(i)}+U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i})+\lambda_i^{-1}U_{ii|\mathrm{pa}(i)}\right\}\right]\lambda_i^{-\alpha_i/2}. \tag{7}$$

Thus $\left(\lambda_i, \beta_{\mathrm{pa}(i)}\right)$ for $i = 1,\ldots,p$ are mutually independent. To complete the proof we first integrate out $\beta_{\mathrm{pa}(i)}$ to obtain the marginal density of λ_i. In (7) each exponential terms as a function of $\beta_{\mathrm{pa}(i)}$ is a unnormalized multivariate normal density, which shows the marginal density of λ_i is

$$\begin{aligned}&\int_{\mathbb{R}^{\mathrm{pa}(i),i}}\exp\left\{-\frac{1}{2}\lambda_i^{-1}(\beta_{\mathrm{pa}(i)}+U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i})^\top U_{\mathrm{pa}(i)}(\beta_{\mathrm{pa}(i)}-U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i})+\lambda_i^{-1}U_{ii|\mathrm{pa}(i)}\right\}\lambda_i^{-\alpha_i/2}d(\beta_{\mathrm{pa}(i)})\\ &\propto\exp\left\{-\frac{1}{2}\lambda_i^{-1}U_{ii|\mathrm{pa}(i)}\right\}\prod_{i\in V}\lambda^{-\alpha_i/2+\mathrm{pa}_i/2}.\end{aligned} \tag{8}$$

The above shows that

$$\lambda_i \sim IG\left(\frac{\alpha_i}{2}-\frac{\mathrm{pa}_i}{2}-1, \frac{1}{2}U_{ii|\mathrm{pa}(i)}\right).$$

By dividing the i-th factor of (7) by the marginal density of λ_i we conclude that

$$\beta_{\mathrm{pa}(i)} \mid \lambda_i \sim \mathrm{N}_{\mathrm{pa}_i}\left(-U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i}, \lambda_i U_{\mathrm{pa}(i)}^{-1}\right).$$

The same result holds for $(D_{ii}, L_{\mathrm{pa}(i),i}),\ i = 1,2,\ldots,p$ as specified in the statement of Theorem 2.2. □

The converse part of Theorem 2.2 is left to the reader.

Corollary 2.1. *Let $\mathcal{D}$ be an arbitrary DAG and suppose $(L,D) \sim \pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$. Then the density of L with respect to $dL = \prod_{i=1}^p dL_{\mathrm{pa}(i),i}$ is*

$$\prod_{i=1}^p c_i\left\{\frac{1}{2}U_{ii|\mathrm{pa}(i)}+(L_{\mathrm{pa}(i),i}+U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i})^\top U_{\mathrm{pa}(i)}(L_{\mathrm{pa}(i),i}+U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i})\right\}^{-\alpha_i/2+1},$$

where

$$c_i = \frac{\det(U_{\mathrm{pa}(i)})^{1/2}(U_{ii|\mathrm{pa}(i)})^{\alpha_i/2-\mathrm{pa}_i/2-1}\Gamma\left(\frac{\alpha_i}{2}-1\right)}{2^{\alpha_i/2-1}\pi^{\mathrm{pa}_i/2}\Gamma\left(\frac{\alpha_i}{2}-\frac{\mathrm{pa}_i}{2}-1\right)}. \tag{9}$$

Proof. Using Theorem 2.2 we have

$$\begin{aligned}&\int\frac{1}{(2\pi)^{\mathrm{pa}_i/2}\det(D_{ii}U_{\mathrm{pa}(i)}^{-1})^{1/2}}\exp\left\{(L_{\mathrm{pa}(i),i}+U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i})^\top(D_{ii}^{-1}U_{\mathrm{pa}(i)})(L_{\mathrm{pa}(i),i}+U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i})\right\}\\ &\times\frac{(1/2U_{ii|\mathrm{pa}(i)})^{\alpha_i/2-\mathrm{pa}_i/2-1}}{\Gamma\left(\frac{\alpha_i}{2}-\frac{\mathrm{pa}_i}{2}-1\right)}D_{ii}^{-\alpha_i/2+\mathrm{pa}_i/2}\exp\left\{-1/2U_{ii|\mathrm{pa}(i)}D_{ii}^{-1}\right\}dD_{ii}\\ =&\frac{\det(U_{\mathrm{pa}(i)})^{1/2}(U_{ii|\mathrm{pa}(i)})^{\alpha_i/2-\mathrm{pa}_i/2-1}}{2^{\alpha_i/2-1}\pi^{\mathrm{pa}_i/2}\Gamma\left(\frac{\alpha_i}{2}-\frac{\mathrm{pa}_i}{2}-1\right)}\int D_{ii}^{-\alpha_i/2}\exp\left(-u_iD_{ii}^{-1}\right)dD_{ii}\\ =&\frac{\det(U_{\mathrm{pa}(i)})^{1/2}(U_{ii|\mathrm{pa}(i)})^{\alpha_i/2-\mathrm{pa}_i/2-1}}{2^{\alpha_i/2-1}\pi^{\mathrm{pa}_i/2}\Gamma\left(\frac{\alpha_i}{2}-\frac{\mathrm{pa}_i}{2}-1\right)}\times\frac{\Gamma\left(\frac{\alpha_i}{2}-1\right)}{u_i^{\alpha_i/2-1}},\end{aligned}$$

where $u_i = \frac{1}{2}U_{ii|\mathrm{pa}(i)} + (L_{\mathrm{pa}(i),i} + U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i})^\top U_{\mathrm{pa}(i)}(L_{\mathrm{pa}(i),i} + U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i})$. Therefore the density of $L_{\mathrm{pa}(i),i}$is

$$c_i \left(1/2U_{ii|\mathrm{pa}(i)} + (L_{\mathrm{pa}(i),i} + U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i})^\top U_{\mathrm{pa}(i)}(L_{\mathrm{pa}(i),i} + U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i})\right)^{-\alpha_i/2+1}. \tag{10}$$

By Theorem 2.2, $L_{\mathrm{pa}(i),i}$ are mutually independent, hence the form of the density in the statement of the corollary is immediate from the above calculations. The parameters corresponding to the t-distribution follow by comparing the density in (10) to the functional form of the density of the multivariate t-distribution. □

2.3 The posterior distribution

Proposition 2.3. *Let $\mathcal{D}$ be a DAG and let $x_1, \ldots, x_n$ be i.i.d. observations from $\mathrm{N}_p(\mathbf{0}, (L^{-1})^\top DL^{-1})$, where $(D, L) \in \Theta_\mathcal{D}$. Let $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ denote the empirical covariance matrix. If the prior distribution on (D, L) is $\pi^{\Theta_\mathcal{D}}_{U,\alpha}$, then the posterior distribution of (D, L) is $\pi_{\widetilde{U},\widetilde{\alpha}^{\Theta_\mathcal{D}}}$, where $\widetilde{U} = nS + U$ and $\widetilde{\alpha} = (n + \alpha_1, n + \alpha_2, \ldots, n + \alpha_p)$.*

Proof. The likelihood of the data is given as follows:

$$L\left((L, D)|\, x_1, \ldots, x_n\right) = \frac{1}{(\sqrt{2\pi})^{np}} \exp\left[-\frac{1}{2}\mathrm{tr}\left\{LD^{-1}L^\top(nS)\right\}\right] \det(D)^{-n/2}.$$

When using $\pi^{\Theta_\mathcal{D}}_{U,\alpha}$ as the prior for (D, L), the posterior distribution of (D, L) given the random sample $(x_1, \ldots, x_n)$ is

$$\pi^{\Theta_\mathcal{D}}_{U,\alpha}(L, D|\, x_1, \ldots, x_n) \propto \exp\left[-\frac{1}{2}\mathrm{tr}\left\{LD^{-1}L^\top(nS + U)\right\}\right] \prod_{i=1}^p D_{ii}^{-n/2-\alpha_i/2}, \quad (D, L) \in \Theta_\mathcal{D}. \tag{11}$$

Hence the functional form of the posterior density is the same as that of the prior density, that is,

$$\pi^{\Theta_\mathcal{D}}_{U,\alpha}(\cdot \mid x_1, \ldots, x_n) = \pi_{\widetilde{U},\widetilde{\alpha}}(\cdot),$$

where $\widetilde{U} = nS + U$ and $\widetilde{\alpha} = (\alpha_1 + n, \ldots, \alpha_p + n)$. □

Remark 2.1. *The case of non-zero mean (that is, when $x_1, \ldots, x_n$ are i.i.d. observations from $\mathrm{N}_p(\mu, \Sigma)$ with $\mu \neq 0$) can be handled with a similar manner, because the sample covariance matrix S is a sufficient statistic for Σ and $nS \sim \mathrm{W}_p(n - 1, \Sigma)$.*

2.4 The Laplace transform

We compute the Laplace transform of $\pi^{\Xi_\mathcal{D}}_{U,\alpha}$ by exploiting the results established in Theorem 2.2. First a preliminary result on the Laplace transform of a Gaussian inverse-gamma distribution is required.

Lemma 2.4. *Suppose (g, X) is a random variable with Gaussian inverse-gamma distribution:*

$$X \mid g \sim \mathrm{N}_p(\mu, g\Psi), \quad \mu \in \mathbb{R}^p, \Psi \in \mathrm{PD}_p(\mathbb{R});$$

$$g \sim \mathrm{IG}(\nu, \eta).$$

Then the Laplace transform of (g, X) *at* $(\xi, u) \in \mathbb{R}_+ \times \mathbb{R}_+^p$ *is*

$$\frac{2}{\Gamma(\nu)} \exp\left(u^\top \mu\right) \left(\eta(\xi - \frac{1}{2} u^\top \Psi u)\right)^{\nu/2} K_\nu \left(2\sqrt{\eta(\xi - \frac{1}{2} u^\top \Psi u)}\right),$$

where $K_\nu(\cdot)$ *is the modified Bessel function of the second type and* $\xi - \frac{1}{2} u^\top \Psi u$ *is assumed to be positive.*

Proof. By definition, the Laplace transform of (g, X) at $(\xi, u) \in \mathbb{R} \times \mathbb{R}^p$ is

$$\begin{aligned}
&\int \exp\left\{-(g\xi + u^\top x)\right\} d\mathrm{N}_p(\mu, g\Psi)(x) d\mathrm{IG}(\nu, \eta)(g) \\
=&\int \exp\left(-g\xi\right) \left\{ \int \exp\left(-u^\top x\right) d\mathrm{N}_p(\mu, g\Psi)(x) \right\} d\mathrm{IG}(\nu, \eta)(g) \\
=&\int \exp\left(-g\xi\right) \exp\left(-u^\top \mu + \frac{1}{2} g u^\top \Psi u\right) d\mathrm{IG}(\nu, \eta)(g) \\
=&\int \exp\left(-g\xi\right) \exp\left(-u^\top \mu + \frac{1}{2} g u^\top \Psi u\right) \left(\frac{\eta^\nu}{\Gamma(\nu)} \exp\left(-\eta g^{-1}\right) g^{-\nu-1}\right) dg \\
=&\frac{\eta^\nu}{\Gamma(\nu)} \exp\left(-u^\top \mu\right) \int \exp\left\{-(\xi - \frac{1}{2} u^\top \Psi u) g - \eta g^{-1}\right\} g^{-\nu-1} dg \\
=&\frac{2\eta^\nu}{\Gamma(\nu)} \exp\left(-u^\top \mu\right) \left(\frac{\xi - \frac{1}{2} u^\top \Psi u}{\eta}\right)^{\nu/2} K_\nu \left(2\sqrt{\eta(\xi - \frac{1}{2} u^\top \Psi u)}\right) \\
=&\frac{2}{\Gamma(\nu)} \exp\left(-u^\top \mu\right) \left(\eta(\xi - \frac{1}{2} u^\top \Psi u)\right)^{\nu/2} K_\nu \left(2\sqrt{\eta(\xi - \frac{1}{2} u^\top \Psi u)}\right).
\end{aligned}$$

The integral above is computed using the fact that the Laplace transform of $\mathrm{N}_p(\mu, g\Psi)$ at u is equal to $\exp\left(-u^\top \mu + \frac{1}{2} g u^\top \Psi u\right)$. For computing the integral with respect to dg we use (9.42) in (Temme, 1996, page 235).

□

Proposition 2.5. *The Laplace transform of* $\pi_{U,\alpha}^{\Xi_\mathcal{D}}$ *at a typical point* $\times_{i=1}^p (\xi_i, z_{\mathrm{pa}(i),i}) \in \Xi_\mathcal{D}$ *is*

$$\mathcal{L}_{\Xi_\mathcal{D}}\left(\times_{i=1}^p (\xi_i, z_{\mathrm{pa}(i),i})\right) = 2^p \prod_{i=1}^p \frac{1}{\Gamma(r_i)} \exp\left(z_{\mathrm{pa}(i),i}^\top \mu_{\mathrm{pa}(i),i}\right) \left(\eta_i(\xi_i - \frac{1}{2} z_{\mathrm{pa}(i),i}^\top \Psi_{\mathrm{pa}(i)} z_{\mathrm{pa}(i),i})\right)^{r_i/2} \tag{12}$$

$$\times K_{r_i}\left(2\sqrt{\eta_i(\xi_i - \frac{1}{2} z_{\mathrm{pa}(i),i}^\top \Psi_{\mathrm{pa}(i)} z_{\mathrm{pa}(i),i})}\right), \tag{13}$$

where $r_i = \frac{\alpha_i}{2} - \frac{\mathrm{pa}_i}{2} - 1$, $\eta_i = \frac{1}{2} U_{ii|\mathrm{pa}(i)}$, $\mu_{\mathrm{pa}(i),i} = -U_{\mathrm{pa}(i)}^{-1} U_{\mathrm{pa}(i),i}$, $\Psi_{\mathrm{pa}(i)} = U_{\mathrm{pa}(i)}^{-1}$, *and* $\xi_i - \frac{1}{2} z_{\mathrm{pa}(i),i}^\top \Psi_{\mathrm{pa}(i)} z_{\mathrm{pa}(i),i}$ *are assumed to be positive for each* i.

