arXiv:1109.4371v6 [math.ST] 23 Aug 2026

High dimensional Bayesian inference for Gaussian directed acyclic graph models

Emanuel Ben-David [*1], Tianxi Li [†2], Hélène Massam[‡3], and Bala Rajaratnam [§4]

[1]Department of Statistics, Columbia University
[2]Department of Statistics, University of Michigan
[3]Department of Mathematics and Statistics, York University
[4]Department of Statistics, Stanford University

Corrected working preprint v6.2, August 2026

Abstract

We study centered Gaussian models Markov with respect to a directed acyclic graph (DAG) whose vertices have a fixed parent ordering. We construct a conjugate family on the modified Cholesky parameters, with one shape parameter per vertex, and derive its induced distributions on incomplete covariance and precision coordinates. The distribution is proper exactly when $\alpha_i > |\mathrm{pa}(i)| + 2$ for every vertex, and its nodewise conditional-variance and regression parameters are independent across vertices. This factorization gives a closed-form normalizing constant, conjugate updating, marginal likelihoods, and explicit full-matrix posterior means. We distinguish these posterior means from nonlinear completions of incomplete-coordinate means. We also distinguish the transformed Cholesky-coordinate mode from modes defined using covariance or precision coordinates. The model-selection procedure searches only over DAGs compatible with the specified ordering. Historical simulation and data examples illustrate the method; their evidentiary limitations and reproducibility requirements are stated explicitly.

[*]bendavid@stat.columbia.edu
[†]tianxili@umich.edu
[‡]massamh@mathstat.yorku.ca
[§]brajarat@stanford.edu

Revision note. This working correction preserves the model and main construction of the 2015 arXiv version while repairing non-integrable displayed kernels, sign inconsistencies, covariance maps, posterior-mean and mode interpretations, and the Hausdorff-measure example. Version 6.1 also removes an invalid scalar density attributed to the inverse type-II Wishart and replaces the asserted family identity by the precise block-factorization relationship supported by [17]. Empirical claims that cannot be reproduced from the public artifacts are labeled as historically reported rather than independently verified. Version 6.2 replaces every externally rendered DAG figure, including all three molecular-network panels, by native TikZ source embedded in this document.

1 Introduction

Priors for Gaussian distributions Markov with respect to a DAG have a long history beginning with the compatible parameter priors studied by Geiger and Heckerman [10, 9]. Dawid and Lauritzen [6] introduced the hyper inverse Wishart for decomposable undirected graphs, and Letac and Massam [17] developed the richer type-I and type-II Wishart families on the corresponding cones. For a perfect orientation of a decomposable graph, Theorem 4.4 of [17] gives a block inverse-Wishart/matrix-normal factorization of the inverse type-II law. This is the decomposable-graph analogue of the vertexwise conditional-variance/regression factorization below.

The relationship is a specialization, not an unrestricted identity of families. The type-II shapes are attached to clique and separator blocks and satisfy admissibility constraints, whereas the present construction uses one free shape parameter per vertex. Compatible tied choices recover the classical Wishart and hyper-inverse-Wishart special cases, and certain perfect-DAG specializations correspond to inverse type-II laws. Arbitrary vertexwise DAG–Wishart shapes need not admit a type-II parameterization. Our construction therefore starts directly on modified-Cholesky coordinates for an arbitrary DAG and derives the covariance and precision images afterward. The resulting nodewise factorization gives closed-form normalizing constants and posterior moments.

For a non-perfect DAG, the covariance and precision parameter spaces are typically curved submanifolds of the ambient symmetric-matrix space and therefore have ambient Lebesgue measure zero. A distribution on either space is nevertheless well defined as a pushforward measure; what fails is the existence of an ambient-Lebesgue density. We use projections onto Euclidean spaces of functionally independent covariance and precision entries, following the completion results in [2]. This construction yields tractable coordinate densities. A completion of a coordinatewise expectation must not, however, be confused with the expectation of the completed random matrix; Sections 4 and 5 give the full posterior-mean formulas.

The remainder is organized as follows. Section 2 fixes the Gaussian-DAG and modified-Cholesky notation. Section 3 constructs the DAG–Wishart on $\Theta_{\mathcal{D}}$. Sections 4

and 5 derive its images on incomplete precision and covariance coordinates, including the corrected full-matrix moments. Section 6 preserves the fixed-order simulation and data examples while separating reproducible facts from historical numerical reports. Section 7 states the main limitations; Supplemental Sections A–D provide proofs, geometry, and algorithms.

2 Preliminaries

A brief summary of graph theory, associated Markov and other properties required for analyzing DAG models is given in Supplemental section A.

2.1 Gaussian DAG models

Let V be a set with p elements. For any $a, b \subseteq V$ [1] let $\mathbb{R}^{a\times b}$ denote the real linear space of functions $A \equiv \left((i,j) \mapsto A_{ij}\right) : a \times b \to \mathbb{R}$. Each element of $\mathbb{R}^{a\times b}$ is called an $|a| \times |b|$ matrix. In particular, we define the space of symmetric matrices $S_a(\mathbb{R}) = \left\{A \in \mathbb{R}^{a\times a} : A_{ij} = A_{ji}\ ,\ \text{for every } i, j \in a\right\}$, and the set of positive definite matrices $\mathrm{PD}_a(\mathbb{R}) = \{A \in S_a(\mathbb{R}) :\ x^\top A x > 0,\ \text{for every}\ \ x \in \mathbb{R}^a \setminus \{0\}\}$. Now let $\Sigma > 0$ denote $\Sigma \in \mathrm{PD}_p(\mathbb{R})$. For a partition a, b of V, consider the corresponding block partitioning of Σ as follows.

$$\Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix},$$

where $\Sigma_{aa} = (\Sigma_{ij})_{i,j\in a} \in \mathrm{PD}_a(\mathbb{R})$, $\Sigma_{bb} = (\Sigma_{ij})_{i,j\in b} \in \mathrm{PD}_b(\mathbb{R})$, $\Sigma_{ab} = (\Sigma_{ij})_{i\in a, j\in b} \in \mathbb{R}^{a\times b}$ and $\Sigma_{ba} = \Sigma_{ab}^\top$. The Schur complement of the sub-matrix Σ_{aa} is defined as $\Sigma_{bb|a} = \Sigma_{bb} - \Sigma_{ba}(\Sigma_{aa})^{-1}\Sigma_{ab}$.

Remark 2.1. Throughout this paper, we shall in general suppress the notation for a principal submatrix Σ_{aa} and refer to it as Σ_a. We shall also use the convention Σ_a^{-1} for $(\Sigma_{aa})^{-1}$ and Σ^a for $(\Sigma^{-1})_{aa}$.

In this paper we focus on multivariate Gaussian distributions which obey the directed Markov property with respect to a DAG $\mathcal{D}$. From now on and unless otherwise stated, we shall always assume without loss of generality that $\mathcal{D} = (V, E)$ is a DAG given in a parent ordering[2], i.e., the vertices are labeled $1, 2, \ldots, p$, and $i \to j$ implies that $i > j$. A Gaussian DAG model (or Gaussian Bayesian network) over $\mathcal{D}$, denoted by $\mathcal{N}(\mathcal{D})$, is the statistical model that consists of all multivariate Gaussian distributions

[1] Lowercase letters denote subsets of V.

[2] We emphasize here that unlike in the decomposable precision graph setting or the covariance graph setting (where the existence of an ordering is important either for the perfect order of cliques and separators, or to preserve zeros), existence of such an ordering is not necessary in the DAG setting, since a parent ordering is always available for a DAG.

$N_p(\mu, \Sigma)$ obeying the ordered directed Markov property with respect to $\mathcal{D}$. Therefore, $\mathbf{x} \sim N_p(\mu, \Sigma) \in \mathscr{N}(\mathcal{D}) \implies x_i \perp\!\!\!\perp \mathbf{x}_{\{i+1,\ldots,p\}\setminus pa(i)} | \mathbf{x}_{pa(i)}$ for each i.

Remark 2.2. Note that $N_p(\mu, \Sigma) \in \mathscr{N}(\mathcal{D})$ if and only if $N_p(0, \Sigma) \in \mathscr{N}(\mathcal{D})$. Therefore, without loss of generality, we shall only consider centered Gaussian distributions

$$\left\{N_p(0, \Sigma) : \Sigma \in PD_{\mathcal{D}}\right\} \subseteq \mathscr{N}(\mathcal{D}).$$

For convenience, with a slight abuse of notation, we shall still denote

$$\mathscr{N}(\mathcal{D}) = \left\{N_p(0, \Sigma) : \Sigma \in PD_{\mathcal{D}}\right\}.$$

The Gaussian distributions in $\mathscr{N}(\mathcal{D})$ are naturally parametrized by the elements of

$$PD_{\mathcal{D}} = \left\{\Sigma \succ 0 : N_p(0, \Sigma) \in \mathscr{N}(\mathcal{D})\right\} \text{ or } P_{\mathcal{D}} = \left\{\Sigma^{-1} \succ 0 : N_p(0, \Sigma) \in \mathscr{N}(\mathcal{D})\right\}.$$

These sets are referred to as the space of covariance matrices and the space of precision matrices. A precision matrix in $P_{\mathcal{D}}$ is usually denoted by Ω. Similarly, for an undirected graph $\mathcal{G} = (V, \mathscr{E})$ we define $\mathscr{N}(\mathcal{G})$ as the set of multivariate Gaussian distributions obeying the (undirected) Markov property with respect to $\mathcal{G}$. In this model the corresponding parameter spaces are the space of covariance matrices $PD_{\mathcal{G}} = \left\{\Sigma : N_p(0, \Sigma) \in \mathscr{N}(\mathcal{G})\right\}$ and the space of precision matrices $P_{\mathcal{G}} = \left\{\Omega : \Omega^{-1} \in PD_{\mathcal{G}}\right\}$. Note that, for us, $PD_{\mathcal{D}}$ and $P_{\mathcal{D}}$ are parameter spaces of primary interest as they arise naturally in the parameterization of Gaussian densities. However, in order to develop multi-shape parameters Wishart priors on these spaces, which is the main purpose of this paper,we begin with the more natural and more convenient Cholesky type parameterization of $\mathscr{N}(\mathcal{D})$ that we discuss in the next subsection.

2.2 Cholesky parametrizations of Gaussian DAG models

Consider a Gaussian DAG distribution $N_p(0, \Sigma) \in \mathscr{N}(\mathcal{D})$. Let $\mathcal{L}_{\mathcal{D}}$ denote the set of lower triangular matrices with unit diagonal and $L_{ij} = 0$ if $i \notin pa(j)$, and let $\mathcal{D}_+^p$ denote the set of strictly positive diagonal matrices in $\mathbb{R}^{p\times p}$. Then $\Sigma^{-1} \in P_{\mathcal{D}}$ if and only if there is a unique $(D, L) \in \mathcal{D}_+^p \times \mathcal{L}_{\mathcal{D}}$ such that

$$\Omega = \Sigma^{-1} = LD^{-1}L^\top, \qquad \Sigma = L^{-\top}DL^{-1}.$$

The first identity is the modified Cholesky decomposition of Ω. We call $\Theta_{\mathcal{D}} = \mathcal{D}_+^p \times \mathcal{L}_{\mathcal{D}}$ the Cholesky space of $\mathcal{D}$, and

$$\left\{N_p(0, L^{-\top}DL^{-1}) : (D, L) \in \Theta_{\mathcal{D}}\right\} \equiv \mathscr{N}(\mathcal{D})$$

the Cholesky parametrization of $\mathscr{N}(\mathcal{D})$.

We can also obtain a variant of this parameterization, in vector form, from the recursive factorization property of the Gaussian densities in $\mathscr{N}(\mathcal{D})$ (see Supplemental section A subsection 1.3 for details). First, let us recall the following notation from [1].
Notation. For each $i \in V$ let

$$\begin{array}{lll} \prec i \succ = pa(i) & [i \succ = \{i\} \times pa(i) & \prec i] = pa(i) \times \{i\}, \\ \not\prec i \not\succ = \{j : j \succ i\} \setminus pa(i) & [i \not\succ = \{i\} \times \not\prec i \not\succ & \prec i \not\succ = \prec i \succ \times \not\prec i \not\succ \\ \preceq i \succeq = fa(i) & & \end{array}$$

By applying the directed factorization property (DF) of $\mathrm{N_p}(0, \Sigma) \in \mathscr{N}(\mathcal{D})$ we have

$$\begin{aligned} d\mathrm{N_p}(0,\Sigma)(\mathrm{x}) &= \prod_{i \in V} d\mathrm{N}(\mu_{\mathrm{i}|\mathrm{pa(i)}}, \Sigma_{\mathrm{ii}|\mathrm{pa(i)}})(\mathrm{x_i}|\mathrm{x_{pa_i}}) \\ &= \prod_{i \in V} d\mathrm{N}(\Sigma_{[\mathrm{i}\succ}\Sigma^{-1}_{\prec \mathrm{i}\succ}\mathrm{x}_{\prec \mathrm{i}\succ}, \Sigma_{\mathrm{ii}|\prec \mathrm{i}\succ})(\mathrm{x_i}), \end{aligned} \tag{1}$$

for each $x = (x_i)_{i \in V} \in \mathbb{R}^p$. Note that $\mathrm{N}(\Sigma_{[\mathrm{i}\succ}\Sigma^{-1}_{\prec \mathrm{i}\succ}\mathrm{x}_{\prec \mathrm{i}\succ}, \Sigma_{\mathrm{ii}|\prec \mathrm{i}\succ})(\mathrm{x_i})$ is the conditional distribution of $\mathbf{x}_i|\mathbf{x}_{\prec i\succ} = x_{\prec i\succ}$. Moreover, $\Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ}$ is the regression coefficient of $\mathbf{x}_i$ in the regression of $\mathbf{x}_i$ on $\mathbf{x}_{\prec i\succ}$, and $\Sigma_{ii|\prec i\succ}$ is the conditional variance of $\mathbf{x}_i|\mathbf{x}_{\prec i\succ} = x_{\prec i\succ}$. Furthermore, using the exact functional form of the densities of the Gaussian distributions in (1), we obtain the following equation.

$$\mathrm{tr}(\Sigma^{-1} x x^\top) = \sum_{i \in V} \mathrm{tr}\left(\Sigma^{-1}_{ii|\prec i\succ}(x_i - \Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ}x_{\prec i\succ})(x_i - \Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ}x_{\prec i\succ})^\top\right) \tag{2}$$

It is shown in [1] that $\Sigma \in \mathrm{PD}_{\mathcal{D}}$ if and only if $\Sigma \succ 0$ and satisfies (2) for all $x \in \mathbb{R}^p$. On the other hand, by the parent ordered Markov property of $\mathbf{x} \sim \mathrm{N_p}(0, \Sigma)$ we have $\Sigma \in \mathrm{PD}_{\mathcal{D}}$ if $i \perp\!\!\!\perp \not\prec i \not\succ \mid \prec i \succ$ (or equivalently $i \perp\!\!\!\perp \{i+1, \ldots, p\} \setminus pa(i)|pa(i)$). Hence another characterization given by [1] for $\Sigma \in \mathrm{PD}_{\mathcal{D}}$ is that $\Sigma \succ 0$ and

$$\Sigma_{[i\not\succ} = \Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ}\Sigma_{\prec i\not\succ}, \text{ for every } i \in V. \tag{3}$$

We use two related vectors and do not interchange their signs:

$$\ell_i := L_{\prec i]}, \qquad \beta_i := -\ell_i = \Sigma^{-1}_{\prec i\succ}\Sigma_{\prec i]}.$$

Thus ℓ_i is the free Cholesky column and β_i is the regression coefficient of X_i on $X_{\mathrm{pa}(i)}$. Defining $\Xi_{\mathcal{D}} = \times_{i \in V}\left(\mathbb{R}_+ \times \mathbb{R}^{\prec i]}\right)$, it can be shown that the mapping

$$\Pi_{\mathcal{D}} \equiv \left(\Sigma \mapsto \times_{i \in V}(\Sigma_{ii|\prec i\succ}, \Sigma^{-1}_{\prec i\succ}\Sigma_{\prec i]})\right) : \mathrm{PD}_{\mathcal{D}} \to \Xi_{\mathcal{D}} \tag{4}$$

is a diffeomorphism. In order to construct the inverse of this mapping let $\times_{i \in V}(\lambda_i, \beta_{\prec i]})$ denote a typical element in $\Xi_{\mathcal{D}}$, with the convention that $\beta_{\prec i]} = 0$ whenever $\prec i \succ =$

$pa(i) = \emptyset$. Using (3), the corresponding Σ can be recursively constructed starting from the largest index p, by setting

$$\begin{cases} i) & \Sigma_{ii} = \lambda_i + \beta_{<i]}^\top \Sigma_{<i>} \beta_{<i]}; \\ ii) & \Sigma_{<i]} = \Sigma_{<i>} \beta_{<i]}; \\ iii) & \Sigma_{[i\not>} = \Sigma_{[i>} \Sigma_{<i>}^{-1} \Sigma_{<i\not>}. \end{cases} \tag{5}$$

The reader is referred to [1] for greater detail, where in addition, it is shown that the inverse mapping above yields a positive definite matrix in $\mathrm{PD}_p(\mathbb{R})$, and consequently in $\mathrm{PD}_{\mathcal{D}}$. The mapping $\Pi_{\mathcal{D}}$ in (4) gives another parametrization of $\mathscr{N}(\mathcal{D})$ in terms of the elements $\times_{i\in V}(\lambda_i, \beta_{<i]}) \in \Xi_{\mathcal{D}}$. One can show that for each $i \in V$, $L_{<i]} = -\Sigma_{<i>}^{-1} \Sigma_{<i]}$ and $D_{ii} = \Sigma_{ii|<i>}$, therefore each $\times_{i\in V}(\lambda_i, \beta_{<i]}) \in \Xi_{\mathcal{D}}$ is, essentially, a vectorized form of a $(D, L) \in \Theta_{\mathcal{D}}$.

