Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions

Dan Geiger and David Heckerman
Microsoft Research, Bldg 31
Redmond 98052-6399, WA
dang@cs.technion.ac.il *, heckerma@microsoft.com

Abstract

We show that the only parameter prior for complete Gaussian DAG models that satisfies global parameter independence, complete model equivalence, and some weak regularity assumptions, is the normal-Wishart distribution. Our analysis is based on the following new characterization of the Wishart distribution: let W be an $n \times n$, $n \geq 3$, positive-definite symmetric matrix of random variables and $f(W)$ be a pdf of W. Then, f(W) is a Wishart distribution if and only if $W_{11} - W_{12}W_{22}^{-1}W'_{12}$ is independent of $\{W_{12}, W_{22}\}$ for every block partitioning $W_{11}, W_{12}, W'_{12}, W_{22}$ of W. Similar characterizations of the normal and normal-Wishart distributions are provided as well. We also show how to construct a prior for every DAG model over **X** from the prior of a single regression model.

1   Introduction

Directed Acyclic Graphical (DAG) models have increasing number of applications in Statistics (Spiegelhalter, Dawid, Lauritzen, and Cowell, 1993) as well as in Decision Analysis and Artificial Intelligence (Heckerman, Mamdani, Wellman, 1995b; Howard and Matheson, 1981; Pearl, 1988). A DAG model $m = (s, \mathcal{F}_s)$ for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ each associated with a set of possible values D_i, respectively, is a set of joint probability distributions for $D_1 \times \cdots \times D_n$ specified via two components: a structure s and a set of local distribution families $\mathcal{F}_s$. The structure s for **X** is a directed graph with no directed cycles (i.e., a Directed Acyclic Graph) having for every variable X_i in **X** a node labeled X_i with parents labeled by $\mathbf{Pa}_i^m$. The structure s represents the set of conditional independence assertions, and only these conditional independence assertions, which are implied by a factorization of a joint distribution for **X** given by $p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\mathbf{pa}_i^m)$, where **x** is a value for **X** (an

n-tuple) and x_i is a value for X_i. When x_i has no incoming arcs in m (no parents), $p(x_i|\mathbf{pa}_i^m)$ stands for $p(x_i)$. The local distributions are the n conditional and marginal probability distributions that constitute the factorization of $p(\mathbf{x})$. Each such distribution belongs to the specified family of allowable probability distributions $\mathcal{F}_s$. A DAG model is often called a *Bayesian network*, although the later name sometimes refers to a specific joint probability distribution that factorizes according to a DAG, and not, as we mean herein, a set of joint distributions each factorizing acccording to the same DAG. A DAG model is *complete* if it has no missing arcs. Note that any two complete DAG models for **X** encode the same assertions of conditional independence, namely none.

In this paper, we assume that each local distribution is selected from a family $\mathcal{F}_s$ which depends on a finite set of parameters $\theta_m \in \Theta_m$ (a parametric family). The parameters for a local distribution is a set of real numbers that completely determine the functional form of $p(x_i|\mathbf{pa}_i^m)$ when x_i has parents and of $p(x_i)$ when x_i has no parents. We denote by m^h the model hypothesis that the true joint probability distribution of **X** is perfectly represented by a structure s of a DAG model m with local distributions from $\mathcal{F}_s$, namely, that the joint probability distribution satisfies only the conditional independence assertions implied by this factorization and none other. Consequently, the true joint distribution for a DAG model m is given by,

$$p(\mathbf{x}|\theta_m, m^h) = \prod_{i=1}^{n} p(x_i|\mathbf{pa}_i^m, \theta_i, m^h) \qquad (1)$$

where $\mathbf{y} = \{x_i\}_{X_i \in \mathbf{Y}}$ denotes a value of $\mathbf{Y} \subseteq \mathbf{X}$ and $\theta_1, \ldots \theta_n$ are subsets of θ_m. Whereas in a general formulation of DAG models, the subsets $\{\theta_i\}_{i=1}^{n}$ could possibly overlap allowing several local distribution to have common parameters, in this paper, we shall shortly exclude this possibility (Assumption 5). Note that θ_m denotes the union of $\theta_1, \ldots, \theta_n$ for a DAG model m.

We consider the Bayesian approach when the parameters θ_m and the model hypothesis m^h are uncertain but the parametric families are known. Given data $d = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, a random sample from $p(\mathbf{x}|\theta_m, m^h)$ where θ_m and m^h are the true parameters and model

hypothesis, respectively, we can compute the posterior probability of a model hypothesis m^h using

$$p(m^h|d) = c\, p(m^h)\, p(d|m^h) = \qquad (2)$$

$$c\, p(m^h) \int p(d|\theta_m, m^h)\, p(\theta_m|m^h)\, d\theta_m$$

where c is a normalization constant. We can then select a DAG model that has a high posterior probability or average several good models for prediction.

The problem of selecting an appropriate DAG model, or sets of DAG models, given data, posses a serious computational challenge, because the number of DAG models grows faster than exponential in n. Methods for searching through the space of model structures are discussed (e.g.) by Cooper and Herskovits (1992), Heckerman, Geiger, and Chickering (1995a), and Friedman and Goldszmidt (1997).

From a statistical viewpoint, an important question which needs to be addressed is how to specify the quantities $p(m^h)$, $p(d|\theta_m, m^h)$, $p(\theta_m|m^h)$, needed for evaluating $p(m^h|d)$ for every DAG model m that could conceivably be considered by a search algorithm. Buntine (1994) and Heckerman et al. (1995a) discuss methods for specifying the priors $p(m^h)$ via a small number of direct assessments. Geiger and Heckerman (1994) and Heckerman and Geiger (1995) develop practical methods for assigning parameter priors $p(\theta_m|m^h)$ to every candidate DAG model m via a small number of direct assessments. Another relevant paper is by Dawid and Lauritzen (1993) who discuss the notion of hyper and meta Markov laws.

The contributions of this paper are twofold: A methodology for specifying parameter priors for Gausian DAG models using a prior for a single regression model (Section 2). An analysis of complete Gaussian DAG models which shows that the only parameter prior that satisfies our assumptions is the normal-Wishart distribution (Section 3).

The analysis is based on the following new characterization of the Wishart, normal, and normal-Wishart distributions.

Theorem *Let W be an $n \times n$, $n \geq 3$, positive-definite symmetric matrix of real random variables such that no entry in W is zero, μ be a an n-dimensional vector of random variables, $f_W(W)$ be a pdf of W, $f_\mu(\mu)$ be a pdf of μ, and $f_{\mu,W}(\mu, W)$ be a pdf of $\{\mu, W\}$. Then, $f_W(W)$ is a Wishart distribution, $f_\mu(\mu)$ is a normal distribution, and $f_{\mu,W}(\mu, W)$ is a normal-Wishart distribution if and only if global parameter independence holds for unknown W, unknown μ, or unknown $\{\mu, W\}$, respectively.*

The assumption of global parameter independence is expressed differently for each of the three cases treated by this theorem and the proof follows from Theorems 6, 8 and 9, respectively, proven in Section 3. It should be noted that a single principle, global parameter independence, is used to characterize three different distributions. In Section 4, we compare these characterizations to a recent characterization of the Dirichlet distribution (Geiger and Heckerman, 1997; Járai , 1998) and conjecture that the later characterization uses a redundant assumption (local parameter independence)—that is, global parameter independence may also characterize the Dirichlet distribution. The Dirichlet, normal, Wishart, and normal-Wishart distributions are the conjugate distributions for the standard multivariate exponential families.

