High dimensional Bayesian inference for Gaussian directed acyclic graph models

Emanuel Ben-David [*] Tianxi Li [†] Hélène Massam [‡]
Bala Rajaratnam [§]

US Census Bureau, University of Michigan, York University, Stanford University

Abstract

In this paper, we consider Gaussian models Markov with respect to an arbitrary directed acyclic graph with a known ordering of the vertices. We first construct a family of conjugate priors for the Cholesky parametrization of the covariance matrix of such models. This family has as many shape parameters as the number of vertices, and naturally extends the work of Geiger and Heckerman (2002). From these distributions, we derive prior distributions for the covariance and precision parameters of the Gaussian directed acyclic graph models. Our work thus extends to arbitrary directed acyclic graphs the works of Dawid and Lauritzen (1993) and Letac and Massam (2007) for Gaussian models Markov with respect to a decomposable graph. For this reason, we call our distributions DAG-Wisharts. These distributions possess the strong hyper Markov properties and thus allow for explicit estimation of the covariance and precision parameters, regardless of the dimension of the problem. They also allow for model selection and covariance estimation in the space of directed acyclic graph models with a known ordering of the vertices. We demonstrate via several numerical examples that the proposed method scales well to high-dimensions.

1 Introduction

The priors on the parameters of a normal distribution Markov with respect to a directed acyclic graph now have a long history which starts with Geiger and Heckerman (2002).

[*]e.h.bendavid@gmail.com
[†]tianxili@umich.edu
[‡]massamh@mathstat.yorku.ca
[§]brajaratnam01@gmail.com

Such distributions have been derived from some types of (inverse) Wishart distributions and we thus call them the DAG-Wishart priors (DAG is a standard acronym for a directed acyclic graph). The different steps in this history are marked by an increase in flexibility in the shape of the prior. In Geiger and Heckerman (2002), the prior is derived from the Wishart distribution which has only one shape parameter. Dawid and Lauritzen (1993) introduced the hyper inverse Wishart distribution which is the equivalent of the inverse Wishart but for the incomplete covariance matrix which corresponds to the free parameters of a Gaussian distribution Markov with respect to a decomposable graph. The hyper inverse Wishart in Dawid and Lauritzen (1993) is actually equivalent to the DAG-Wishart defined in Geiger and Heckerman (2002) but for the restricted class of so-called perfect directed acyclic graphs, those that are Markov equivalent to decomposable graphs. The hyper inverse Wishart still has only one shape parameter. For decomposable graphs, Letac and Massam (2007) introduced a generalization of the hyper inverse Wishart, denoted the $IW_{P_\mathcal{G}}$ which has $k+1$ shape parameters, where k is the number of cliques. This distribution thus offers greater flexibility than the hyper inverse Wishart.

In this paper $\mathcal{D}$ always denotes a directed acyclic graph where we assume that the ordering of the vertices are known. We introduce a DAG-Wishart which is analogous to the $IW_{P_\mathcal{G}}$ but introduces yet more flexibility in the choice of multiple shape parameters. The hyper inverse Wishart and the $IW_{P_\mathcal{G}}$ Wishart were derived from the Wishart. In this paper, we proceed in the other direction, we start by defining the multiple shape parameter DAG-Wishart on a convenient space, with one shape parameter for each vertex, and then fold it back into a Wishart-type distribution for the incomplete covariance matrix corresponding to the parametrization of the Gaussian distribution Markov with respect to $\mathcal{D}$. An advantage of the DAG-Wishart distributions proposed in this paper is that, when we use them as priors, high dimensional posterior analysis is readily amenable mainly because these distributions possess strong directed hyper Markov properties, which in turn result in closed form solutions for their posterior moments and marginal likelihoods.

The main difficulty in achieving this goal is that when a directed acyclic graph is no longer perfect, defining distributions on the space of covariance or precision matrices is, in a sense, an ill-defined problem, as these spaces are curved manifolds, and thus no distribution defined on them has red a density with respect to the Lebesgue measure. Consequently, tools for posterior inference on these spaces are not immediately available. For this reason, we need to identify isomorphisms between these two spaces and new spaces which are the projections of covariance and precision matrices onto Euclidean spaces. These are termed the space of incomplete covariance and precision matrices and correspond, respectively, to functionally independent elements of the covariance and precision matrices of Gaussian directed acyclic graph models. Given an incomplete matrix in the space defined by a directed acyclic graph $\mathcal{D}$, we rely on results and algorithms for com-

pletion given in Ben-David and Rajaratnam (2012) to obtain the corresponding unique covariance and precision matrices of the corresponding Gaussian directed acyclic graph model. Therefore, with our approach we develop a unified framework for Gaussian directed acyclic graph models that naturally extends to general directed acyclic graphs the recent methodological contributions by Letac and Massam (2007) and others (Rajaratnam et al., 2008) valid only for decomposable Gaussian graphical models, that is perfect directed acyclic graphs. We also use the DAG-Wishart approach to develop a Bayesian methodology for model selection and covariance estimation that can scale better than any other Bayesian methods that we are aware of. Model selection is undertaken within the class of directed acyclic graphs when the order of the vertices is given. Our Bayesian model selection which is based on the marginal likelihood score uses the Lasso-DAG method (Shojaie and Michailidis, 2010) for various levels of the penalty parameter as possible starting points. It then explores the space of Gaussian directed acyclic graph models further with an improved version of the stochastic shot-gun search (SSS) of Jones et al. (2005). Our method is thus a hybrid version of these two principled approaches. More importantly, our proposed approach is able to overcome the computational challenges of the Bayesian model selection problem and is shown to perform extremely well.

2 Preliminaries

2.1 Gaussian directed acyclic graph models

For a set V, let $|V|$ denote the cardinality of V. Let $\mathbb{R}^V$ and $\mathbb{R}^{V\times V}$ denote respectively the linear spaces of $|V|$-dimensional vecorts $x = (x_i \mid i \in V)$ and $|V| \times |V|$ real matrices $A = (A_{ij})_{i,j\in V}$. The spaces of $|V| \times |V|$ symmetric and positive definite matrices are respectively denoted by $S_V(\mathbb{R})$ and $PD_V(\mathbb{R})$. When $V = \{1, 2, \ldots, p\}$ aforementioned spaces are denoted by $\mathbb{R}^p$, $\mathbb{R}^{p\times p}$, $S_p(\mathbb{R})$ and $PD_p(\mathbb{R})$. A positive definite matrix is sometimes denoted by $\Sigma \succ 0$. For $a, b \subseteq V$, let x_a denote the subvector $(x_i \mid i \in a)$ and let A_{ab} denote the $|a| \times |b|$ submatrix $(A_{ij})_{i\in a, j\in b} \in \mathbb{R}^{a\times b}$. When $b = V \setminus a$ the Schur complement of Σ_{aa} is defined as $\Sigma_{bb|a} = \Sigma_{bb} - \Sigma_{ba}(\Sigma_{aa})^{-1}\Sigma_{ab}$.