3 The DAG-Wishart distribution on $\Theta_{\mathcal{D}}$

This section introduces a multi-shape family on the Cholesky space $\Theta_{\mathcal{D}}$ by generalizing the modified-Cholesky image of an ordinary Wishart law. The family is defined for every DAG. Its relationship to the inverse type-II Wishart is through compatible perfect-DAG specializations and the parallel directed-hyper-Markov factorization, not through equality for arbitrary vertexwise shapes.

3.1 DAG-Wishart densities

Let us start with a natural course that will lead us to the general form of the multi-shape parameter DAG-Wishart distributions on the Cholesky space $\Theta_{\mathcal{D}}$ with the desired properties. We begin with the classical Wishart distribution. Let us consider $\Sigma^{-1} \sim \mathrm{W_p}(\eta, \mathrm{U})$ as a prior for the precision parameter of the full Gaussian model $\{\mathrm{N_p}(0, \Sigma) : \Sigma \succ 0\}$. Note that this model corresponds to the saturated Gaussian DAG model $\mathscr{N}(\mathcal{D})$, i.e., when $\mathcal{D}$ is a complete DAG with p vertices (see Figure 1). Consider the mapping $\Sigma^{-1} \mapsto (D, L)$, where (D, L) is the Cholesky factorization of Σ^{-1}. This mapping transforms the Wishart distribution $\mathrm{W_p}(\eta, \mathrm{U})$ to a distribution on $\Theta_{\mathcal{D}}$ with density proportional to

$$\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left(LD^{-1}L^\top\right)U\right)\right\} \prod_{i=1}^{p} D_{ii}^{-\frac{\alpha_i}{2}}, \tag{6}$$

with $\alpha_i = \eta + p - 2i + 3$. Indeed, for the complete parent ordering, the absolute Jacobian of $(D, L) \mapsto LD^{-1}L^\top$ is $\prod_{i=1}^{p} D_{ii}^{-(p-i+2)}$. Although the α_i's in (6) look like multiple shape parameters, they are functions of the single Wishart shape η. We now free these shape parameters and use the same kernel for an arbitrary DAG.

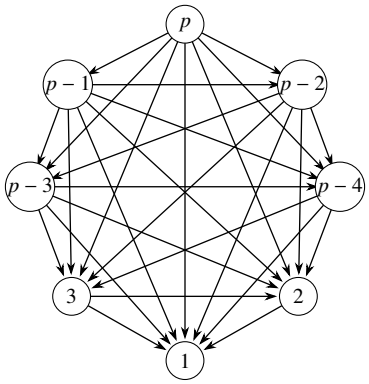

Figure 1: Schematic of a complete DAG in the parent ordering. The eight displayed vertices illustrate the saturated orientation; intermediate labels are suppressed when $p > 8$.

For context, let $\mathcal{G}$ be decomposable and let $\mathcal{D}$ be a perfect orientation compatible with a perfect clique order. Theorem 4.4 of [17] states that the inverse type-II Wishart is strong directed hyper Markov: its residual covariance blocks have inverse-Wishart laws, the associated regression blocks have conditional matrix-normal laws, and the listed block pairs are mutually independent. The theorem is intrinsically matrix-valued. It does not yield the scalar quadratic-form density printed in earlier drafts of this manuscript, and we do not use it to prove any result for a non-perfect DAG. The vertexwise analogue follows directly below by completing the square in the Cholesky density.

Let $pa_i = |\mathrm{pa}(i)|$ and

$$a_i = \frac{\alpha_i - pa_i}{2} - 1, \qquad U_{ii|<i>} = U_{ii} - U_{[i>}U_{<i>}^{-1}U_{<i]}.$$

For an empty parent set, determinants below equal one and the associated vectors and quadratic forms are absent.

Definition 3.1. For $U \in \mathrm{PD}_p(\mathbb{R})$ and $\alpha_i > pa_i + 2$ for every i, the DAG–Wishart distribution on $\Theta_{\mathcal{D}}$ has density

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

with respect to $\prod_i dD_{ii} \prod_i \prod_{j\in\mathrm{pa}(i)} dL_{ji}$, where

$$z_{\mathcal{D}}(U, \alpha) = \prod_{i=1}^{p} 2^{\alpha_i/2-1}\pi^{pa_i/2}\Gamma(a_i)\det(U_{<i>})^{-1/2}U_{ii|<i>}^{-a_i}. \tag{7}$$

Equivalently, using $\det(U_{\preceq i\succeq}) = \det(U_{<i>})U_{ii|<i>}$, this is the determinant-ratio expression used in earlier versions.

Proposition 3.2. *The kernel in Definition 3.1 is integrable if and only if $\alpha_i > pa_i + 2$ for all i. Under the normalized law, the pairs (D_{ii}, β_i), $i = 1, \dots, p$, are mutually independent and*

$$D_{ii} \sim IG\left(a_i, \frac{1}{2}U_{ii|<i>}\right), \tag{8}$$

$$\beta_i \mid D_{ii} \sim \mathrm{N}_{pa_i}\left(U_{<i>}^{-1}U_{<i]}, D_{ii}U_{<i>}^{-1}\right). \tag{9}$$

Here $IG(a, b)$ has density $b^a\Gamma(a)^{-1}x^{-a-1}\exp(-b/x)$, $x > 0$. Equivalently, $\ell_i \mid D_{ii} \sim \mathrm{N}_{pa_i}(-U_{<i>}^{-1}U_{<i]}, D_{ii}U_{<i>}^{-1})$.

Proof. Since $L_{<i]} = -\beta_i$, completing the square gives

$$\mathrm{tr}(LD^{-1}L^\top U) = \sum_{i=1}^{p} D_{ii}^{-1}\left\{U_{ii|<i>} + (\beta_i - U_{<i>}^{-1}U_{<i]})^\top U_{<i>}(\beta_i - U_{<i>}^{-1}U_{<i]})\right\}.$$

The density therefore factorizes by vertex. Gaussian integration over β_i, followed by inverse-gamma integration over D_{ii}, gives (7); the latter integral is finite exactly when $a_i > 0$. The two conditional laws and mutual independence follow from the same factorization. □

This nodewise factorization is the directed hyper-Markov property used below. Conjugacy follows from the Gaussian likelihood. If

$$Y_r \overset{\text{iid}}{\sim} \mathrm{N}_p(0, L^{-\top}DL^{-1}), \qquad T = \sum_{r=1}^{n} Y_r Y_r^\top = nS,$$

then

$$(D, L) \mid Y_{1:n} \sim \pi^{\Theta_{\mathcal{D}}}_{U+T,\, \alpha+n\mathbf{1}}.$$

Remark 3.3. For fixed α, allowing every $U \succ 0$ is generally nonidentifiable: the density depends only on the family blocks used in the nodewise factorization. Restricting U to $\mathrm{PD}_{\mathcal{D}}$, or equivalently parameterizing it by its free entries U^E, restores identifiability. This restricted family is a full regular exponential family when $\mathcal{D}$ is perfect and has a curved representation in the larger moral-graph natural-parameter space otherwise; Supplemental Section B gives the precise statement.

4 The DAG-Wishart distribution on the space of incomplete precision matrices

The preceding section defined the DAG–Wishart $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ in modified-Cholesky coordinates. We now derive its pushforward to incomplete precision coordinates, extending the DAG analogue of the type-II-Wishart viewpoint beyond decomposable graphs [17].

4.1 Motivation and notation

To follow in the tradition of Wishart-type conjugate priors, we derive the precision and covariance images of $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ for $\mathrm{N}_p(0,\Sigma) \in \mathscr{N}(\mathcal{D})$ under the mappings

$$\left((D,L) \mapsto LD^{-1}L^{\top}\right) : \Theta_{\mathcal{D}} \to \mathrm{P}_{\mathcal{D}} \tag{10}$$

$$\left((D,L) \mapsto \left(LD^{-1}L^{\top}\right)^{-1}\right) : \Theta_{\mathcal{D}} \to \mathrm{PD}_{\mathcal{D}}. \tag{11}$$

Viewed inside the full space $\mathrm{S}_p(\mathbb{R})$, every lower-dimensional DAG model has ambient Lebesgue measure zero. Its intrinsic geometry is more informative: for a perfect DAG, the precision model is open in $\mathrm{Z}_{\mathcal{D}}$, whereas for a non-perfect DAG it is a curved subset of the larger moral-graph space $\mathrm{Z}_{\mathcal{D}^m}$. Analogous intrinsic-coordinate issues arise for covariance matrices. One can derive a precision density with respect to Hausdorff measure, but even for small non-perfect DAGs its area factor complicates posterior analysis; Supplemental section C gives the precise construction and an example.

To overcome this problem, we follow what was done for the hyper inverse Wishart in [16] or for the type I Wishart in [17] and we work with the projections of $\mathrm{P}_{\mathcal{D}}$ and $\mathrm{PD}_{\mathcal{D}}$ onto the Euclidean space that only retain the functionally independent elements of the precision and covariance matrices of Gaussian DAG models.

The projected spaces, as we shall see, are subsets of incomplete matrices, which we call the incomplete precision space $\mathrm{R}_{\mathcal{D}}$ and the incomplete covariance space $\mathrm{S}_{\mathcal{D}}$, respectively. The precise definitions are as follows.

Definition 4.1. Let $\mathcal{D} = (V,E)$ be a DAG[3] and $\mathcal{D}^u = (V,E^u)$ its undirected version.

(*a*) 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, for $i \neq j$, $A_{ij} = 0$ whenever neither (i,j) nor (j,i) belongs to E. Thus the allowed off-diagonal positions are the edges of the skeleton $\mathcal{D}^u$. Because E includes the loops, the dimension of $\mathrm{Z}_{\mathcal{D}}$ is $|E|$.

[3]Note an important convention here that the edge set E contains all the loops (see Supplemental section A for details).

(*b*) Let $\mathrm{I}_{\mathcal{D}}$ denote the real linear space of symmetric functions $\Gamma = \left(\Gamma_{ij}\right)_{(i,j)\in E^u}$, i.e., $\Gamma_{ij} = \Gamma_{ji} \in \mathbb{R}$ for each $(i, j) \in E$. An element $\Gamma \in \mathrm{I}_{\mathcal{D}}$ is called a (symmetric) $\mathcal{D}$-incomplete matrix, and can be considered as a matrix in $\mathrm{S}_p(\mathbb{R})$ where only the entries corresponding to the edges of $\mathcal{D}$ are specified and the rest are unspecified. The projection mapping from $\mathrm{S}_p(\mathbb{R})$ onto $\mathrm{I}_{\mathcal{D}}$ is denoted by $A \mapsto A^E$

(*c*) For $\Gamma \in \mathrm{I}_{\mathcal{D}}$ let $(\Gamma)^0$ denote the $|V| \times |V|$ matrix

$$(\Gamma)^0_{ij} = \begin{cases} \Gamma_{ij} & \text{if } (i, j) \in E^u, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $(\Gamma)^0$ fills or completes the unspecified positions with zeros to obtain a full matrix in $\mathrm{Z}_{\mathcal{D}}$. For each clique c of $\mathcal{D}$ the restriction of Γ on c, denoted by Γ_c, is a full matrix. Moreover, Γ is uniquely determined by the blocks of matrices $(\Gamma_c : c \in \mathscr{C}_{\mathcal{D}})$, where $\mathscr{C}_{\mathcal{D}}$ denotes the set of cliques of $\mathcal{D}^u$.

(*d*) Let $\mathrm{Q}_{\mathcal{D}} \subseteq \mathrm{I}_{\mathcal{D}}$ denote the set of $\mathcal{D}$-incomplete matrices $\Gamma \in \mathrm{I}_{\mathcal{D}}$ such that Γ_c is positive definite for each clique $c \in \mathscr{C}_{\mathcal{D}}$. Each element of $\mathrm{Q}_{\mathcal{D}}$ is said to be a partially positive definite matrix over $\mathcal{D}$.

(*e*) Let $\mathcal{A} \subseteq \mathrm{S}_p(\mathbb{R})$. We say that a $\mathcal{D}$-incomplete matrix $\Gamma \in \mathrm{I}_{\mathcal{D}}$ can be completed in $\mathcal{A}$ if there exists a matrix $A \in \mathcal{A}$ such that $A_{ij} = \Gamma_{ij}$ for each $(i, j) \in E$, i.e., $A^E = \Gamma$. We refer to A as a completion of Γ in $\mathcal{A}$.

(*f*) The space of incomplete precision matrices over $\mathcal{D}$, denoted by $\mathrm{R}_{\mathcal{D}}$, is the set of $\Gamma \in \mathrm{I}_{\mathcal{D}}$ that can be completed in the space of precision matrices $\mathrm{P}_{\mathcal{D}}$.

(*g*) The space of incomplete covariance matrices over $\mathcal{D}$, denoted by $\mathrm{S}_{\mathcal{D}}$, is the set of $\Upsilon \in \mathrm{I}_{\mathcal{D}}$ that can be completed in the space of covariance matrices $\mathrm{PD}_{\mathcal{D}}$.

Remark 4.2. If $\mathcal{A}$ is the set of positive definite matrices $\mathrm{PD}_p(\mathbb{R})$, then the completion in $\mathcal{A}$ reduces to the standard definition of positive definite completion [11]. We shall consider below the positive definite completion of partially positive precision/covariance matrices that correspond to DAGs (vs. those that correspond to undirected graphs as in [11]). Note that an incomplete matrix $\Gamma \in \mathrm{I}_{\mathcal{D}}$ has a positive definite completion only if $\Gamma \in \mathrm{Q}_{\mathcal{D}}$, i.e., it is partially positive definite over $\mathcal{D}$.

4.2 The space of incomplete precision matrices

We first recall the definition of $\mathrm{R}_{\mathcal{D}}$ in Definition 4.1 and the following result from [2].

Proposition 4.3. *[2] Let* Γ *be a* $\mathcal{D}$*-partial matrix in* $\mathrm{I}_{\mathcal{D}}$*. If* $\Gamma_{11} \neq 0$*, then*

(*a*) *Almost everywhere (with respect to the Lebesgue measure on* $\mathrm{I}_{\mathcal{D}}$*), there exist a unique lower triangular matrix* $L \in \mathcal{L}_{\mathcal{D}}$ *and a unique diagonal matrix* $\Lambda \in \mathbb{R}^{p\times p}$ *such that* $\widehat{\Gamma} = L\Lambda L^{\top}$ *is a completion of* Γ.

(*b*) *The matrix* $\widehat{\Gamma}$ *is the unique positive definite completion of* Γ *in* $\mathrm{P}_{\mathcal{D}}$ *if and only if the diagonal entries of* Λ *are all strictly positive.*

Proposition 4.3 is of interest to us, because it explicitly shows that without loss of generality every precision matrix $\Omega \in \mathrm{P}_{\mathcal{D}}$ can be represented by a $\mathcal{D}$-incomplete matrix which only consists of the free parameters of Ω, i.e., Ω^E. The rest, entries corresponding to the missing edges of the DAG, can be discarded, as whenever needed they can be obtained from Ω^E according to a constructive completion procedure given by the proof of Proposition 4.3. We re-formalize this as follows.