Proof. Let $\times_{i=1}^{p}(\lambda_i, \beta_{\mathrm{pa}(i)}) \sim \pi_{U,\alpha}^{\Xi_{\mathcal{D}}}$. Theorem 2.2 implies that the finite sequence of random variables $(\lambda_i, \beta_{\mathrm{pa}(i)})$ are independent and each has a Gaussian inverse-gamma distribution as given by (4) and (5). It therefore suffices to compute the Laplace transform of each random vector $(\lambda_i, \beta_{\mathrm{pa}(i)})$ individually. The Laplace transform of $\pi_{U,\alpha}^{\Xi_{\mathcal{D}}}$ now follows immediately from Lemma 2.4. □

We now give the Laplace transform of $\pi_{U,\alpha}^{\Theta}$.

Corollary 2.2. *The Laplace transform of* $\pi_{U,\alpha}^{\Theta}$ *at* $(\Lambda, Z) \in \Theta_{\mathcal{D}}$ *is*

$$\left(\frac{2}{e}\right)^p \prod_{i=1}^{p} \left\{ \frac{1}{\Gamma(r_i)} \exp\left(z_{\mathrm{pa}(i),i}^{\top} \mu_{\mathrm{pa}(i),i}\right) \left(\eta_i(\xi_i - \frac{1}{2} z_{\mathrm{pa}(i),i}^{\top} \Psi_{\mathrm{pa}(i)} z_{\mathrm{pa}(i),i})\right)^{r_i/2} K_{r_i}\left(2\sqrt{\eta_i(\xi_i - \frac{1}{2} z_{\mathrm{pa}(i),i}^{\top} \Psi_{\mathrm{pa}(i)} z_{\mathrm{pa}(i),i})}\right) \right\}$$

Proof. By definition, the Laplace transform of $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ at $(\Lambda, Z) \in \Theta_{\mathcal{D}}$ is

$$\mathcal{L}_{\Theta_{\mathcal{D}}}(\Lambda, Z) = \int \exp\left[-\mathrm{tr}(\Lambda D^{\top}) - \mathrm{tr}(ZL^{\top})\right] \pi_{U,\alpha}^{\Theta_{\mathcal{D}}}(D, L) dD dL.$$

Now under the change of variable $\phi : \Theta_{\mathcal{D}} \to \Xi_{\mathcal{D}}$ defined in (6) and the fact that

$$\mathrm{tr}(\Lambda D^{\top}) + \mathrm{tr}(ZL^{\top}) = \sum_{i}^{p} D_{ii}\Lambda_{ii} + \sum_{i=1}^{p} \left(1 + L_{\mathrm{pa}(i),i}^{\top} Z_{\mathrm{pa}(i),i}\right)$$

we obtain

$$\begin{aligned}\mathcal{L}_{\Theta_{\mathcal{D}}}(\Lambda, Z) &= \int \exp\left\{-\sum_{i}^{p} D_{ii}\Lambda_{ii} - \sum_{i=1}^{p}\left(1 + L_{\mathrm{pa}(i),i}^{\top} Z_{\mathrm{pa}(i),i}\right)\right\} \pi_{U,\alpha}^{\Xi_{\mathcal{D}}}(\times_{i=1}^{p}(D_{ii}, L_{\mathrm{pa}(i),i})) \prod_{i=1}^{p} dD_{ii} dL_{\mathrm{pa}(i),i} \\ &= e^{-p} \mathcal{L}_{\Xi_{\mathcal{D}}}(\times_{i=1}^{n}(\Lambda_{ii}, Z_{\mathrm{pa}(i),i})).\end{aligned}$$

□

2.5 The expectation

We now compute the expected values of our priors. First some necessary notation is introduced: Suppose $a, b \subseteq V$ and $A \in \mathbb{R}^{a \times b}$ a matrix of size $|a| \times |b|$. Then define $(A)^0 \in \mathbb{R}^{V \times V}$ by

$$(A)_{ij}^{0} = \begin{cases} A_{ij} & i \in a, j \in b \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, if $L_{\mathrm{pa}(i),i}$ is a vector in $\mathbb{R}^{\mathrm{pa}(i),i}$, then we consider

$$\begin{pmatrix} 1 \\ L_{\mathrm{pa}(i),i} \end{pmatrix}$$

as a vector in $\mathbb{R}^{\mathrm{fa}(i)}$ with 1 in ii position.

Now recall from 2.1 that $L_{\mathrm{pa}(i),i}$ has a multivariate t-distribution. This result readily allows us to compute the mean and covariance of the random elements of L. They are

$$E\left(L_{\mathrm{pa}(i),i}\right) = -U^{-1}_{\mathrm{pa}(i)}\,U_{\mathrm{pa}(i),i} \text{ and } Var(L_{\mathrm{pa}(i),i}) = \frac{\nu_i^2}{2\nu_i - 4}\,U_{ii|\mathrm{pa}(i)}\,U^{-1}_{\mathrm{pa}(i)}.$$

Consequently, if $A = \{1, i_2, \ldots, i_r\} \subseteq V$ is the set of vertices i such that $pa(i) \neq \emptyset$, then $E\left(\times_{i\in A} L_{\mathrm{pa}(i),i}\right) = -\times_{i\in A}\, U^{-1}_{\mathrm{pa}(i)}\,U_{\mathrm{pa}(i),i}$. This can be expressed in matrix form as follows:

$$E\left(L\right) = E\left(\sum_{j=1}^{p}\left(L_{\cdot j}\right)^0\right) = \sum_{j=1}^{p}\begin{pmatrix}1\\ E\left(L_{\mathrm{pa}(i),i}\right)\end{pmatrix}^0 = \sum_{j=1}^{p}\begin{pmatrix}1\\ -U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i}\end{pmatrix}^0.$$

The expression for $Var(\times_{i\in A} L_{\mathrm{pa}(i),i})$ is the block diagonal matrix

$$\begin{pmatrix}
\frac{\nu_1^2}{2\nu_1-4}U_{11|\mathrm{pa}(1)}U^{-1}_{\mathrm{pa}(1)} & 0 & \cdots & 0\\
0 & \frac{\nu_{i_2}^2}{2\nu_{i_2}-4}U_{i_2i_2|\mathrm{pa}(i_2)}U^{-1}_{\mathrm{pa}(i_2)} & & \\
\vdots & & \ddots & \\
0 & & & \frac{\nu_{i_r}^2}{2\nu_{i_r}-4}U_{i_ri_r|\mathrm{pa}(i_r)}U^{-1}_{\mathrm{pa}(i_r)}
\end{pmatrix}.$$

The expected value of D can also be easily computed using the result in (4). Under the Cholesky decomposition parametrization we have $E\left(D\right) = \mathrm{Diag}\left(\dfrac{U_{ii|\mathrm{pa}(i)}}{\alpha_i - \mathrm{pa}_i - 4} : i \in V\right)$.

2.6 The posterior mode

The posterior mode of $\pi^{\Xi_{\mathcal{D}}}_{U,\alpha}$ is often a useful quantity in Bayesian inference. Let us first compute the mode of $\pi^{\Xi_{\mathcal{D}}}_{U,\alpha}$. From (7), the density of $\pi^{\Xi_{\mathcal{D}}}_{U,\alpha}$ is proportional to

$$\exp\left\{-\frac{1}{2}\sum_{i\in V}\lambda_i^{-1}(\beta_{\mathrm{pa}(i)} + U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i})^\top U_{\mathrm{pa}(i)}(\beta_{\mathrm{pa}(i)} + U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i})\right\}\exp\left(-\frac{1}{2}\lambda_i^{-1}U_{ii|\mathrm{pa}(i)}\right)\prod_{i\in V}\lambda_i^{-\alpha_i/2}.$$

For each λ_i, $\exp\left\{-\frac{1}{2}\lambda_i^{-1}(\beta_{\mathrm{pa}(i)} + U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i})^\top U_{\mathrm{pa}(i)}(\beta_{\mathrm{pa}(i)} + U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i})\right\}$ is maximized at

$$\beta_{\mathrm{pa}(i)} = -U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i}.$$

Notice that $\exp\left(-\frac{1}{2}\lambda_i^{-1}U_{ii|\mathrm{pa}(i)}\right)\prod_{i\in V}\lambda_i^{-\alpha_i/2}$ corresponds to the inverse-gamma distribution

$$IG\left(\frac{\alpha_i}{2} - 1, \frac{1}{2}U_{ii|\mathrm{pa}(i)}\right)$$

and thus its mode is equal to $\dfrac{U_{ii|\mathrm{pa}(i)}}{\alpha_i}$. Combining the results above we have the mode of $\pi^{\Xi_{\mathcal{D}}}_{U,\alpha}$ is

$$\times_{i=1}^{p}\left(\frac{U_{ii|\mathrm{pa}(i)}}{\alpha_i}, -U^{-1}_{\mathrm{pa}(i)}U_{\mathrm{pa}(i),i}\right).$$

Under other parametrizations the posterior modes can be computed with similar calculation.

2.7 The Jacobians of the mappings

To derive the density of $\pi_{U,\alpha}^{\mathrm{R}_\mathcal{D}}$ we need to compute the Jacobian of the mapping

$$\psi \equiv \left((L,D) \mapsto \left(LD^{-1}L^\top\right)^E\right) : \Theta_\mathcal{D} \to \mathrm{R}_\mathcal{D}.$$

The Jacobian of ψ is a variant of similar transformations found in (Roverato, 2000) and (**?**). For completeness we still compute this Jacobian in the following lemma.

Lemma 2.6. *The Jacobian of the mapping* $\psi : ((D,L) \mapsto \left(LD^{-1}L^\top\right)^E$ *is* $\prod_{j=1}^p D_{jj}^{-(\mathrm{pa}_j+2)}$.

Proof. Let $\Upsilon \in \mathrm{R}_\mathcal{D}$, and $(D,L) \in \Theta_\mathcal{D}$ such that $\widehat{\Upsilon} = LD^{-1}L^\top$ is the completion of Υ in $\mathrm{P}_\mathcal{D}$. For each $(i,j) \in E$,

$$\Upsilon_{ij} = (LD^{-1}L^\top)_{ij} = \sum_{k=1}^p L_{ik}L_{jk}D_{kk}^{-1} = \sum_{k=1}^j L_{ik}L_{jk}D_{kk}^{-1}, \tag{14}$$

since L is lower triangular. Now from (14) and the fact that $L_{jj} = 1$, for each j, we conclude that

$$\frac{\partial}{\partial L_{ij}}(LD^{-1}L^\top)_{ij} = D_{jj}^{-1}, \ \ (i,j) \in E, \qquad \frac{\partial}{\partial D_{ii}}(LD^{-1}L^\top)_{ii} = -D_{ii}^{-2}, \ \ i = 1,2,\cdots,p.$$

Arrange the entries of $(D,L) \in \Theta_\mathcal{D}$ as D_{11}, $\{L_{2k} : (2,k) \in E, 1 \le k < 2\}$, D_{22}, $\{L_{3k} : (3,k) \in E, 1 \le k < 3\}$, …, $D_{p-1,p-1}$, $\{L_{pk} : (p,k) \in E, 1 \le k < p\}$, D_{pp}, and the entries of $\Upsilon \in \mathrm{R}_\mathcal{D}$ as Υ_{11}, $\{\Upsilon_{2k} : (2,k) \in E, 1 \le k < 2\}$, Υ_{22}, $\{\Upsilon_{3k} : (3,k) \in E, 1 \le k < 3\}$, …, $\Upsilon_{p-1,p-1}$, $\{\Upsilon_{pk} : (p,k) \in E, 1 \le k < p\}$, Υ_{pp}. From (14) it is easily seen that Υ_{ij} depends on

$$\{L_{jk} : (j,k) \in E, 1 \le k < j\}, \ \{L_{ik} : (i,k) \in E, 1 \le k < j\} \text{ and } \{D_{kk}, 1 \le k \le j\}.$$

Hence it is clear that Υ_{ij} is functionally independent of elements of $\Theta_\mathcal{D}$ that follow it in the arrangement described above. Hence the gradient matrix of ψ (with this arrangement) is a lower triangular matrix, and the Jacobian of ψ is therefore

$$\prod_{i=1}^p \left(\prod_{j \in ch(i)} D_{jj}^{-1} \right) \prod_{i=1}^p D_{ii}^{-2}.$$

It follows from the expression above that the Jacobian of ψ is

$$\prod_{j=1}^p D_{jj}^{-(\mathrm{pa}_j+2)}.$$

□

To derive the density of $\pi_{U,\alpha}^{\mathrm{S}_\mathcal{D}}$ we need to compute the following Jacobian.

Lemma 2.7. *The Jacobian of the mapping* $(\Sigma^{-E} \mapsto \Sigma^E) : \mathrm{R}_{\mathcal{D}} \to \mathrm{S}_{\mathcal{D}}$ *is*

$$\prod_{i=1}^{p} \frac{\det \Sigma_{\mathrm{fa}(i)}^{\mathrm{pa}_i+2}}{\det \Sigma_{\mathrm{pa}(i)}^{\mathrm{pa}_i+1}}.$$

Proof. The mapping $\Sigma^{-E} \mapsto \Sigma^E$ can be written as the composition of the two mappings

$$(\Sigma^{-E} \mapsto \times_{i=1}^{p}(\Sigma_{ii|\mathrm{pa}(i)}, \Sigma_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i}) : \mathrm{R}_{\mathcal{D}} \to \Theta_{\mathcal{D}};$$
$$(\times_{i=1}^{p}(\Sigma_{ii|\mathrm{pa}(i)}, \Sigma_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i}) \mapsto \Sigma^E) : \Theta_{\mathcal{D}} \to \mathrm{S}_{\mathcal{D}}.$$

It is easy to check that the Jacobian of the first mapping is the same as the Jacobian of the inverse of the mapping $\psi : (L, D) \mapsto (LD^{-1}L^\top)^E$ in Lemma 2.6 and is therefore equal to $\prod_{i=1}^{p} \Sigma_{ii|\mathrm{pa}(i)}^{\mathrm{pa}_i+2}$.