Convention 2.1 *Throughout the paper we shall also use the following conventions:* $A_a = A_{aa}$, $A_a^{-1} = (A_{aa})^{-1}$, $A_\emptyset = 1$, *and when* $a \neq \emptyset$, $A_{a,\emptyset}^\top = A_{\emptyset,a} = 0 \in \mathbb{R}^a$.

Let $\mathcal{D} = (V, E)$ be a directed acyclic graph. From now on we assume that the vertices are labeled $1, 2, \ldots, p$. A vertex i is a parent of $j \in V$, denoted by $i \to j$, if $(i, j) \in E$. The set of parents of j and its cardinality are denoted by $pa(j)$ and pa_j, respectively. The family of j, denoted by $fa(j)$, is $pa(j) \cup \{j\}$ and its cardinality is denoted by fa_j. The set of ancestors of a vertex j, denoted by $an(j)$, is the set of those vertices $i \neq j$ such that there is a directed path $i \to \cdots \to j$ (note that paths of length 0 are not allowed). Similarly,

the set of descendants of a vertex $i \neq j$, denoted by $de(i)$, is the set of those vertices j such that $i \to \cdots \to j$. The set of non-descendants of i is $nd(i) = V \setminus (de(i) \cup \{i\})$. A set $A \subseteq V$ is called ancestral when A contains the parents of its members. In this paper, unless otherwise stated, we shall always assume without loss of generality that the ordering of the vertices of $\mathcal{D}$ is parent ordering, that is, $i \to j$ implies that $i > j$.

Let $X = (X_1, \ldots, X_p)^\top$ and $x = (x_1, \ldots, x_p)^\top$ denote a random vector and its observed value in $\mathbb{R}^p$, respectively. A Gaussian directed acyclic graph model (or Gaussian Bayesian network) over $\mathcal{D}$, denoted by $\mathscr{N}(\mathcal{D})$, is the statistical model that consists of all multivariate Gaussian distributions $N_p(0, \Sigma)$ obeying the ordered directed Markov property with respect to $\mathcal{D}$, that is, $X \sim N_p(0, \Sigma) \in \mathscr{N}(\mathcal{D}) \implies X_i \perp\!\!\!\perp X_{pr(i)} \mid X_{pa(i)}$ for each i, where $pr(i) = \{i+1, \ldots, p\} \setminus pa(i)$ is the set of the predecessors without the parents of i. Note that $N_p(\mu, \Sigma) \in \mathscr{N}(\mathcal{D})$ if and only if $N_p(0, \Sigma) \in \mathscr{N}(\mathcal{D})$ (see Andersson and Perlman (1998) for a simple proof). Let $PD_{\mathcal{D}}$ denote the space of covariance matrices $\{\Sigma \succ 0 : \ N_p(0, \Sigma) \in \mathscr{N}(\mathcal{D})\}$ and let $P_{\mathcal{D}}$ denote the space of precision matrices $\{\Sigma^{-1} \succ 0 : \ N_p(0, \Sigma) \in \mathscr{N}(\mathcal{D})]$. A precision matrix in $P_{\mathcal{D}}$ is usually denoted by Ω. It is clear that the Gaussian distributions in $\mathscr{N}(\mathcal{D})$ can be parametrized by the elements of $PD_{\mathcal{D}}$ or $P_{\mathcal{D}}$.

For an undirected graph $\mathcal{G}$, $\mathscr{N}(\mathcal{G})$ is defined similarly 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 $PD_{\mathcal{G}} = \{\Sigma : N_p(0, \Sigma) \in \mathscr{N}(\mathcal{G})\}$ and the space of precision matrices $P_{\mathcal{G}} = \{\Omega : \Omega^{-1} \in PD_{\mathcal{G}}\}$. Note that, for us, $PD_{\mathcal{D}}$ and $P_{\mathcal{D}}$ are parameter spaces of primary interest as they arise naturally in the parametrization of Gaussian densities. However, in order to develop multiple shape parameter Wishart priors on these spaces, which is the main theoretical purpose of this paper,we begin with the more natural and more convenient Cholesky type parametrization of $\mathscr{N}(\mathcal{D})$ that we discuss in the next subsection.

2.2 Cholesky parametrizations of Gaussian directed acyclic graph models

Consider a Gaussian distribution $N_p(0, \Sigma) \in \mathscr{N}(\mathcal{D})$. It is a well-known fact that the structure of $\mathcal{D}$ is reflected in the Cholesky decomposition of the precision matrix Σ^{-1}. A precise explanation is as follows. Let $\mathcal{L}_{\mathcal{D}}$ denote the set of lower triangular matrices with unit diagonals and $L_{ij} = 0$ if $i \notin pa(j)$, and let $\mathcal{D}_+^p$ denote the set of strictly positive diagonal $p \times p$ matrices. Then $\Sigma^{-1} \in P_{\mathcal{D}}$ if and only if there exist $L \in \mathcal{L}_{\mathcal{D}}$ and $D \in \mathcal{D}_+^p$ such that $\Sigma^{-1} = LD^{-1}L^\top$. The latter decomposition of $\Omega = \Sigma^{-1}$ is called the modified Cholesky decomposition of Ω. We call $\Theta_{\mathcal{D}} = \mathcal{D}_+^p \times \mathcal{L}_{\mathcal{D}}$ the Cholesky space of $\mathcal{D}$, the

pair $(D, L) \in \Theta_{\mathcal{D}}$ a Cholesky parameter, and $\left\{ N_p(0, \left(L^\top\right)^{-1} DL^{-1}) : \ (D, L) \in \Theta_{\mathcal{D}} \right\} \equiv \mathcal{N}(\mathcal{D})$ as the Cholesky parametrization of $\mathcal{N}(\mathcal{D})$.
By applying the directed factorization property (DF) of $N_p(0, \Sigma) \in \mathcal{N}(\mathcal{D})$ we have, for $x \in \mathbb{R}^p$,

$$dN_p(0, \Sigma)(x) = \prod_{i \in V} dN(\Sigma_{i,pa(i)} \Sigma^{-1}_{pa(i)} x_{pa(i)}, \Sigma_{ii|pa(i)})(x_i \mid x_{pa(i)}), \tag{1}$$

where $N(\Sigma_{i,pa(i)} \Sigma^{-1}_{pa(i)} x_{pa(i)}, \Sigma_{ii|pa(i)})(x_i \mid x_{pa(i)})$ is the conditional distribution of X_i given $X_{pa(i)}$. Note that $\Sigma_{i,pa(i)} \Sigma^{-1}_{pa(i)}$ is the regression coefficient of X_i in the regression of X_i on $X_{pa(i)}$, and $\Sigma_{ii|pa(i)}$ is the conditional variance of $X_i \mid X_{pa(i)} = x_{pa(i)}$. One can easily show that