Corollary 4.4. *The projection* $\Omega \mapsto \Omega^E : \mathrm{P}_{\mathcal{D}} \to \mathrm{R}_{\mathcal{D}}$ *is a diffeomorphism, with inverse* $\Upsilon \mapsto \widehat{\Upsilon}$. *In particular,* $\mathrm{R}_{\mathcal{D}}$ *is open in* $\mathrm{I}_{\mathcal{D}} \cong \mathbb{R}^{|E|}$.

Proof. The projection is linear. Proposition 4.3 gives a unique inverse on $\mathrm{R}_{\mathcal{D}}$. Recovering the modified Cholesky pivots and free entries, and then the omitted precision entries, uses rational operations whose pivot denominators are strictly positive on $\mathrm{R}_{\mathcal{D}}$. Thus the inverse is smooth. Equivalently, the coordinate map in (12) is bijective and has the nonzero Jacobian in Lemma 4.5; the inverse function theorem gives openness and a smooth local inverse everywhere. Uniqueness makes those local inverses the global completion map. □

4.3 The DAG-Wishart distribution on $\mathrm{R}_{\mathcal{D}}$

In light of Corollary 4.4 we identify $\mathrm{P}_{\mathcal{D}}$ with $R_{\mathcal{D}}$ through the bijection $\Omega \mapsto \Omega^E$. Note that $\mathrm{R}_{\mathcal{D}}$, unlike $\mathrm{P}_{\mathcal{D}}$, is open in its affine support $\mathrm{I}_{\mathcal{D}}$ and, as a consequence of Corollary 4.4, diffeomorphic to $\Theta_{\mathcal{D}}$. Recall that we refer to $\mathrm{R}_{\mathcal{D}}$ as the space of incomplete precision matrices over $\mathcal{D}$. Now let $\pi^{\mathrm{R}_{\mathcal{D}}}_{U,\alpha}$ denote the image of $\pi^{\Theta_{\mathcal{D}}}_{U,\alpha}$ under the mapping

$$\psi \equiv \left((L, D) \mapsto \left(LD^{-1}L^{t}\right)^{E}\right) : \Theta_{\mathcal{D}} \to \mathrm{R}_{\mathcal{D}} \tag{12}$$

Since $\mathrm{R}_{\mathcal{D}}$ is an open subset of the Euclidean space $\mathbb{R}^{|E|}$, the distribution $\pi^{\mathrm{R}_{\mathcal{D}}}_{U,\alpha}$ has a density with respect to the Lebesgue measure on $\mathrm{R}_{\mathcal{D}}$. Hence, in light of the diffeomorphism $\Omega \mapsto \Omega^E$, in both a natural and practical sense, we define $\pi^{\mathrm{R}_{\mathcal{D}}}_{U,\alpha}$ as the DAG-Wishart distribution on the space of incomplete precision matrices $\mathrm{R}_{\mathcal{D}}$. To derive the density of $\pi^{\mathrm{R}_{\mathcal{D}}}_{U,\alpha}$ we need to compute the Jacobian of the mapping ψ in (12). The Jacobian of ψ is a variant of similar transformations found in [21, 15]. For completeness we still compute this Jacobian in the following lemma. The proof is given in Supplemental section B subsection 2.7.

Lemma 4.5. *[21, 15] The Jacobian of the mapping* $\psi : (D, L) \mapsto \left(LD^{-1}L^\top\right)^E$ *is* $\prod_{j=1}^{p} D_{jj}^{-(pa_j+2)}$.

We now proceed to express the density of $\pi_{U,\alpha}^{\mathrm{R}_\mathcal{D}}$ and some of its properties. The proofs are immediate results of Lemma 4.5 and the iterative construction of $\pi_{U,\alpha}^{\Theta_\mathcal{D}}$.

Theorem 4.6. *Let* Υ *be the image of* $(L, D) \sim \pi_{U,\alpha}^{\Theta_\mathcal{D}}$ *under the mapping* ψ*. Then*

a) The density of $\Upsilon \sim \pi_{U,\alpha}^{\mathrm{R}_\mathcal{D}}$ *with respect to the standard Lebesgue measure on* $\mathrm{R}_\mathcal{D}$ *is given by*

$$z_\mathcal{D}(U, \alpha)^{-1} \exp\left\{-\frac{1}{2}\operatorname{tr}(\widehat{\Upsilon}U)\right\} \prod_{i=1}^{p} D_{ii}^{-\frac{1}{2}\alpha_i + pa_i + 2},$$

where $D_{ii} = \left(\widehat{\Upsilon}^{-1}\right)_{ii|<i>}$ *is explicitly a function of* Υ *and* $z_\mathcal{D}$ *is defined in* (7)*.*

b) For every symmetric K *such that* $U + 2K \succ 0$*, the completion-based Laplace functional is*

$$E\{\exp[-\operatorname{tr}(K\Omega)]\} = \frac{z_\mathcal{D}(U + 2K, \alpha)}{z_\mathcal{D}(U, \alpha)}.$$

To state the ordinary Euclidean transform without double counting, define

$$\langle H, \Upsilon\rangle_E = \sum_i H_{ii}\Upsilon_{ii} + \sum_{\substack{i>j \\ (i,j)\in E}} H_{ij}\Upsilon_{ij},$$

and let K_H *have diagonal entries* $(K_H)_{ii} = H_{ii}$*, edge entries* $(K_H)_{ij} = (K_H)_{ji} = H_{ij}/2$*, and zeros elsewhere. If* $U + 2K_H \succ 0$*, then*

$$E\{\exp[-\langle H, \Upsilon\rangle_E]\} = \frac{z_\mathcal{D}(U + 2K_H, \alpha)}{z_\mathcal{D}(U, \alpha)}.$$

The factors of one half compensate for the two appearances of each off-diagonal entry in a symmetric trace. A general K *in the first display can additionally evaluate completed entries.*

c) The full precision-matrix mean, not merely the mean of its incomplete coordinates, is

$$\mathbb{E}(\Omega) = \sum_{j=1}^{p}(\alpha_j - pa_j - 2)\left(U_{\leq j\geq}^{-1}\right)^0 - \sum_{j=1}^{p}(\alpha_j - pa_j - 3)\left(U_{<j>}^{-1}\right)^0. \qquad (13)$$

Consequently, $\mathbb{E}(\Upsilon) = \mathbb{E}(\Omega)^E$.

Proof. Part (a) is the change-of-variables formula using Lemma 4.5; part (b) follows by combining the Laplace kernel with the parameter U. For part (c), write $\Omega = \sum_j D_{jj}^{-1} v_j v_j^\top$, where v_j has entry one at j, entries $\ell_j = -\beta_j$ at $\mathrm{pa}(j)$, and zero elsewhere. Proposition 3.2 gives

$$E(D_{jj}^{-1}) = \frac{\alpha_j - pa_j - 2}{U_{jj|<j>}}.$$

Taking the first two conditional moments of ℓ_j and applying the block-inverse identity for $U_{\leq j\geq}^{-1}$ yields the jth summand in (13). Summing over vertices proves the claim. □

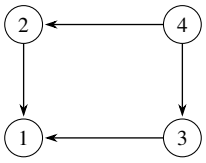

Figure 2: A four-node DAG whose skeleton is a 4-cycle.

Example 4.7. Let $\mathcal{D}$ be the DAG given by Figure 2. Then the DAG-Wishart density $\pi_{U,\alpha}^{\mathrm{R}_\mathcal{D}}$ is given by

$$\pi_{U,\alpha}^{\mathrm{R}_\mathcal{D}}(\Upsilon) = z_\mathcal{D}(U,\alpha)^{-1} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\widehat{\Upsilon} U\right)\right\} D_{11}(\Upsilon)^{-\frac{1}{2}\alpha_1+4} D_{22}(\Upsilon)^{-\frac{1}{2}\alpha_2+3} D_{33}(\Upsilon)^{-\frac{1}{2}\alpha_3+3} D_{44}(\Upsilon)^{-\frac{1}{2}\alpha_4+2},$$

where, using Proposition 4.3, $\widehat{\Upsilon}$ and $D_{ii}(\Upsilon)$ are computed as follows.

$$\widehat{\Upsilon} = \begin{pmatrix} \Upsilon_{11} & \Upsilon_{21} & \Upsilon_{31} & 0 \\ \Upsilon_{21} & \Upsilon_{22} & \dfrac{\Upsilon_{21}\Upsilon_{31}}{\Upsilon_{11}} & \Upsilon_{42} \\ \Upsilon_{31} & \dfrac{\Upsilon_{21}\Upsilon_{31}}{\Upsilon_{11}} & \Upsilon_{33} & \Upsilon_{43} \\ 0 & \Upsilon_{42} & \Upsilon_{43} & \Upsilon_{44} \end{pmatrix}$$

$$D_{11} = \Upsilon_{11}^{-1},$$
$$D_{22} = \left(\Upsilon_{22} - \frac{\Upsilon_{21}^2}{\Upsilon_{11}}\right)^{-1}, \qquad D_{33} = \left(\Upsilon_{33} - \frac{\Upsilon_{31}^2}{\Upsilon_{11}}\right)^{-1},$$
$$D_{44} = \left(\widehat{\Upsilon}_{44} - \widehat{\Upsilon}_{4,\{1,2,3\}} \widehat{\Upsilon}_{\{1,2,3\}}^{-1} \widehat{\Upsilon}_{\{1,2,3\},4}\right)^{-1}.$$

5 The inverse DAG-Wishart distribution on the space of incomplete covariance matrices

We now derive the covariance-coordinate image of the DAG–Wishart, which parallels the hyper-inverse Wishart and the inverse type-II Wishart $\mathrm{IW}_{\mathrm{P}_\mathcal{G}}$. We first define incomplete covariance matrices and recall the completion results from [2] needed to construct this image.

5.1 The space of incomplete covariance matrices

Recall that $\mathrm{PD}_\mathcal{D}$ is the space of covariance matrices for the Gaussian DAG model $\mathcal{N}(\mathcal{D})$, the elements of which, according to (3), can be characterized as:

$$\Sigma \in \mathrm{PD}_\mathcal{D} \iff \Sigma \succ 0 \;\;\&\;\; \Sigma_{[i\not\succ} = \Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ}\Sigma_{\prec i\not\succ}, \text{ for each } i \in V. \tag{14}$$

The above characterization allows us to identify $\mathrm{PD}_\mathcal{D}$ with the functionally independent elements of Σ. The following proposition is a key ingredient in this identification.

Proposition 5.1. *[2] Let $\Gamma \in \mathrm{Q}_\mathcal{D}$, then*

1. *There exists a completion process of polynomial complexity that can determine whether Γ can be completed in $\mathrm{PD}_\mathcal{D}$;*

2. *If a completion exists, this completion is unique and can be determined constructively using the following process:*

   i) Set $\Sigma_{ij} = \Gamma_{ij}$ for each $(i, j) \in E$ and set $j = p$.

   ii) If $j > 1$, then set $j = j - 1$ and proceed to the next step, otherwise Σ is successfully completed.

   iii) If $\Sigma_{\preceq j\succeq} \succ 0$, then proceed[4] to the next step, otherwise the completion in $\mathrm{PD}_\mathcal{D}$ does not exist.

   iv) If $\Sigma_{\not\prec j]}$ is non-empty, then set $\Sigma_{\not\prec j]} = \Sigma_{\not\prec j\succ}\Sigma^{-1}_{\prec j\succ}\Sigma_{\prec j]}$, $\Sigma_{[j\not\succ} = \Sigma^t_{\not\prec j]}$ and return to step (2).

Remark 5.2. Note once more that the procedure in Proposition 5.1 itself determines if Γ can be completed in $\mathrm{PD}_\mathcal{D}$. It is clear from Step *(iii)* above that the necessary and sufficient condition for the existence of a positive definite completion is that, for each $j \in V$, the covariance sub-matrix $\Sigma_{\preceq j\succeq} \succ 0$ and not just $\Sigma_{\prec j\succ} \succ 0$. Furthermore, the completion procedure in Proposition 5.1 can terminate midway.

[4] Note that for each j, the submatrix $\Sigma_{\preceq j\succeq}$ is fully determined by step *(ii)*

From Definition 4.1 recall that $\mathrm{S}_{\mathcal{D}}$ denotes the set of $\Gamma \in \mathrm{I}_{\mathcal{D}}$ that can be completed in $\mathrm{PD}_{\mathcal{D}}$. We call this set the space of incomplete covariance matrices over $\mathcal{D}$. The next corollary formalizes the fact that $\mathrm{S}_{\mathcal{D}}$ can be identified with $\mathrm{PD}_{\mathcal{D}}$. Its proof is immediate from Proposition 5.1 above.

Corollary 5.3. *The projection $\Sigma \mapsto \Sigma^E : \mathrm{PD}_{\mathcal{D}} \to \mathrm{S}_{\mathcal{D}}$ is a diffeomorphism with inverse $\Gamma \mapsto \widetilde{\Gamma}$, where $\widetilde{\Gamma}$ is the completion constructed in Proposition 5.1. In particular, $\mathrm{S}_{\mathcal{D}}$ is open in $\mathrm{I}_{\mathcal{D}} \cong \mathbb{R}^{|E|}$.*

Proof. The projection is linear and Proposition 5.1 gives a unique inverse. At step j, the completion uses $\Sigma^{-1}_{<j>}$; its determinant is positive throughout $\mathrm{S}_{\mathcal{D}}$. Every completed entry is therefore a rational, hence smooth, function of the incomplete coordinates. The Jacobian in Lemma 5.5 is strictly positive, so the inverse function theorem also gives openness and smoothness of the inverse. □

Remark 5.4. Suppose $\mathcal{D}$ is perfect. Then $\mathrm{PD}_{\mathcal{D}}$ is identical to $\mathrm{PD}_{\mathcal{D}^{\mathrm{u}}}$ and, therefore, by the completion result in Grone et al. [11], every incomplete matrix in $\mathrm{Q}_{\mathcal{D}}$ can be completed in $\mathrm{PD}_{\mathcal{D}}$. Hence for $\mathcal{D}$ perfect, $\mathrm{S}_{\mathcal{D}}$ and $\mathrm{Q}_{\mathcal{D}}$ are identical.

5.2 The inverse DAG-Wishart distribution on $\mathrm{S}_{\mathcal{D}}$

Let $\pi^{\mathrm{S}_{\mathcal{D}}}_{U,\alpha}$ denote the image of $\pi^{\Theta_{\mathcal{D}}}_{U,\alpha}$ under

$$(D, L) \longmapsto (L^{-\top} D L^{-1})^E : \Theta_{\mathcal{D}} \longrightarrow \mathrm{S}_{\mathcal{D}}.$$

In parallel to our notation $\pi^{\mathrm{R}_{\mathcal{D}}}_{U,\alpha}$, we call this the inverse DAG–Wishart distribution on incomplete covariance coordinates. To derive its Lebesgue density, we first compute the Jacobian of $(\Sigma^E \mapsto \Sigma^{-E}) : \mathrm{S}_{\mathcal{D}} \to \mathrm{R}_{\mathcal{D}}$, where $\Sigma^{-E} = (\Sigma^{-1})^E$ and Σ is the unique completion of Σ^E in $\mathrm{PD}_{\mathcal{D}}$.

Lemma 5.5. *Let $\mathcal{D} = (V, E)$ be an arbitrary DAG, then the Jacobian of the mapping $(\Sigma^{-E} \mapsto \Sigma^E) : \mathrm{R}_{\mathcal{D}} \to \mathrm{S}_{\mathcal{D}}$ is given by $\prod_{i=1}^{p} \dfrac{\det \Sigma^{pa_i+2}_{\leq i \geq}}{\det \Sigma^{pa_i+1}_{<i>}}$.*

Proof. Write the mapping $\Sigma^{-E} \mapsto \Sigma^E$ as the composition

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

The first map is the inverse of the free-coordinate version of ψ in Lemma 4.5; its Jacobian is

$$\prod_{i=1}^{p} \lambda_i^{pa_i+2}, \qquad \lambda_i = \Sigma_{ii|<i>}.$$

For the second map, it is simplest to calculate the inverse $\Sigma^E \mapsto (\lambda_i, \beta_i)_{i=1}^p$. Remove vertex 1 and apply induction to the ancestral subgraph on $\{2, \ldots, p\}$. Conditional on that subgraph, the remaining block map is

$$(\Sigma_{11}, \Sigma_{<1]}) \longmapsto \left(\Sigma_{11|<1>}, \Sigma_{<1>}^{-1}\Sigma_{<1]}\right),$$

whose block-triangular derivative has determinant $\det(\Sigma_{<1>})^{-1}$. Hence

$$J\{\Sigma^E \mapsto (\lambda_i, \beta_i)_{i=1}^p\} = \prod_{i=1}^p \det(\Sigma_{<i>})^{-1},$$

and the forward map from $\Xi_{\mathcal{D}}$ to $\mathrm{S}_{\mathcal{D}}$ has Jacobian $\prod_i \det(\Sigma_{<i>})$. Multiplying the two forward Jacobians and using $\lambda_i = \det(\Sigma_{\leq i\geq})/\det(\Sigma_{<i>})$ gives

$$\prod_{i=1}^p \frac{\det(\Sigma_{\leq i\geq})^{pa_i+2}}{\det(\Sigma_{<i>})^{pa_i+1}},$$

as claimed. □

We now proceed to state the functional form of the density of $\pi_{U,\alpha}^{\mathrm{S}_{\mathcal{D}}}$ with respect to Lebesgue measure.