By mathematical induction we compute the Jacobian of the second mapping. Assume that the Jacobian of the mapping

$$(\times_{i=1}^{p}(\Sigma_{ii|\mathrm{pa}(i)}, \Sigma_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i}) \mapsto \Sigma^E) : \Theta_{\mathcal{D}} \to \mathrm{S}_{\mathcal{D}}$$

is equal to $\prod_{i=1}^{|V|} \det(\Sigma_{\mathrm{pa}(i)})$ for any DAG $\mathcal{D}$ with $|V| < p$. We will show that the result holds true for $|V| = p$. The case $p = 1$ is trivial. So assume that $p \geq 2$. Let $\mathcal{D}_{[1]}$ be the induced subgraph of $\mathcal{D}$ with the vertex set $V_{[1]} = V \setminus [1]$ and the edge set $E_{[1]}$. Since $V_{[1]}$ is an ancestral subset of V, if Σ^E belongs to $\mathrm{S}_{\mathcal{D}}$, then $\Sigma^{E_{[1]}}$, the projection of Σ on $I_{G_{[1]}}$, is an element of $\mathrm{S}_{\mathcal{D}_{[1]}}$. Furthermore, the positive definite completion of $\Sigma^{E_{[1]}}$ in $\mathrm{PD}_{\mathcal{D}_{[1]}}$ is the principal sub-matrix $\Sigma_{V_{[1]}}$. The observations above follow from the recursive nature of the completion process. We now decompose the inverse mapping $\Sigma^E \mapsto \times_{i=1}^{p}(\Sigma_{ii|\mathrm{pa}(i)}, \Sigma_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i})$ as

$$\begin{array}{ccccc} \mathrm{S}_{\mathcal{D}} & \to & \mathbb{R}_+ \times \mathbb{R}^{\mathrm{pa}(1),1} \times \mathrm{S}_{\mathcal{D}_{[1]}} & \to & \mathbb{R}_+ \times \mathbb{R}^{\mathrm{pa}(1),1} \times \Theta_{\mathcal{D}_{[1]}} = \Theta_{\mathcal{D}} \\ \Sigma^E & \mapsto & (\Sigma_{11|\mathrm{pa}(1)}, \Sigma_{\mathrm{pa}(1)}^{-1}\Sigma_{\mathrm{pa}(1),1}, \Sigma^{E_{[1]}}) & \mapsto & \left(\Sigma_{11|\mathrm{pa}(1)}, \Sigma_{\mathrm{pa}(1)}^{-1}\Sigma_{\mathrm{pa}(1),1}, \times_{i=2}^{p}(\Sigma_{ii|\mathrm{pa}(i)}, \Sigma_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i})\right) \end{array}$$

By the inductive hypothesis the Jacobian of the second mapping,

$$(\Sigma_{11|\mathrm{pa}(1)}, \Sigma_{\mathrm{pa}(1)}^{-1}\Sigma_{\mathrm{pa}(1),1}, \Sigma^{E_{[1]}}) \mapsto \left(\Sigma_{11|\mathrm{pa}(1)}, \Sigma_{\mathrm{pa}(1)}^{-1}\Sigma_{\mathrm{pa}(1),1}, \times_{i=2}^{p}(\Sigma_{ii|\mathrm{pa}(i)}, \Sigma_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i})\right),$$

is equal to $\prod_{i=2}^{p} \det(\Sigma_{\mathrm{pa}(i)})^{-1}$. Hence it suffices to prove that the Jacobian of the first mapping, $\Sigma^E = (\Sigma_{11}, \Sigma_{\mathrm{pa}(1),1}, \Sigma^{E_{[1]}}) \mapsto (\Sigma_{11|\mathrm{pa}(1)}, \Sigma_{\mathrm{pa}(1)}^{-1}\Sigma_{\mathrm{pa}(1),1}, \Sigma^{E_{[1]}})$ is $\det(\Sigma_{\mathrm{pa}(1)})^{-1}$. This follows by noting that the Jacobian matrix of this mapping is the lower triangular

$$\begin{pmatrix} I & 0 & 0 \\ * & \Sigma_{\mathrm{pa}(1)}^{-1} & 0 \\ * & * & 1 \end{pmatrix}$$

The results now follows by induction.

□

2.8 The DAG-Wishart distributions as a curved exponential family

We now analyze the $\pi^{\mathrm{R}_\mathcal{D}}_{U,\alpha}$ as a class of distributions in their own right. First let $\mathrm{Z}_\mathcal{D} \subseteq \mathbb{R}^{p\times p}$ denote the real linear space of $p \times p$ symmetric matrices A such that $A_{ij} = A_{ji} = 0$ if (i, j) is not in E. Notice that the dimension of $\mathrm{Z}_\mathcal{D}$ is $|E|$. Let α a given vector in $\mathbb{R}^p$ such that $\alpha_i > \mathrm{pa}_i + 2$, for each i. Now consider the family of DAG-Wishart distributions $\left[\pi^{\mathrm{R}_\mathcal{D}}_{U,\alpha} : U \in \mathrm{PD}_\mathcal{D}\right]$. Recall that $\mathcal{S}_\mathcal{D}$ is the image of $\mathrm{PD}_\mathcal{D}$ under the projection $U \mapsto U^E$. Since $\mathcal{S}_\mathcal{D}$ is isomorphic to $\mathrm{PD}_\mathcal{D}$, it is more natural to parametrize this family of distributions as $\left[\pi^{\mathrm{R}_\mathcal{D}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{D}\right]$. It is easy to check that this is an identifiable parametrization, that is if $\pi^{\mathrm{R}_\mathcal{D}}_{U_1^E,\alpha}$ is a.s. equal to $\pi^{\mathrm{R}_\mathcal{D}}_{U_1^E,\alpha}$, then $U_1^E = U_2^E$. The following lemma formalizes these points.

Lemma 2.8. *If $\mathcal{D}$ is a perfect DAG, then the family of distributions*

$$\left\{\pi^{\mathrm{R}_\mathcal{D}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{D}\right\}, \text{ or equivalently } \left\{\pi^{\mathrm{P}_\mathcal{D}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{D}\right\}$$

is a general exponential family. If $\mathcal{D}$ is not perfect, then $\left\{\pi^{\mathrm{R}_\mathcal{D}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{D}\right\}$ is no longer a general exponential family but a curved exponential family.

Proof. Let $t : \mathrm{R}_\mathcal{D} \to \mathrm{Z}_\mathcal{D}$ be the embedding $\Upsilon \mapsto (\Upsilon)^0$ and let $\eta : \mathrm{S}_\mathcal{D} \to \mathrm{Z}_\mathcal{D}$ be the embedding $U^E \mapsto \left(U^E\right)^0$. Then $\mathrm{tr}(\widehat{\Upsilon}U)$ is equal to the inner product of $(\Upsilon)^0$ and $\left(U^E\right)^0$ in Euclidean space $\mathrm{Z}_\mathcal{D}$. Under these embedding mappings $\mathrm{R}_\mathcal{D}$ and $\mathrm{S}_\mathcal{D}$ are open subsets of $\mathrm{Z}_\mathcal{D}$. Therefore $\left\{\pi^{\mathrm{P}_\mathcal{D}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{D}\right\}$, is a general exponential family.

Now if $\mathcal{D}$ is not perfect, the expression $\mathrm{tr}(\widehat{\Upsilon}U)$ not only depends on the entries in position ij where i, j are adjacent in $\mathcal{D}$, but also on a position ij where there exists an immorality $i \to k \leftarrow j$. Therefore, $\mathrm{tr}(\widehat{\Upsilon}U)$ is not equal to $\mathrm{tr}\left((\Upsilon)^0\left(U^E\right)^0\right)$, the inner product of $(\Upsilon)^0$ and $\left(U^E\right)^0$ in $\mathrm{Z}_\mathcal{D}$. However, clearly, $\mathrm{tr}(\widehat{\Upsilon}U)$ is the inner product of the projection of $\widehat{\Upsilon}$ and U in Euclidean space $\mathrm{Z}_{\mathcal{D}^\mathrm{m}}$, which has higher dimension than $|E|$. Hence when $\mathcal{D}$ is not perfect $\left\{\pi^{\mathrm{R}_\mathcal{D}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{D}\right\}$ is no longer an exponential family, but only a curved exponential family.

□

The proof of Lemma 2.8 shows that for any non-perfect DAG $\mathcal{D}$, the family of

$$\left\{\pi^{\mathrm{R}_\mathcal{D}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{D}\right\} \subsetneq \left[\pi^{\mathrm{R}_\mathcal{D}}_{U,\alpha} : U \in \mathrm{PD}_p(\mathbb{R})\right].$$

On the other hand, when $\mathcal{D}$ is perfect $\left\{\pi^{\mathrm{R}_\mathcal{D}}_{U,\alpha} : U \in \mathrm{PD}_p(\mathbb{R})\right\}$ is identical to $\left[\pi^{\mathrm{R}_\mathcal{D}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{D}\right\}$.

2.9 The inverse DAG-Wishart distribution for homogeneous DAGs

We now show that the class of $\pi^{\mathrm{S}_\mathcal{D}}_{U,\alpha}$ contains the sub-class of distributions introduced by (Khare and Rajaratnam, 2011) in the context of Gaussian covariance graph models. In the process we also demonstrate that for a special class of DAGs, the functional form of the density of the $\pi^{\mathrm{S}_\mathcal{D}}_{U,\alpha}$ can be simplified. A Gaussian covariance graph model over an undirected graph $\mathcal{G} = (V, \mathscr{E})$, denoted by $\mathscr{N}(\mathcal{G}_{\mathrm{cov}})$, is defined as

Definition 2.1. *Let $\mathrm{PD}_{\mathcal{G}_{cov}}$ denote the set of positive definite matrices $\Sigma \in \mathrm{PD}_p(\mathbb{R})$ such that $\Sigma_{ij} = 0$ whenever $i \not\sim_{\mathcal{G}} j$, that is when i and j are not neighbors. Then the Gaussian covariance graph model over $\mathcal{G}$ is defined by $\mathscr{N}(\mathcal{G}_{\mathrm{cov}}) = \{\mathrm{N}_p(0,\Sigma) : \Sigma \in \mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}\}$.*

A formal comparison between the DAG-Wishart priors introduced in this paper and the Wishart priors introduced in Khare and Rajaratnam (2011) requires a few technical definitions.

Definition 2.2. *a) A DAG $\mathcal{D}$ is called homogeneous of type I if it is transitive (that is $i \to j \to k$ implies that $i \to k$), and perfect. A DAG $\mathcal{D}$ is called a homogeneous of type II if it is transitive and does not contain any induced subgraph of the form $j \leftarrow i \to k$.*

b) An undirected graph $\mathcal{G} = (V, \mathscr{E})$ is called homogeneous if $i \sim_{\mathcal{G}} j \implies \mathrm{ne}(i) \cup \{i\} \subseteq \mathrm{ne}(j) \cup \{j\}$ or $\mathrm{ne}(j) \cup \{j\} \subseteq \mathrm{ne}(i) \cup \{i\}$, for every $i, j \in V$.

Equivalently, a graph $\mathcal{G}$ is said to be homogeneous if it is decomposable and does not contain the A_4 path as an induced subgraph. The reader is referred to Letac and Massam (2007) for further details on homogeneous graphs.

If $\mathcal{D}$ is homogeneous of either types, then $\mathcal{D}^u$ is homogeneous. On the other hand, if $\mathcal{G} = (V, \mathscr{E})$ is homogeneous, then one can construct a homogeneous DAG of type I or II that is a DAG version of $\mathcal{G}$. This can be achieved by using the Hasse tree associated with the undirected homogeneous graph. To obtain a DAG of type I. Reversing the orientation (that is redirecting all the arrows to the root of the tree) will yield a homogeneous DAG of type II. More precisely we shall now show an example that constructs a DAG version that is homogeneous of type II. Let $\mathcal{D}$ be a directed version of $\mathcal{G}$ obtained by directing each edge $i \sim_{\mathcal{G}} j$ to a directed edge $i \to j$ if $\mathrm{ne}(i) \cup \{i\} \subsetneq \mathrm{ne}(j) \cup \{j\}$, or $j \to i$ if $\mathrm{ne}(j) \cup \{j\} \subsetneq \mathrm{ne}(i) \cup \{i\}$. If $\mathrm{ne}(i) \cup \{i\} = \mathrm{ne}(j) \cup \{j\}$, an arbitrary direction is chosen. From Definition 2.2 one can check that $\mathcal{D}$ is a transitive DAG and it does not contain any induced subgraph of the form $j \leftarrow i \to k$. In general, it can be shown that if $\mathcal{D}$ is homogeneous of type II and a DAG version of $\mathcal{G}$, then $\mathscr{N}(\mathcal{D})$ is identical to the Gaussian covariance model $\mathscr{N}(\mathcal{G}_{\mathrm{cov}})$ in the sense that $\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}} = \mathrm{PD}_{\mathcal{D}}$ (Pearl and Wermuth, 1994). It is also evident, from the Markov equivalence of perfect DAGs and decomposable graphs, that for a homogeneous DAG $\mathcal{D}$ of type I which is a DAG version of $\mathcal{G}$, we have $\mathrm{PD}_{\mathcal{G}} = \mathrm{PD}_{\mathcal{D}}$.