2   Priors for DAG models

In this section we provide a novel presentation of our previous results in (Geiger and Heckerman, 1994; Heckerman and Geiger, 1995). We have sharpenned the assumptions involved in learning DAG models with no hidden variables from complete data. As a result, we show that a prior for one regression model dictates, under our assumptions, the prior for all Gaussian DAG models over the same variables. Our new presentation, which uses matrix notation for expressing independence of parameters of Gaussian DAG models, enables us to prove the characterization theorems in the next section.

This section is organized as follows: A methodology for specifying parameter priors for many structures using a few direct assessments (Section 2.1). A formula that computes the marginal likelihood for every dag model (Section 2.2). A specialization of this formula to an efficient computation for Gaussian DAG models (Section 2.3).

2.1   The Construction of Parameter Priors

We start by presenting a set of assumptions that simplify the assessment of parameter priors and a method of assessing these priors. The assumptions are as follows:

Assumption 1 (Complete model equivalence)
Let $m_1 = (s_1, \mathcal{F}_{s1})$ be a complete DAG model for a set of variables $\mathbf{X}$. The family $\mathcal{F}_{s2}$ of every complete DAG model $m_2 = (s_2, \mathcal{F}_{s2})$ for $\mathbf{X}$ is such that m_1 and m_2 represent the same set of joint probability distributions.

We explain this assumption by providing an example where it fails. Suppose the set of variables $\mathbf{X} = \{X_1, X_2, X_3\}$ consists of three variables each with possible values $\{x_i, \overline{x}_i\}$, respectively, and s_1 is the complete structure with arcs $X_1 \rightarrow X_2$, $X_1 \rightarrow X_3$, and $X_2 \rightarrow X_3$. Suppose further, that the local distributions $\mathcal{F}_{s1}$ of model m_1 are restricted to the sigmoid function

$$p(x_i|\mathbf{pa}_i^m, \theta_i, m^h) = \frac{1}{1 + \exp\left\{a_i + \sum_{x_j \in \mathbf{pa}_i^m} b_{ji} x_j\right\}}$$

where $\theta_1 = \{a_1\}$, $\theta_2 = \{a_2, b_{12}\}$, and $\theta_3 = \{a_3, b_{13}, b_{23}\}$.

Consider now a second complete model m_2 for $\mathbf{X} = \{X_1, X_2, X_3\}$ whose structure consists of the arcs

$X_1 \to X_2$, $X_1 \to X_3$, and $X_3 \to X_2$. Assumption 1 asserts that the families of local distributions for m_1 and m_2 are such that the set of joint distributions for **X** represented by these two complete models is the same. In this example, however, if we specify the local families for m_2 by also restricting them to be sigmoid functions, the two models will represent different sets of joint distributions over $\{X_1, X_2, X_3\}$. Hence, Assumption 1 will be violated. Using Bayes rule one can always determine a set of local distribution families that will satisfy Assumption 1, however, their functional form will usually involve an integral (and will often violate Assumption 5 below). An important exception is discussed in Section 2.3.

Our definition of m^h, that the true joint pdf of a set of variables **X** is perfectly represented by m, and Assumption 1, which says that two complete models represent the same set of joint pdfs for **X**, imply that for two complete models $m_1^h = m_2^h$. This is a strong assumption. It implies that $p(\theta_{m2}|m_2^h) = p(\theta_{m2}|m_1^h)$ because two complete models represent the same set of distributions. It also implies $p(d|m_1^h) = p(d|m_2^h)$ which says that the marginal likelihood for two complete DAG models is the same for every data set, or equivalently, that complete DAG models cannot be distinguished by data. Obviousely, in the example with the sigmoid functions, the two models can be distinguished by data because they do not represent the same set of joint distributions. [1]

Assumption 2 (Regularity) *For every two complete DAG models m_1 and m_2 for **X** there exists a one-to-one mapping f_{12} between the parameters θ_{m1} of m_1 and the parameters θ_{m2} of m_2 such that the likelihoods satisfy $p(\mathbf{x}|\theta_{m1}, m_1^h) = p(\mathbf{x}|\theta_{m2}, m_2^h)$ where $\theta_{m2} = f_{1,2}(\theta_{m1})$. The Jacobian $|\partial\theta_{m1}/\partial\theta_{m2}|$ exists and is non-zero for all values of Θ_{m1}.*

Assumption 2 implies $p(\theta_{m2}|m_1^h) = \left|\frac{\partial\theta_{m1}}{\partial\theta_{m2}}\right| p(\theta_{m1}|m_1^h)$ where $\theta_{m2} = f_{1,2}(\theta_{m1})$. Furthermore, due to Assumption 1, $p(\theta_{m2}|m_2^h) = p(\theta_{m2}|m_1^h)$, and thus

$$p(\theta_{m2}|m_2^h) = \left|\frac{\partial\theta_{m1}}{\partial\theta_{m2}}\right| p(\theta_{m1}|m_1^h). \qquad (3)$$

---

[1] A technical point worth mentioning here is our use of the term variable and its relationship to the standard definition of a random variable. A *continuous random variable* X, according to most probability text books, is a function $X : \Omega \to R$ such that $\{w|X(w) \le x\} \in \mathcal{A}$ where $\mathcal{A}$ is a σ-field of subsets of Ω and Ω is a sample space of a probability space $(\Omega, \mathcal{A}, P)$ and where P is a fixed probability measure. A *discrete random variable* is a function $X : \Omega \to D$ where D is a discrete set such that $\{w|X(w) = x_i\} \in \mathcal{A}$ for every $x_i \in D$ where $\mathcal{A}$ is a σ-field and Ω is a sample space of a probability space $(\Omega, \mathcal{A}, P)$. We use the term *variable*, as common to much of the literature on DAG models, to mean a function $X_i : A \to D_i$, where A is a σ-field of subsets of Ω, parallel to the usual definition of a random variable, but without fixing a specific probability measure P. A model m for a set of variables **X**, (and a DAG model in particular), is simply a set of probability measures on the Cartesian product $\times_i D_i$. Once a particular probability measure from m is picked, a variable in our sense becomes a random variable in the usual sense.