$$L_{pa(i),i} = -\Sigma^{-1}_{pa(i)} \Sigma_{pa(i),i} \text{ and } D_{ii} = \Sigma_{ii|pa(i)} \ \text{ for each } i \in V \tag{2}$$

and that the mapping

$$\Pi_{\mathcal{D}} \equiv \left(\Sigma \mapsto \times_{i \in V} (\Sigma_{ii|pa(i)}, \Sigma^{-1}_{pa(i)} \Sigma_{pa(i),i}) \right) : PD_{\mathcal{D}} \to \Xi_{\mathcal{D}} = \times_{i \in V} \left(\mathbb{R}_+ \times \mathbb{R}^{pa(i)} \right) \tag{3}$$

is a bijection. In order to construct the inverse of this mapping let $\times_{i \in V} (\lambda_i, \beta_{pa(i)})$ denote a typical element in $\Xi_{\mathcal{D}}$, with the convention that $\beta_{pa(i)} = 0$ whenever $pa(i) = \emptyset$. Using the fact (see Andersson and Perlman (1998)) that

$$\Sigma_{i,pr(i)} = \Sigma_{i,pa(i)} \Sigma^{-1}_{pa(i)} \Sigma_{pa(i),pr(i)}, \text{ for every } i \in V,$$

the covariance matrix Σ can be recursively constructed starting from the largest index p, by setting i) $\Sigma_{ii} = \lambda_i + \beta^\top_{pa(i)} \Sigma_{pa(i)} \beta_{pa(i)}$; ii) $\Sigma_{pa(i),i} = \Sigma_{pa(i)} \beta_{pa(i)}$ and iii) $\Sigma_{i,pr(i)} = \Sigma_{i,pa(i)} \Sigma^{-1}_{pa(i)} \Sigma_{pa(i),pr(i)}$.

3 The multiple-shape parameter DAG-Wishart distribution on Cholesky Space

The main goal of this section is to introduce a new family of multiple shape parameter distributions on the Cholesky space $\Theta_{\mathcal{D}}$ as a natural generalization of the distribution of the Cholesky factor of a Wishart random matrix. The distributions we are going to define now are multiple shape parameter distributions, defined for all directed acyclic graphs, which are extensions of the traditional Wishart priors studied in Geiger and Heckerman (2013, 2002) and the Wishart $W_{P_{\mathcal{G}}}$ defined in Letac and Massam (2007). We will also explore, in this section, some of the important properties of these distributions.

3.1 DAG-Wishart density

In this subsection we explain how we are led to the definition of the DAG-Wishart on the Cholesky space for an arbitrary directed acyclic graph.
(a) Suppose $\mathcal{D}$ is a complete directed acyclic graph. Then $P_{\mathcal{D}}$ is the space of positive definite matrices and the classical Wishart $W_p(\eta, U)$ is defined on this space. We can show that if $\Sigma^{-1} \sim W_p(\eta, U)$, then under the mapping $\Sigma^{-1} \mapsto (D, L)$, where (D, L) is the modified Cholesky factorization of Σ^{-1}, the image of $W_p(\eta, U)$ on $\Theta_{\mathcal{D}}$ is a density proportional to

$$\exp\left[-\frac{1}{2}tr\left(\left(LD^{-1}L^{\top}\right)U\right)\right]\prod_{i=1}^{p}D_{ii}^{-\frac{\alpha_i}{2}}, \tag{4}$$

with $\alpha_i = \eta + p - 2i + 3$. Note that by using Eq. (2) this density can be written as

$$\exp\left[-\frac{1}{2}\sum_{i=1}^{p}\Sigma^{-1}_{ii|pa(i)}(\Sigma^{-1}_{pa(i)}\Sigma_{pa(i),i} - U^{-1}_{pa(i)}U_{pa(i),i})^{\top}U_{pa(i)}(\Sigma^{-1}_{pa(i)}\Sigma_{pa(i),i} - U^{-1}_{pa(i)}U_{pa(i),i})\right]$$
$$\prod_{i=1}^{p}\exp\left[\Sigma^{-1}_{ii|pa(i)}U_{ii|pa(i)}\right]\Sigma^{-\frac{1}{2}\alpha_i}_{ii|pa(i)}. \tag{5}$$

In Eq. (4) and Eq. (5) the $\alpha_i,\ i = 1, \ldots, p$ appear as multiple shape parameters, however, since they are all functions of the one original shape parameter η, there is still just one shape parameter.

(b) Next suppose $\mathcal{D}$ is a perfect directed acyclic graph. Then $P_{\mathcal{D}}$ is identical to $P_{\mathcal{G}}$, where $\mathcal{G}$ is the undirected version of $\mathcal{D}$ and therefore decomposable. Let k be the number of the cliques of $\mathcal{G}$. In this case we can consider the Wishart $W_{P_{\mathcal{G}}}(U, \eta)$ defined in Letac and Massam (2007), with multiple shape parameter $\eta \in \mathbb{R}^k$, as a distribution on $P_{\mathcal{D}}$. By Lemma 5.1 in Ben-David and Rajaratnam (2014) there exists a perfect order $(C_1, C_2, \ldots, C_k)$ of the cliques of $\mathcal{G}$ which respects the vertex numbering in $\mathcal{D}$, that is, there exists a perfect order of the cliques of $\mathcal{G}$ such that the histories are ancestral in $\mathcal{D}$. We recall that the histories, residuals and minimal separators are defined as $H_1 = \emptyset,\ \ H_\nu = C_1 \cup \cdots \cup C_\nu$, $R_\nu = C_\nu \setminus H_{\nu-1}$ and $S_\nu = H_{\nu-1} \cap C_\nu$ for $\nu = 2, \ldots, k$. If we use the convention $R_0 = S_2, R_1 = C_1 \setminus C_2, S_0 = \emptyset$ and $S_1 = S_2$ and Convention 2.1, then by Theorem 4.4 in Letac and Massam (2007), under the mapping $\Sigma^{-1} \mapsto \times_{\nu=0}^{k}\left(\Sigma_{R_\nu|S_\nu}, -\Sigma^{-1}_{S_\nu}\Sigma_{S_\nu R_\nu}\right)$, $W_{P_{\mathcal{G}}}(U, \eta)$ is transformed to a density proportional

to

$$\exp\left[-\frac{1}{2}\sum_{\nu=0}^{k}\left(\Sigma_{R_\nu|S_\nu}^{-1}(\Sigma_{S_\nu}^{-1}\Sigma_{S_\nu R_\nu}-U_{S_\nu}^{-1}U_{S_\nu R_\nu}\right)^\top U_{S_\nu}\left(\Sigma_{S_\nu}^{-1}\Sigma_{S_\nu R_\nu}-U_{S_\nu}^{-1}U_{S_\nu R_\nu}\right)\right]$$
$$\times\prod_{\nu=0}^{k}\exp\left(-\frac{1}{2}\Sigma_{R_\nu|S_\nu}^{-1}U_{R_\nu|S_\nu}\right)\Sigma_{R_\nu|S_\nu}^{-\frac{1}{2}\eta_\nu}. \tag{6}$$