Proposition 2.9. *Let $\mathcal{D} = (V, E)$ be homogeneous of either type I or II and let $\mathcal{G} = (V, \mathscr{E})$ be a homogeneous graph.*

(i) The density of $\pi_{U,\alpha}^{\mathrm{S}_{\mathcal{D}}}$ is
$z_{\mathcal{D}}(U,\alpha)^{-1} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma(\Gamma)^{-1}U)\right\} \prod_{i=1}^{p} \Sigma_{ii|\mathrm{pa}(i)}^{-\frac{1}{2}(\alpha_i + 2ch_i(\mathcal{D}))}$, *where $ch_i(\mathcal{D}) = |\mathrm{ch}_{\mathcal{D}}(i)|$.*

(ii) If $\mathcal{D}$ is of type II and a DAG version of $\mathcal{G}$, then the open cone $\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}$ can be identified with $\mathrm{S}_{\mathcal{D}}$ via the bijective mapping

$$\left(\Gamma \mapsto (\Gamma)^0 = \Sigma(\Gamma)\right) : \mathrm{S}_{\mathcal{D}} \to \mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}. \tag{15}$$

Let $\pi_{U,\alpha}^{\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}}$ denote the probability image of the $\pi_{U,\alpha}^{\mathrm{S}_{\mathcal{D}}}$ under the mapping in (15). Then the density of $\pi_{U,\alpha}^{\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}}$ with respect to Lebesgue measure is the expression in part(a) above.

Proof. (i) It suffices to prove that for every $\Sigma \in \mathrm{PD}_{\mathcal{D}}$,

$$\prod_{i \in V} \det(\Sigma_{\mathrm{pa}(i)}) = \prod_{i \in V} \Sigma_{ii|\mathrm{pa}(i)}^{ch_i(\mathcal{D})}. \tag{16}$$

1. Suppose that $\mathcal{D}$ is homogeneous of type I. We shall first show that for every $i \in V$

$$\det(\Sigma_{\mathrm{pa}(i)}) = \prod_{\ell \in \mathrm{pa}(i)} \Sigma_{\ell\ell|\mathrm{pa}(\ell)}. \tag{17}$$

If $\mathrm{pa}(i) = \emptyset$ for some i, then by our convention $\det(\Sigma_{\mathrm{pa}(i)}) = 1$ and $\Sigma_{\ell\ell|\mathrm{pa}(\ell)} = 1$ for any $\ell \in \mathrm{pa}(i)$ and therefore (17) holds. Now let ℓ_0 be the smallest integer in $\mathrm{pa}(i)$. One then can easily check that since $\mathcal{D}$ is both transitive and perfect we have $\mathrm{pa}(i) = \{\ell_0\} \cup \mathrm{pa}(\ell_0)$. From this we write $\det(\Sigma_{\mathrm{pa}(i)}) = \Sigma_{\ell_0\ell_0|\mathrm{pa}(\ell_0)} \det(\Sigma_{\mathrm{pa}(\ell_0)})$. Now by repeating this procedure we obtain the result in (17). Finally we write

$$\prod_{i \in V} \det(\Sigma_{\mathrm{pa}(i)}) = \prod_{i \in V} \prod_{\ell \in \mathrm{pa}(i)} \Sigma_{\ell\ell|\mathrm{pa}(\ell)} = \prod_{i \in V} \Sigma_{ii|\mathrm{pa}(i)}^{ch_i(\mathcal{D})}.$$

2. Suppose $\mathcal{D}$ is homogeneous of type II. We shall proceed by mathematical induction. It is clear that (16) holds when $p = |V| = 1$. Now by the inductive hypothesis assume that (16) holds for every homogeneous DAG of type II, connected or disconnected, with fewer vertices than $p = |V|$. Using the inductive hypothesis we shall show that (16) will also hold for $\mathcal{D}$ with p vertices. Given $\Sigma \in \mathrm{PD}_{\mathcal{D}}$, consider two possible cases:

   – The DAG $\mathcal{D}$ is connected. Let $\mathcal{D}_{[1]}$ be the induced DAG on $V \setminus \{1\}$. It is clear that $\mathcal{D}_{[1]}$ is also homogeneous of type II and therefore by the induction hypothesis

$$\prod_{i=2}^{p} \det(\Psi_{\mathrm{pa}(i)}) = \prod_{i=2}^{p} \Psi_{ii|\mathrm{pa}(i)}^{ch_i(\mathcal{D}_{[1]})},$$

   where $\Psi = \Sigma_{V\setminus\{1\}}$. Notice that $\mathcal{D}_{[1]}$ is an ancestral subgraph of $\mathcal{D}$ and hence $\mathrm{fa}_{\mathcal{D}_{[1]}}(i) = \mathrm{fa}_{\mathcal{D}}(i)$ for each $i = 2, \ldots, p$. Thus $\Psi_{\mathrm{pa}(i)} = \Sigma_{\mathrm{pa}(i)}$ and $\Psi_{ii|\mathrm{pa}(i)} = \Sigma_{ii|\mathrm{pa}(i)}$. All together these imply that $\prod_{i=2}^{p} \det(\Sigma_{\mathrm{pa}(i)}) = \prod_{i=2}^{p} \Sigma_{ii|\mathrm{pa}(i)}^{ch_i(\mathcal{D}_{[1]})}$. Now we claim that $\mathrm{fa}_{\mathcal{D}}(1) = V$. Assume to the contrary that $V \setminus \mathrm{fa}_{\mathcal{D}}(1) \neq \emptyset$. Since $\mathcal{D}$ is connected, this implies that there exist vertices $i \in \mathrm{fa}_{\mathcal{D}}(1)$ and $j \in V \setminus \mathrm{fa}_{\mathcal{D}}(1)$ such that i, j are adjacent in $\mathcal{D}$. But this implies $j \to i \to 1$ or $j \leftarrow i \to 1$. By definition these induced subgraphs cannot occur in $\mathcal{D}$. Thus $\mathrm{fa}(1) = V$ and therefore we have $\det(\Sigma_{\mathrm{pa}(1)}) = \Sigma_{11|\mathrm{pa}(1)}^{-1} \det(\Sigma) = \prod_{i=2}^{p} \Sigma_{ii|\mathrm{pa}(i)}$. Also the fact that $\mathrm{fa}_{\mathcal{D}}(1) = V$ implies that for each $i \in V \setminus \{1\}$ we have $\mathrm{ch}_i(\mathcal{D}_{[1]}) = \mathrm{ch}_i(\mathcal{D}) - 1$. Therefore

$$\prod_{i \in V} \det(\Sigma_{\mathrm{pa}(i)}) = \det(\Sigma_{\mathrm{pa}(1)}) \prod_{i=2}^{p} \det(\Sigma_{\mathrm{pa}(i)}) = \prod_{i=2}^{p} \Sigma_{ii|\mathrm{pa}(i)} \prod_{i=2}^{p} \Sigma_{ii|\mathrm{pa}(i)}^{ch_i(\mathcal{D}_{[1]})} = \prod_{i \in V} \Sigma_{ii|\mathrm{pa}(i)}^{ch_i(\mathcal{D})}.$$

   – The DAG $\mathcal{D}$ is disconnected. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ denote respectively the induced subgraphs of $\mathcal{D}$ on $\mathrm{fa}_D(1)$ and $V \setminus \mathrm{fa}_{\mathcal{D}}(1)$. It is clear that $\mathcal{D}_1$ and $\mathcal{D}_2$ are both homogeneous of type II. In addition it is also easily verified that they are ancestral. Now let $\Psi = \Sigma_{\mathrm{fa}(1)} \in \mathrm{PD}_{\mathcal{D}_1}$ and $\Psi' = \Sigma_{V\setminus\mathrm{fa}_{\mathcal{D}}(1)} \in \mathrm{PD}_{\mathcal{D}_2}$. The induction hypothesis and the fact that $\mathcal{D}_1$ and $\mathcal{D}_2$ are disjoint together imply that

$$\prod_{i \in V} \det(\Sigma_{\mathrm{pa}(i)}) = \prod_{i \in \mathrm{fa}_{\mathcal{D}}(1)} \det(\Sigma_{\mathrm{pa}(i)}) \prod_{i \in V\setminus\mathrm{fa}_{\mathcal{D}}(1)} \det(\Sigma_{\mathrm{pa}(i)})$$

$$
\begin{aligned}
&= \prod_{i\in \mathrm{fa}_{\mathcal{D}}(1)} \det(\Psi_{\mathrm{pa}(i)}) \prod_{i\in V\setminus \mathrm{fa}_{\mathcal{D}}(1)} \det(\Psi'_{\mathrm{pa}(i)}) \\
&= \prod_{i\in \mathrm{fa}_{\mathcal{D}}(1)} \det(\Psi_{ii|\mathrm{pa}(i)})^{ch_i(\mathcal{D}_1)} \prod_{i\in V\setminus \mathrm{fa}_{\mathcal{D}}(1)} \det(\Psi'_{ii|\mathrm{pa}(i)})^{ch_i(\mathcal{D}_2)} \\
&= \prod_{i\in \mathrm{fa}_{\mathcal{D}}(1)} \det(\Sigma_{ii|\mathrm{pa}(i)})^{ch_i(\mathcal{D})} \prod_{i\in V\setminus \mathrm{fa}_{\mathcal{D}}(1)} \det(\Sigma'_{ii|\mathrm{pa}(i)})^{ch_i(\mathcal{D})} \\
&= \prod_{i\in V} \det(\Sigma_{ii|\mathrm{pa}(i)})^{ch_i(\mathcal{D})}.
\end{aligned}
$$

(ii) One can show that the mapping (15) is a diffeomorphism and its Jacobian is equal to 1. Therefore the functional form of the density $\pi_{U,\alpha}^{\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}}$ with respect to Lebesgue measure is the same as $\pi_{U,\alpha}^{\mathrm{S}_{\mathcal{D}}}$ that was given in Proposition 2.9.

□

Remark 2.2. *For a homogeneous graph $\mathcal{G}$ the distribution $\pi_{U,\alpha}^{\mathrm{PD}_{\mathcal{G}_{cov}}}$ with the associated density derived in Proposition (2.9) coincides with the distribution introduced by (Khare and Rajaratnam, 2011).*

2.10 Equivalent Gaussian DAG models and corresponding DAG-Wishart distributions

In this section we ask whether equivalent DAG models yield the same inverse DAG-Wishart distributions. We demonstrate that a particular sub-class of the DAG-Wishart distributions yield the exactly the same prior on equivalent Gaussian DAG models. We show that this sub-class coincides with that of (Geiger and Heckerman, 2002), thus making a connection between the DAG-Wishart priors and those of Geiger and Heckerman.

(Geiger and Heckerman, 2002) develop a methodology for assigning priors $\pi(\theta \mid \mathcal{D})$ to a DAG model $\mathscr{M}(\mathcal{D})$. Their proposed methodology for assigning priors is based on the following five assumptions: 1) Complete model equivalence, 2) Regularity conditions, 3) Likelihood modularity, 4) Prior Modularity, and 5) Global parameter independence. For this, (Geiger and Heckerman, 2002) extend a given DAG model by filling in all the missing arrows that respect the give parent ordering to obtain a complete DAG model. Now at the level of the complete DAG model, a Wishart prior can be assigned. Thereafter, (Geiger and Heckerman, 2002) introduce a mechanism to reduce this prior to the original (non-complete) DAG model so that the above five conditions are met. Two main consequences of the above assumptions are as follows.

a) For each DAG model $\mathscr{M}(\mathcal{D})$, the prior $\pi(\theta \mid \mathcal{D})$ is uniquely determined by a specific prior for an arbitrary complete DAG model (Geiger and Heckerman, 2002, Theorem 1).

b) If two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent, and X is a directed Markov field with respect to $\mathcal{D}_1$, and therefore with respect to $\mathcal{D}_2$, then $\pi(\theta \mid \mathcal{D}_2) = \pi(\theta \mid \mathcal{D}_1)$.

For a Gaussian DAG model $\mathscr{N}(\mathcal{D})$ Geiger and Heckerman's priors are essentially defined on the Cholesky space $\Theta_{\mathcal{D}}$ originating from a classical Wishart prior for a complete DAG model. For complete Gaussian DAG models, (Geiger and Heckerman, 2002) also show that under the aforementioned assumptions the induced prior for the inverse-covariance matrix is always the Wishart distribution.

Now a relevant question pertains to how Geiger and Heckerman's priors compare with the DAG-Wishart priors $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$. To this end, let us first examine the functional form of the assigned prior $\pi(\theta \mid \mathcal{D})$ for $\mathscr{N}(\mathcal{D})$ under Geiger and Heckerman's methodology. Because of global parameter independence assumption, under the D-parametrization of $\mathscr{N}(\mathcal{D})$ it suffices to assign a prior to each local parameter $\theta_i = (\Sigma_{ii|\mathrm{pa}(i)}, \Sigma_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i})$. Now fix i and let $\widetilde{\mathcal{D}}_i$ be a complete DAG such that $\mathrm{pa}_{\widetilde{\mathcal{D}}_i}(i) = \mathrm{pa}_{\mathcal{D}}(i)$. By (Geiger and Heckerman, 2002)

$$\pi(\theta_i \mid \mathcal{D}) = \pi(\theta_i \mid \widetilde{\mathcal{D}}_i). \tag{18}$$

Let $\mathrm{W}_p(\eta, U)$ be the Wishart prior for the invariance covariance matrix in the complete Gaussian DAG model $\mathscr{N}(\widetilde{\mathcal{D}}_i)$. Recall that in this case the Wishart distribution is the only prior that satisfies Geiger and Heckerman's assumptions (in contrast to our generalized DAG-Wishart priors with multiple shape parameters). We are now in a position to embark on a formal comparison between the DAG-Wishart distributions and Geiger and Heckerman's priors.