Assumption 3 (Likelihood Modularity) *For every two DAG models m_1 and m_2 for **X** such that X_i has the same parents in m_1 and m_2, the local distributions for x_i in both models are the same, namely, $p(x_i|\mathbf{pa}_i^m, \theta_i, m_1^h) = p(x_i|\mathbf{pa}_i^m, \theta_i, m_2^h)$ for all $X_i \in \mathbf{X}$.*

Assumption 4 (Prior Modularity) *For every two DAG models m_1 and m_2 for **X** such that X_i has the same parents in m_1 and m_2, $p(\theta_i|m_1^h) = p(\theta_i|m_2^h)$.*

Assumption 5 (Global Parameter Independence) *For every DAG model m for **X**, $p(\theta_m|m^h) = \prod_{i=1}^n p(\theta_i|m^h)$.*

The likelihood and prior modularity assumptions have been used implicitly in the work of (e.g.) Cooper and Herskovits (1992), Spiegelhalter et al. (1993), and Buntine (1994). Heckerman et al. (1995a) made Assumption 4 explicit in the context of discrete variables under the name parameter modularity. Spiegelhalter and Lauritzen (1990) introduced Assumption 5 in the context of DAG models under the name global independence. Assumption 5 excludes the possibility that two local distributions would share a common parameter.

The assumptions we have made lead to the following significant implication: When we specify a parameter prior $p(\theta_{mc}|m_c^h)$ for one complete DAG model m_c, we also implicitly specify a prior $p(\theta_m|m^h)$ for any DAG model m among the super exponentially many possible DAG models. Consequently, we have a framework in which a manageable number of direct assessments leads to all the priors needed to search the model space. In the rest of this section, we explicate how all parameter priors are determined by the one elicited prior. In Section 2.3, we show how to elicit the one needed prior $p(\theta_{mc}|m_c^h)$ under specific distributional assumptions.

Due to the complete model equivalence and regularity assumptions, we can compute $p(\theta_{mc}|m_c^h)$ for one complete model for **X** from the prior of another complete model for **X**. In so doing, we are merely performing coordinate transformations between parameters for different variable orderings in the factorization of the joint likelihood (Eq. 3). Thus by specifying parameter prior for one complete model, we have implicitly specified a prior for every complete model.

It remains to examine how the prior $p(\theta_m|m^h)$ is computed for an incomplete DAG model m for **X**. Due to global parameter independence we have $p(\theta_m|m^h) = \prod_{i=1}^n p(\theta_i|m^h)$ and therefore it suffices to examine each of the n terms separately. To compute $p(\theta_i|m^h)$, we identify a complete DAG model m_{ci} such that $\mathbf{Pa}_i^m = \mathbf{Pa}_i^{m_{ci}}$. The prior $p(\theta_{mci}|m_{ci}^h)$ is obtained from $p(\theta_{mc}|m_c^h)$, as we have shown for every pair of complete DAG models. Now, global parameter independence states that $p(\theta_{mc}|m_{ci}^h)$ can be written as a product $\prod_{i=1}^n p(\theta_i|m_{ci}^h)$, and therefore, $p(\theta_i|m_{ci}^h)$ is available. Finally, due to prior modularity $p(\theta_i|m^h)$ is equal to $p(\theta_i|m_{ci}^h)$.

The following theorem summarizes this discussion.

Theorem 1 *Given Assumptions 1 through 5, the parameter prior $p(\theta_m|m^h)$ for every DAG model m is determined by a specified parameter prior $p(\theta_{mc}|m_c^h)$ for an arbitrary complete DAG model m_c.*

Theorem 1 shows that once we specify the parameter prior for one complete DAG model all other priors can be generated automatically and need not be specified manually. Consequently, together with Eq. 2 and due to the fact that also likelihoods can be generated automatically in a similar fashion, we have a manageable methodology to automate the computation of $p(d|m^h)$ for any DAG model of $\mathbf{X}$ which is being considered by a search algorithm as a candidate model. Next we show how this computation can be done implicitly without actually computing the priors and likelihoods.

2.2 Computation of the Marginal Likelihood for Complete Data

For a given $\mathbf{X}$, consider a DAG model m and a complete random sample d. Assuming global parameter independence, the parameters remain independent given complete data. That is,

$$p(\theta_m|d,m^h) = \prod_{i=1}^{n} p(\theta_i|d,m^h) \tag{4}$$

In addition, assuming global parameter independence, likelihood modularity, and prior modularity, the parameters remain modular given complete data. In particular, if X_i has the same parents in s_1 and s_2, then

$$p(\theta_i|d,m_1^h) = p(\theta_i|d,m_2^h) \tag{5}$$

Also, for any $\mathbf{Y} \subseteq \mathbf{X}$, define $d^{\mathbf{Y}}$ to be the random sample d restricted to observations of $\mathbf{Y}$. For example, if $\mathbf{X} = \{X_1, X_2, X_3\}$, $\mathbf{Y} = \{X_1, X_2\}$, and $d = \{\mathbf{x}_1 = \{x_{11}, x_{12}, x_{13}\}, \mathbf{x}_2 = \{x_{21}, x_{22}, x_{23}\}\}$, then we have $d^{\mathbf{Y}} = \{\{x_{11}, x_{12}\}, \{x_{21}, x_{22}\}\}$. Let $\mathbf{Y}$ be a subset of $\mathbf{X}$, and s_c be a complete structure for any ordering where the variables in $\mathbf{Y}$ come first. Then, assuming global parameter independence and likelihood modularity, it is not difficult to show that

$$p(\mathbf{Y}|d,m_c^h) = p(\mathbf{Y}|d^{\mathbf{Y}}, m_c^h) \tag{6}$$

Given these observations, we can compute the marginal likelihood as follows.

Theorem 2 *Given any complete DAG model m_c for $\mathbf{X}$, any DAG model m for $\mathbf{X}$, and any complete random sample d, Assumptions 1 through 5 imply*

$$p(d|m^h) = \prod_{i=1}^{n} \frac{p(d^{\mathbf{Pa}_i \cup \{X_i\}}|m_c^h)}{p(d^{\mathbf{Pa}_i}|m_c^h)} \tag{7}$$

Proof: From the rules of probability, we have

$$p(d|m^h) = \prod_{l=1}^{m} \int p(\mathbf{x}_l|\theta_m, m^h) \, p(\theta_m|d_l, m^h) \, d\theta_m \tag{8}$$

where $d_l = \{\mathbf{x}_1, \ldots, \mathbf{x}_{l-1}\}$. Using Equations 1 and 4 to rewrite the first and second terms in the integral, respectively, we obtain

$$p(d|m^h) = \prod_{l=1}^{m} \int \prod_{i=1}^{n} p(x_{il}|\mathbf{pa}_{il}, \theta_i, m^h) \, p(\theta_i|d_l, m^h) \, d\theta_m$$

where x_{il} is the value of X_i in the l-th data point.