This density has $k+1$ free shape parameters, one for each block $C_1 \setminus S_2$, S_2, $R_2, \ldots, R_k$. If we further split these blocks according to the parent ordering of the vertices in $\mathcal{D}$, then we can show that the density in Eq. (6) is transformed to a density on Θ_D that has the same form as in Eq. (5) and, consequently, Eq. (4), and in which each α_i is a function of η_ν's (therefore there are still $k+1$ free parameters). Note that the obtained density on $\Theta_\mathcal{D}$ retains the directed strong hyper Markov property of the $W_{P_\mathcal{G}}$ in Theorem 4.4 of Letac and Massam (2007), in the sense that

$$D_{ii} = \Sigma_{ii|pa(i)} \sim IG(\frac{\alpha_i}{2}-\frac{pa_i}{2}-1, \frac{1}{2}U_{ii|pa(i)}), \text{ and} \tag{7}$$
$$\Sigma_{pa(i)}^{-1}\Sigma_{pa(i),i} \mid D_{ii} \sim N_{pa_i}\left(U_{pa(i)}^{-1}U_{pa(i),i}, D_{ii}\,U_{pa(i)}^{-1}\right) \tag{8}$$

and $\{D_{ii},\ \Sigma_{pa(i)}^{-1}\Sigma_{pa(i)},\ i=1,\ldots,p\}$ are independent.

(c) Now let $\mathcal{D}$ be an arbitrary directed acyclic graph. In light of (a) and (b), to form our DAG-Wishart on $\Theta_\mathcal{D}$, we consider the density given in (5) but with all p parameters α_i being free, thus one shape parameter for each vertex of the graph, and take its image under the mapping $\times_{i\in V}\left(\lambda_i, \beta_{pa(i)}\right) \mapsto (D, L)$. We thus define the DAG-Wishart distribution $\pi_{U,\alpha}^{\Theta_\mathcal{D}}$ to be the distribution with density on the Cholesky space $\Theta_\mathcal{D}$ proportional to Eq. (4) and with normalizing constant that we are about to compute. We do so by multiple integration of the non-normalized density in Eq. (5) and take advantage of the strong directed hyper Markov property manifested by Eq. (7) and Eq. (8). The calculation yields:

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

for $(D, L) \in \Theta_\mathcal{D}$ and

$$z_\mathcal{D}(U,\alpha) = \prod_{i=1}^{p}\frac{\Gamma\left(\frac{\alpha_i}{2}-\frac{pa_i}{2}-1\right)2^{\frac{\alpha_i}{2}-1}(\sqrt{\pi})^{pa_i}\det(U_{pa(i)})^{\frac{\alpha_i}{2}-\frac{pa_i}{2}-\frac{3}{2}}}{\det(U_{fa(i)})^{\frac{\alpha_i}{2}-\frac{pa_i}{2}-1}}. \tag{9}$$

Note that $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ is a conjugate prior for $\mathcal{M}(\mathcal{D})$. More precisely, if the prior distribution on (D, L) is $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ and $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n$ is an independent and identically distributed sample from $N_p(\mathbf{0}, (L^\top)^{-1} D L^{-1})$, then the posterior distribution of (D, L) is given by $\pi_{\widetilde{U},\widetilde{\alpha}}^{\Theta_{\mathcal{D}}}$, where $S = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^\top$ denotes the empirical covariance matrix, $\widetilde{U} = nS + U$ and $\widetilde{\alpha} = (n + \alpha_1, n + \alpha_2, \cdots, n + \alpha_p)$. As mentioned above, $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ is strong directed hyper Markov with properties (7) and (8). We would like to emphasize here that a distribution of type Eq. (5) cannot be derived from the Type II Wishart distribution in Letac and Massam (2007) when $\mathcal{D}$ is arbitrary because $W_{P_{\mathcal{G}}}$ is derived as the natural exponential family generated by an appropriate measure on $P_{\mathcal{G}}$, a machinery which cannot be employed if directed acyclic graphs are not perfect. It should also be noted that Bayesian inference for models of the form given in (4), for a given order of the vertices, can be done using the Gibbs sampling approach in **?**. However, with that approach, the analytic expressions for the marginal likelihood is not available since, in that paper, there is no equivalent to (9).

Remark 3.1 *Note that, since the mapping $U \mapsto \pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ is not one-to-one, the parametrization of $\pi_{U,\alpha}^{\theta_{\mathcal{D}}}$ by $U \succ 0$ is not identifiable unless $\mathcal{D}$ is perfect. However, if the parameter set is restricted to $PD_{\mathcal{D}}$, then the parametrization is identifiable. As a parametrized model, $\{\pi_{U,\alpha}^{\Theta_{\mathcal{D}}} : U \in PD_{\mathcal{D}}\}$ is in general a curved exponential family for an arbitrary $\mathcal{D}$, and a natural exponential family if and only if $\mathcal{D}$ is perfect (see supplemental section 2.7 for details).*

Remark 3.2 *In supplemental section 2, we demonstrate that a particular sub-class of DAG-Wisharts yields the same prior on Markov equivalent directed acyclic graph models. In fact we show that this sub-class coincides with that of Geiger and Heckerman (2002), thus making a connection between the DAG-Wishart priors and those of Geiger and Heckerman.*

Remark 3.3 *As we have seen above, the shape parameters are added to the sample size in the posterior distribution and thus these shape parameters can be used to reflect the different degrees of confidence in the strength of the regression relationships. The regression interpretation and the multiple shape parameters allow for prior beliefs to be flexibly incorporated in the same manner as with standard regression.*

Remark 3.4 *We tabulate the properties of the various recently introduced Wishart distributions used in Gaussian graphical models in supplemental section 2.11.*

Let us now illustrate the functional form of the DAG-Wishart density on a specific directed acyclic graph.

Figure 1: Directed acyclic graph studied in example 4.1.