Corollary 5.6. *Let $\Sigma \sim \pi_{U,\alpha}^{\mathrm{PD}_{\mathcal{D}}}$ and let $\Gamma = \Sigma^E$, i.e., Σ is the completion of Γ in $\mathrm{PD}_{\mathcal{D}}$. Then the density of $\Gamma \sim \pi_{U,\alpha}^{\mathrm{S}_{\mathcal{D}}}$ with respect to Lebesgue measure is given by*

$$z_{\mathcal{D}}(U,\alpha)^{-1} \exp\left\{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}U)\right\} \prod_{i=1}^p \frac{\det \Sigma_{\leq i\geq}^{-\frac{1}{2}\alpha_i}}{\det \Sigma_{<i>}^{-\frac{1}{2}\alpha_i+1}}. \tag{15}$$

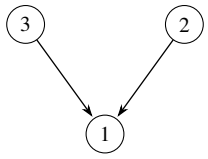

Figure 3: The collider DAG studied in Example 5.7.

Example 5.7. Consider the DAG $\mathcal{D}$ given in Figure 3. Then the inverse DAG-Wishart on $\mathcal{D}$ is given by

$$\pi_{U,\alpha}^{\mathrm{S}_{\mathcal{D}}}(\Gamma) = z_{\mathcal{D}}(U,\alpha)^{-1} \exp\left\{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}U)\right\} D_{11}^{-\frac{1}{2}\alpha_1} D_{22}^{-\frac{1}{2}\alpha_2} D_{33}^{-\frac{1}{2}\alpha_3} \det(\Sigma_{<1>})^{-1},$$

where Σ, the completion of Γ, is simply computed as

$$\Sigma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} \\ \Gamma_{21} & \Gamma_{22} & 0 \\ \Gamma_{31} & 0 & \Gamma_{33} \end{pmatrix}.$$

Remark 5.8. For a decomposable graph, compatible perfect-DAG specializations of $\pi^{S_{\mathcal{D}}}_{U,\alpha}$ have the same block conditional structure as the inverse type-II Wishart of [17]. This statement does not identify the unrestricted vertexwise family with the entire type-II family; their shape parameterizations and admissibility constraints differ. For the homogeneous covariance-graph subclass, the specialization described in Supplemental Section B agrees with the inverse Wishart construction of Khare and Rajaratnam [15].

5.3 Properties of the inverse DAG-Wishart distributions

One of the main useful properties of the inverse DAG–Wishart $\pi^{S_{\mathcal{D}}}_{U,\alpha}$ is its nodewise directed hyper-Markov factorization. For an arbitrary DAG this follows directly from Proposition 3.2 under the diffeomorphism in Corollary 5.3; it is not inferred from the decomposable-graph Theorem 4.4 of [17]. For a perfect DAG, the two results have parallel block conditional interpretations.

Theorem 5.9. *If* $\Sigma^E \sim \pi^{S_{\mathcal{D}}}_{U,\alpha}$, *then*
i) $\left\{(\Sigma_{ii|<i>}, \Sigma^{-1}_{<i>}\Sigma_{<i]} : i \in V\right\}$ *are mutually independent; thus* $\pi^{S_{\mathcal{D}}}_{U,\alpha}$ *has the stated strong directed hyper-Markov factorization.*
ii) The distribution of $\Sigma_{ii|<i>}$ *and* $\Sigma^{-1}_{<i>}\Sigma_{<i]}|\Sigma_{ii|<i>}$ *are, respectively, given by*

$$\Sigma_{ii|<i>} \sim IG(\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1, \frac{1}{2}U_{ii|<i>}), \text{ and} \tag{16}$$

$$\Sigma^{-1}_{<i>}\Sigma_{<i]}|\Sigma_{ii|<i>} \sim \mathrm{N}_{\mathrm{pa_i}}(\mathrm{U}^{-1}_{<\mathrm{i}>}\mathrm{U}_{<\mathrm{i}]}, \Sigma_{\mathrm{ii}|<\mathrm{i}>}\mathrm{U}^{-1}_{<\mathrm{i}>}). \tag{17}$$

Proof. Under the covariance map, $D_{ii} = \Sigma_{ii|<i>}$ and $\beta_i = \Sigma^{-1}_{<i>}\Sigma_{<i]} = -L_{<i]}$. The mutual independence and the two displayed laws are therefore exactly Proposition 3.2 expressed in covariance coordinates. □

The next result computes the expectation of the *full* covariance matrix, including entries that are not retained in Σ^E.

Proposition 5.10. *Suppose* $\Sigma = L^{-\top}DL^{-1}$ *is induced by* $\pi^{\Theta_{\mathcal{D}}}_{U,\alpha}$, *and assume* $\alpha_i > pa_i + 4$ *for every i. Define*

$$m_i = U^{-1}_{<i>}U_{<i]}, \qquad r_i = \frac{U_{ii|<i>}}{\alpha_i - pa_i - 4}, \qquad B_i = r_iU^{-1}_{<i>} + m_im_i^{\top}.$$

For an empty parent set, take m_i *and* B_i *to be empty and* $r_i = U_{ii}/(\alpha_i - 4)$. *The matrix* $M = \mathbb{E}(\Sigma)$ *is obtained for* $i = p, p-1, \ldots, 1$ *as follows:*

$$M_{ji} = M_{ij} = M_{j,<i>}m_i, \qquad j = i+1, \ldots, p, \tag{18}$$

$$M_{ii} = r_i + \mathrm{tr}(M_{<i>}B_i)\,. \tag{19}$$

When $\mathrm{pa}(i) = \emptyset$, *the first expression is zero and the trace term is absent. In particular,* $\mathbb{E}(\Sigma^E) = M^E$.

Proof. The recursive Gaussian representation associated with the modified Cholesky factor is

$$X_i = \beta_i^\top X_{\mathrm{pa}(i)} + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}(0, D_{ii}),$$

with ε_i independent of the higher-index variables. The parameters (D_{ii}, β_i) are independent of the parameters that determine $\Sigma_{\{i+1,\dots,p\}}$. Proposition 3.2 gives

$$E(D_{ii}) = r_i, \qquad E(\beta_i) = m_i, \qquad E(\beta_i \beta_i^\top) = B_i.$$

Taking expectations in $\Sigma_{ji} = \Sigma_{j,\mathrm{pa}(i)}\beta_i$ and $\Sigma_{ii} = D_{ii} + \beta_i^\top \Sigma_{\mathrm{pa}(i)}\beta_i$ proves (18)–(19). □

Remark 5.11. The completion map is nonlinear. Consequently, the DAG completion of $\mathbb{E}(\Sigma^E)$ is generally not $\mathbb{E}(\Sigma)$; likewise, completing $\mathbb{E}(\Omega^E)$ generally does not give $\mathbb{E}(\Omega)$. Equations (13) and (18)–(19) must be used when a full-matrix posterior mean is required.

6 Simulation study and applications to real data

This section records the numerical studies associated with the original preprint and states exactly what they can support. The structure search is over DAGs compatible with a *fixed, supplied parent ordering*; it is not unrestricted DAG learning and it does not estimate the ordering. The public companion archive contains the principal R functions and some driver scripts, but not the saved simulation outputs, the Sachs input used for the figure, or the call-center data. The tables below are therefore historical point estimates unless explicitly identified as reproduced. They illustrate feasibility and motivate further evaluation; they do not establish a general scaling rate or uniform risk dominance.

6.1 Bayesian model selection via DAG-Wishart prior

Let $\mathfrak{D}_\prec$ be the set of DAGs compatible with the fixed ordering. For $\mathcal{D} \in \mathfrak{D}_\prec$,

$$p(\mathcal{D} \mid X) \propto p(X \mid \mathcal{D}) p(\mathcal{D}).$$

For observations modeled with known zero mean and $T = X^\top X$, conjugacy gives the exact marginal likelihood

$$p(X \mid \mathcal{D}) = (2\pi)^{-np/2} \frac{z_{\mathcal{D}}(U + T, \alpha + n\mathbf{1})}{z_{\mathcal{D}}(U, \alpha)}. \tag{20}$$

The original experiments used a uniform prior over $\mathfrak{D}_\prec$. Because each of the $p(p-1)/2$ admissible edges can then be present or absent independently, this prior has edge-inclusion probability $1/2$ and is not sparsity-favoring. A transparent alternative is to

write $e(\mathcal{D}) = |E(\mathcal{D})| - p$ for the number of non-loop edges and use

$$p(\mathcal{D} \mid \rho) \propto \rho^{e(\mathcal{D})}(1-\rho)^{p(p-1)/2-e(\mathcal{D})},$$

with fixed $\rho < 1/2$ or a stated beta hyperprior. The historical numbers below retain the uniform graph prior so that their target is not changed retrospectively.

The search heuristic combines stochastic shotgun search (SSS) [13] with starting graphs from the LassoDAG regularization path [23]. It approximates the maximizer of the posterior score; it does not sample from the exact graph posterior. The DAG-W procedure is specified below.

Algorithm 6.1 (DAG-W)**.** Assume the following are given: the standardized data matrix X, the hyper-parameters α, U and the maximum iteration number M. Estimate N models corresponding to different points on the 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 differ from $\mathcal{D}_0$ by one admissible edge toggle. Evaluate their log posterior scores $s_1, \ldots, s_{N_1}$ using (20) and the stated graph prior, and update the retained list $\mathcal{L}^{(k)}$.

   (b) Sample the next graph from the current candidates with probability

   $$p_i = \frac{\exp\{\gamma(s_i - s_{\max})\}}{\sum_j \exp\{\gamma(s_j - s_{\max})\}},$$

   where γ is an annealing parameter. Set the sampled graph as the new $\mathcal{D}_0$.

   (c) Return to Step 1-(a).

2. Combine the retained lists $\mathcal{L}^{(k)}$, $k = 1, \ldots, N$.

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

The initial models come from different points on the LassoDAG regularization path. In [23], the penalty parameter τ_i for the Lasso problem at node i is

$$\tau_i = 2\frac{Z^*_{\frac{\kappa}{2p(i-1)}}}{\sqrt{n}}, \tag{21}$$

where Z^*_q denotes the $(1-q)$th quantile of the standard normal distribution; [23] recommends $\kappa = 0.1$. The historical comparison uses that setting for LassoDAG.

The historical model-selection study used $U = I$ and $\alpha_i = c\, pa_i + b$, with $c = 1$ and $b = 3$. This satisfies properness but implies

$$D_{ii} \sim IG(1/2, 1/2), \qquad \beta_i \text{ has a multivariate } t \text{ law with one degree of freedom.}$$

Thus the local regression prior is Cauchy-tailed and neither $E(D_{ii})$ nor $E(\beta_i)$ exists marginally; the setting should not be described through prior means. The study used $N = 16$ initial states: fifteen values $\kappa = (k/15)^4 p$, $k = 1, \ldots, 15$, and the LassoDAG recommendation $\kappa = 0.1$. Values $\kappa > 1$ fall outside the false-positive-control interpretation in [23] and were used only to diversify starting graphs. The reported search settings were $M = 100$, $N_1 = 30$, and $\gamma = 0.5$, with reduced search budgets for the largest dimensions.

The driver code uses `pcalg::randomDAG` with independent admissible-edge probability 0.01 and regression weights sampled between 0.2 and 0.8. It fixes $n = 100$ and considers $p = 50, 100, 200, 500, 1000, 1500, 2000$. For $p \geq 500$, the manuscript reports nine starts and at most 50 search steps per start. The public archive does not contain a $p = 2000$ driver or any saved output. Its $p = 1000$ and $p = 1500$ drivers call an absent `PerfEval2.R`, request 16 starts and 100 search iterations rather than the reduced manuscript budgets, and use only two replications per parallel invocation. Package versions, hardware, elapsed times, Monte Carlo standard errors, and a complete mapping from scripts to table rows were not preserved. These mismatches and omissions prevent exact reproduction of Table 1 from the archive alone.

There is also a likelihood mismatch in the archived search path. The selection routine centers and scales the observations using sample statistics, after which the marginal-score routine increments α by n. Under the flat unknown-mean analysis in Remark 7.6, centering instead yields the update $\alpha + (n-1)\mathbf{1}$. Thus the historical graph scores should be understood as scores from a data-dependent standardized, zero-mean working likelihood, not as the corrected unknown-mean marginal likelihood.

Sensitivity and specificity are reported in the historical table. With approximately one true edge per 100 admissible edges, specificity near one can still correspond to many false discoveries, so precision (positive predictive value), MCC, structural Hamming distance, and uncertainty should also be reported in a new experiment. For example, using the $p = 2000$ averages as plug-in rates gives roughly 3,886 true and 11,082 false positives for DAG-W (precision about 26%), versus 1,977 true and 6,531 false positives for LassoDAG (precision about 23%). This calculation is only an interpretation of averaged rates, not a reconstruction of the missing replicate-level results.

6.2 Covariance Estimation Performance

We now consider covariance and precision estimation *conditional on the true DAG and its ordering*; graph-selection uncertainty is not included in this experiment. Following [20], the historical code uses a coordinatewise squared-error loss

$$L_2(M, \widehat{M}) = \sum_{(i,j)\in E} (M_{ij} - \widehat{M}_{ij})^2,$$

	LassoDAG		DAG-W	
p	Sensitivity	Specificity	Sensitivity	Specificity
50	0.6156	1.0000	0.7828	0.9980
100	0.4826	~ 1	0.7524	0.9977
200	0.3969	~ 1	0.7405	0.9975
500	0.2497	~ 1	0.6517	0.9982
1000	0.1748	0.9991	0.4248	0.9971
1500	0.1226	0.9981	0.2672	0.9962
2000	0.0989	0.9967	0.1944	0.9944

Table 1: Historically reported average sensitivity and specificity for $n = 100$ and admissible-edge probability 0.01. Replicate-level outputs and Monte Carlo uncertainty are unavailable in the public archive; the table is not independently reproduced in this revision.

where E includes the diagonal under the convention of Definition 4.1. This is not the full Frobenius loss. The full-matrix Stein loss used in the code is

$$L_1(\widehat{M}, M) = \operatorname{tr}(\widehat{M}M^{-1}) - \log(\det(\widehat{M}M^{-1})) - p.$$

Let

$$\bar{\Sigma} = E(\Sigma \mid X), \qquad \bar{\Omega} = E(\Omega \mid X),$$

computed from Proposition 5.10 and (13) with the posterior parameters. Their edge projections are Bayes actions under the coordinatewise squared loss. Under the displayed full Stein loss, the Bayes actions are

$$\delta_{\Sigma}^{\text{Stein}} = \bar{\Omega}^{-1}, \qquad \delta_{\Omega}^{\text{Stein}} = \bar{\Sigma}^{-1}.$$

Neither action is generally DAG-constrained, because expectations and inversion do not preserve the nonlinear DAG parameter spaces.

The historical precision routine first computes $\bar{\Omega}^E$ and then applies the DAG completion map; denote this graph-constrained representative by $\widehat{\Omega}_{\text{coord}} = C_{\mathcal{D}}(\bar{\Omega}^E)$. It has the correct posterior-mean edge coordinates and is therefore valid for the coordinate loss, but it is not $\bar{\Omega}$ and is not the full-Stein Bayes action. The third reported estimator transforms the joint posterior mode in (D, L) coordinates into covariance and precision matrices. We denote these plug-ins by $\widetilde{\Sigma}_{\text{chol}}$ and $\widetilde{\Omega}_{\text{chol}}$; they are not called MAP estimators because modes change under nonlinear reparameterization. Supplemental Section D gives the corrected algorithms.

The historical study uses $\alpha_i = c\,pa_i + 3$ and $U = uI$, with the values shown in the table, and simulates $p = 500$ with admissible-edge probability 0.01. The driver specifies 15 parallel batches of 20 replications per setting, but the saved outputs are

absent, so execution and aggregation cannot be verified. Table 2 is retained as a historical report of relative changes from the graph-constrained MLE. It contains no Monte Carlo standard errors or intervals, and its Stein-loss rows evaluate the listed procedures rather than establishing that each is a Bayes action.