Using likelihood modularity and Equation 5, we get

$$p(d|m^h) = \prod_{l=1}^{m} \int \prod_{i=1}^{n} p(x_{il}|\mathbf{pa}_{il}, \theta_i, m_{ci}^h) \, p(\theta_i|d_l, m_{ci}^h) \, d\theta_m \tag{9}$$

where s_{ci} is a complete structure with variable ordering $\mathbf{Pa}_i$, X_i followed by the remaining variables. Decomposing the integral over θ_m into integrals over the individual parameter sets θ_i, and performing the integrations, we have

$$p(d|m^h) = \prod_{l=1}^{m} \prod_{i=1}^{n} p(x_{il}|\mathbf{pa}_{il}, d_l, m_{ci}^h)$$

Using Equation 6, we obtain

$$
\begin{aligned}
p(d|m^h) &= \prod_{l=1}^{m} \prod_{i=1}^{n} \frac{p(x_{il}, \mathbf{pa}_{il}|d_l, m_{ci}^h)}{p(\mathbf{pa}_{il}|d_l, m_{ci}^h)} \\
&= \prod_{l=1}^{m} \prod_{i=1}^{n} \frac{p(x_{il}, \mathbf{pa}_{il}|d_l^{\mathbf{Pa}_i \cup \{X_i\}}, m_{ci}^h)}{p(\mathbf{pa}_{il}|d_l^{\mathbf{Pa}_i}, m_{ci}^h)} \\
&= \prod_{i=1}^{n} \frac{p(d^{\mathbf{Pa}_i \cup \{X_i\}}|m_{ci}^h)}{p(d^{\mathbf{Pa}_i}|m_{ci}^h)}
\end{aligned} \tag{10}
$$

By the likelihood modularity, complete model equivalence, and regularity assumptions, we have that $p(d|m_{ci}^h) = p(d|m_c^h), i = 1, \ldots, n$. Consequently, for any subset $\mathbf{Y}$ of $\mathbf{X}$, we obtain $p(d^{\mathbf{Y}}|m_{ci}^h) = p(d^{\mathbf{Y}}|m_c^h)$ by summing over the variables in $d^{\mathbf{X} \setminus \mathbf{Y}}$. Consequently, using Equation 10, we get Equation 7. $\square$

An important feature of the formula for marginal likelihood (Equation 7), which we now demonstrate, is that two DAG models that represent the same assertions of conditional independence have the same marginal likelihood. We say that two structures for $\mathbf{X}$ are *independence equivalent* if they represent the same assertions of conditional independence. Independence equivalence is an equivalence relation, and induces a set of equivalence classes over the possible structures for $\mathbf{X}$.

Verma and Pearl (1990) provide a simple characterization of independence-equivalent structures using the concept of a v-structure. Given a structure s, a *v-structure* in s is an ordered node triple (X_i, X_j, X_k) where s contains the arcs $X_i \rightarrow X_j$ and $X_j \leftarrow X_k$, and there is no arc between X_i and X_k in either direction.

Verma and Pearl show that two structures for **X** are independence equivalent if and only if they have identical edges and identical v-structures. This characterization makes it easy to identify independence equivalent structures.

An alternative characterization by Chickering (1995) is useful for proving our claim that independence equivalent structures have the same marginal likelihood. An *arc reversal* is a transformation from one structure to another, in which a single arc between two nodes is reversed. An arc between two nodes is said to be *covered* if those two nodes would have the same parents if the arc were removed.

Theorem 3 (Chickering, 1995)

Two structures for **X** *are independence equivalent if and only if there exists a set of covered arc reversals that transform one structure into the other.*

A proof of this theorem can also be found in (Heckerman et al., 1995a). We are ready to prove our claim.

Theorem 4 *Given Assumptions 1 through 5, every two independence equivalent DAG models have the same marginal likelihood.*

Proof: Theorem 3 implies that we can restrict the proof to two DAG models that differ by a single covered arc. Say the arc is between X_i and X_j and that the joint parents of X_i and X_j are denoted by π. For these two models, Equation 7 differs only in terms i and j. For both models the product of these terms is $p(d^{\pi \cup \{X_i, X_j\}} | m_c^h)/p(d^\pi | m_c^h)$. $\square$

The conclusions of Theorems 2 and 4 are not justified when our assumptions are violated. In the example of the sigmoid functions, discussed in the previous subsection, the structures s_1 and s_2 differ by the reversal of a covered arc between X_2 and X_3, but, given that all local distribution families are sigmoid, there are certain joint likelihoods that can be represented by one structure, but not the other, and so their marginal likelihood is different.

2.3 Gaussian Directed Acyclic Graphical Models

We now apply the methodology of previous sections to Gaussian DAG models. A Gaussian DAG model is a DAG model as defined by Eq 1, where each variable $X_i \in \mathbf{X}$ is continuous, and each local likelihood is the linear regression model

$$p(x_i | \mathbf{pa}_i, \theta_i, m^h) = N(x_i | m_i + \sum_{x_j \in \mathbf{pa}_i} b_{ji} x_j, 1/v_i) \quad (11)$$

where $N(x_i | \mu, \tau)$ is a normal distribution with mean μ and precision $\tau > 0$. Given this form, a missing arc from X_j to X_i implies that $b_{ji} = 0$ in the complete DAG model. The local parameters are given by $\theta_i = (m_i, b_i, v_i)$, where b_i is the column vector $(b_{1i}, \ldots, b_{i-1,i})$.

For Gaussian DAG models, the joint likelihood $p(\mathbf{x} | \theta_m, m^h)$ obtained from Eqs 1 and 11 is an n-dimensional multivariate normal distribution with mean μ and symmetric positive definite precision matrix W,

$$p(\mathbf{x} | \theta_m, m^h) = \prod_{i=1}^{n} p(x_i | \mathbf{pa}_i^m, \theta_i, m^h) = N(\mathbf{x} | \mu, W).$$

For a complete model m_c with ordering $(X_1, \ldots, X_n)$ there is a one-to-one mapping between $\theta_{mc} = \bigcup_{i=1}^{n} \theta_i$ where $\theta_i = (m_i, b_i, v_i)$ and $\{\mu, W\}$ which has a nowhere singular Jacobian matrix. Consequently, assigning a prior for the parameters of one complete model induces a parameter prior, via the change of variables formula, for $\{\mu, W\}$ and in turn, induces a parameter prior for every complete model. Any such induced parameter prior must satisfy, according to our assumptions, global parameter independence. Not many prior distributions satisfy such a requirement. In fact, in the next section we show that the parameter prior $p(\mu, W | m_c^h)$ must be a normal-Wishart distribution.