Ex 3.1 *Consider $\mathcal{D}$ given in Figure 1. The DAG-Wishart density $\pi_{U,\alpha}^{\Theta_\mathcal{D}}$ on the Cholesky space $\Theta_\mathcal{D}$ is given as follows:*

$$\begin{aligned}\pi_{U,\alpha}^{\Theta_\mathcal{D}}(D,L) &= \frac{1}{z_\mathcal{D}(U,\alpha)}\exp\left[-\frac{1}{2}D_{33}^{-1}((L_{31},L_{32})^\top - U_{pa(3)}^{-1}U_{pa(3),3})^\top U_{pa(3)}((L_{31},L_{32})^\top - U_{pa(3)}^{-1}U_{pa(3),3})\right] \\ &= \exp\left[D_{11}^{-1}U_{11}\right]D_{11}^{-\frac{1}{2}\alpha_2} \times \exp\left[D_{22}^{-1}U_{22}\right]D_{22}^{-\frac{1}{2}\alpha_2} \times \exp\left[D_{33}^{-1}U_{33|pa(3)}\right]D_{33}^{-\frac{1}{2}\alpha_3}.\end{aligned}$$

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

4.1 Motivation

In the previous section we introduced the DAG-Wishart distribution $\pi_{U,\alpha}^{\Theta_\mathcal{D}}$ on the Cholesky space $\Theta_\mathcal{D}$. We proceed now to define, for general directed acyclic graphs, an analog of the $W_{P_\mathcal{G}}$ and its inverse, the $IW_{P_\mathcal{G}}$, which are only defined for decomposable graphs. This is motivated by the fact that Gaussian distributions are more naturally parametrized over the covariance or precision matrices. Moreover the Wishart types distribution introduced by Letac and Massam (2007), similar to the classical Wishart or the hyper Wisharts, are defined on the space of covariance or precision matrices. Therefore we would also like to derive DAG-Wisharts for the covariance and precision matrices of $N(0,\Sigma) \in \mathcal{N}_(\mathcal{D})$. Formally this requires to derive the image of the $\pi_{U,\alpha}^{\Theta_\mathcal{D}}$ distribution under the mappings

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

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

From a purely mathematical or theoretical point of view, one can derive the corresponding densities on $P_\mathcal{D}$ and $PD_\mathcal{D}$ with respect to Hausdorff measure. But even for the simplest directed acyclic graphs, the Hausdorff density is not amenable to posterior analysis (see supplemental section 3 for a more detailed discussion of this approach). To overcome this problem, we follow what was done for the hyper inverse Wishart in Lauritzen (1996) or for the type I Wishart in Letac and Massam (2007) and we work with the projections of $P_\mathcal{D}$ and $PD_\mathcal{D}$ onto the Euclidean space that only retain the functionally independent elements of the precision and covariance matrices of Gaussian directed acyclic graph models. The projected spaces, as we shall precisely define later, are subsets of incomplete matrices, which

we call the spaces of incomplete precision and covariance matrices, respectively. The results in this section are also important from a practical point of view. Indeed, in Theorem 4.2 and Proposition 4.5, we derive the analytic expression of the expected value of the projection Ω^E and Σ^E of the precision and covariance matrices respectively. With these analytic expressions, we are able to compute the risk functions obtained with various estimates of Ω and Σ, the MLE and the Bayes estimator, and show the superior performance of our Bayes estimators in supplemental section 4.

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

For an element $\Omega \in P(\mathcal{D})$, let Ω^E denote $(\Omega_{ij} : (i,j) \in E^u)$, where E^u is the edge set of the undirected version of $\mathcal{D}$, and let $R_{\mathcal{D}}$ denote the set of all such Ω^E. We call Ω^E an incomplete precision matrix, where only the entries along the edges of $\mathcal{D}$ are specified, and $R_{\mathcal{D}}$ the space of incomplete precision matrices. Note that $R_{\mathcal{D}}$ is simply the image space of the projection mapping $\Omega \mapsto \Omega^E$. It follows from the definition that if Υ is an incomplete precision matrix in $R_{\mathcal{D}}$, then there is a precision matrix $\Omega \in P_{\mathcal{D}}$, called a positive definite completion of Υ in $P_{\mathcal{D}}$, such that $\Omega^E = \Upsilon$. In fact by Proposition 3.5 in Ben-David and Rajaratnam (2012), a positive definite completion of Υ in $P_{\mathcal{D}}$ is unique and can be explicitly computed in polynomial time. In other words, the mapping $\Omega \mapsto \Omega^E$ is a bijection and has an explicit inverse $\Upsilon \mapsto \Omega$. Hence the space of precision matrices $P_{\mathcal{D}}$ can be naturally identified with the space of incomplete precision matrices $R_{\mathcal{D}}$. Note that $R_{\mathcal{D}}$, unlike $P_{\mathcal{D}}$, is open in its affine support. Now let $\pi_{U,\alpha}^{R_{\mathcal{D}}}$ denote the image of $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ under the mapping

$$\psi \equiv \left((L,D) \mapsto \left(LD^{-1}L^{\top}\right)^E\right) : \Theta_{\mathcal{D}} \to R_{\mathcal{D}}. \tag{12}$$

Since $R_{\mathcal{D}}$ is open in its affine support, the distribution $\pi_{U,\alpha}^{R_{\mathcal{D}}}$ has a density with respect to the Lebesgue measure on $R_{\mathcal{D}}$. Thus in light of the bijection $\Omega \mapsto \Omega^E$, in both a natural and practical sense, we define $\pi_{U,\alpha}^{R_{\mathcal{D}}}$ as the DAG-Wishart distribution on the space of incomplete precision matrices $R_{\mathcal{D}}$. To derive the density of $\pi_{U,\alpha}^{R_{\mathcal{D}}}$ we need to compute the Jacobian of the mapping ψ in Eq. (12). The Jacobian of ψ is a variant of similar transformations found in Roverato (2000); Khare and Rajaratnam (2011). For completeness we still compute this Jacobian in the following lemma. The proof is given in supplemental section 2.7.

Lemma 4.1 *(Roverato, 2000; Khare and Rajaratnam, 2011) 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 $\pi_{U,\alpha}^{R_{\mathcal{D}}}$ and some of its properties. The proofs are immediate results of Lemma 4.1 and the iterative construction of $\pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$.