Lemma 2.10. *Let $\widetilde{\mathcal{D}}_i$ be a complete DAG as above and let $\Sigma^{-1} \sim \mathrm{W}_p(\eta, U)$. Then*

1) $\Sigma_{ii|\mathrm{pa}(i)} \sim \mathrm{IG}(\frac{\eta-p}{2} - \mathrm{pa}_i - \frac{3}{2}, \frac{1}{2}U_{ii|\mathrm{pa}(i)})$,

2) $\Sigma_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i} \mid \Sigma_{ii|\mathrm{pa}(i)} \sim \mathrm{N}_{\mathrm{pa}_i}(U_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i}, \Sigma_{ii|\mathrm{pa}(i)}U_{\mathrm{pa}(i)})$.

Proof. The Jacobian of the mapping

$$\left(\Sigma^{-1} \mapsto \times_{j=1}^{p}(\Sigma_{jj|\mathrm{pa}(j)}, \Sigma_{\mathrm{pa}(j)}^{-1}\Sigma_{\mathrm{pa}(j),j})\right) : \mathrm{P}_{\widetilde{\mathcal{D}}_i} \to \Theta_{\widetilde{\mathcal{D}}_i} \tag{19}$$

is equal to $\prod_{j=1}^{p} \Sigma_{jj|\mathrm{pa}(j)}^{|\mathrm{pa}_{\widetilde{\mathcal{D}}_i}(j)|+2}$. Now the image of $\mathrm{W}_p(\eta, U)$ under the mapping in (19) is the $\pi_{U,\alpha}^{\Theta_{\widetilde{\mathcal{D}}_i}}$, where $\alpha_j = \eta - p - |\mathrm{pa}_{\widetilde{\mathcal{D}}_i}(j)| - 3$. □

Corollary 2.3. *Let $\eta > 0$. If we set $\alpha_j = \eta - p - \mathrm{pa}_j - 3$, then the $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ is the Geiger and Heckerman's prior and therefore $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ is the same for any DAG Markov equivalent to $\mathcal{D}$.*

Proof. This follows directly from (18), Lemma 2.10 above and the fact that

$$\Sigma_{ii|\mathrm{pa}(i)} \sim IG(\frac{\alpha_i}{2} - \frac{\mathrm{pa}_i}{2} - 1, \frac{1}{2}U_{ii|\mathrm{pa}(i)}), \text{ and}$$
$$\Sigma_{\mathrm{pa}(i)}^{-1}\Sigma_{\mathrm{pa}(i),i} \mid D_{ii} \sim \mathrm{N}_{\mathrm{pa}_i}\left(U_{\mathrm{pa}(i)}^{-1}U_{\mathrm{pa}(i),i}, D_{ii}U_{\mathrm{pa}(i)}^{-1}\right)$$

□

Remark 2.3. *By Corollary 2.3 the class of Geiger and Heckerman's priors are a subclass of the multiple shape parameter $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$. Appropriately choosing the shape parameters can thus lead to the same priors for Markov equivalent DAGs. The class of DAG-Wishart priors are therefore sufficiently flexible so as to accommodate equivalent DAG models. The Geiger and Heckerman's priors are useful when the DAG models are simply treated as probabilistic models encoding a set of conditional independences. For these priors, the shape parameter is a scalar and thus they are an immediate extension of the hyper-Wishart distribution of (Dawid and Lauritzen, 1993). On the other hand, the flexible multiple shape parameter α_i's present in the DAG-Wishart distribution allows one to incorporate different prior knowledge at the level of each variable, even if the two corresponding DAGs encode the same conditional independences.*

2.11 Summary of Wishart distributions

Table 1 summarizes the properties of the various multiple shape parameter Wishart distributions that have been recently introduced to the statistics literature for use in Gaussian graphical models. One can see on this table that the DAG-Wishart distributions introduced in this paper are applicable in all generality - and not just when the graph is perfect, or equivalently, decomposable. The ability to specify the induced Wishart distributions and posterior moments for arbitrary graphs is especially useful.

	DAG			UG			COVG		
	ALL	P	H	ND	D	H	ND	D	H
Conjugacy property	✔	✔	✔	✔	✔	✔	✗	✔	✔
Normalizing constant in closed form	✔	✔	✔	✗	✔	✔	✗	✗	✔
Posterior moments in closed from	✔	✔	✔	✗	✔	✔	✗	✗	✔
Posterior mode in closed from	✔	✔	✔	✗	✔	✔	✗	✗	✔
Hyper Markov properties	✔	✔	✔	✗	✔	✔	✗	✗	✔
Tractable sampling from the distribution	✔	✔	✔	✗	✔	✔	✗	✔	✔

Table 1: Properties of Wishart distributions for the three classes of Gaussian graphical models. Abbreviations. ND: Non-decomposable, D/P: Decomposable/Perfect, H: Homogeneous.

3 The DAG-Wishart distribution on $\mathrm{P}_{\mathcal{D}}$ and its density with respect to Hausdorff measure

3.1 Introduction

In this section we consider a general approach for defining the DAG-Wishart distribution directly on the space of precision matrices $\mathrm{P}_{\mathcal{D}}$, for an arbitrary DAG $\mathcal{D}$. As a probability distribution, we can consider $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{D}}}$, the image of the $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ under the mapping $\left((D,L)\mapsto LD^{-1}L^{\top}\right):\Theta_{\mathcal{D}}\to\mathrm{P}_{\mathcal{D}}$ as the DAG-Wishart distribution on $\mathrm{P}_{\mathcal{D}}$. When $\mathcal{D}$ is perfect the space of precision matrices $\mathrm{P}_{\mathcal{D}}$ can be naturally identified with $\mathrm{R}_{\mathcal{D}}$, and therefore $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{D}}}$ can be identified with $\pi_{(U,\alpha)}^{\mathrm{R}_{\mathcal{D}}}$ and hence has a density with respect to Lebesgue measure on $\mathbb{R}^{|E|}$. This is due to the fact that in this case the space $\mathrm{P}_{\mathcal{D}}$ is an open subset of $\mathrm{Z}_{\mathcal{D}}\cong\mathbb{R}^{|E|}$. However,when $\mathcal{D}$ is not a perfect DAG several complications arise, mainly because the space $\mathrm{P}_{\mathcal{D}}$ is a curved manifold that has Lebesgue measure zero in any Euclidean vector space containing it. This implies that $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{D}}}$ does not have a density with respect to Lebesgue measure. In theory a solution to this problem requires deriving the density of $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{D}}}$ with respect to Hausdorff measure. This section elaborates on this topic in much detail.

3.2 A discussion about density functions on $\mathrm{P}_{\mathcal{D}}$

In this section we undertake a measure theoretic analysis of the space $\mathrm{P}_{\mathcal{D}}$ when $\mathcal{D}$ is not perfect. Lemma 1.1 implies the following: $\mathrm{P}_{\mathcal{D}} \subset \mathrm{P}_{\mathcal{D}^{\mathrm{m}}} \subset \mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$. Thus if $\mathcal{D} = (V, E)$ is not perfect, then $\mathrm{P}_{\mathcal{D}}$ has Lebesgue measure zero in any Euclidean vector space containing it. The next lemma gives a formal proof of this assertion.

Lemma 3.1. *Suppose $\mathcal{D} = (V, E)$ is a non-perfect DAG and* E *a Euclidean space containing* $\mathrm{P}_{\mathcal{D}}$. *Then* E *contains* $\mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$. *Moreover,* $\mathrm{P}_{\mathcal{D}}$ *has Lebesgue measure zero in* E.

Proof. For each $(i, j) \in E^{\mathrm{m}}$ with $j \leq i$ let us define the elementary symmetric matrix $\widetilde{E}^{(ij)} \in \mathrm{S}_p(\mathbb{R})$ as

$$\widetilde{E}^{(ij)}_{uv} = \begin{cases} 1 & \text{if } \{u, v\} = \{i, j\}, \\ 0 & \text{otherwise.} \end{cases}$$

The set of $\widetilde{E}^{(ij)}$ forms a basis of $\mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$. Since E contains $\mathrm{Z}_{\mathcal{D}} \supset \{\widetilde{E}^{(ij)} : (i, j) \in E\}$, it suffices to prove that E contains the rest of $\widetilde{E}^{(ij)}$. For this, let (i, j) be in $E^{\mathrm{m}} \setminus E$ with $i > j$. Thus there exists $k < j < i$ such that $i \to k \leftarrow j$. Let $L^{(ij)}$ denote the lower triangular matrix

$$L^{(ij)}_{uv} = \begin{cases} 1 & \text{if } (u, v) = (i, k), \\ 1 & \text{if } (u, v) = (j, k), \\ 1 & \text{if } u = v, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathrm{P}_{\mathcal{D}} \ni L^{(ij)}(L^{(ij)})^{\top} = T + 2\widetilde{E}^{(ij)}$, for some $T \in \mathrm{Z}_{\mathcal{D}}$. Therefore $\widetilde{E}^{(ij)} \in E$ and $\mathrm{P}_{\mathcal{D}} \subset V \Rightarrow \mathrm{Z}_{\mathcal{D}^{\mathrm{m}}} \subset E$, thus $\mathrm{P}_{\mathcal{D}} \subset \mathrm{Z}_{\mathcal{D}^{\mathrm{m}}} \subset E$.

Since $\mathrm{P}_{\mathcal{D}}$ is a manifold of dimension $|E|$ diffeomorphic to $\Theta_{\mathcal{D}}$, it is an open subset of Euclidean space of dimension $|E|$. Furthermore, the dimension of $\mathrm{Z}_{\mathcal{D}^{\mathrm{m}}} = |E^{\mathrm{m}}| < |E|$. Thus any Euclidean space that contains $\mathrm{P}_{\mathcal{D}}$ has dimension strictly larger than $|E|$. Consequently, $\mathrm{P}_{\mathcal{D}}$ has Lebesgue measure zero in any Euclidean space that contains it. □

Lemma 3.1 implies when $\mathcal{D}$ is not perfect $\pi^{\mathrm{P}_{\mathcal{D}}}_{U,\alpha}$ has no density with respect to Lebesgue measure.

3.3 The density of $\pi^{\mathrm{P}_{\mathcal{D}}}_{U,\alpha}$ with respect to Hausdorff measure

We now derive the density of $\pi^{\mathrm{P}_{\mathcal{D}}}_{U,\alpha}$ with respect to Hausdorff measure. The reader is referred to (Billingsley, 1979) for background on Hausdorff measures. Let $\Delta_{\mathcal{D}}$ denote the set of (D, L), where $D \in R^{p \times p}$ is a diagonal matrix and $L \in \mathcal{L}_{\mathcal{D}}$. The set $\Delta_{\mathcal{D}}$ is a real linear space of dimension $|E|$ with the following scalar product and sum operation.

1. $\lambda(D, L) = (\lambda D, \lambda L)$, for all $\lambda \in R$;
2. $(D', L') + (D'', L'') = (D, L)$, where $D = (D' + D'')$, and L is a lower triangular matrix with $L_{ij} = L'_{ij} + L''_{ij}$ if $i \neq j$ and $L_{ii} = 1$.

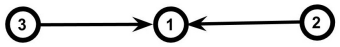

Figure 1: DAG-Wishart density with respect to a Hausdorff measure

Now $\Theta_{\mathcal{D}}$ is open in $\Delta_{\mathcal{D}}$, $\mathrm{P}_{\mathcal{D}} \subseteq \mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$ and $\psi : \Theta_{\mathcal{D}} :\to \mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$ satisfies the conditions of Theorem 19.3 in (Billingsley, 1979). Thus it has a density $\pi^{\mathrm{P}_{\mathcal{D}}}_{U,\alpha}$ with respect to the $|E|$-dimensional Hausdorff measure on $\mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$. To obtain an explicit expression for $\mathrm{J}(\psi(D,L))$ we need to compute the matrix of partial derivatives $\frac{\partial \psi_{kl}}{\partial D_{ii}}$ and $\frac{\partial \psi_{kl}}{\partial L_{ij}}$. We order the coordinates of $\Delta_{\mathcal{D}}$ as follows: D_{11}, L_{21} if $(2,1) \in E, D_{22}, L_{31}$ if $(3,1) \in E, L_{32}$ if $(3,2) \in E, \ldots, D_{(p-1)(p-1)}, L_{pl}, l = 1, \ldots (p-1)$ if $(p,l) \in E, D_{pp}$. Likewise, we order the coordinates of $\mathrm{Z}_{\mathcal{D}^{\mathrm{m}}} \cong \mathbb{R}^{|E|} \times \mathbb{R}^{|\mathscr{I}|}$, where $\mathscr{I} = E^{\mathrm{m}} \setminus E$, by ordering first the positions $(k,l) \in E$ as above, in their entirety, and then we order the positions $(k,l) \in \mathscr{I}$ according to their lexicographical order. The latter positions correspond to immoralities. These partial derivatives can be computed as follows:

$$\frac{\partial (L D^{-1} L^{\top})_{kl}}{\partial D_{ii}} = -D_{ii}^{-2} L_{ki} L_{li} \tag{20}$$

$$\frac{\partial (L D^{-1} L^{\top})_{kl}}{\partial L_{ij}} = \delta_{ik} D_{jj}^{-1} L_{lj} + \delta_{il} D_{jj}^{-1} L_{kj}, \tag{21}$$

where δ_{uv} is the Kronecker delta function. Using (20) and (21) we partition the Jacobian matrix $D\psi(D,L)$, considered as a mapping from $\mathbb{R}^{|E|}$ to $\mathbb{R}^{|E|} \times \mathbb{R}^{|\mathscr{I}|}$, into two blocks of matrices $A_\psi = D\psi(L,D)_{EE}$ of size $|E| \times |E|$ and $C_\psi = D\psi(D,L)_{\mathscr{I}E}$ of size $|\mathscr{I}| \times |E|$, respectively. The matrix A_ψ is the same as the Jacobian matrix from Lemma 2.6, and C_ψ is the last $|\mathscr{I}|$-th rows of the Jacobian matrix $D\psi(D,L)$, with each row of C_ψ being the partial derivatives obtained by (20) and (21) for $(k,l) \in \mathscr{I}$ and $(i,j) \in E$. Finally, we calculate the Jacobian of ψ as follows.

$$\begin{aligned}
J\psi(D,L) &= \det\left(\begin{pmatrix} A_\psi^\top & \vdots & C_\psi^\top \end{pmatrix} \begin{pmatrix} A_\psi \\ \cdots \\ C_\psi \end{pmatrix} \right)^{1/2} \\
&= \sqrt{\det(A_\psi^\top A_\psi + C_\psi^\top C_\psi)} \\
&= |\det(A_\psi)| \sqrt{\det(I + A_\psi^{-t} C_\psi^\top C_\psi A_\psi^{-1})} \\
&= \prod_{j=1}^{p} D_{jj}^{-(\mathrm{pa}_j + 2)} \sqrt{\det(I + A_\psi^{-t} C_\psi^\top C_\psi A_\psi^{-1})}.
\end{aligned}$$

Therefore we have proved the following.