For now we proceed by simply choosing $p(\mu, W | m_c^h)$ to be a normal-Wishart distribution. In particular, $p(\mu | W, m_c^h)$ is a multivariate-normal distribution with mean ν and precision matrix $\alpha_\mu W$ ($\alpha_\mu > 0$); and $p(W | m_c^h)$ is a Wishart distribution, given by,

$$p(W | m_c^h) = c(n, \alpha) |T|^{\alpha/2} |W|^{(\alpha-n-1)/2} e^{-1/2 tr\{TW\}} \quad (12)$$

with α degrees of freedom ($\alpha > n - 1$) and a positive-definite parametric matrix T and where $c(n, \alpha)$ is a normalization constant given by

$$c(n, \alpha) = \left[2^{\alpha n/2} \pi^{n(n-1)/4} \prod_{i=1}^{n} \Gamma\left(\frac{\alpha + 1 - i}{2}\right) \right]^{-1} \quad (13)$$

(e.g., DeGroot, 1970, p. 57).

This choise satisfies global parameter independence due to the following well known theorem.

Define a block partitioning $\{W_{11}, W_{12}, W_{12}', W_{22}\}$ of an n by n matrix W to be *compatible* with a partitioning μ_1, μ_2 of an n dimensional vector μ, if the indices of the rows that correspond to block W_{11} are the same as the indices of the terms that constitute μ_1 and similarly for W_{22} and μ_2.

Theorem 5 *If $f(\mu, W)$ is an n dimensional normal-Wishart distribution, $n \geq 2$, with parameters ν, α_μ, α, and T, then $\{\mu_1, W_{11} - W_{12} W_{22}^{-1} W_{12}'\}$ is independent of $\{\mu_2 - W_{22}^{-1} W_{12}' \mu_1, W_{12}, W_{22}\}$ for every partitioning μ_1, μ_2 of μ where W_{11}, W_{12}, W_{12}', W_{22} is a block partitioning of W compatible with the partitioning μ_1, μ_2. Furthermore, the pdf of $\{\mu_1, W_{11} - W_{12} W_{22}^{-1} W_{12}'\}$ is normal-Wishart with parameters ν_1, α_μ, $T_{11} - T_{12} T_{22}^{-1} T_{12}'$, and $\alpha - n + l$ where T_{11}, T_{12}, T_{12}', T_{22} is a compatible block partitioning of T, ν_1, ν_2 is a compatible partitioning of ν, and l is the size of the vector ν_1.*

The proof of Theorem 5 requires a change of variables from (μ, W) to $(\mu_1, \mu_2 - W_{22}^{-1} W_{12}' \mu_1)$ and $(W_{11} -$

$W_{12}W_{22}^{-1}W_{12}'$, W_{12}, W_{22}). Press carries out these computations for the Wishart distribution (1971, p. 117-119). Standard changes are needed to obtain the claim for the normal-Wishart distribution.

To see why the independence conditions in Theorem 5 imply global parameter independence, consider the partitioning in which the first block contains the first $n - 1$ coordinates which correspond to $X_1, \ldots, X_{n-1}$ while the second block contains the last coordinate which corresponds to X_n. For this partitioning, $b_n = -W_{22}^{-1}W_{12}'$, $v_n = W_{22}^{-1}$, and $m_n = \mu_2 - W_{22}^{-1}W_{12}'\mu_1$. Furthermore, $((W^{-1})_{11})^{-1} = W_{11} - W_{12}W_{22}^{-1}W_{12}'$ is the precision matrix associated with $X_1, \ldots, X_{n-1}$. Consequently, $\{m_n, b_n, v_n\}$ is independent of $\{\mu_1, ((W^{-1})_{11})^{-1}\}$. We now recursively repeat this argument with $\{\mu_1, ((W^{-1})_{11})^{-1}\}$ instead of $\{\mu, W\}$, to obtain global parameter independence. The converse, namely that global parameter independence implies the independence conditions in Theorem 5, is established similarly.

Our choise of prior implies that the posterior $p(\mu, W|d, m_c^h)$ is also a normal-Wishart distribution (DeGroot, 1970, p. 178). In particular, $p(\mu|W, d, m_c^h)$ is multivariate normal with mean vector ν' given by

$$\nu' = \frac{\alpha_\mu \nu + m\overline{\mathbf{x}}_m}{\alpha_\mu + m} \qquad (14)$$

and precision matrix $(\alpha_\mu + m)W$, where $\overline{\mathbf{x}}_m$ is the sample mean of d, and $p(W|d, m_c^h)$ is a Wishart distribution with $\alpha + m$ degrees of freedom and parametric matrix R given by

$$R = T + S_m + \frac{\alpha_\mu m}{\alpha_\mu + m}(\nu - \overline{\mathbf{x}}_m)(\nu - \overline{\mathbf{x}}_m)' \qquad (15)$$

where $S_m = \sum_{i=1}^m (\mathbf{x}_i - \overline{\mathbf{x}}_m)(\mathbf{x}_i - \overline{\mathbf{x}}_m)'$. From these equations, we see that α_μ and α can be thought of as equivalent sample sizes for μ and W, respectively.

According to Theorem 5, if $p(\mu, W|m_c^h)$ is a normal-Wishart distribution with the parameters given by the theorem, then $p(\mu_\mathbf{Y}, ((W^{-1})_\mathbf{YY})^{-1}|m_c^h)$ is also a normal-Wishart distribution with parameters $\nu_\mathbf{Y}$, α_μ, $T_\mathbf{Y} = ((T^{-1})_\mathbf{YY})^{-1}$ and $\alpha' = \alpha - n + l$, where $\mathbf{Y}$ is a subset of l coordinates. Thus, applying standard formulas pertaining to t-distributions (e.g., DeGroot, 1970, p. 179-180), we obtain the terms in Equation 7:

$$p(d^\mathbf{Y}|m_c^h) = (2\pi)^{-lm/2} \cdot \qquad (16)$$

$$\left(\frac{\alpha_\mu}{\alpha_\mu + m}\right)^{l/2} \frac{c(l, \alpha')}{c(l, \alpha' + m)}|T_\mathbf{Y}|^{\frac{\alpha'}{2}}|R_\mathbf{Y}|^{-\frac{\alpha' + m}{2}}$$

where $R_\mathbf{Y} = ((R^{-1})_\mathbf{YY})^{-1}$ is the posterior parametric matrix restricted to the $\mathbf{Y}$ coordinates.

We have just shown how to compute the marginal likelihood for Gaussian DAG models given the direct assessment of a parameter prior $p(\mu, W|m_c^h)$ for one complete model. The task of assessing a parameter prior for one complete Gaussian DAG model is

equivalent, in general, to assessing priors for the parameters of a set of n linear regression models (due to Equation 11). However, to satisfy global parameter independence, the prior for the linear regression model for X_n given $X_1, \ldots, X_{n-1}$ determines the priors for the linear coefficients and variances in all the linear regression models that define a complete Gaussian model. In particular, $1/v_n$ has a one dimensional Wishart pdf $W(1/v_n \mid \alpha + n - 1, T_{22} - T_{12}'T_{11}^{-1}T_{12})$ (i.e., a gamma distribution), and b_n has a normal pdf $N(b_n \mid T_{11}^{-1}T_{12}, T_{22}/v_n)$. Consequently, the degrees of freedom α and the parametric matrix T, which completely specify the Wishart prior distribution, are determined by the normal-gamma prior for one regression model. Kadane et al. (1980) address in detail the assessment of such a normal-gamma prior for a linear regression model and their method applies herein with no needed changes. The relationships between this elicited prior and the priors for the other $n - 1$ linear regression models can be used to check consistency of the elicited prior. Finally, a normal prior for the means of $X_1, \ldots, X_n$ is assessed separately and it requires only the assessment of a vector of means along with an equivalent sample size α_μ.