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

Table 2: Historically reported relative change from the constrained MLE when estimating Ω, with $p = 500$. Positive values indicate lower loss. The procedures are the DAG completion of posterior-mean precision coordinates, the full-Stein action $\bar{\Sigma}^{-1}$, and the transformed Cholesky-mode plug-in. Saved outputs and Monte Carlo uncertainty are unavailable.

The historical results are highly sensitive to the hyperparameters and to the generating edge probability. In particular, Supplemental Section D reports severe deterioration at edge probability 0.02; there is no uniform dominance over the MLE. The supplement also contains one Gaussian-contamination experiment. Better performance than the MLE in that single design is evidence about shrinkage under that contamination, not a robustness theorem or a general guarantee under misspecification.

6.3 Historical molecular-network example

The original analysis used the data of [22], containing measurements of $p = 11$ proteins and phospholipids on $n = 7466$ cells, and imposed an ordering derived from

the network reported in that study. That reported network is a scientific benchmark, not known ground truth; accordingly, the terms true and false positive below mean agreement or disagreement with that benchmark, not validated causal truth. The exact input data and preprocessing object used to produce Figure 4 are not present in the companion archive, so the analysis is not independently reproduced here.

With $\kappa = 0.1$ for LassoDAG and $b = 3, c = 1$ for DAG-W, the historically reported rates are 78.95% sensitivity and 52.78% specificity for LassoDAG, and 94.74% sensitivity and 47.22% specificity for DAG-W. Relative to a 19-edge benchmark among 55 ordered pairs, these rates correspond to approximately 15 benchmark edges plus 17 additional edges for LassoDAG (precision 46.9%) and 18 benchmark edges plus 19 additional edges for DAG-W (precision 48.6%). Both selected graphs are therefore dense. The result shows a sensitivity–false-discovery tradeoff under the assumed ordering; it does not by itself validate additional direct or indirect molecular mechanisms.

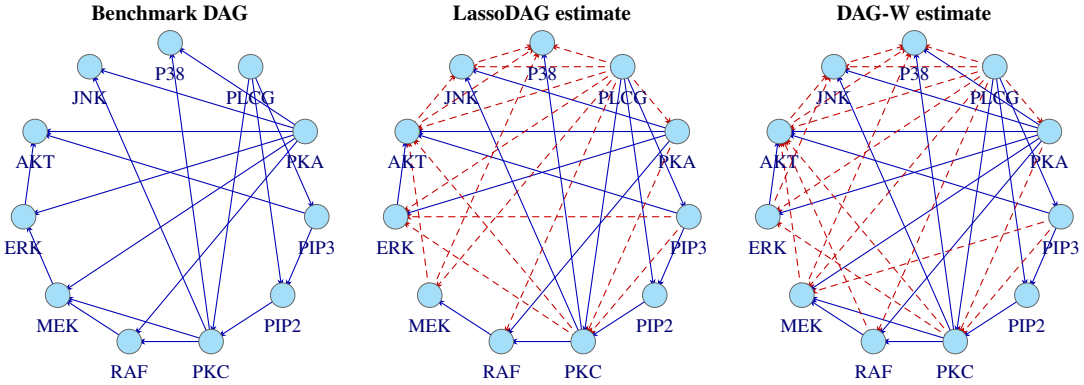

Figure 4: Historical selected graphs compared with the network benchmark from [22]. Blue edges agree with the benchmark and red edges do not; colors do not establish causal truth. The analysis could not be reproduced from the public archive because the exact input object is absent.

Supplemental Section D also preserves a historical call-center prediction example based on data used in [3] and [20]. Its R script is available, but the external data file and saved fitted model are not. The numerical comparison is therefore not independently verified in this revision.

7 Closing remarks

The DAG–Wishart construction is valid for every DAG once an ordering is fixed. Its main tractable objects are the nodewise Cholesky parameters: their independence

yields the exact normalizing constant, conjugate posterior, marginal likelihood, simulation scheme, and the full posterior means in (13) and (18)–(19). The induced Lebesgue densities on incomplete coordinates are also explicit.

Three limitations are essential. First, nonlinear completion does not commute with expectation. Second, a mode computed with respect to Lebesgue measure in Cholesky coordinates is not automatically a covariance- or precision-coordinate MAP. Third, the model-selection procedure in this paper assumes the ordering and therefore does not resolve Markov-equivalence or unknown-order learning.

Subsequent work established high-dimensional graph-selection and estimation consistency under explicit assumptions [5]. Other work showed that arbitrary DAG–Wishart hyperparameters can assign different marginal likelihoods to Markov-equivalent DAGs and constructed compatible specifications [19]. Those results should be consulted when the inferential target extends beyond the fixed-order model class studied here.

The numerical examples preserved in this preprint are historical evidence. A new empirical release should include immutable input data, package and hardware manifests, all seeds, replicate-level outputs, uncertainty intervals, run times, and comparisons that report false discoveries as well as sensitivity.

References

[1] Steen A. Andersson and Michael D. Perlman. Normal linear regression models with recursive graphical Markov structure. *J. Multivariate Anal.*, 66(2):133–187, 1998.

[2] Emanuel Ben-David and Bala Rajaratnam. Positive definite completion problems for Bayesian networks. *SIAM J. Matrix Anal. Appl.*, 33(2):617–638, 2012.

[3] Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008.

[4] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, New York-Chichester-Brisbane, 1979. Wiley Series in Probability and Mathematical Statistics.

[5] Xuan Cao, Kshitij Khare, and Malay Ghosh. Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Ann. Statist.*, 47(1):319–348, 2019. doi:10.1214/18-AOS1689.

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

[7] Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science*, 23(2):151–178, 05 2008.

[8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[9] Dan Geiger and David Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.*, 30(5):1412–1440, 2002.

[10] Dan Geiger and David Heckerman. Learning Gaussian networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 235–243, 1994. Reissued as arXiv:1302.6808.

[11] Robert Grone, Charles R. Johnson, Eduardo M. de Sá, and Henry Wolkowicz. Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.*, 58:109–124, 1984.

[12] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.*, 13:2409–2464, 2012.

[13] Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West. Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.*, 20(4):388–400, 2005.

[14] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11):1–26, 5 2012.

[15] Kshitij Khare and Bala Rajaratnam. Wishart distributions for decomposable covariance graph models. *Ann. Statist.*, 39(1):514–555, 2011.

[16] Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. Oxford Science Publications.

[17] Gérard Letac and Hélène Massam. Wishart distributions for decomposable graphs. *Ann. Statist.*, 35(3):1278–1323, 2007.

[18] Judea Pearl and Nanny Wermuth. When can association graphs admit a causal interpretation? In P. Cheeseman and R.W. Oldford, editors, *Selecting Models from Data*, volume 89 of *Lecture Notes in Statistics*, pages 205–214. Springer New York, 1994.

[19] Stefano Peluso and Guido Consonni. Compatible priors for model selection of high-dimensional Gaussian DAGs. *Electron. J. Stat.*, 14(2):4110–4132, 2020. doi:10.1214/20-EJS1768.

[20] Bala Rajaratnam, Hélène Massam, and Carlos M. Carvalho. Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.*, 36(6):2818–2849, 2008.

[21] Alberto Roverato. Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*, 87(1):99–112, 2000.

[22] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[23] Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

[24] Nico M. Temme. *Special functions : an introduction to the classical functions of mathematical physics*. J. Wiley & sons, New York, 1996.

[25] Nanny Wermuth. Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.*, 75(372):963–972, 1980.

Supplemental Section A: Graph theory and Gaussian DAG preliminaries

Graph theoretic notation and terminology

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, we include all loops in E, although figures omit them.

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 subgraph 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 i to j is a sequence $i_0 = i, \ldots, i_n = j$ such that each consecutive pair is adjacent; it is simple when $i_0, \ldots, i_n$ are distinct. In a directed graph, the path is directed from i to j when $i_{\nu-1} \to i_\nu$ for every $\nu = 1, \ldots, n$. We say that i leads to j, written $i \longmapsto j$, when such a directed path exists. A graph is connected when every distinct pair is joined by a path in its underlying undirected graph. A cycle of length $n \geq 2$ is a sequence $i_0, i_1, \ldots, i_{n-1}, i_n = i_0$ in which $i_0, \ldots, i_{n-1}$ are distinct and consecutive vertices are adjacent; it is directed when every edge is oriented $i_{\nu-1} \to i_\nu$.

An undirected graph $\mathcal{G} = (V, \mathscr{E})$ is decomposable (or chordal) if it has no induced cycle of length at least four. A directed graph $\mathcal{D} = (V, E)$ is acyclic when it has no directed cycle apart from the notational loops. Its skeleton $\mathcal{D}^{\mathrm{u}}$ is obtained by replacing directed edges by undirected edges. An immorality is an induced configuration $i \to k \leftarrow j$ with nonadjacent parents i and j. The moral graph $\mathcal{D}^{\mathrm{m}}$ is obtained by joining every pair of co-parents and then dropping all arrowheads. A DAG is perfect when it has no immoralities, equivalently when the parents of every vertex are pairwise adjacent. Every decomposable undirected graph admits a perfect acyclic orientation [16].

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 $i \longmapsto j$. Similarly, the set of descendants of a vertex i, denoted by $\mathrm{de}(i)$, is the set of those vertices j such that $i \longmapsto j$. 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 $\mathrm{An}(B)$.

Markov properties for DAG models

Let V be a finite set of indices and $(X_i)_{i\in V}$ 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 DAG $\mathcal{D}$ 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 w.r.t. $\mathcal{D}$, 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

$$\int k^i(y_i, x_{pa(i)})d\mu_i(y_i) = 1, \quad \forall i \in V,$$

and P has a density p, w.r.t. the product measure $\mu = \otimes_{i\in V}\mu_i$, given by

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

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

Lemma 7.1. *[16] If P admits a recursive factorization w.r.t. the directed graph $\mathcal{D}$, then it also admits a factorization w.r.t. the undirected graph $\mathcal{D}^{\mathrm{m}}$, and, consequently, obeys the global Markov property[5] w.r.t. $\mathcal{D}^{\mathrm{m}}$.*

Proof. Note that for each vertex $i \in V$ the set $fa(i)$ is a complete subset of $\mathcal{D}^{\mathrm{m}}$. Thus if we define $\psi_{fa(i)}(x_{fa(i)}) = k^i(x_i, x_{pa(i)})$, then $p(x) = \prod_{i\in V} p(x_i|x_{pa(i)}) = \prod_{i\in V} k^i(x_i, x_{pa(i)}) = \prod_{i\in V} \psi_{fa(i)}(x_{fa(i)})$. Therefore, P admits a factorization w.r.t. $\mathcal{D}^{\mathrm{m}}$ and by proposition 3.8 in [16] it also obeys the global Markov property w.r.t. $\mathcal{D}^{\mathrm{m}}$. □

Another direct implication of the DF property is that if P admits a recursive factorization w.r.t. $\mathcal{D}$, then, for each ancestral set A, the marginal distribution P_A admits a recursive factorization w.r.t. the induced graph $\mathcal{D}_A$. Combining this result with

[5] see [16] for definition.

Lemma 7.1 we obtain the following: if P admits a recursive factorization w.r.t. $\mathcal{D}$, then $A \perp\!\!\!\perp B|S$ $[P]$ 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: it recovers 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) w.r.t. $\mathcal{D}$ if for each $i \in V$

$$i \perp\!\!\!\perp (nd(i) \setminus pa(i)) \mid pa(i).$$

For a given DAG $\mathcal{D}$, a parent ordering relabels the vertices as $1, \ldots, |V|$ so that $pa(i) \subseteq \{i+1, \ldots, |V|\}$ for every i. Every DAG has at least one such ordering, although it need not be unique. We say that P obeys the parent-ordered Markov property (PO) with respect to $\mathcal{D}$ if, for every vertex i,

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

If P has a density with respect to μ, then the four directed Markov properties DF, DG, DL, and PO are equivalent under the usual regularity conditions [16].

Linear recursive properties of Gaussian DAGs

Let $\mathbf{x} = \left(x_1, \ldots, x_p\right)^\top$ 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:

$$\begin{array}{rcl} x_1 - \beta_{12}x_2 - \beta_{13}x_3 - \cdots - \beta_{1p}x_p &=& \epsilon_1 \\ x_2 - \beta_{23}x_3 - \cdots - \beta_{2p}x_p &=& \epsilon_2 \\ &\vdots& \\ x_p &=& \epsilon_p \end{array} \quad \text{or equivalently} \quad \begin{array}{rcl} x_1 &=& \beta_{12}x_2 + \beta_{13}x_3 + \cdots + \beta_{1p}x_p + \epsilon_1 \\ x_2 &=& \beta_{23}x_3 + \cdots + \beta_{2p}x_p + \epsilon_2 \\ &\vdots& \\ x_p &=& \epsilon_p, \end{array}$$

where β_{ij} is the regression coefficient of x_j $(j > i)$ in the regression of x_i on its predecessors. It is zero when $j \notin \mathrm{pa}(i)$. The residuals ϵ_i are mutually independent centered normal variables with variance $\sigma^2_{ii|pa(i)}$. The equations can be written $B\mathbf{x} = \boldsymbol{\epsilon}$, where B is upper triangular:

$$B = \begin{pmatrix} 1 & -\beta_{12} & \ldots & -\beta_{1p} \\ 0 & 1 & \ldots & -\beta_{2p} \\ 0 & \ldots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}.$$

From this we obtain:

$$\begin{aligned}
&\mathbb{V}ar[B\mathbf{x}] = \mathbb{V}ar[\epsilon] \\
\Rightarrow &B\Sigma B^\top = diag(\sigma^2_{1|pa(1)}, \ldots, \sigma^2_{p-1|pa(p-1)}, \sigma^2_{pp}) =: D \\
\Rightarrow &\Sigma = B^{-1}D(B^\top)^{-1} \\
\Rightarrow &\Sigma^{-1} = B^\top D^{-1} B. \qquad (22)
\end{aligned}$$

Thus $L = B^\top$ gives $\Sigma^{-1} = LD^{-1}L^\top$, with $L_{\mathrm{pa}(i),i} = -\beta_i$, exactly as specified in Section 2.2. The Gaussian distribution is Markov with respect to $\mathcal{D}$ precisely when the corresponding non-parent entries of L vanish [25].