Theorem 3.2. *Let A_ψ, C_ψ be defined as the block matrices in partitioning of the (Hausdorff) Jacobian matrix of ψ above. Then the density of $\pi^{\mathrm{P}_{\mathcal{D}}}_{U,\alpha}$ with respect to Hausdorff measure $\mathcal{H}^{|E|}$ on $\mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$ is*

$$z_{\mathcal{D}}(U,\alpha)^{-1} \exp\left\{ -\frac{1}{2} \mathrm{tr}(\Omega U) \right\} \prod_{i=1}^{p} D_{ii}^{-\frac{1}{2}\alpha_i + \mathrm{pa}_i + 2} \det(I + A_\psi^{-t} C_\psi^\top C_\psi A_\psi^{-1})^{-1/2}. \tag{22}$$

Example 3.1. *Consider DAG $\mathcal{D}$ given in Figure 1. The Jacobian matrix corresponding to (20) and (21) are given as follows:*

$$M_\psi = \begin{pmatrix} -D_{11}^{-2} & 0 & 0 & 0 & 0 \\ -L_{21}D_{11}^{2} & D_{11}^{-1} & 0 & 0 & 0 \\ -L_{21}^{2}D_{11}^{-2} & 2L_{21}D_{11}^{-1} & -D_{22}^{-2} & 0 & 0 \\ -L_{31}D_{11}^{-2} & 0 & 0 & D_{11}^{-1} & 0 \\ -L_{31}^{2}D_{11}^{-2} & 0 & 0 & 2L_{31}D_{11}^{-1} & 0 \\ -L_{21}L_{31}D_{11}^{-2} & L_{31}D_{11}^{-1} & 0 & L_{21}D_{11}^{-1} & -D_{33}^{-2} \end{pmatrix}$$

By computing $\det(M_\psi^\top M_\psi)$ *we obtain*

$$J\psi(D,L) = D_{11}^{-4}D_{22}^{-2}D_{33}^{-2}\left(L_{31}^4 + 4L_{31}^2 + 1\right)^{1/2}.$$

Thus the density of $\pi_{U,\alpha}^{\mathrm{P}_\mathcal{D}}$ *with respect to* $\mathcal{H}^5$ *on* $\mathbb{R}^6$ *is*

$$z_\mathcal{D}(U,\alpha)^{-1}\exp\left\{-\frac{1}{2}\mathrm{tr}(\Omega U)\right\} D_{11}^{-\alpha_1/2+4}D_{22}^{-\alpha_2/2+2}D_{33}^{-\alpha_3/2+2}\left(L_{31}^4 + 4L_{31}^2 + 1\right)^{-1/2},$$

where D_{ii} *and* L_{ij} *are considered as functions of* Ω*.*

4 Computational algorithms and more related results

4.1 DAG model selection stochastic search evaluation

In this section, we demonstrate the effectiveness of our algorithm for model selection in the space of DAGs. The target is to find the graph that admits the largest posterior (marginal) probability under the Bayesian model. This an NP-hard problem, because it requires computing the marginal likelihood for $2^{p(p-1)/2}$ graphs. Our `DAG-W` algorithm on the other hand uses stochastic search along the model selection path given by the `lassoDAG` (Shojaie and Michailidis, 2010) for this purpose. In this section, we demonstrate that our algorithm can identify the true MAP graph with high probability but with significantly fewer enumerations. For completeness, we give the details of the algorithm again.

Algorithm 4.1 (`DAG-W`)**.** *Assume the following are given: the standardized data matrix X, the hyperparameters α, U and the maximum iteration number M. Estimate N models corresponding to different points on `lassoDAG` regularization path, labeled as $\mathcal{D}^{(k)}, k = 1, \cdots, N$. Then for each $k = 1, 2, \cdots, N$, do the following.*

1. *Let $\mathcal{D}_0 = \mathcal{D}^{(k)}$. Until the maximum iteration number M is achieved:*
   (a) *Select N_1 graphs that are one edge away from $\mathcal{D}_0$. Evaluate the log-posterior scores $s_1, \cdots, s_{N_1}$ for each of these graphs, according to the DAG-Wishart prior/posterior. Record all of these graphs and scores as a list $\mathcal{L}^{(k)}$.*
   (b) *Sample the next graph from the current graph list with probability $p_i \propto \exp(s_i)^\gamma$, where γ is an annealing parameter. Take the sampled graph $\mathcal{D}_{new}$ as $\mathcal{D}_0$.*
   (c) *Return to Step 1-(a).*

2. *Collect/Assemble all the* $\mathcal{L}^{(k)}, k = 1, \cdots, N$.

3. *Return the graph with the largest score as the selected model.*

Since `lassoDAG` path is potentially close to the MAP graph, we narrow our search for graphs near the path to reduce the number of graphs to be evaluated and thus ease the burden of computation. We use two metrics for evaluating the performance of `DAG-W` algorithm. Let $\hat{p}$ be the marginal likelihood of the estimated graph and let $\tilde{p}$ be the largest marginal likelihood of all graphs. Define

$$\zeta_g = Pr(\hat{p} = \tilde{p}).$$

and

$$\chi_g = E[\hat{p}/\tilde{p}].$$

So ζ_g measures the proportion that our stochastic search does find the global optimal while χ_g measures on average, how close is our stochastic search probability to $\tilde{p}$.

We use the number of variables $p = 7$, such that we can still calculate $\tilde{p}$ within reasonable amount of time, by enumerating all of the $2^{21} = 2097152$ potential graphs. In the paper, our numerical examples always have $p \geq 50$. Thus to be fair, here we reduce our searching numbers dramatically. There are three searching parameters we have to specify in the algorithm. The first one is the number of starting points along `lassoDAG` path N, which we set to 3. We take $\kappa = 0.1, 1.5, 3$ in the equation (13) of the paper as the initial model from `lassoDAG` . The second parameter N_1 is the number of new graphs to be generated in each searching step. The third parameter M is the number to search from each initial graph from `lassoDAG` and the third one is N_1. Thus the total number of graphs to search is at most $3 \times N_1 \times M$. We will evaluate the searching efficiency of `DAG-W` by using a sequence of N_1 and M on simulated DAG model with different sparsity and sample sizes. All numerical results in this section are based on average of 100 independent replications.

In the first setting, we fix the sample size to be $n = 50$ and change the sparsity of the graph by setting the probability of having each edge independently to be $\rho = 0.1, 0.2, 0.4$ respectively. Table 2 shows the results of estimate of ζ_g and χ_g. Here we fix $M = 20$ and varying N_1 from 1 to 10. The total number of graphs to evaluate is from 60 to 600 which is between 2.9×10^{-5} to 2.9×10^{-4} of full enumeration. It can be seen that the performance is better for sparser graphs though the difference between $\rho = 0.1$ and $\rho = 0.2$ is small. When $\rho = 0.4$, which is already a very dense graph in general, the sparse assumption of `lassoDAG` is broken and it fails to provide good initializations, thus the performance of the stochastic search also drops. From the values of ζ_g we can see that in all sparsity levels, having $N_1 \geq 8$ is enough to achieve almost 100% successful rate. On the other hand, χ_g indicates that the graph our algorithm selects typically has a posterior probability that is very close to $\tilde{p}$.

Table 3 shows the performance when we fix $N_1 = 5$ and vary M in $\{2, 4, 6, \cdots, 30\}$. Similar patterns are observed. The performance of the algorithm is better when the underlying graph is sparser. It turns out the impact of increasing M is smaller than that of increasing N_1 as increasing M from 2 to 30 only slightly increase the searching efficiency.

In the second settings, we investigate the searching efficiency when one has different sample sizes. Here we use $n = 100, 50, 25$ respectively and the sparsity is fixed to be 0.1. All the other configurations are the same as before.

	Sparsity ρ	$\rho = 0.1$		$\rho = 0.2$		$\rho = 0.4$	
N_1	Search Proportion	ζ_g	χ_g	ζ_g	χ_g	ζ_g	χ_g
1	0.3×10^{-4}	0.93	0.985	0.96	0.991	0.94	0.972
2	0.6×10^{-4}	0.95	0.989	0.97	0.992	0.94	0.972
3	0.9×10^{-4}	0.96	0.995	0.97	0.991	0.96	0.979
4	1.1×10^{-4}	0.98	0.997	0.99	0.998	0.98	0.988
5	1.4×10^{-4}	0.99	0.998	0.97	0.993	0.97	0.986
6	1.7×10^{-4}	0.99	0.998	1.00	1.000	0.99	0.996
7	2.0×10^{-4}	0.99	0.998	0.99	0.995	0.99	0.996
8	2.3×10^{-4}	1.00	1.000	1.00	1.000	0.99	0.995
9	2.6×10^{-4}	1.00	1.000	1.00	1.000	1.00	1.000
10	2.9×10^{-4}	1.00	1.000	1.00	1.000	1.00	1.000

Table 2: Performance metrics of `DAG-W` when $n = 50, p = 7$. Here we fix $M = 20$ and vary N_1 from 1 to 10.

	Sparsity ρ	$\rho = 0.1$		$\rho = 0.2$		$\rho = 0.4$	
M	Search Proportion	ζ_g	χ_g	ζ_g	χ_g	ζ_g	χ_g
2	2.86×10^{-5}	0.97	0.996	0.96	0.993	0.95	0.978
4	5.72×10^{-5}	0.97	0.996	0.98	0.998	0.95	0.978
6	8.58×10^{-5}	0.98	0.991	0.97	0.993	0.96	0.984
8	1.14×10^{-4}	0.96	0.989	0.98	0.998	0.96	0.984
10	1.43×10^{-4}	0.99	0.998	0.97	0.993	0.96	0.982
12	1.72×10^{-4}	0.97	0.996	0.97	0.993	0.96	0.984
14	2.00×10^{-4}	0.97	0.996	0.97	0.993	0.95	0.978
16	2.28×10^{-4}	0.98	0.997	0.97	0.993	0.97	0.988
18	2.57×10^{-4}	0.99	0.998	1.00	1.000	0.97	0.986
20	2.86×10^{-4}	0.99	0.998	0.97	0.993	0.97	0.986
22	3.15×10^{-4}	0.99	0.998	0.97	0.993	0.95	0.978
24	3.43×10^{-4}	0.99	0.998	0.97	0.993	0.96	0.984
26	3.72×10^{-4}	0.99	0.998	0.97	0.993	0.96	0.984
28	4.00×10^{-4}	0.99	0.998	0.97	0.993	0.95	0.978
30	4.29×10^{-4}	0.99	0.998	0.97	0.994	0.95	0.978

Table 3: Performance metrics of `DAG-W` when $n = 50, p = 7$. Here we fix $N_1 = 5$ and vary M from 2 to 30.

	Sample Size	$n = 100$		$n = 50$		$n = 25$	
N_1	Search Proportion	ζ_g	χ_g	ζ_g	χ_g	ζ_g	χ_g
1	0.3×10^{-4}	0.99	0.996	0.93	0.985	0.87	0.952
2	0.6×10^{-4}	0.99	0.996	0.95	0.989	0.89	0.960
3	0.9×10^{-4}	0.99	0.996	0.96	0.995	0.93	0.974
4	1.1×10^{-4}	0.99	0.996	0.98	0.997	0.98	0.996
5	1.4×10^{-4}	1.00	1.000	0.99	0.998	0.95	0.977
6	1.7×10^{-4}	1.00	1.000	0.99	0.998	1.00	1.000
7	2.0×10^{-4}	1.00	1.000	0.99	0.998	1.00	1.000
8	2.3×10^{-4}	0.99	0.996	1.00	1.000	0.99	0.998
9	2.6×10^{-4}	1.00	1.000	1.00	1.000	1.00	1.000
10	2.9×10^{-4}	1.00	1.000	1.00	1.000	1.00	1.000

Table 4: Performance metrics of DAG-W when one varies the sample size for $n = 100, 50, 25$ and the sparsity level is 0.1. Here we fix $M = 20$ and vary N_1 from 1 to 10.