Theorem 4.2 *Let Ω^E be the image of $(L, D) \sim \pi_{U,\alpha}^{\Theta_{\mathcal{D}}}$ under the mapping ψ. Then*

a) The density of $\Omega^E \sim \pi_{U,\alpha}^{R_{\mathcal{D}}}$ with respect to the standard Lebesgue measure on $R_{\mathcal{D}}$ is given by

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

where Ω is the unique positive completion of Ω^E in $P_{\mathcal{D}}$, $D_{ii} = (\Omega^{-1})_{ii|pa(i)}$, which is explicitly a function of Ω^E, and $z_{\mathcal{D}}(U,\alpha)$ as defined in Eq. (9)*.*

b) The Laplace transform of $\pi_{U,\alpha}^{R_{\mathcal{D}}}$ at $K^E \in R_{\mathcal{D}}$ is given by $\mathcal{L}_{R_{\mathcal{D}}}(K^E) = \dfrac{z_{\mathcal{D}}(2K+U,\alpha)}{z_{\mathcal{D}}(U,\alpha)}$ where K is the completion of K^E in $P_{\mathcal{D}}$.

c) $\mathbb{E}\left(\Omega^E\right) = \left(\sum_{j=1}^{p}(\alpha_j - pa_j - 2)\left(U_{fa(j)}^{-1}\right)^0 - \sum_{j=1}^{p}(\alpha_j - pa_j - 3)\left(U_{pa(j)}^{-1}\right)^0\right)^E$, where $\left(U_{fa(j)}^{-1}\right)^0$ and $\left(U_{pa(j)}^{-1}\right)^0$ denote the $|V| \times |V|$ matrices obtained by inserting zeros in the positions not in $fa(j)$ and $pa(j)$, respectively.

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

In this subsection, we shall define the distribution that corresponds to the hyper-inverse Wishart or more generally the inverse Type II Wishart $IW_{P_{\mathcal{G}}}$. First we introduce the space of incomplete covariance matrices. For a covariance matrix $\Sigma \in PD_{\mathcal{D}}$, the incomplete covariance matrix Σ^E is $(\Sigma_{ij} : (i,j) \in E^u)$, and the space of incomplete covariance matrices, denoted by $S_{\mathcal{D}}$ is the set of all Σ^E such that $\Sigma \in PD_{\mathcal{D}}$. Here we note that $S_{\mathcal{D}}$ is the projection space of the projection mapping $\Sigma \mapsto \Sigma^E$. By Proposition 3.6 in Ben-David and Rajaratnam (2012) for each $\Gamma \in PD_{\mathcal{D}}$ there a unique and explicitly (polynomial-time) computable positive definite matrix in $\Sigma \in PD_{\mathcal{D}}$ such that $\Sigma^E = \Gamma$. Note that Σ is simply the positive definite completion of Γ in $PD_{\mathcal{D}}$. Now since the mapping $\Sigma \mapsto \Sigma^E$ is a bijection, we can identify the space of covariance matrices $PD_{\mathcal{D}}$ with the space of incomplete covariance matrices $S_{\mathcal{D}}$.

Let $\pi_{U,\alpha}^{S_\mathcal{D}}$ denote the image of $\pi_{U,\alpha}^{\Theta_\mathcal{D}}$ under the mapping $(D, L) \mapsto (L^{-\top} D L^{-1})^E : \Theta_\mathcal{D} \to S_\mathcal{D}$, where $L^{-\top} = \left(L^\top\right)^{-1}$. In parallel to our notation $\pi_{U,\alpha}^{R_\mathcal{D}}$, we will denote the inverse DAG-Wishart distribution on the space of incomplete covariance matrices $S_\mathcal{D}$ as $\pi_{U,\alpha}^{S_\mathcal{D}}$. Next we shall derive the density of this distribution with respect to the Lebesgue measure. First we compute the Jacobian of the mapping $(\Sigma^E \mapsto \Sigma^{-E}) : S_\mathcal{D} \to R_\mathcal{D}$, where $\Sigma^{-E} = (\Sigma^{-1})^E$ and Σ is the completion of Σ^E in $PD_\mathcal{D}$ (see supplemental section 2.8 for proof).

Lemma 4.3 *The Jacobian of the mapping $(\Sigma^{-E} \mapsto \Sigma^E) : R_\mathcal{D} \to S_\mathcal{D}$ is given by $\prod_{i=1}^{p} \frac{\det \Sigma_{fa(i)}^{pa_i+2}}{\det \Sigma_{pa(i)}^{pa_i+1}}$.*

Ex 4.1 *Consider $\mathcal{D}$ given in Figure 1. Then the inverse DAG-Wishart on $\mathcal{D}$ is given by*

$$\pi_{U,\alpha}^{S_\mathcal{D}}(\Sigma^E) = z_\mathcal{D}(U,\alpha)^{-1} \exp\left[-\frac{1}{2} tr(\Sigma^{-1} U)\right] D_{11}^{-\frac{\alpha_1}{2}} D_{22}^{-\frac{\alpha_2+2}{2}} D_{33}^{-\frac{\alpha_3+2}{2}},$$

where Σ, the completion of Σ^E, is obtained by setting $\Sigma_{23} = 0$ since $X_2 \perp\!\!\!\perp X_3$.

Remark 4.1 *We remind the reader that for a decomposable graph $\mathcal{G}$ the $IW_{P_\mathcal{G}}$ in Letac and Massam (2007) is a variant of $\pi_{U,\alpha}^{S_\mathcal{D}}$ for a perfect directed acyclic graph version of $\mathcal{G}$. Furthermore, in the setting of Gaussian covariance graph models, the inverse Wishart distribution introduced by Khare and Rajaratnam (2011) for a homogeneous graph $\mathcal{G}$ is an equivalent form of $\pi_{U,\alpha}^{S_\mathcal{D}}$ for a transitive and perfect version $\mathcal{D}$ of $\mathcal{G}$. The proof of this result is rather technical and is given in supplemental section 2.9.*

4.4 Properties of the inverse DAG-Wishart distributions

One of the main useful property of the inverse DAG-Wishart $\pi_{U,\alpha}^{S_\mathcal{D}}$ for an arbitrary $\mathcal{D}$ is its strong directed hyper Markov property. As discussed in section 3.1, this follows directly from Theorem 4.4 in Letac and Massam (2007) but is generalized to arbitrary directed acyclic graphs. The precise statement of the strong directed hyper Markov property for $\pi_{U,\alpha}^{S_\mathcal{D}}$ is as follows.

Theorem 4.4 *If $\Sigma^E \sim \pi_{U,\alpha}^{S_\mathcal{D}}$, then*

$i)$ $\left((\Sigma_{ii|pa(i)}, \Sigma_{pa(i)}^{-1}\Sigma_{pa(i),i} : i \in V\right)$ are mutually independent and therefore $\pi_{U,\alpha}^{S_\mathcal{D}}$ has strong directed hyper Markov property.

$ii)$ The distribution of $\Sigma_{ii|pa(i)}$ and $\Sigma_{pa(i)}^{-1}\Sigma_{pa(i),i} \mid \Sigma_{ii|pa(i)}$ are as in Eq. (7) and Eq. (8) of section 3.

We can also evaluate the expected value under $\pi_{U,\alpha}^{S_{\mathcal{D}}}$. The process for computing this quantity, given in the following proposition, is the exact equivalent of Theorem 3.1 in Rajaratnam et al. (2008) but now generalized so it is applicable to any directed acyclic graphs.