Supplemental Section B: Properties of the class of DAG-Wishart distributions

7.1 Deriving the closed form expression for the DAG-Wishart $\pi^{\Theta_\mathcal{D}}_{U,\alpha}$

Theorem 7.2. *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 $pa_i := |pa(i)|$. Then,*

$$\int_{\Theta_\mathcal{D}} \exp\{-\frac{1}{2}\operatorname{tr}(LD^{-1}L^tU)\} \prod_{i=1}^p D_{ii}^{-\frac{1}{2}\alpha_i} dLdD < \infty$$

if and only if

$$\alpha_i > pa_i + 2 \quad \forall i = 1, \ldots, p.$$

Furthermore, in this case

$$z_\mathcal{D}(U,\alpha) = \prod_{i=1}^p \frac{\Gamma\left(\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1\right) 2^{\frac{\alpha_i}{2}-1} (\sqrt{\pi})^{pa_i} \det(U_{<i>})^{\frac{\alpha_i}{2}-\frac{pa_i}{2}-\frac{3}{2}}}{\det(U_{\leq i\geq})^{\frac{\alpha_i}{2}-\frac{pa_i}{2}-1}}. \qquad (23)$$

Proof. Let us first simplify the expression by integrating out the terms involving D_{ii}'s.

$$
\begin{aligned}
& \int \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left(LD^{-1}L^\top\right)U\right)\right\}\prod_{i=1}^{p} D_{ii}^{-\frac{1}{2}\alpha_i} dLdD \\
= & \int \exp\left\{-\frac{1}{2}\operatorname{tr}\left(D^{-1}\left(L^\top UL\right)\right)\right\}\prod_{i=1}^{p} D_{ii}^{-\frac{1}{2}\alpha_i} dLdD \\
= & \int \exp\left\{-\frac{1}{2}\sum_{i=1}^{p} D_{ii}^{-1}(L^\top UL)_{ii}\right\}\prod_{i=1}^{p} D_{ii}^{-\frac{1}{2}\alpha_i} dDdL \\
= & \int\left(\prod_{i=1}^{p}\int \exp\left\{-\frac{1}{2}D_{ii}^{-1}(L^\top UL)_{ii}\right\} D_{ii}^{-\frac{1}{2}\alpha_i} dD_{ii}\right) dL \\
= & \int\prod_{i=1}^{p}\frac{\Gamma\left(\frac{\alpha_i}{2}-1\right)2^{\frac{\alpha_i}{2}-1}}{\left((L^\top UL)_{ii}\right)^{\frac{\alpha_i}{2}-1}} dL \qquad (\text{ iff } \alpha_i > 2\ \forall\ i = 1, 2, \cdots, p) \\
= & \int\prod_{i=1}^{p}\frac{\Gamma\left(\frac{\alpha_i}{2}-1\right)2^{\frac{\alpha_i}{2}-1}}{\left((L_{\cdot i})^\top UL_{\cdot i}\right)^{\frac{\alpha_i}{2}-1}} dL \\
= & \int\prod_{i=1}^{p}\frac{\Gamma\left(\frac{\alpha_i}{2}-1\right)2^{\frac{\alpha_i}{2}-1}}{\left(\begin{pmatrix}1 & L_{\prec i]}^\top\end{pmatrix}\begin{pmatrix}U_{ii} & U_{[i\succ}\\ U_{\prec i]} & U_{\prec i\succ}\end{pmatrix}\begin{pmatrix}1\\ L_{\prec i]}\end{pmatrix}\right)^{\frac{\alpha_i}{2}-1}} dL \\
= & \prod_{i=1}^{p}\int_{\mathbb{R}^{pa_i}}\frac{\Gamma\left(\frac{\alpha_i}{2}-1\right)2^{\frac{\alpha_i}{2}-1}}{\left(\begin{pmatrix}1 & L_{\prec i]}^\top\end{pmatrix}\begin{pmatrix}U_{ii} & U_{[i\succ}\\ U_{\prec i]} & U_{\prec i\succ}\end{pmatrix}\begin{pmatrix}1\\ L_{\prec i]}\end{pmatrix}\right)^{\frac{\alpha_i}{2}-1}} dL_{\prec i]}. \qquad \text{eqn(A)}
\end{aligned}
$$

We now show how in general one can evaluate an integral of the form

$$\int_{\mathbb{R}^d}\frac{d\mathbf{x}}{\left(\begin{pmatrix}1 & \mathbf{x}^\top\end{pmatrix}\begin{pmatrix}a & \mathbf{b}^\top\\ \mathbf{b} & A\end{pmatrix}\begin{pmatrix}1\\ \mathbf{x}\end{pmatrix}\right)^{\gamma}},$$

where the block matrix, formed by $a \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^d$, and a $d \times d$ matrix A, is positive definite. To simplify the integral, proceed in two steps.
1) By [7, page 16],

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

By repeated application, we can generalize the above formula to

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

2) Now consider the general integral

$$\int_{\mathbb{R}^d} \frac{d\mathbf{x}}{\left(\begin{pmatrix}1 & \mathbf{x}^\top\end{pmatrix}\begin{pmatrix}a & \mathbf{b}^\top \\ \mathbf{b} & A\end{pmatrix}\begin{pmatrix}1 \\ \mathbf{x}\end{pmatrix}\right)^\gamma}.$$

Making the linear transformation $\mathbf{y} = A^{\frac{1}{2}}\mathbf{x} + A^{-\frac{1}{2}}\mathbf{b}$ it follows that for $\gamma > \frac{d}{2}$,

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{d\mathbf{x}}{\left(\begin{pmatrix}1 & \mathbf{x}^\top\end{pmatrix}\begin{pmatrix}a & \mathbf{b}^\top \\ \mathbf{b} & A\end{pmatrix}\begin{pmatrix}1 \\ \mathbf{x}\end{pmatrix}\right)^\gamma} &= \frac{1}{\det(A)^{\frac{1}{2}}}\int_{\mathbb{R}^d} \frac{1}{(\mathbf{y}^\top\mathbf{y} + a - \mathbf{b}^\top A^{-1}\mathbf{b})^\gamma} d\mathbf{y} \\ &= \frac{(\sqrt{\pi})^d \Gamma\left(\gamma - \frac{d}{2}\right)}{\Gamma(\gamma)\det(A)^{\frac{1}{2}}\,(a - \mathbf{b}^\top A^{-1}\mathbf{b})^{\gamma - \frac{d}{2}}}. \qquad (24)\end{aligned}$$

Applying the result from (24) to the desired integral in (A) we obtain

$$\begin{aligned}z_{\mathcal{D}}(U,\alpha) &= \prod_{i=1}^{p}\int_{\mathbb{R}^{pa_i}} \frac{\Gamma\left(\frac{\alpha_i}{2} - 1\right)2^{\frac{\alpha_i}{2}-1}}{\left(\begin{pmatrix}1 & L_{<i]}^\top\end{pmatrix}\begin{pmatrix}U_{ii} & U_{[i>} \\ U_{<i]} & U_{<i>}\end{pmatrix}\begin{pmatrix}1 \\ L_{<i]}\end{pmatrix}\right)^{\frac{\alpha_i}{2}-1}} dL_{<i]} \\ &= \prod_{i=1}^{p} \frac{\Gamma\left(\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1\right)2^{\frac{\alpha_i}{2}-1}(\sqrt{\pi})^{pa_i}\det(U_{<i>})^{\frac{\alpha_i}{2}-\frac{pa_i}{2}-\frac{3}{2}}}{\det(U_{\preceq i\succeq})^{\frac{\alpha_i}{2}-\frac{pa_i}{2}-1}},\end{aligned}$$

where $\det(U_{<i>}) = 1$ whenever $pa(i) = \emptyset$. It is easily seen that $z_{\mathcal{D}}(U,\alpha)$ is finite iff $\alpha_i > pa_i + 2$ for each $i = 1,\ldots,p$. □

The hyper-Markov properties of the DAG-Wishart

Theorem 7.3. *Let $\mathcal{D}$ be an arbitrary DAG and $(D,L) \sim \pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$. Then $\{(D_{ii}, L_{<i]}) : i = 1,\ldots,p\}$ are mutually independent. Moreover,*

$$D_{ii} \sim IG(\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1, \frac{1}{2}U_{ii|<i>}), \text{ and} \qquad (25)$$

$$L_{<i]}|D_{ii} \sim \mathrm{N}_{pa_i}(-U_{<i>}^{-1}U_{<i]}, D_{ii}U_{<i>}^{-1}). \qquad (26)$$

Proof. The coordinate map

$$\phi : (D,L) \longmapsto \underset{i\in V}{\times}(D_{ii}, \ell_i), \qquad \ell_i = L_{<i]}, \qquad (27)$$

has unit Jacobian. It is a vectorization of the Cholesky parameters, not of the regression coefficients; the latter are $\beta_i = -\ell_i$. Cyclicity of trace and completion of the square give

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

Therefore, the density in the coordinates $(\lambda_i, \ell_i) = (D_{ii}, L_{<i]})$ is

$$z_{\mathcal{D}}(U,\alpha)^{-1}\prod_{i\in V}\left[\lambda_i^{-\alpha_i/2}\exp\left\{-\frac{1}{2\lambda_i}\Big(U_{ii|<i>} + (\ell_i + U_{<i>}^{-1}U_{<i]})^\top U_{<i>}(\ell_i + U_{<i>}^{-1}U_{<i]})\Big)\right\}\right]. \tag{28}$$

The product proves mutual independence across vertices. The ith factor is the product of the inverse-gamma density in (25) and the conditional Gaussian density in (26), including their normalizing constants. This proves both stated laws. □

Corollary 7.4. *Let $(L, D) \sim \pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ and set $\nu_i = \alpha_i - pa_i - 2 > 0$. The columns $L_{<i]}$ are mutually independent and*

$$L_{<i]} \sim t_{\nu_i}\left(-U_{<i>}^{-1}U_{<i]}, \frac{U_{ii|<i>}}{\nu_i}U_{<i>}^{-1}\right).$$

Equivalently, their joint density is

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

where

$$c_i = \frac{\det(U_{<i>})^{1/2}(U_{ii|<i>})^{\alpha_i/2-pa_i/2-1}\Gamma(\alpha_i/2-1)}{2^{\alpha_i/2-1}\pi^{pa_i/2}\Gamma(\alpha_i/2 - pa_i/2 - 1)}. \tag{29}$$

Proof. Multiply the conditional Gaussian density in (26) by the inverse-gamma density in (25) and integrate D_{ii} over $(0,\infty)$. The integral is $\Gamma(\alpha_i/2-1)u_i^{-\alpha_i/2+1}$, where

$$u_i = \frac{1}{2}\left\{U_{ii|<i>} + (L_{<i]} + U_{<i>}^{-1}U_{<i]})^\top U_{<i>}(L_{<i]} + U_{<i>}^{-1}U_{<i]})\right\}.$$

Collecting constants gives (29); comparison with the multivariate t density gives the stated degrees of freedom and scale. □

The Posterior distribution of the DAG-Wishart

Proposition 7.5. *Let $\mathcal{D}$ be an arbitrary DAG and let $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ be an i.i.d. sample from $\mathrm{N_p}(\mathbf{0}, (\mathrm{L}^{-1})^\top \mathrm{D}\mathrm{L}^{-1})$, where $(D, L) \in \Theta_{\mathcal{D}}$. Let $S = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i \mathbf{Y}_i^\top$ denote the empirical covariance matrix. If the prior distribution on (D, L) is $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$, then the posterior distribution of (D, L) is given by $\pi_{\widetilde{U},\widetilde{\alpha}}^{\Theta_{\mathcal{D}}}$, where $\widetilde{U} = nS + U$ and $\widetilde{\alpha} = (n+\alpha_1, n+\alpha_2, \cdots, n+\alpha_p)$.*

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

$$f(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n \mid L, D) = \frac{1}{(\sqrt{2\pi})^{np}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(LD^{-1}L^\top(nS)\right)\right\} \det(D)^{-\frac{1}{2}n}.$$

When using $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ as the prior for (D, L), the posterior distribution of (D, L) given the data $(\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n)$ is given by

$$\begin{aligned}\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}(L, D \mid \mathbf{Y}_{1:n}) \propto \exp\left\{-\frac{1}{2}\operatorname{tr}\left(LD^{-1}L^\top(nS + U)\right)\right\} \\ \times \prod_{i=1}^{p} D_{ii}^{-(n+\alpha_i)/2}, \qquad (D, L) \in \Theta_{\mathcal{D}}.\end{aligned} \tag{30}$$

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

$$\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}(\cdot \mid \mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n) = \pi_{\widetilde{U},\widetilde{\alpha}}(\cdot),$$

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

Remark 7.6. If $Y_r \overset{\text{iid}}{\sim} \mathrm{N}_p(\mu, \Sigma)$ and the prior for μ is the independent flat prior $\pi(\mu) \propto 1$, integrate μ out and define

$$\bar{Y} = n^{-1}\sum_{r=1}^{n} Y_r, \qquad T_c = \sum_{r=1}^{n}(Y_r - \bar{Y})(Y_r - \bar{Y})^\top.$$

Then $T_c \sim W_p(n-1, \Sigma)$, independently of $\bar{Y}$, and the posterior Cholesky law is

$$\pi_{U+T_c,\, \alpha+(n-1)\mathbf{1}}^{\Theta_{\mathcal{D}}}.$$

Equivalently, if $S_c = T_c/(n-1)$, then $(n-1)S_c \sim W_p(n-1, \Sigma)$. This update depends on the stated flat prior; a different prior or conditioning treatment for μ must be derived separately.

The Laplace transform of the DAG-Wishart

We start with computing the Laplace transform of $\pi^{\Xi_{\mathcal{D}}}_{U,\alpha}$ by exploiting the results established in Theorem 7.3. First a preliminary result on the Laplace transform of a Gaussian inverse Gamma distribution is required.

Lemma 7.7. *Suppose* $(\lambda, \mathbf{x})$ *is a random variable with Gaussian-inverse gamma distribution:*

$$\begin{aligned} \mathbf{x}|\lambda &\sim \mathrm{N}_p(\mu, \lambda\Psi), \quad \mu \in \mathbb{R}^p, \Psi \in \mathrm{PD}_p(\mathbb{R}); \\ \lambda &\sim \mathrm{IG}(\nu, \eta). \end{aligned}$$

For $u \in \mathbb{R}^p$*, the Laplace transform* $E[\exp\{-\xi\lambda - u^\top \mathbf{x}\}]$ *is finite exactly when* $\xi \geq \frac{1}{2}u^\top \Psi u$*. When the inequality is strict, it equals*

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

where K_ν *is the modified Bessel function of the second kind. At equality, the continuous limit is* $\exp(-u^\top \mu)$*.*

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

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

Note that in computing the integral above we have used the fact that the Laplace transform of $\mathrm{N}_p(\mu, \lambda\Psi)$ at u is equal to $\exp\{-u^\top \mu + \frac{1}{2}\lambda u^\top \Psi u\}$. For computing the integral w.r.t. $d\lambda$ we use the Equation (9.42) in [24, page 235].

□

Proposition 7.8. *At test arguments* $\times_{i=1}^{p}(\xi_i, z_{\prec i]})$ *satisfying* $\xi_i > \frac{1}{2} z_{\prec i]}^{\top} \Psi_{\prec i\succ} z_{\prec i]}$, *the Laplace transform of* $\pi_{U,\alpha}^{\Xi_{\mathcal{D}}}$ *is*

$$\mathcal{L}_{\Xi_{\mathcal{D}}}(\times_{i=1}^{p}(\xi_i, z_{\prec i]})) := 2^p \prod_{i=1}^{p} \frac{1}{\Gamma(r_i)} \exp\{-z_{\prec i]}^{\top} \mu_{\prec i]}\} \left(\eta_i(\xi_i - \frac{1}{2} z_{\prec i]}^{\top} \Psi_{\prec i\succ} z_{\prec i]})\right)^{\frac{1}{2} r_i} \tag{31}$$

$$\times K_{r_i}\left(2\sqrt{\eta_i(\xi_i - \frac{1}{2} z_{\prec i]}^{\top} \Psi_{\prec i\succ} z_{\prec i]})}\right), \tag{32}$$

where $r_i = \frac{\alpha_i}{2} - \frac{pa_i}{2} - 1$, $\eta_i = \frac{1}{2} U_{ii|\prec i\succ}$, $\mu_{\prec i]} = -U_{\prec i\succ}^{-1} U_{\prec i]}$, *and* $\Psi_{\prec i\succ} = U_{\prec i\succ}^{-1}$. *Boundary arguments with equality are obtained by taking the corresponding continuous limits.*

Proof. Let $\times_{i=1}^{p}(\lambda_i, \ell_i) \sim \pi_{U,\alpha}^{\Xi_{\mathcal{D}}}$. Theorem 7.3 makes these Gaussian–inverse-gamma pairs independent, so the result is the product of Lemma 7.7 applied with the displayed parameters. □

Because the coordinate map ϕ in (27) has unit Jacobian, the preceding proposition is also the joint Laplace transform of the *free* Cholesky coordinates $\{D_{ii}, L_{\prec i]} : i = 1, \ldots, p\}$. The fixed unit diagonal of L is not included as a random coordinate. A matrix-trace version can be obtained by setting $\xi_i = \Lambda_{ii}$ and taking $z_{\prec i]}$ from the free entries of the test matrix; any deterministic diagonal contribution must be included separately.

7.2 Expected Cholesky parameters

Corollary 7.4 gives $\nu_i = \alpha_i - pa_i - 2$ degrees of freedom for the marginal t distribution of $L_{\prec i]}$. Consequently, if $\alpha_i > pa_i + 3$,

$$E(L_{\prec i]}) = -U_{\prec i\succ}^{-1} U_{\prec i]}.$$

If the stronger condition $\alpha_i > pa_i + 4$ holds, then

$$\mathrm{Var}(L_{\prec i]}) = \frac{U_{ii|\prec i\succ}}{\alpha_i - pa_i - 4} U_{\prec i\succ}^{-1}, \tag{33}$$

$$E(D_{ii}) = \frac{U_{ii|\prec i\succ}}{\alpha_i - pa_i - 4}. \tag{34}$$

The Cholesky columns are mutually independent, so their joint covariance is block diagonal with the blocks in (33). The diagonal of L is deterministically one, and all non-parent entries are deterministically zero. These existence conditions are stronger than the properness condition $\alpha_i > pa_i + 2$.