Table 4 shows the performance when we fix $M = 20$ and increase N_1 from 1 to 10. As before, increasing N_1 is very effective in improving the search efficiency. As expected, the searching is more efficient when the sample size is larger. Even when there are only 25 samples, using $N_1 > 6$ is able to achieve almost perfect searching accuracy. Table 5 shows the performance when we fix $N_1 = 5$ and increase M from 1 to 10. Increasing sample size makes the problem easier and gives better searching efficiency. Similar as in the case of varying sparsity, the gain in accuracy by increasing M is smaller than that by increasing N_1 though the accuracy is slightly better for larger M on average.

	Sample Size	$n = 100$		$n = 50$		$n = 25$	
M	Search Proportion	ζ_g	χ_g	ζ_g	χ_g	ζ_g	χ_g
2	2.86×10^{-5}	0.99	0.996	0.97	0.996	0.92	0.966
4	5.72×10^{-5}	0.99	0.996	0.97	0.996	0.96	0.985
6	8.58×10^{-5}	0.99	0.996	0.98	0.991	0.94	0.971
8	1.14×10^{-4}	0.99	0.996	0.96	0.989	0.96	0.985
10	1.43×10^{-4}	1.00	1.000	0.99	0.998	0.96	0.983
12	1.72×10^{-4}	0.99	0.996	0.97	0.996	0.96	0.979
14	2.00×10^{-4}	1.00	1.000	0.97	0.996	0.97	0.991
16	2.28×10^{-4}	1.00	1.000	0.98	0.997	0.97	0.989
18	2.57×10^{-4}	1.00	1.000	0.99	0.998	0.97	0.990
20	2.86×10^{-4}	1.00	1.000	0.99	0.998	0.95	0.977
22	3.15×10^{-4}	1.00	1.000	0.99	0.998	0.96	0.982
24	3.43×10^{-4}	1.00	1.000	0.99	0.998	0.95	0.977
26	3.72×10^{-4}	1.00	1.000	0.99	0.998	0.95	0.977
28	4.00×10^{-4}	1.00	1.000	0.99	0.998	0.95	0.977
30	4.29×10^{-4}	1.00	1.000	0.99	0.998	0.95	0.977

Table 5: Performance metrics of DAG-W when one varies the sample size for $n = 100, 50, 25$ and the sparsity level is 0.1. Here we fix $N_1 = 5$ and vary M from 2 to 30.

4.2 Completion algorithms for computing the ML and MAP estimators

Algorithm 4.2 (Maximum Likelihood)**.** *Let S denote the sample covariance matrix for n i.i.d. observations and assume the sample size $n \geq \max\{\mathrm{pa}_i + 1 : i \in V\}$. For each i set*

$$\lambda_i = S_{ii|\mathrm{pa}(i)} \in \mathbb{R}_+ \quad \textit{and} \quad \beta_{\mathrm{pa}(i)} = S^{-1}_{\mathrm{pa}(i)} S_{\mathrm{pa}(i),i} \in \mathbb{R}^{\mathrm{pa}_i}.$$

Notice that $\lambda_i = S_{ii}$ whenever $\mathrm{pa}(i) = \emptyset$. For $i = p, p-1, \cdots 1$:

1. *Initialize $\widehat{\Sigma}_{ii} = \lambda_i$ for each i such that $\mathrm{pa}(i) = \emptyset$ (in particular for $i = p$);*
2. *set $\widehat{\Sigma}_{\mathrm{pa}(i),i} = \widehat{\Sigma}_{\mathrm{pa}(i)} \beta_{\mathrm{pa}(i)}$ if $\mathrm{pa}(i) \neq \emptyset$;*
3. *set $\widehat{\Sigma}_{ii} = \lambda_i + \beta^\top_{\mathrm{pa}(i)} \widehat{\Sigma}_{\mathrm{pa}(i)} \beta_{\mathrm{pa}(i)}$ if $\mathrm{pa}(i) \neq \emptyset$;*
4. *set $\widehat{\Sigma}_{\mathrm{pr}(i),i} = \widehat{\Sigma}_{\mathrm{pr}(i),\mathrm{pa}(i)} \widehat{\Sigma}^{-1}_{\mathrm{pa}(i)} \widehat{\Sigma}_{\mathrm{pa}(i),i}$ if $\mathrm{pa}(i) \neq \emptyset$, otherwise set $\widehat{\Sigma}_{\mathrm{pr}(i),i} = 0$.*

For the precision matrix Ω, the maximum likelihood estimator (MLE) denoted by $\hat{\Omega}$ is the inverse of $\hat{\Sigma}$.

Another commonly used Bayes estimator is the maximum a posteriori probability (MAP) estimator which can be obtained from the DAG-Wishart prior. Since the DAG-Wishart prior is a conjugate prior for Gaussian DAG model, by slightly modifying Algorithm 4.2 we can compute the MAP estimator for Σ, denoted by Σ_{MAP} as follows:

Algorithm 4.3 (Posterior Mode)**.** *Let S denote the sample covariance matrix obtained from n i.i.d. observations and assume the sample size $n \geq \max\{\mathrm{pa}_i + 1 : i \in V\}$.*

Initialization: For $i = 1, \cdots, p$, set $\lambda_i = \frac{(nS+U)_{ii|\mathrm{pa}(i)}}{\alpha_i + n}$, $\beta_i = -((nS+U)_{\mathrm{pa}(i)})^{-1}(nS+U)_{\mathrm{pa}(i),i}$. By default, $\lambda_i = \frac{(nS+U)_{ii}}{\alpha_i+n}$, $\beta_i = 0$ whenever $\mathrm{pa}(i) = \emptyset$.

Compute: For $i = p, p-1, \cdots, 1$, do step 1-4 in Algorithm 4.2.

Now Ω_{MAP}, the MAP estimator of Ω, is the inverse of Σ_{MAP}.

4.3 Covariance estimation performance

We now consider the problem of estimating covariance and precision matrices for data generated from a Gaussian DAG model. As in (Rajaratnam et al., 2008), we measure the accuracy of the estimators using two losses: the modified squared error loss and Stein's loss. The modified squared error loss, restricted to the functionally independent elements of covariance or precision matrix, is $L_2(M, \widehat{M}) = \sum_{(i \to j) \in E} (M_{ij} - \widehat{M}_{ij})^2$, where M is the true covariance or inverse covariance matrix and $\widehat{M}$ is its estimator. Stein's loss is a commonly used loss function and is $L_1(\widehat{M}, M) = \mathrm{tr}(\widehat{M} M^{-1}) - \log(\det(\widehat{M} M^{-1})) - p$.

Since the DAG Wishart distribution is a conjugate prior, to compute $\widehat{\Omega}_{\mathrm{BAYES}}$ and $\widehat{\Sigma}_{\mathrm{BAYES}}$, first we use Theorem 4.2, part (c), and Proposition 4.5 of the main paper to compute $\widehat{\Omega}^E_{\mathrm{BAYES}}$ and $\widehat{\Sigma}^E_{\mathrm{BAYES}}$. Then we use the completion process in Ben-David and Rajaratnam (2012) to complete these incomplete matrices and obtain $\widehat{\Omega}_{\mathrm{BAYES}}$ and $\widehat{\Sigma}_{\mathrm{BAYES}}$. The specific algorithms for computing the MLE and MAP estimators are described in

previous section. A similar proof as in (Rajaratnam et al., 2008) shows that $(\widehat{\Omega}_{\text{BAYES}})^{-1}$ and $(\widehat{\Sigma}_{\text{BAYES}})^{-1}$ are the Bayes estimates under Stein's loss.

We use the same data generating procedures as in the previous section. For the DAG-Wishart prior, we need $\alpha_i > \text{pa}_i + 2$. Here we choose the shape parameter as $\alpha_i = c \cdot \text{pa}_i + 3$, where $c = 2.5, 3, 3.5$. In addition, the scale parameter is chosen as $U = I(u) = u \cdot I$ for $u = 2.5, 3, 3.5$. For conciseness, we only show the performance of the estimators of the precision matrix Ω. The results for the estimation of Σ are also provided below. Table 6 shows the estimation performance as the relative improvement over the ML estimate given by the three Bayesian estimates for $p = 500$ and different sample sizes. The best improvement settings under each performance measure and sample size are shown by bold characters. As expected, the advantage of the Bayes estimators is more significant when the sample size n is small. We see, in particular, that when $n = 30$, the Bayes estimator can achieve up to more than 80% reduction for L_2 loss and also close to 50% reduction for L_1 loss. Moreover, it can be seen that different estimators are preferable under the two loss functions.

		n=30		n=50		n=100	
(c, U)	Estimator	L_1	L_2	L_1	L_2	L_1	L_2
$(2.5, I(3))$	$\widehat{\Omega}_{\text{BAYES}}$	41.8%	77.9%	26.8%	56.5%	14.2%	29.8%
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	45.8%	60.2%	29.8%	30.7%	15.9%	3.6%
	$\widehat{\Omega}_{\text{MAP}}$	38.7%	**82.0%**	23.9%	**63.0%**	12.3%	37.9%
$(3, I(3))$	$\widehat{\Omega}_{\text{BAYES}}$	39.2%	80.5%	24.7%	60.5%	12.9%	34.6%
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	47.4%	65.9%	31.1%	39.9%	16.7%	13.8%
	$\widehat{\Omega}_{\text{MAP}}$	34.4%	81.5%	20.1%	62.3%	9.7%	37.7%
$(3.5, I(3))$	$\widehat{\Omega}_{\text{BAYES}}$	35.9%	81.9%	21.9%	62.8%	11.1%	37.4%
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	**47.9%**	70.1%	**31.6%**	47.6%	**17.1%**	22.3%
	$\widehat{\Omega}_{\text{MAP}}$	29.5%	79.9%	15.7%	59.7%	6.7%	35.5%
$(3, I(2.5))$	$\widehat{\Omega}_{\text{BAYES}}$	34.5%	81.9%	20.1%	**63.0%**	10.3%	**38.6%**
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	47.8%	72.4%	31.3%	51.0%	16.8%	27.1%
	$\widehat{\Omega}_{\text{MAP}}$	26.6%	77.2%	13.2%	55.9%	5.1%	32.7%
$(3, I(3.5))$	$\widehat{\Omega}_{\text{BAYES}}$	42.9%	77.0%	27.0%	54.2%	14.4%	25.8%
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	45.6%	59.0%	29.6%	27.3%	15.7%	-2.6%
	$\widehat{\Omega}_{\text{MAP}}$	39.6%	81.9%	24.9%	62.6%	13.0%	36.5%

Table 6: The relative improvement given by Bayes estimators over the MLE when estimating Ω using L_1 and L_2 losses with dimension $p = 500$ and sample sizes $n = 30, 50, 100$.

Using different hyperparameters can result in very different performances. The choice of hyperparameters for the prior is context-specific. Here $c = 3$ and $u = 3$ seem to be a good pair of hyperparameters for estimating both Ω and Σ for our specific $p = 500$ and edge proportion 0.01. However, this might be not a good choice for other cases.

The performance of the estimators of Σ is shown in Table 7. A before, $p = 500$ and the random graph edge proportion is 0.01. Choosing the hyperparameter $(c, U) = (3.5, 3I)$ gives the best estimator among the other choices. The differences however are not very large. Using the Bayes estimator $\widehat{\Sigma}_{\text{BAYES}}$ is preferable

under L_2 loss and $(\widehat{\Omega}_{\text{BAYES}})^{-1}$ is the best under L_1 loss. As expected when the sample size is small, the risk reductions given by the Bayes estimators is more significant than in larger sample sizes.

		n=30		n=50		n=100	
(c, U)	Estimator	L_1	L_2	L_1	L_2	L_1	L_2
$(2.5, I(3))$	$\widehat{\Sigma}_{\text{BAYES}}$	-9.8%	4.6%	-8.0%	1.4%	-5.1%	0.2%
	$(\widehat{\Omega}_{\text{BAYES}})^{-1}$	27.9%	-113.2%	17.4%	-100.2%	8.8%	-68.7%
	Σ_{MAP}	27.4%	-32.0%	17.4%	-27.0%	9.1%	-16.4%
$(3, I(3))$	$\widehat{\Sigma}_{\text{BAYES}}$	1.0%	10.5%	-1.2%	5.4%	-1.4%	2.6%
	$(\widehat{\Omega}_{\text{BAYES}})^{-1}$	30.1%	-130.6%	19.2%	-115.6%	10.0%	-79.4%
	Σ_{MAP}	27.1%	-45.2%	17.0%	-38.2%	8.8%	-23.6%
$(3.5, I(3))$	$\widehat{\Sigma}_{\text{BAYES}}$	7.6%	**12.3%**	4.0%	**6.8%**	1.6%	**3.4%**
	$(\widehat{\Omega}_{\text{BAYES}})^{-1}$	**31.0%**	-148.1%	**19.9%**	-131.7%	**10.5%**	-90.8%
	Σ_{MAP}	26.0%	-58.8%	15.9%	-50.1%	8.0%	-31.7%
$(3, I(2.5))$	$\widehat{\Sigma}_{\text{BAYES}}$	7.9%	11.8%	4.5%	6.2%	2.0%	2.9%
	$(\widehat{\Omega}_{\text{BAYES}})^{-1}$	30.8%	-141.6%	19.7%	-124.3%	10.4%	-84.7%
	Σ_{MAP}	24.6%	-49.2%	14.8%	-42.1%	7.3%	-26.6
$(3, I(3.5))$	$\widehat{\Sigma}_{\text{BAYES}}$	-9.8%	8.2%	-8.7%	4.0%	-5.9%	1.8%
	$(\widehat{\Omega}_{\text{BAYES}})^{-1}$	27.4%	-120.3%	16.9%	-107.6%	8.4%	-74.5%
	Σ_{MAP}	27.4%	-41.6%	17.4%	-34.6%	9.1%	-21.0%

Table 7: The relative improvements over MLE on L_1 and L_2 losses brought by Bayes estimators when estimating Σ. Here we fix $p = 500$.