Our method for constructing parameter priors for many DAG models from a prior for one regression model has recently been applied to analyses of data in the domain of image compression (Thiesson et al., 1998). Our method also provides a suitable Bayesian alternative for many of the examples discussed in (Spirtes et al., 1993).

3 Characterization of Several Probability Distributions

We now characterize the Wishart distribution as the only pdf that satisfies global parameter independence for an unknown precision matrix W with $n \geq 3$ coordinates (Theorem 6). This theorem is phrased and proven in a terminology that relates to known facts about the Wishart distribution. We proceed with similar characterizations of the normal and normal-Wishart distributions (Theorems 8 and 9).

Theorem 6 *Let W be an $n \times n$, $n \geq 3$, positive-definite symmetric matrix of random variables and $f(W)$ be a pdf of W. Then, $f(W)$ is a Wishart distribution if and only if $W_{11} - W_{12}W_{22}^{-1}W_{12}'$ is independent of $\{W_{12}, W_{22}\}$ for every block partitioning $W_{11}, W_{12}, W_{12}', W_{22}$ of W.*

Proof: That $W_{11} - W_{12}W_{22}^{-1}W_{12}'$ is independent of $\{W_{12}, W_{22}\}$ whenever $f(W)$ is a Wishart distribution is a well known fact (Press 1971, p. 117-119). It is also expressed by Theorem 5. The other direction is proven by induction on n. The base case $n = 3$ is treated at the end.

The pdf of W can be written in $n!$ orderings. In particular, due to the assumed independence conditions, we have the following equality:

$$f(W) = f_1(W_{11} - W_{12}W_{22}^{-1}W_{12}')f_{2|1}(W_{22}, W_{12})$$
$$= f_2(W_{22} - W_{12}'W_{11}^{-1}W_{12})f_{1|2}(W_{11}, W_{12}) \qquad (17)$$

where a subscripted f denotes a pdf. Since $n > 3$, we can divide the indices of W into three non-empty sets a, b and c such that b includes at least two indices. We now group a and b to form a block and b and c to form a block. For each of the two cases, let W_{11} be the block consisting of the indices in $\{a, b\}$ or $\{b, c\}$, respectively, and W_{22} be the block consisting of the indices of c or a, respectively. By the induction hypothesis, and since the independence conditions on W can be shown to hold for any block W_{11} of W, we conclude that $f_1(V)$ is a Wishart distribution $W(V \mid \alpha_1, T_1)$ and $f_2(V)$ is a Wishart distribution $W(V \mid \alpha_2, T_2)$. Consequently, the pdf of the block corresponding to the indices in b is a Wishart distribution, and from the two alternative ways by which this pdf can be formed, it follows that $\alpha_1 - l_1 = \alpha_2 - l_2$, where l_i is the number of indices in block i (Press, 1971, Theorem 5.1.4). Thus,

$$c_1 |W_{11.2}|^\beta e^{tr\{T_1 W_{11.2}\}} f_{2|1}(W_{22}, W_{12}) =$$
$$c_2 |W_{22.1}|^\beta e^{tr\{T_2 W_{22.1}\}} f_{1|2}(W_{11}, W_{12}) \quad (18)$$

where c_1 and c_2 are normalizing constants, $\beta = (\alpha_1 - l_1 - 1)/2$, $W_{11.2} = W_{11} - W_{12} W_{22}^{-1} W_{12}'$, and $W_{22.1} = W_{22} - W_{12}' W_{11}^{-1} W_{12}$. Define

$$F_{2|1}(W_{22}, W_{12}) = \quad (19)$$
$$c_1 f_{2|1}(W_{22}, W_{12}) / |W_{22}|^\beta e^{tr\{T_2 W_{22} + T_1(W_{12} W_{22}^{-1} W_{12}')\}}$$

$$F_{1|2}(W_{11}, W_{12}) = \quad (20)$$
$$c_2 f_{1|2}(W_{11}, W_{12}) / |W_{11}|^\beta e^{tr\{T_1 W_{11} + T_2(W_{12}' W_{11}^{-1} W_{12})\}},$$

substitute into Equation 18, and obtain, using $|W_{11} - W_{12} W_{22}^{-1} W_{12}'||W_{22}| = |W|$, that $F_{2|1}(W_{22}, W_{12}) = F_{1|2}(W_{11}, W_{12})$. Consequently, $F_{2|1}$ and $F_{1|2}$ are functions only of W_{12} and thus, using Equation 19, we obtain

$$f(W) = |W|^\beta e^{tr\{T_1 W_{11} + T_2 W_{22}\}} H(W_{12}) \quad (21)$$

for some function H.

To show that $f(W)$ is Wishart we must find the form of H. Considering the three possible pairs of blocks formed with the sets of indices a, b, and c, Equation 21 can be rewritten as follows.

$$f(W) = |W|^{\beta_1} e^{tr\{T_{aa} W_{aa} + T_{bb} W_{bb} + T_{cc} W_{cc}\}} \cdot \quad (22)$$
$$e^{2tr\{T_{ab}' W_{ab} + T_{ac}' W_{ac} + T_{bc}' W_{bc}\}} H_1(W_{ac}, W_{bc})$$

$$f(W) = |W|^{\beta_2} e^{tr\{S_{aa} W_{aa} + S_{bb} W_{bb} + S_{cc} W_{cc}\}} \cdot \quad (23)$$
$$e^{2tr\{S_{ab}' W_{ab} + S_{ac}' W_{ac} + S_{bc}' W_{bc}\}} H_2(W_{ab}, W_{bc})$$

$$f(W) = |W|^{\beta_3} e^{tr\{R_{aa} W_{aa} + R_{bb} R_{bb} + T_{cc} W_{cc}\}} \cdot \quad (24)$$
$$e^{2tr\{R_{ab}' W_{ab} + R_{ac}' W_{ac} + R_{bc}' W_{bc}\}} H_3(W_{ab}, W_{ac})$$

By setting $W_{ab} = W_{ac} = W_{bc} = 0$, we get $\beta_1 = \beta_2 = \beta_3$ and $T_{ii} = S_{ii} = R_{ii}$, for $i = a, b, c$. By comparing Equations 22 and 23 we obtain

$$e^{2tr\{(T_{ac}' - S_{ac}') W_{ac}\}} H_1(W_{ac}, W_{bc}) = \quad (25)$$
$$e^{2tr\{(S_{ab}' - T_{ab}') W_{ab} + (S_{bc}' - T_{bc}') W_{bc}\}} H_2(W_{ab}, W_{bc})$$

Each side of this equation must be a function only of W_{bc}. We denote this function by H_{12}. Hence,

$$H_1(W_{ac}, W_{bc}) = H_{12}(W_{bc}) e^{2tr\{(S_{ac}' - T_{ac}') W_{ac}\}}$$

and by symmetric arguments, comparing Equations 22 and 24,

$$H_1(W_{ac}, W_{bc}) = H_{13}(W_{ac}) e^{2tr\{(R_{bc}' - T_{bc}') W_{bc}\}}$$

Thus, $H_{12}(W_{bc})$ is proportional to $e^{2tr\{(R_{bc}' - T_{bc}') W_{bc}\}}$ and so $f(W)$ is a Wishart distribution, as claimed.