7.3 Mode in Cholesky coordinates

We compute the joint density mode with respect to Lebesgue measure on the free coordinates $(D_{ii}, L_{\prec i]})$. This base measure is part of the definition: a nonlinear transformation of this point is not, in general, a mode under the transformed coordinate measure. From (28), the density is proportional to

$$\prod_{i\in V} \lambda_i^{-\alpha_i/2} \exp\left\{-\frac{1}{2\lambda_i}\left[U_{ii|<i>} + (\ell_i + U_{<i>}^{-1}U_{\prec i]})^\top U_{<i>}(\ell_i + U_{<i>}^{-1}U_{\prec i])\right]\right\}.$$

For each i, first maximizing over ℓ_i and then differentiating with respect to λ_i gives the unique joint mode

$$\bigtimes_{i=1}^{p}\left(\frac{U_{ii|<i>}}{\alpha_i}, -U_{<i>}^{-1}U_{\prec i]}\right).$$

The regression-coordinate component is the negative of the displayed Cholesky component.

Proposition 7.9. *Let $Y_1, \ldots, Y_n$ be i.i.d. centered Gaussian observations and set $T = nS = \sum_r Y_r Y_r^\top$. Under the prior $\pi_{U,\alpha}^{\Xi_\mathcal{D}}$, the posterior Cholesky-coordinate mode is*

$$\bigtimes_{i=1}^{p}\left(\frac{(T+U)_{ii|<i>}}{\alpha_i + n}, -(T+U)_{<i>}^{-1}(T+U)_{\prec i]}\right).$$

Its regression coefficient is $(T+U)_{<i>}^{-1}(T+U)_{\prec i]}$. Completing these local parameters gives a transformed Cholesky-mode plug-in, not a Lebesgue-density MAP in covariance or precision coordinates.

The Jacobian of the mapping $(L, D) \mapsto \left(LD^{-1}L^t\right)^E$

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

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

The Jacobian of ψ is a variant of similar transformations found in [21, 15]. For completeness we still compute this Jacobian in the following lemma.

Lemma 7.10. *The Jacobian of the mapping $\psi : ((D, L) \mapsto \left(LD^{-1}L^t\right)^E$ is $\prod_{j=1}^{p} D_{jj}^{-(pa_j+2)}$.*

Proof. Let $\Upsilon \in \mathrm{R}_\mathcal{D}$, and $(D, L) \in \Theta_\mathcal{D}$ such that $\widehat{\Upsilon} = LD^{-1}L^\top$. Note that for each $(i, j) \in E$,

$$\Upsilon_{ij} = (LD^{-1}L^\top)_{ij} = \sum_{k=1}^{p} L_{ik}L_{jk}D_{kk}^{-1} = \sum_{k=1}^{j} L_{ik}L_{jk}D_{kk}^{-1}, \tag{35}$$

since L is lower triangular. Now from (35) it follows by noting that $L_{jj} = 1, \forall\, j$,

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

Arrange the entries of $(D, L) \in \Theta_{\mathcal{D}}$ as D_{11}, $\{L_{2k} : (2,k) \in E, 1 \leq k < 2\}$, D_{22}, $\{L_{3k} : (3,k) \in E, 1 \leq k < 3\}$, …, $D_{p-1,p-1}$, $\left\{L_{pk} : \; (p,k) \in E, 1 \leq k < p\right\}, D_{pp}$, and the entries of $\Upsilon \in \mathrm{R}_{\mathcal{D}}$ as Υ_{11}, $\{\Upsilon_{2k} : \; (2,k) \in E, 1 \leq k < 2\}$, Υ_{22}, $\{\Upsilon_{3k} : \; (3,k) \in E, 1 \leq k < 3\}$, …, $\Upsilon_{p-1,p-1}$, $\left\{\Upsilon_{pk} : \; (p,k) \in E, 1 \leq k < p\right\}$, Υ_{pp}. From (35) it is easily seen that Υ_{ij} depends on

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

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 given as

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

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

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

□

Exponential-family representation

Fix $\alpha_i > pa_i + 2$. We consider the identifiable family

$$\left\{\pi_{U,\alpha}^{R_{\mathcal{D}}} : U \in \mathrm{PD}_{\mathcal{D}}\right\},$$

equivalently parameterized by $U^E \in S_{\mathcal{D}}$ and the unique DAG completion U. Thus, if $\pi_{U_1,\alpha}^{R_{\mathcal{D}}} = \pi_{U_2,\alpha}^{R_{\mathcal{D}}}$ almost everywhere, then $U_1^E = U_2^E$. The ambient exponential-family representation explains the distinction between perfect and non-perfect DAGs.

Lemma 7.11. *Let $\mathcal{D}$ be DAG and let α be given. If $\mathcal{D}$ is perfect, then the Wishart family*

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

is a full regular exponential family in its natural affine support. If $\mathcal{D}$ is not perfect, then $\{\pi^{\mathrm{R}_{\mathcal{D}}}_{U^E,\alpha} : U^E \in \mathrm{S}_{\mathcal{D}}\}$ has a curved exponential-family representation in the larger moral-graph natural-parameter space.

Proof. If $\mathcal{D}$ is perfect, its moral graph adds no edges and every completed precision matrix has zeros outside E. Hence

$$\mathrm{tr}(\widehat{\Upsilon} U) = \langle \Upsilon^0, (U^E)^0 \rangle_{Z_{\mathcal{D}}},$$

and U^E ranges over an open subset of $Z_{\mathcal{D}}$. This gives a full regular exponential family.

For a non-perfect DAG, moralization adds at least one edge. The same density can be embedded in the exponential family on $Z_{\mathcal{D}^{\mathrm{m}}}$ with canonical statistic $(\widehat{\Upsilon})^{E^{\mathrm{m}}}$ and natural parameter $-\frac{1}{2}U^{E^{\mathrm{m}}}$. The constraint $U \in \mathrm{PD}_{\mathcal{D}}$ makes the added moral-edge coordinates smooth, nonlinear functions of the $|E|$ free coordinates U^E. The resulting natural-parameter set is therefore a $|E|$-dimensional curved submanifold of the higher-dimensional moral-graph parameter space. □

For non-perfect $\mathcal{D}$, this DAG-constrained family is a proper subfamily of the ambient family obtained by allowing an arbitrary positive-definite U.

The inverse DAG-Wishart for homogeneous DAGs

We next show that the inverse DAG–Wishart family $\pi^{\mathrm{S}_{\mathcal{D}}}_{U,\alpha}$ contains the inverse Wishart subclass introduced by Khare and Rajaratnam [15] for Gaussian covariance graph models. For the associated special DAGs, the density simplifies considerably. Recall that a Gaussian covariance graph model over an undirected graph $\mathcal{G} = (V, \mathscr{E})$ is denoted by $\mathscr{N}(\mathcal{G}_{\mathrm{cov}})$ and is defined as follows.

Definition 7.12. 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 $i \not\sim_{\mathcal{G}} j$, i.e., when i and j are not neighbors. Then the Gaussian covariance graph model over $\mathcal{G}$ is defined by $\mathscr{N}(\mathcal{G}_{\mathrm{cov}}) = \left\{\mathrm{N}_p(0, \Sigma) : \Sigma \in \mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}\right\}$.

A formal comparison between the DAG Wishart priors introduced in this paper and the covariance Wishart priors introduced in [15] requires a few technical definitions.

Definition 7.13. *a)* A DAG $\mathcal{D}$ is called a homogeneous DAG of type I if it is transitive (i.e., $i \to j \to k$ implies that $i \to k$), and perfect. A DAG $\mathcal{D}$ is called a homogeneous DAG 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 $i \sim_{\mathcal{G}} j \Longrightarrow \mathrm{ne}(i) \cup \{i\} \subseteq \mathrm{ne}(j) \cup \{j\}$ 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 [17] for further details on homogeneous graphs.

Note that if $\mathcal{D}$ is a homogeneous DAG 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 homogeneous (undirected) graph and using the given orientation to obtain a DAG of type I. Reversing the orientation (i.e., redirecting all the arrows to the root of the tree) will yield a 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 7.13 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 a homogeneous DAG 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 $\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}} = \mathrm{PD}_{\mathcal{D}}$ (see [18] for instance for more details.) 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 7.14. *Let $\mathcal{D} = (V, E)$ be a homogeneous DAG of either type I or II and let $\mathcal{G} = (V, \mathscr{E})$ be a homogeneous graph.*

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

b) *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*

$$\left(\Gamma \mapsto (\Gamma)^0 = \Sigma(\Gamma)\right) : \mathrm{S}_{\mathcal{D}} \to \mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}. \tag{36}$$

Let $\pi_{U,\alpha}^{\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}}$ denote the probability image of the inverse DAG Wishart $\pi_{U,\alpha}^{\mathrm{S}_{\mathcal{D}}}$ under the mapping in (36)*. Then the density of $\pi_{U,\alpha}^{\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}}$ w.r.t. Lebesgue measure is given by the expression in part(a) above.*

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

$$\prod_{i \in V} \det(\Sigma_{<i>}) = \prod_{i \in V} \Sigma_{ii|<i>}^{ch_i(\mathcal{D})}. \tag{37}$$

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

$$\det(\Sigma_{<i>}) = \prod_{\ell \in \mathrm{pa}(i)} \Sigma_{\ell\ell|<\ell>}. \tag{38}$$

If $\mathrm{pa}(i) = \emptyset$ for some i, then by our convention $\det(\Sigma_{<i>}) = 1$ and $\Sigma_{\ell\ell|<\ell>} = 1$ for any $\ell \in \mathrm{pa}(i)$ and therefore (38) 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 $\det(\Sigma_{<i>}) = \Sigma_{\ell_0\ell_0|<\ell_0>} \det(\Sigma_{<\ell_0>})$. Now by repeating this procedure we obtain the result in (38). Finally we write

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

2) Suppose $\mathcal{D}$ is homogeneous of type II. We shall proceed by induction. It is clear that (37) holds when $p = |V| = 1$. Now by the inductive hypothesis assume that (37) 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 (37) will also hold for $\mathcal{D}$ with p vertices. Now let $\Sigma \in \mathrm{PD}_{\mathcal{D}}$ be given.

Case 1) Suppose that $\mathcal{D}$ is connected. Let $\mathcal{D}_{[1]}$ be the induced DAG on $V \setminus \{1\}$. It is clear that $\mathcal{D}_{[1]}$ is a homogeneous DAG of type II and therefore by the induction hypothesis $\prod_{i=2}^{p} \det(\Psi_{<i>}) = \prod_{i=2}^{p} \Psi_{ii|<i>}^{ch_i(\mathcal{D}_{[1]})}$, where $\Psi = \Sigma_{V\setminus\{1\}}$. Note 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 $i = 2, \ldots, p$ and consequently $\Psi_{<i>} = \Sigma_{<i>}$ and $\Psi_{ii|<i>} = \Sigma_{ii|<i>}$. All together these imply that $\prod_{i=2}^{p} \det(\Sigma_{<i>}) = \prod_{i=2}^{p} \Sigma_{ii|<i>}^{ch_i(\mathcal{D}_{[1]})}$. 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 $\preceq 1 \succeq = V$ and therefore we have $\det(\Sigma_{<1>}) = \Sigma_{11|<1>}^{-1} \det(\Sigma) = \prod_{i=2}^{p} \Sigma_{ii|<i>}$. 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

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

Case 2) Suppose $\mathcal{D}$ 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 $\Psi = \Sigma_{\preceq 1 \succeq} \in \mathrm{PD}_{\mathcal{D}_1}$ and $\Psi' = \Sigma_{V\setminus \mathrm{fa}_{\mathcal{D}}(1)} \in \mathrm{PD}_{\mathcal{D}_2}$. Now applying the

induction hypothesis and the fact that $\mathcal{D}_1$ and $\mathcal{D}_2$ are disjoint we have:

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

b) It is clear that the mapping in (36) is a diffeomorphism and the Jacobian of this mapping is 1. Thus the functional form of the density $\pi^{\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}}_{U,\alpha}$ w.r.t. Lebesgue measure is same as $\pi^{\mathrm{S}_{\mathcal{D}}}_{U,\alpha}$ given by Proposition 7.14. □

Remark 7.15. For a homogeneous graph $\mathcal{G}$, the distribution $\pi^{\mathrm{PD}_{\mathcal{G}_{\mathrm{cov}}}}_{U,\alpha}$ with the density in Proposition 7.14 coincides with the covariance Wishart family introduced by Khare and Rajaratnam [15].

Supplemental Section C: The DAG-Wishart on $\mathrm{P}_{\mathcal{D}}$ and its density w.r.t. Hausdorff measure

Introduction

In this section we define the DAG–Wishart distribution directly on the precision-matrix model $\mathrm{P}_{\mathcal{D}}$ for an arbitrary DAG $\mathcal{D}$. At the level of probability measures, $\pi^{\mathrm{P}_{\mathcal{D}}}_{U,\alpha}$ is simply the pushforward of $\pi^{\Theta_{\mathcal{D}}}_{U,\alpha}$ under $(D, L) \mapsto LD^{-1}L^{\top}$. If $\mathcal{D}$ is perfect, then $\mathrm{P}_{\mathcal{D}}$ is an open subset of $\mathrm{Z}_{\mathcal{D}} \cong \mathbb{R}^{|E|}$, so this pushforward has an ordinary Lebesgue density. For a non-perfect DAG, however, $\mathrm{P}_{\mathcal{D}}$ is an $|E|$-dimensional embedded manifold in the larger space $\mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$ and therefore has ambient Lebesgue measure zero. The appropriate reference measure is instead the $|E|$-dimensional Hausdorff measure induced by a specified Euclidean metric.

Lebesgue measure of $\mathrm{P}_{\mathcal{D}}$

Lemma 7.1 gives $\mathrm{P}_{\mathcal{D}} \subset \mathrm{P}_{\mathcal{D}^{\mathrm{m}}} \subset \mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$. For a non-perfect DAG, the model has Lebesgue measure zero in every ambient linear subspace of $\mathrm{S}_p(\mathbb{R})$ that contains it. The next lemma makes the dimension argument precise.

Lemma 7.16. *Suppose $\mathcal{D} = (V, E)$ is a non-perfect DAG and $\mathbb{E}$ is a linear subspace of $\mathrm{S}_p(\mathbb{R})$ containing $\mathrm{P}_\mathcal{D}$. Then $\mathbb{E}$ contains $\mathrm{Z}_{\mathcal{D}^\mathrm{m}}$. Consequently, $\mathrm{P}_\mathcal{D}$ has Lebesgue measure zero in $\mathbb{E}$.*

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

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

The matrices $\widetilde{E}^{(ij)}$ form a basis of $\mathrm{Z}_{\mathcal{D}^\mathrm{m}}$. First, $\mathbb{E}$ contains $\mathrm{Z}_\mathcal{D}$. Indeed, positive diagonal precision matrices give the diagonal directions by taking linear combinations, and, for each directed edge $i \to j$, matrices of the form $(I + te_ie_j^\top)(I + te_ie_j^\top)^\top$, together with the already obtained diagonal directions, give $\widetilde{E}^{(ij)}$. It remains to obtain the moral edges. Let $(i, j) \in E^\mathrm{m} \setminus E$, with $i > j$. Then some $k < j < i$ satisfies $i \to k \leftarrow j$. Define the lower triangular matrix $L^{(ij)} \in \mathcal{L}_\mathcal{D}$ by

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

Then $L^{(ij)}(L^{(ij)})^\top = T + \widetilde{E}^{(ij)}$ for some $T \in \mathrm{Z}_\mathcal{D}$. Hence $\widetilde{E}^{(ij)} \in \mathbb{E}$ and $\mathrm{Z}_{\mathcal{D}^\mathrm{m}} \subseteq \mathbb{E}$.

The model $\mathrm{P}_\mathcal{D}$ is a smooth manifold of dimension $|E|$, whereas $\dim(\mathrm{Z}_{\mathcal{D}^\mathrm{m}}) = |E^\mathrm{m}| > |E|$ for a non-perfect DAG. Thus $\dim(\mathbb{E}) > |E|$, and the embedded $|E|$-dimensional manifold $\mathrm{P}_\mathcal{D}$ has Lebesgue measure zero in $\mathbb{E}$. □

Consequently, Lemma 7.16 implies that if $\mathcal{D}$ is non-perfect then $\pi^{\mathrm{P}_\mathcal{D}}_{U,\alpha}$ has no density w.r.t. Lebesgue measure.