Proposition 4.5 *Suppose* $\Sigma^E \sim \pi_{U,\alpha}^{S_{\mathcal{D}}}$, *with* $\alpha > pa_i + 4$. *Then the expected value of* Σ^E *can be recursively computed by the following steps for* $i = p-1, p-2, \ldots, 1$.

$(i)\ \mathbb{E}\left(\Sigma_{pp}\right) = \dfrac{U_{pp}}{\alpha_p - 4}$, $(ii)\ \mathbb{E}\left(\Sigma_{pa(i),i}\right) = -\mathbb{E}\left(\Sigma_{pa(i)}\right) U_{pa(i)}^{-1} U_{pa(i),i}$ *and*

$$(iii)\ \mathbb{E}\left(\Sigma_{ii}\right) = \frac{U_{ii|pa(i)}}{\alpha_i - pa_i - 4} + tr\left[\mathbb{E}\left(\Sigma_{pa(i)}\right)\left(\frac{U_{ii|pa(i)}\, U_{pa(i)}^{-1}}{\alpha_i - pa_i - 4} + U_{pa(i)}^{-1} U_{pa(i),i} U_{i,pa(i)}\, U_{pa(i)}^{-1}\right)\right].$$

5 Simulation study and Applications to real data

We will now illustrate the use of our DAG-Wishart distributions by applying them to two problems in modern high dimensional statistical inference: Bayesian model selection in the space of Gaussian directed acyclic graph models with a given order of the vertices and parameter estimation. Our Bayesian model selection method based on the DAG-Wishart prior uses a closed form marginal likelihood, and to our knowledge it thus is more scalable than previous Bayesian approaches (in our examples, we illustrate the model selection of graphs with as high as $p = 2000$). Our estimation of the covariance and precision matrices corresponding to Gaussian directed acyclic graph models uses the closed form solutions for the estimates of the precision and covariance matrices due to Proposition 4.5 and the conjugacy of the DAG-Wishart. The ease of implementation and scalability for model selection is illustrated using simulated data and also the real molecular network data.

5.1 Bayesian model selection via DAG-Wishart prior

In many applications, the graph structure is unknown beforehand and estimating an underlying graph is an important contemporary problem. In this section, we illustrate how to apply the DAG-Wishart priors to model selection problems.

Assuming a uniform prior on the space of all graphs on p vertices, we want to compute the marginal likelihood $p(X \mid \mathcal{D}) = \int f(X \mid \Sigma, \mathcal{D})\pi(\Sigma \mid \mathcal{D})d\Sigma$ for each model $\mathcal{D}$. The marginal likelihood can be computed in closed form for our flexible DAG-Wishart priors. In Bayesian model selection, the goal is to identify the $\mathcal{D}$ that gives the largest marginal likelihood. Such procedure requires enumerating all $2^{\binom{p}{2}}$ graphs and is NP hard. Instead, we propose a powerful stochastic searching strategy to identify the "best" graph with high accuracy and in polynomial time. Our approach is an improved version of the stochastic shot-gun search (SSS) of Jones et al. (2005) coupled with the LassoDAG method in Shojaie and Michailidis (2010). Our model selection algorithm, DAG-W, is specified below:

Algorithm 5.1 (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) *Randomly select N_1 graphs that are one edge away from $\mathcal{D}_0$. Evaluate the* log *posterior scores $s_1, \cdots, s_{N_1}$ for each of these graphs, according to the DAG-Wishart prior/posterior. Record all of these graphs and scores as a list $\mathcal{L}^{(k)}$.*
   (b) *Sample the next graph from the current graph list with probability $p_i \propto \exp(s_i)^\gamma$, where γ is an annealing parameter. Take the sampled graph $\mathcal{D}_{new}$ as $\mathcal{D}_0$.*
   (c) *Return to Step 1-(a).*
2. *Collect/Assemble all the $\mathcal{L}^{(k)}, k = 1, \cdots, N$.*
3. *Return the graph with the largest score as the selected model.*

The intuition behind the algorithm DAG-W is as follows: though LassoDAG does not in general yield the $\mathcal{D}$ with the highest marginal likelihood, the model selection path of LassoDAG should explore the model space effectively. Therefore instead of evaluating the marginal likelihood of all possible graphs, one can narrow the searching area to the graphs close to the LassoDAG path. As indicated in the algorithm above, we take the various models corresponding to different penalty parameter values on the LassoDAG regularization path as starting points of our model search. Supplemental section 4.1 includes our numerical evaluation of DAG-W in the setting of $p = 7$. In this setting, we show that our algorithm is able to identify the highest marginal likelihood graph with very high probability (≥ 0.99) by only searching a negligible proportion of the full model space. Thus our algorithm is a computationally feasible solver of the NP hard problem with very high accuracy. In Shojaie and Michailidis (2010), the penalty parameter τ_i for the Lasso problem of node i is set to

$$\tau_i = 2\frac{Z^*_{\frac{\kappa}{2p(i-1)}}}{\sqrt{n}}, \tag{13}$$

where in general Z^*_q denotes the $(1-q)$th quantile of standard normal distribution and $\kappa = 0.1$ is the recommended value in Shojaie and Michailidis (2010). Here we use the same setup as in Shojaie and Michailidis (2010) to evaluate and compare the performance of the LassoDAG to our DAG-W algorithm. More details are as follows.

The scale parameter U of the DAG-Wishart is taken to be the identity matrix. We constrain the shape parameters to be $c \cdot pa_i + b$ such that $c \cdot pa_i + b > pa_i + 2$. In particular, we take $b = 3, c = 1$ in model selection as this seems to give reasonably good model selection results in all of our evaluation tasks (with different p, n and sparsity). We set $N = 16$ in Algorithm 5.1 for our Bayesian model selection: 15 initial states were chosen by taking $\kappa = (k/15)^4 p, k = 1 \cdots 15$ in Eq. (13) and the sixteenth state was selected using the LassoDAG recommendation[1] $\kappa = 0.1$. Furthermore, we take $M = 100$, $N_1 = 30$ and $\gamma = 0.5$.

The data is generated by the random directed acyclic graph generator in the R-package `pcalg` ((Kalisch et al., 2012; Hauser and Bühlmann, 2012)). In our evaluation, we specify the edge proportion (sparsity) to be 0.01 in generating the directed acyclic graph and the edge regression weights are uniformly sampled between 0.2 and 0.8. The reader is referred to `pcalg` documentation for details about the model generating procedure. Fixing $n = 100$, we check the model selection performance when the edge proportion is 0.01 and $p = 50$, 100, 200, 500, 1000, 1500 and 2000[2]. The performance is measured by two competing measurements: **sensitivity** and **specificity**, which are frequently used in model selection tasks (see Baldi et al. (2000)). Sensitivity is used to measure the proportion of true edges discovered while specificity is used to measure the proportion of the null edges that are correctly excluded.