It remains to examine the case $n = 3$. We first assume $n = 2$ in which case $f(W)$ is not necessarily a Wishart distribution. In the full version of this paper (Submitted to Annals of Statistics) we show that given the independence conditions for two coordinates, f must have the form

$$f(W) = c|W|^\beta e^{tr\{TW\}} H(W_{12}) \quad (26)$$

where H is an arbitrary function, and that the marginal distributions of W_{11} and W_{22} are one dimensional Wishart distributions. The proof rests on techniques from the theory of functional equations (Aczél, 1966) and results from (Járai, 1986, 1998). A weaker proof, under some regularity conditions, can be found in (Geiger and Heckerman, 1998).

We now treat the case $n = 3$ using these assertions about the case $n = 2$. Starting with Equation 17, and proceeding with blocks a, b, c each containing exactly one coordinate, we get, due to the given independence conditions for two coordinates, that f_1 has the form given by Equation 26, and that f_2 is a one dimensional Wishart distribution. Proceeding parallel to Equations 18 through 20, we obtain,

$$H(a_{12} - b_1^2 b_2^2 / W_{22}) F_{2|1}(W_{22}, W_{12}) = F_{1|2}(W_{11}, W_{12}) \quad (27)$$

where (b_1, b_2) is the matrix W_{12}, a_{12} is the off-diagonal element of W_{11}, $a_{12} - b_1^2 b_2^2 / W_{22}$ is the off diagonal element of $W_{11} - W_{12} W_{22}^{-1} W_{12}'$, and W_{22} is a 1×1 matrix. Note that the right hand side depends on W_{11} only through a_{12}. Let b_1 and b_2 be fixed, $y = b_1^2 b_2^2 / W_{22}$, and $x = a_{12}$. Also let $F(t) = F_{2|1}(b_1^2 b_2^2 / t, (b_1, b_2))$ and $G(a_{12}) = F_{1|2}(W_{11}, (b_1, b_2))$. We can now rewrite Equation 27 as $H(x - y) F(y) = G(x)$. Now set $z = x - y$, and obtain for every $y, z > 0$

$$H(z) F(y) = G(y + z) \quad (28)$$

the only measurable solution of which for H is $H(z) = c e^{bz}$ (e.g., Aczél, 1966).

Substituting this form of H into Equation 26, we see that W_{11} has a two dimensional Wishart distribution. Recall that W_{22} has a one dimensional Wishart distribution. We can now apply the induction step starting form Equation 18 and prove the Theorem for $n = 3$. □

We now treat the situation when only the means are unknown, characterizing the normal distribution. The

two dimensional case turns out to be covered by the Skitovich-Darmois theorem (e.g., Kagan, Linnik, and Rao (1973)).

Theorem 7 (Skitovich-Darmois) *Let $z_1, \ldots, z_k$ be independent random variables and $\alpha_i, \beta_i, 1 < i < k$, be constant coefficients. If $L_1 = \sum \alpha_i z_i$ is independent of $L_2 = \sum \beta_i z_i$, then each z_i for which $\alpha_i \beta_i \neq 0$ is normal.*

The Skitovich-Darmois theorem is used in the proof of the base case of our next characterization. Several generalizations of the Skitovich-Darmois theorem are described in Kagan et al. (1973).

Theorem 8 *Let W be an $n \times n$, $n \geq 2$, positive-definite symmetric matrix of real random variables such that no entry in W is zero, μ be an n-dimensional vector of random variables, and $f(\mu)$ be a pdf of μ. Then, $f(\mu)$ is an n dimensional normal distribution $N(\mu|\eta, \gamma W)$ where $\gamma > 0$ if and only if μ_1 is independent of $\mu_2 + W_{22}^{-1} W_{12}' \mu_1$ for every partitioning μ_1, μ_2 of μ where $W_{11}, W_{12}, W_{12}', W_{22}$ is a block partitioning of W compatible with the partitioning μ_1, μ_2.*

Proof: The two independence conditions, μ_1 independent of $\mu_2 + W_{22}^{-1} W_{12}' \mu_1$ and μ_2 independent of $\mu_1 + W_{11}^{-1} W_{12} \mu_2$, are equivalent to the following functional equation

$$f(\mu) = f_1(\mu_1) f_{2|1}(\mu_2 + W_{22}^{-1} W_{12}' \mu_1) \qquad (29)$$
$$= f_2(\mu_2) f_{1|2}(\mu_1 + W_{11}^{-1} W_{12} \mu_2)$$

where a subscripted f denotes a pdf. We show that the only solution for f that satisfies this equation is the normal distribution. Consequently both the if and only if portions of the theorem will be established.

For $n \geq 3$, we can divide the indices of W into three non-empty sets a, b and c. We group a and b to form a block and b and c to form a block. For each of the two cases, let W_{11} be the block consisting of the indices in $\{a, b\}$ or $\{b, c\}$, respectively, and W_{22} be the block consisting of the indices of c or a, respectively. By the induction hypothesis applied to both cases and marginalization we can assume that $f_1(\mu_1)$ is a normal distribution $N(\mu_1|\eta_1, \gamma_1(W^{-1})_{11})^{-1})$ and that $f_2(\mu_2) = N(\mu_2|\eta_2, \gamma_2(W^{-1})_{22})^{-1})$. Consequently, the pdf of the block corresponding to the indices in b is a normal distribution, and from the two alternative ways by which this pdf can be formed, it follows that $\gamma_1 = \gamma_2$.