4.4 Role of sparsity and robustness to outliers

The previous results correspond to an underlying true graph with degree of sparsity (or edge proportion) equal to 0.01. We also investigate if the same hyperparameters work similarly well for graphs with different sparsity levels. Table 8 shows the relative improvement under different loss functions on graphs with edge proportion 0.005, 0.01, 0.015 and 0.02. All of the results use the configuration $c = 3, u = 3$. It can be seen that the sparsity of the graph is related to the performance of the estimators. In particular, even though $c = 3, u = 3$ constitute a good hyperparameters in the case of edge proportion 0.01, the same configuration does not work as well when the graph generating sparsity is increased to 0.015 or 0.02. In such denser situations, the Bayes estimators give better estimation only when the sample size is small (say, $n = 30$). To achieve better performance, one has to use other hyperparameter configurations. One pattern that is seemingly odd is that as the sample size increases, the difference between the performance of the Bayes estimator and that of the MLE increases. This is unexpected, as when n is larger enough, the performance of the MLE and the Bayes estimator should be essentially the same. The reason could be size that size $n = 100$ is far from being "large enough". All estimators have better estimation as we increase n but in this small range n, the performance of MLE improves more quickly with the increasing sample size. Figure 2 shows the L_2 loss of estimators for Ω for various values of n and when the edge proportion is 0.015.

We now test the robustness of the estimators to outliers. Instead of using purely Gaussian $\mathrm{N}(0, 1)$ in the simulation, we add outliers of $\mathrm{N}(0, 100)$ with probability 0.01 in the simulation. For the sake of comparison,

		n=30		n=50		n=100	
Edge proportion	Estimator	L_1	L_2	L_1	L_2	L_1	L_2
0.005	$\widehat{\Omega}_{\text{BAYES}}$	33.7%	72.7%	21.4%	53.7%	11.3%	32.5%
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	38.9%	68.3%	25.6%	51.0%	13.9%	31.8%
	Ω_{MAP}	22.0%	62.2%	11.8%	41.5%	5.1%	22.3%
0.01	$\widehat{\Omega}_{\text{BAYES}}$	39.2%	80.5%	24.7%	60.5%	12.9%	34.6%
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	47.4%	65.9%	31.1%	39.9%	16.7%	13.8%
	Ω_{MAP}	34.4%	81.5%	20.1%	62.3%	9.7%	37.7%
0.015	$\widehat{\Omega}_{\text{BAYES}}$	45.1%	57.9%	23.2%	-34.5%	8.6%	-244.4%
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	49.1%	45.8%	28.4%	-67.0%	8.3%	-307.8%
	Ω_{MAP}	48.6%	64.3%	26.9%	-14.3%	12.4%	-198.2%
0.02	$\widehat{\Omega}_{\text{BAYES}}$	38.1%	60.0%	-7.0%	-118.1%	-59.1%	-812.6%
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	36.3%	58.7%	-13.3%	-124.7%	-68.4%	-834.4%
	Ω_{MAP}	47.9%	60.8%	6.4%	-113.4%	-44.7%	-795.0%

Table 8: The relative improvement over MLE on L_1 and L_2 losses brought by Bayes estimators when estimating Ω in cases of different edge proportions. In small sample problems, the Bayes estimators is still preferable. But the performance of the MLE improves more quickly when the sample size increases. This indicates that the good hyperparameters for one particular sparsity might not be good if the sparsity changes.

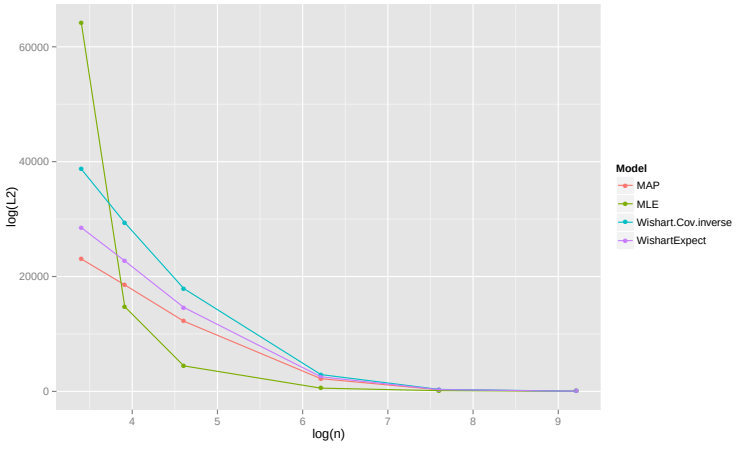

Figure 2: The L_2 loss for the estimators when n increases. Here the sample proportion is 0.015.

we still use the same hyperparameter configurations as in the previous section, but focus on the improvements under such contaminated data.

Table 9 includes the relative improvement of the Bayes estimators over the MLE when outliers are included. As compared with the Gaussian cases, the Bayes estimators have improved except in the case of L_2 loss on Σ, in which the difference is not significant. For instance, when $n = 30, p = 500$, the improvement on L_2 loss for the precision matrix is about 95% when outliers are included, compared with the improvement around 80% in the pure Gaussian case. Thus in general the Bayes estimators are more robust to outliers (or in some sense, misspecification of distributions) than the ML estimators.

		n=30		n=50		n=100	
Target	Estimator	L_1	L_2	L_1	L_2	L_1	L_2
Ω	$\widehat{\Omega}_{\text{BAYES}}$	65.1%	95.0%	50.9%	87.3%	32.2%	69.7%
	$(\widehat{\Sigma}_{\text{BAYES}})^{-1}$	72.1%	95.2%	58.0%	89.7%	37.8%	75.0%
	$\widehat{\Omega}_{\text{MAP}}$	59.7%	93.3%	44.9%	83.3%	27.0%	63.4%
Σ	$\widehat{\Sigma}_{\text{BAYES}}$	26.6%	9.2%	23.1%	4.6%	15.9%	1.6%
	$(\widehat{\Omega}_{\text{BAYES}})^{-1}$	45.1%	-67.5%	34.3%	-40.2%	21.5%	-16.4%
	$\widehat{\Sigma}_{\text{MAP}}$	17.7%	-13.8%	28.6%	-5.4%	17.7%	-1.0%

Table 9: The relative improvement over MLE on L_1 and L_2 losses brought by Bayes estimators under $\mathrm{N}(0, 100)$ outlier with probability 0.01. Here $p = 500$ and $(c, U) = (3, I(3))$.

4.5 Call center data

The call center data comes from a major financial institute in 2002. It contains all the calls to the center in that year. The center was staffed from 7 a.m. each day until midnight. The weekends, holidays and dysfunctional days are excluded, so we have 239 days in total. In each day from 7 a.m. to midnight, we divided the time into 10-min intervals, and the number of calls are denoted by N_{ij} where $i = 1, \cdots 239$ and $j = 1, \cdots, 102$. A transformation $x_{ij} = \sqrt{N_{ij} + 1/4}$ was applied to make the data closer to normal. Taking the 102 counts of time intervals as a vector, the data naturally contains a valid parent ordering, which is the time order. This is because it is not likely future counts could influence past.

The evaluation task here is the same as in (Bickel and Levina, 2008) and (Rajaratnam et al., 2008) and so we follow their description of the problem for this example. The goal is to predict the call counts in the intervals of the second half of the day conditional on the first half day call counts. We used the conditional distribution as the predictor. Let $x_i = (x_i^{(1)}, x_i^{(2)})$ be the ith observation, where $x_i^{(1)} = (x_{i,1}, \cdots, x_{i,51})$ be the first half covariates and $x_i^{(2)} = (x_{i,52}, \cdots, x_{i,102})$ be the second half. Then we partition the mean and covariance matrix to obtain $\mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. The conditional expectation predictor is $\hat{x}^{(2)} = \mu^{(2)} + \Sigma_{21}\Sigma_{11}^{-1}(x^{(1)} - \mu^{(1)})$. We use the first 10 months as the training data (205 observations in total)and the last 2 months (34 observations in total) as the test data. Then we estimate the mean and

covariance from the training data. One baseline method is to use the naive MLE (denoted Naive-MLE), that is the sample covariance matrix ignoring the potential underlying graphical structure. We can also use the constrained MLE after having estimated the graph structure using either `lassoDAG` or `DAG-W` method: these estimates are denoted by `lassoDAG` -MLE and `DAG-W` -MLE respectively. In addition, we include the Bayes estimator $(\widehat{\Omega}_{\text{BAYES}})^{-1}$ given the model selection results from the DAG-W: we denote it as the DAG-W-Precision (here we mainly want to compare the the various model selection procedures in the context of sparse covariance estimation. We acknowledge that better prediction can be achieved by other methods).

The prediction goal is a supervised task and one way to choose the hyperparameters can be cross-validation. However, to emphasize the effectiveness of our recommended hyper-parameter settings, here we use the recommended configurations $\kappa = 0.1$ in `lassoDAG` and $c = 1, b = 3$ in DAG-W. In Figure 3, the average absolute errors for the 51 time intervals are shown, where we define the average absolute error as $E_j = \frac{1}{34}\sum_{i=1}^{34} \mid x_{ij} - \hat{x}_{ij} \mid$ for $j = 1, 2, \cdots 51$.

It can be seen (see Fig 3) that `DAG-W` -MLE is nearly uniformly better than `lassoDAG` -MLE and Naive-MLE, while `DAG-W` -Precision is even better than the other three. Since the covariance estimation method is the same for `lassoDAG` -MLE and `DAG-W` -MLE, the difference indicates the advantage of model selection by `DAG-W` . Also `DAG-W` -Precision involves Bayes shrinkage in addition to the model selection, resulting in additional benefits. Also `lassoDAG` performs better than MLE in most of the time intervals, thus we see that any model selection is better than no model selection at all.

Another way to measure performance is to treat each sample $x_i \in \mathbb{R}^{51}$ as one individual and take the L_2 errors for each day in test set $SSE_i = \|x_i - \hat{x}_i\|^2$. The MSE for each method is then estimated by the average of SSE_i over all i in test set. Table 10 shows the MSE from the three different methods of covariance estimation. It can be seen that the performance of `DAG-W` is much better than that of `lassoDAG` or naive MLE.

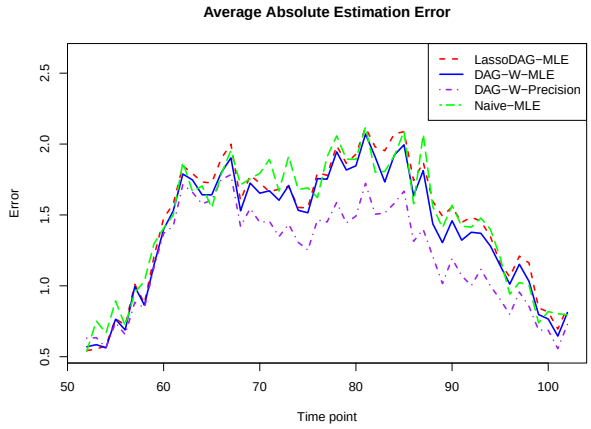

Figure 3: The average absolute errors for all 51 time intervals of second half of the day, from different predictions.

	Naive-MLE	`lassoDAG` -MLE	`DAG-W` -MLE	`DAG-W` -Precision
MSE	172.976	166.138	142.730	123.438

Table 10: Mean squared errors of predictions for the call center data given by different methods.

References

Ben-David, E. and Rajaratnam, B. (2012). "Positive definite completion problems for Bayesian networks." *SIAM J. Matrix Anal. Appl.*, 33(2): 617–638. 26

Bickel, P. J. and Levina, E. (2008). "Regularized estimation of large covariance matrices." *Ann. Statist.*, 36(1): 199–227. 30

Billingsley, P. (1979). *Probability and measure*. John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics. 20, 21

Dawid, P. A. and Lauritzen, S. L. (1993). "Hyper-Markov laws in the statistical analysis of decomposable graphical models." *Ann. Statist.*, 21(3): 1272–1317. 18

Diaconis, P., Khare, K., and Saloff-Coste, L. (2008). "Gibbs Sampling, Exponential Families and Orthogonal Polynomials." *Statistical Science*, 23(2): 151–178. 5

Geiger, D. and Heckerman, D. (2002). "Parameter priors for directed acyclic graphical models and the characterization of several probability distributions." *Ann. Statist.*, 30(5): 1412–1440. 17, 18

Khare, K. and Rajaratnam, B. (2011). "Wishart distributions for decomposable covariance graph models." *Ann. Statist.*, 39(1): 514–555. 14, 15, 17

Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications. 1, 2, 3

Letac, G. and Massam, H. (2007). "Wishart distributions for decomposable graphs." *Ann. Statist.*, 35(3): 1278–1323. 15

Pearl, J. and Wermuth, N. (1994). "When can association graphs admit a causal interpretation?" In Cheeseman, P. and Oldford, R. (eds.), *Selecting Models from Data*, volume 89 of *Lecture Notes in Statistics*, 205–214. Springer New York. 15

Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). "Flexible covariance estimation in graphical Gaussian models." *Ann. Statist.*, 36(6): 2818–2849. 26, 27, 30

Roverato, A. (2000). "Cholesky decomposition of a hyper inverse Wishart matrix." *Biometrika*, 87(1): 99–112. 12

Shojaie, A. and Michailidis, G. (2010). "Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs." *Biometrika*, 97(3): 519–538. 22

Temme, N. M. (1996). *Special functions : an introduction to the classical functions of mathematical physics*. New York: J. Wiley & sons. 9

Wermuth, N. (1980). "Linear recursive equations, covariance selection, and path analysis." *J. Amer. Statist. Assoc.*, 75(372): 963–972. 4