The density of $\pi^{\mathrm{P}_\mathcal{D}}_{U,\alpha}$ w.r.t. Hausdorff measure

We now derive the density of $\pi^{\mathrm{P}_\mathcal{D}}_{U,\alpha}$ with respect to Hausdorff measure. Write $q = |E|$, where the diagonal pairs are included in E, and let

$$\mathcal{O}_\mathcal{D} = (0, \infty)^p \times \mathbb{R}^{q-p}.$$

For $\theta = (d_1, \ldots, d_p, (\ell_{ij})_{(i,j)\in E,\, i>j}) \in \mathcal{O}_\mathcal{D}$, set

$$D(\theta) = \mathrm{diag}(d_1, \ldots, d_p), \qquad L(\theta) = I + \sum_{(i,j)\in E,\, i>j} \ell_{ij} e_i e_j^\top.$$

Thus $\Theta_{\mathcal{D}}$ is identified with the open subset $O_{\mathcal{D}} \subset \mathbb{R}^q$; it is not itself a vector space because D must be positive and L must have unit diagonal. In the target, we identify $\mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}$ with $\mathbb{R}^{|E^{\mathrm{m}}|}$ by retaining one lower-triangular coordinate for each symmetric position. All Euclidean norms and Hausdorff measures below refer to these specified coordinates. Using the Frobenius metric on symmetric matrices would instead weight off-diagonal directions twice in the Gram matrix, so the Jacobian and displayed density would have to be adjusted.

The map

$$\psi : O_{\mathcal{D}} \longrightarrow \mathrm{Z}_{\mathcal{D}^{\mathrm{m}}}, \qquad \psi(\theta) = L(\theta)D(\theta)^{-1}L(\theta)^{\top},$$

is smooth and injective. Order the domain coordinates as D_{11}, L_{21} (when $(2,1) \in E$), D_{22}, L_{31} (when $(3,1) \in E$), and so on. Let $\mathscr{I} = E^{\mathrm{m}} \setminus E$. In the target, order first the positions in E in the corresponding order and then the positions in $\mathscr{I}$ lexicographically. The required partial derivatives are

$$\frac{\partial (LD^{-1}L^{t})_{kl}}{\partial D_{ii}} = -D_{ii}^{-2} L_{ki} L_{li} \tag{39}$$

$$\frac{\partial (LD^{-1}L^{t})_{kl}}{\partial L_{ij}} = \delta_{ik} D_{jj}^{-1} L_{lj} + \delta_{il} D_{jj}^{-1} L_{kj}, \tag{40}$$

where δ_{uv} is the Kronecker delta. Partition $D\psi$ into the $q \times q$ block A_ψ comprising the rows indexed by E and the $|\mathscr{I}| \times q$ block C_ψ comprising the remaining rows. The block A_ψ is the Jacobian from Lemma 4.5; in particular it is nonsingular. Hence ψ is an immersion, and the area formula gives the q-dimensional Jacobian

$$\begin{aligned}
\mathbf{J}\psi(D,L) &= \sqrt{\det((D\psi)^{\top} D\psi)} \\
&= \sqrt{\det(A_\psi^{\top} A_\psi + C_\psi^{\top} C_\psi)} \\
&= |\det(A_\psi)| \sqrt{\det(I + A_\psi^{-\top} C_\psi^{\top} C_\psi A_\psi^{-1})} \\
&= \prod_{j=1}^{p} D_{jj}^{-(pa_j+2)} \sqrt{\det(I + A_\psi^{-\top} C_\psi^{\top} C_\psi A_\psi^{-1})}.
\end{aligned}$$

Therefore we have proved the following.

Theorem 7.17. *Let A_ψ and C_ψ be the derivative blocks defined above. With respect to the restriction to $\mathrm{P}_{\mathcal{D}}$ of the q-dimensional Hausdorff measure induced by the chosen lower-triangular target coordinates, the density of $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{D}}}$ is*

$$z_{\mathcal{D}}(U,\alpha)^{-1} \exp\{-\tfrac{1}{2}\operatorname{tr}(\Omega U)\} \prod_{i=1}^{p} D_{ii}^{-\alpha_i/2 + pa_i + 2} \det(I + A_\psi^{-\top} C_\psi^{\top} C_\psi A_\psi^{-1})^{-1/2}, \tag{41}$$

where D and L are the unique modified-Cholesky coordinates of Ω.

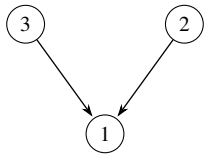

Figure 5: The non-perfect collider DAG $3 \to 1 \leftarrow 2$ used in the Hausdorff-density example.

Example 7.18. For the collider $3 \to 1 \leftarrow 2$ in Figure 5, use

$$\theta = (D_{11}, L_{21}, D_{22}, L_{31}, D_{33}),$$
$$\eta = (\Omega_{11}, \Omega_{21}, \Omega_{22}, \Omega_{31}, \Omega_{33}, \Omega_{32}).$$

Equations (39) and (40) give

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

Consequently,

$$\mathbf{J}\psi(D, L) = D_{11}^{-4} D_{22}^{-2} D_{33}^{-2} \sqrt{(1 + L_{21}^2)(1 + L_{31}^2)}.$$

Thus, under the stated coordinate metric, the density with respect to $\mathcal{H}^5$ on $\mathbb{R}^6$ is

$$z_{\mathcal{D}}(U, \alpha)^{-1} \exp\{-\tfrac{1}{2} \operatorname{tr}(\Omega U)\} D_{11}^{-\alpha_1/2+4} D_{22}^{-\alpha_2/2+2} D_{33}^{-\alpha_3/2+2} \{(1 + L_{21}^2)(1 + L_{31}^2)\}^{-1/2},$$

where the D_{ii} and L_{ij} are the unique functions of Ω defined by the modified Cholesky factorization.

Supplemental Section D: Computational algorithms and more related results

DAG-constrained MLE and Cholesky-mode plug-in

Algorithm 7.19 (Maximum Likelihood)**.** Let $Y_1, \ldots, Y_n$ be independent centered Gaussian observations and $S = n^{-1} \sum_{r=1}^n Y_r Y_r^\top$. Assume $n \geq \max_i(pa_i + 1)$, which makes every displayed Schur complement positive almost surely. (If a mean is first estimated, the corresponding sufficient condition is $n \geq \max_i(pa_i + 2)$.) For each i set

$$\lambda_i = S_{ii|<i>} \in \mathbb{R}_+ \quad \text{and} \quad \boldsymbol{\beta}_i = S_{<i>}^{-1} S_{<i]} \in \mathbb{R}^{pa_i}.$$

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

1. Initialize $\widehat{\Sigma}_{ii} = \lambda_i$ for each i such that $\mathrm{pa}(i) = \emptyset$ (in particular for $i = p$);
2. set $\widehat{\Sigma}_{<i]} = \widehat{\Sigma}_{<i>}\boldsymbol{\beta}_i$ if $\mathrm{pa}(i) \neq \emptyset$;
3. set $\widehat{\Sigma}_{ii} = \lambda_i + \boldsymbol{\beta}_i^{\top}\widehat{\Sigma}_{<i>}\boldsymbol{\beta}_i$ if $\mathrm{pa}(i) \neq \emptyset$;
4. set $\widehat{\Sigma}_{\not<i]} = \widehat{\Sigma}_{\not<i>}\widehat{\Sigma}_{<i>}^{-1}\widehat{\Sigma}_{<i]}$ if $\mathrm{pa}(i) \neq \emptyset$, otherwise set $\widehat{\Sigma}_{\not<i]} = 0$.

The precision-matrix MLE is $\widehat{\Omega} = \widehat{\Sigma}^{-1}$.

Proposition 7.9 gives the joint posterior mode with respect to Lebesgue measure in the free modified-Cholesky coordinates. Applying the same covariance completion to its local residual variances and regression coefficients gives the following plug-in.

Algorithm 7.20 (Transformed Cholesky-mode plug-in)**.** Let $S = n^{-1}\sum_{r=1}^{n} Y_r Y_r^{\top}$. For $i = 1, \ldots, p$, set

$$\lambda_i = \frac{(nS + U)_{ii|<i>}}{\alpha_i + n}, \qquad \boldsymbol{\beta}_i = (nS + U)_{<i>}^{-1}(nS + U)_{<i]}.$$

When $\mathrm{pa}(i) = \emptyset$, use $\lambda_i = (nS + U)_{ii}/(\alpha_i + n)$ and $\boldsymbol{\beta}_i = 0$. For $i = p, p-1, \ldots, 1$, apply steps 1–4 of Algorithm 7.19, writing the result as $\widetilde{\Sigma}_{\mathrm{chol}}$.

Set $\widetilde{\Omega}_{\mathrm{chol}} = \widetilde{\Sigma}_{\mathrm{chol}}^{-1}$. These matrices are transformations of the Cholesky-coordinate mode. They are not, in general, MAP estimators with respect to Lebesgue measure in covariance or precision coordinates.

Historical covariance-estimation results

Table 3 preserves the reported covariance results for $p = 500$ and admissible-edge probability 0.01. The notation is corrected to match Section 6: $\bar{\Sigma}$ is the full posterior mean, $\widehat{\Omega}_{\mathrm{coord}}^{-1}$ is the inverse of the completed posterior-mean precision coordinates, and $\widetilde{\Sigma}_{\mathrm{chol}}$ is the transformed Cholesky-mode plug-in. Positive entries mean lower loss than the constrained MLE; negative entries mean higher loss. The different rankings under L_1 and L_2 illustrate that none of the three procedures uniformly dominates. Because the public archive contains neither saved replicate-level output nor an aggregation script, the values and boldface rankings are historical and have not been independently reproduced; no Monte Carlo uncertainty is available.

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

Table 3: Historically reported relative changes from the constrained MLE when estimating Σ, with $p = 500$. Positive values indicate lower loss. The three rows are the full posterior mean, the inverse of the coordinate-completed precision estimate, and the transformed Cholesky-mode plug-in. Saved outputs and Monte Carlo uncertainty are unavailable.

Sensitivity to edge density and to one contamination design

Table 4 preserves the historical results obtained with $p = 500$, $c = 3$, $u = 3$, and admissible-edge probabilities from 0.005 to 0.02. The results are strongly design-dependent. At edge probability 0.02 and $n = 100$, every listed procedure is much worse than the constrained MLE under the reported L_2 loss, with relative changes between -795.0% and -834.4%. The table therefore refutes any claim of uniform dominance and shows that the chosen hyperparameters do not transfer reliably across edge densities. It does not identify an alternative setting, because that would require a separate, properly replicated tuning study. Figure 6 is the preserved historical loss curve for edge probability 0.015; its numerical source data are absent from the archive.

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

Table 4: Historically reported relative changes from the constrained MLE when estimating Ω at different admissible-edge probabilities, with $p = 500$ and $(c, u) = (3, 3)$. Positive values indicate lower loss and negative values higher loss. The procedures are $\widehat{\Omega}_{\text{coord}}$, $\bar{\Sigma}^{-1}$, and $\widetilde{\Omega}_{\text{chol}}$. Saved outputs and Monte Carlo uncertainty are unavailable.

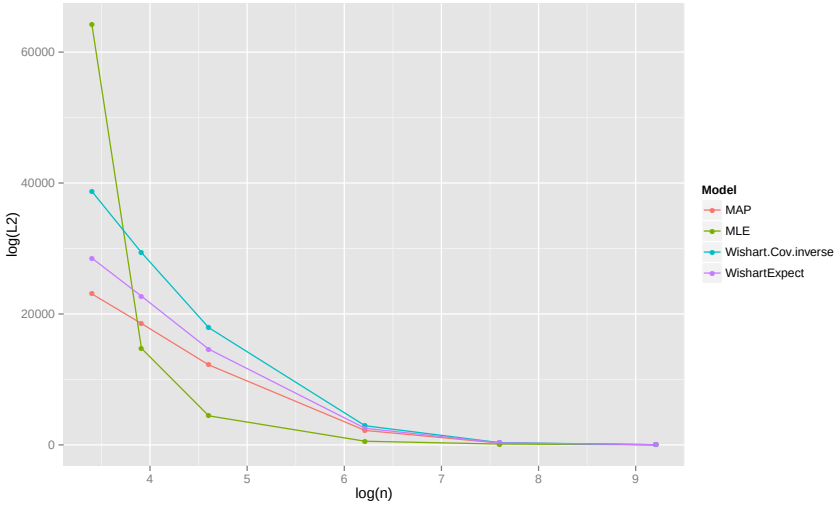

Figure 6: Historical L_2-loss curves as n increases, at admissible-edge probability 0.015. The source values and uncertainty bands are unavailable.

The companion driver also calls `rmvDAG` with the arguments `errDist="mixN100"` and `mix=0.01`, described in the original study as replacing 1% of Gaussian errors by $N(0, 100)$ errors. The generated columns are then centered and standardized before estimation. The exact semantics depend on the unrecorded package version. The driver requests 15 batches of 20 replicates for each sample size, but the resulting `.Rda` file is absent.

Table 5 preserves the reported relative changes for this single contamination mechanism. Several precision-loss changes are larger than in the uncontaminated table, while some covariance L_2 changes remain negative. Without the saved output, Monte Carlo uncertainty, alternative contamination mechanisms, influence analysis, or breakdown calculations, this experiment supports only a design-specific sensitivity comparison. It does not establish robustness or a general guarantee under misspecification.

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

Table 5: Historically reported relative changes from the constrained MLE in the single `mixN100` contamination design, with $p = 500$ and $(c, u) = (3, 3)$. Positive values indicate lower loss. Saved outputs, package versions, and Monte Carlo uncertainty are unavailable; this is not a general robustness assessment.

Historical call-center prediction example

Following the descriptions in [3] and [20], the original analysis reports 239 operating days from a financial call center in 2002, with call counts N_{ij} in 102 ten-minute intervals between 7 a.m. and midnight. It uses the variance-stabilizing transformation $x_{ij} = \sqrt{N_{ij} + 1/4}$. The raw data file is not included in the public archive, so these exclusions, counts, and transformations cannot be checked in this revision. Chronological order is imposed as the admissible parent order. This is a predictive modeling restriction, not evidence that the estimated edges are causal or that latent common causes are absent.

The task is to predict the second 51 intervals from the first 51. Write $x_i = (x_i^{(1)}, x_i^{(2)})$ and partition

$$\mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

For an estimated mean and covariance, the conditional-mean predictor is

$$\widehat{x}^{(2)} = \widehat{\mu}^{(2)} + \widehat{\Sigma}_{21}\widehat{\Sigma}_{11}^{-1}(x^{(1)} - \widehat{\mu}^{(1)}).$$

The reported split uses the first 205 days for training and the last 34 for testing. The four procedures are the unrestricted sample-covariance predictor (Naive-MLE), graph-constrained MLEs after LassoDAG and DAG–Wishart selection, and the inverse $\widehat{\Omega}^{-1}_{\text{coord}}$

after DAG–Wishart selection. The last procedure is called "DAG-W-Precision" in the historical figure, but it is not the posterior mean of Σ.

The script fixes $\kappa = 0.1$ for LassoDAG and $c = 1, b = 3$ for DAG–Wishart; these choices were not selected by cross-validation on the displayed test split. More importantly, after estimating and removing the training mean, the script uses $n\,\mathrm{var}(x)$ and increments α by n. Under the flat-mean analysis in Remark 7.6, the update is instead $(n-1)\,\mathrm{var}(x)$ and $\alpha + (n-1)\mathbf{1}$. Thus the reported coordinate-shrinkage predictor does not implement the corrected unknown-mean posterior. The unavailable raw data and fitted-model object prevent recomputation.

For response interval $k = 1, \ldots, 51$, the plotted average absolute error is

$$E_k = \frac{1}{34} \sum_{i=1}^{34} \left| x_{i,51+k} - \widehat{x}_{i,51+k} \right|.$$

Figure 7 and Table 6 preserve the historically reported point estimates. They have no resampling uncertainty and cannot establish that one graph-selection method is generally superior. The table statistic is the mean, over test days, of the 51-dimensional sum of squared errors,

$$\text{meanSSE} = \frac{1}{34} \sum_{i=1}^{34} \|x_i^{(2)} - \widehat{x}_i^{(2)}\|_2^2,$$

not a per-coordinate mean squared error.

	Naive-MLE	LassoDAG-MLE	DAG-W-MLE	DAG-W-coordinate
Mean SSE	172.976	166.138	142.730	123.438

Table 6: Historically reported mean test-day sums of squared prediction errors for the call-center example. "DAG-W-coordinate" uses $\widehat{\Omega}^{-1}_{\text{coord}}$. These values are not independently verified and do not include uncertainty.

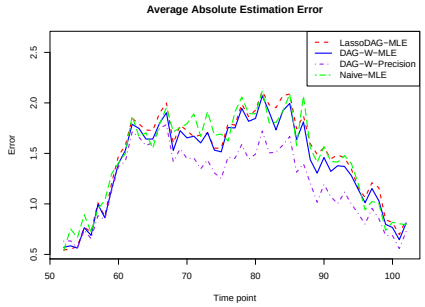

Figure 7: Historical mean absolute errors. "DAG-W-Precision" denotes $\widehat{\Omega}^{-1}_{\text{coord}}$; inputs and uncertainty are unavailable.