Table 1 shows the performance comparison between the Lasso-DAG and DAG-W. Both methods are able to retain very good specificity. The DAG-W gives much better sensitivity with only slightly lower specificity. When p is large, the improvement in sensitivity is more noticeable. In the case of $p = 2000$, the sensitivity given by the DAG-W is more than twice that given by the LassoDAG. One of the main advantages of the DAG-W is in the area of high dimensional biological applications. In such applications gene discoveries which are reliable are important, especially since the gain in sensitivity comes at negligible loss in specificity.

We also demonstrate the efficiency of our DAG-Wishart approach in the context of parameter estimation in high dimensions in supplemental section 4.3. More specifically, we investigate decision theoretic based estimation using two losses, namely Stein's loss and quadratic loss. We consider estimation of both Σ and Ω. The Bayes estimator correspond-

[1] In Shojaie and Michailidis (2010), κ can be used to measure false positive control thus it should be less or equal to 1. Here we do not respect this constraint as our choice turns out to search the model space much better according to our evaluation.

[2] To make it computationally feasible for model selection in such high dimensions, we decrease N from 16 to 9 for problems with $p \geq 500$. And for each initialization points, we only search at most 50 steps ($M = 50$).

	LassoDAG		DAG-W		
p	Sensitivity	Specificity	Sensitivity	Specificity	Timing (min.)
50	0.616	1.000	0.783	0.998	1.21
100	0.483	~ 1	0.752	0.998	2.99
200	0.397	~ 1	0.741	0.998	8.53
500	0.250	~ 1	0.652	0.998	21.90
1000	0.175	0.999	0.425	0.997	66.39
1500	0.123	0.998	0.267	0.996	109.16
2000	0.099	0.997	0.194	0.994	280.56

Table 1: Average performance measurements for different p, when $n = 100$ and edge proportion equal to 0.01. The last column is the average CPU time for one replication of model selection by DAG-W, on Intel Xeon CPU E5606@2.13GHz.

ing to these two loss functions together with the MAP estimator yields three estimators each for the covariance and precision matrix. Our extensive numerical investigations in this regard indicate that estimation using our DAG-Wishart priors yield substantial risk reductions over that of the MLE. We also give guidelines and details regarding choice of hyperparameters, robustness of the DAG-Wishart approach to outliers and the role of sparsity of the graph in estimation performance.

5.2 Real data application: molecular network estimation

In this section, we test our model selection method on the data set of Sachs et al. (2005) which contains $p = 11$ proteins and phospholipids measurements on $n = 7466$ cells. This data set was also used in Shojaie and Michailidis (2010) and Friedman et al. (2008). A directed acyclic graph was established in Sachs et al. (2005) and will be assumed to be the true graph for our purposes. Furthermore, we shall use the established parent order in the following model selection investigation.

The estimated graphs are shown in Figure 2. The blue edges are the correctly discovered ones and the red edges are false discoveries. Again, we set $\kappa = 0.1$ for the LassoDAG and $b = 3, c = 1$ for the DAG-W. LassoDAG gives 78.95% sensitivity with 52.78% specificity, while DAG-W gives 94.74% sensitivity with 47.22% specificity. So DAG-W gains a 15% increase in sensitivity by sacrificing 5% of specificity. Both of the estimations are denser than the one reported by Sachs et al. (2005). Comparing the discoveries of the two models: all of the 15 true discoveries from LassoDAG are also included in the discoveries of DAG-W. The three additional true positive edges from DAG-W are edges

$PKA \to MEK$, $PKA \to P38$ and $PKC \to MEK$. So if the goal is to discover potential associations for future laboratory research, DAG-W is a better choice, since it includes all the discoveries of LassoDAG as a subset, and also finds three other true edges, at the price of two more false discoveries. According to Sachs et al. (2005), the mechanism of edge $PKA \to MEK$ is possibly due to the true molecular influence path $PKA \to Raf_{s621} \to MEK$. Edge $PKA \to P38$ is possibly due to the true molecular influence path $PKA \to MKKs \to P38$. Molecules Raf_{s621} and $MKKs$ however are not measured in the data. Thus the success in detecting indirect influences demonstrates the better sensitivity of DAG-W. On the other hand, there are two distinct influence paths from PKC to MEK, that is, $PKC \to MEK$ and $PKC \to RAF \to MEK$. LassoDAG only detects the latter, which is possibly because the edge effect of $PKC \to RAF \to MEK$ masked that of $PKC \to MEK$. In DAG-W, we are able to discover both of the edges due to better detection sensitivity. We also evaluate our model selection and covariance

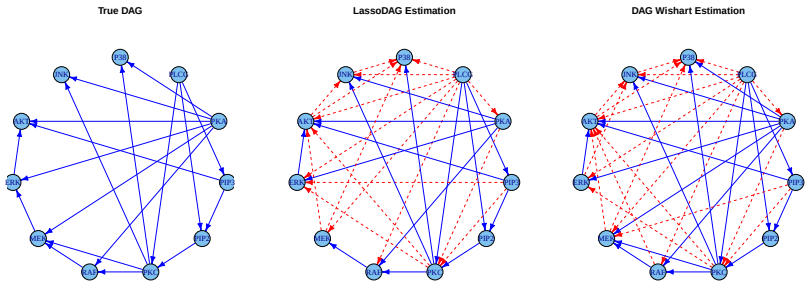

Figure 2: The estimated graphs are compared with the "true" network discovered by Sachs et al. (2005). The solid edges are the correctly discovered ones and the dashed edges are false discoveries.

estimation procedures on the call center data used in Bickel and Levina (2008) and Rajaratnam et al. (2008). The DAG-Wishart model has better performance on that task as well. More details about this example can be found in supplemental section 4.5.

6 Supplementary material

A supplemental section is available at *Biometrika* online and includes some background on directed acyclic graph Markov models, theoretical properties of our new DAG-Wishart

distribution as well as the details of our computational algorithms and results for model selection and estimation.

References

Andersson, S. A. & Perlman, M. D. (1998). Normal linear regression models with recursive graphical Markov structure. *J. Multivariate Anal.*, 66(2):133–187.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424.

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

Ben-David, E. & Rajaratnam, B. (2014). The Letac-Massam conjecture and existence of high dimensional Bayes estimators for Graphical Models. *ArXiv e-prints*.

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

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

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

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

Geiger, D. & Heckerman, D. (2013). Learning gaussian networks. *CoRR*, abs/1302.6808.

Hauser, A. & Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.*, 13:2409–2464.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., & West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.*, 20(4):388–400.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

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

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

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

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

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

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

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