Let $\gamma = \gamma_i$, $i = 1, 2$, and define

$$F_{2|1}(x) = f_{2|1}(x)/N(x|\eta_2 + W_{22}^{-1} W_{12}' \eta_1, \gamma W_{22})$$
$$F_{1|2}(x) = f_{1|2}(x)/N(x|\eta_1 + W_{11}^{-1} W_{12} \eta_2, \gamma W_{11}).$$

By substituting these definitions into Equation 29, substituting the normal form for $f_1(\mu_1)$ and $f_2(\mu_2)$, and canceling on both sides of the equation the term $N(\mu|\eta, \gamma W)$ (which is formed by standard algebra pertaining to quadratic forms (E.g., DeGroot, pp. 55)), we obtain a new functional equation,

$$F_{2|1}(\mu_2 + W_{22}^{-1} W_{12}' \mu_1) = F_{1|2}(\mu_1 + W_{11}^{-1} W_{12} \mu_2).$$

By setting $\mu_2 = -W_{22}^{-1} W_{12}' \mu_1$, we obtain $F_{1|2}((I - (W_{11}^{-1} W_{12})(W_{22}^{-1} W_{12}'))\mu_1) = F_{2|1}(0)$ for every μ_1. Hence, the only solution to this functional equation is $F_{1|2} = F_{2|1} \equiv$ constant. Consequently, $f(\mu) = N(\mu|\eta, \gamma W)$.

It remains to prove the theorem for $n = 2$. Let $z_1 = \mu_1$, $z_2 = \mu_2 + w_{22}^{-1} w_{12} \mu_1$, $L_1 = \mu_1 + w_{11}^{-1} w_{12} \mu_2$, and $L_2 = \mu_2$. By our assumptions z_1 and z_2 are independent and L_1 and L_2 are independent. Furthermore, rewriting L_1 and L_2 in terms of z_1 and z_2, we get, $L_1 = w_{11}^{-1} w_{22}^{-1}(w_{11} w_{22} - w_{12}^2) z_1 + w_{11}^{-1} w_{12} z_2$ and $L_2 = z_2 - w_{22}^{-1} w_{12} z_1$. All linear coefficients in this transformation are non zero due to the fact that W is positive definite and that w_{12} is not zero. Consequently, due to the Skitovich-Darmois theorem, z_1 is normal and z_2 is normal. Furthermore, since z_1 and z_2 are independent, their joint pdf is normal as well. Finally, $\{\mu_1, \mu_2\}$ and $\{z_1, z_2\}$ are related through a non-singular linear transformation and so $\{\mu_1, \mu_2\}$ also have a joint normal distribution $f(\mu) = N(\mu|\eta, A)$ where $A = (a_{ij})$ is a 2×2 precision matrix. Substituting this solution into Equation 29 and comparing the coefficients of μ_1^2, μ_2^2, and $\mu_1 \mu_2$, we obtain $a_{12}/a_{11} = w_{12}/w_{11}$ and $a_{12}/a_{22} = w_{12}/w_{22}$. Thus $A = \gamma W$ where $\gamma > 0$. $\square$

The proofs of Theorems 6 and 8 can be combined to form the following characterization of the normal-Wishart distribution.

Theorem 9 *Let W be an $n \times n$, $n \geq 3$, positive-definite symmetric matrix of real random variables such that no entry in W is zero, μ be an n-dimensional vector of random variables, and $f(\mu, W)$ be a joint pdf of $\{\mu, W\}$. Then, $f(\mu, W)$ is an n dimensional normal-Wishart distribution if and only if $\{\mu_1, W_{11} - W_{12} W_{22}^{-1} W_{12}'\}$ is independent of $\{\mu_2 + W_{22}^{-1} W_{12}' \mu_1, W_{12}, W_{22}\}$ for every partitioning μ_1, μ_2 of μ where $W_{11}, W_{12}, W_{12}', W_{22}$ is a block partitioning of W compatible the partitioning μ_1, μ_2.*

Proof: The two independence conditions, $\{\mu_1, W_{11} - W_{12} W_{22}^{-1} W_{12}'\}$ independent of $\{\mu_2 + W_{22}^{-1} W_{12}' \mu_1, W_{12}, W_{22}\}$ and $\{\mu_2, W_{22} - W_{12}' W_{11}^{-1} W_{12}\}$ independent of $\{\mu_1 + W_{11}^{-1} W_{12} \mu_2, W_{12}', W_{11}\}$, are equivalent to the following functional equation

$$f(\mu, W) = f_1(\mu_1, W_{11.2}) f_{2|1}(\mu_2 + W_{22}^{-1} W_{12}' \mu_1, W_{22}, W_{12})$$
$$= f_2(\mu_2, W_{22.1}) f_{1|2}(\mu_1 + W_{11}^{-1} W_{12} \mu_2, W_{11}, W_{12})$$

where a subscripted f denotes a pdf. We show that the only solution for f that satisfies this functional equation is the normal-Wishart distribution. Setting W to a fixed value yields Equation 29 the solution of which for f is proportional to $N(\mu|\eta, \gamma W)$. Similarly, the solutions for the functions $f_1, f_2, f_{1|2}$, and $f_{2|1}$ are also proportional to normal pdfs. The constants η and γ could potentially change from one value of W to another. However, since η_1 can only be a function of $W_{11} - W_{12} W_{22}^{-1} W_{12}'$ due to the solution for f_1, and since it must also be a function of $\{W_{22}, W_{12}\}$ due to the solution for $f_{2|1}$, it cannot change with W. Simi-

larly η_2 cannot change with W. Substituting this solution into Equation 30 and dividing by the common terms which are equal to $f(\mu|W)$ yields Equation 17 the solution of which for f is a Wishart pdf. □

Note that the conditions set on W in Theorem 9, namely, a positive-definite symmetric matrix of real random variables such that no entry in W is zero, are necessary and sufficient in order for W to be a precision matrix of a complete Gaussian DAG model.

4   Local versus Global Parameter Independence

We have shown that the only pdf for $\{\mu, W\}$ which satisfies global parameter independence, when the number of coordinates is greater than two, is the normal-Wishart distribution. We now discuss additional independence assertions implied by the assumption of global parameter independence.

Definition *Local parameter independence* is the assertion that for every DAG model m for $X_1, \ldots, X_n$, there exists a partition of the parameters of each local distribution into at least two independent sets.

Consider the parameter prior for $\{m_n, b_n, v_n\}$ when the prior for $\{\mu, W\}$ is a normal Wishart as specified by Equations 12 and 13. By a change of variables, we get

$$f_n(m_n, b_n, v_n) = W(1/v_n \mid \alpha + n - 1, T_{22} - T_{12}'T_{11}^{-1}T_{12}) \cdot$$
$$N(b_n \mid T_{11}^{-1}T_{12}, T_{22}/v_n) \cdot N(m_n \mid v_n, \alpha_\mu/v_n)$$

where the first block corresponds to $X_1, \ldots, X_{n-1}$ and the second block corresponds to X_n. We note that the only independence assumption expressed by this product is that m_n and b_n are independent given v_n. However, by standardizing m_n and b_n, namely defining, $m_n^* = (m_n - v_n)/(\alpha_\mu/v_n)^{1/2}$ and $b_n^* = (T_{22}/v_n)^{1/2}(b_n - T_{11}^{-1}T_{12})$, which is well defined because T_{22} is positive definite and $v_n > 0$, we obtain a set of parameters (m_n^*, b_n^*, v_n) which are mutually independent. Furthermore, this mutual independence property holds for every local family and for every Gaussian DAG model over $X_1, \ldots, X_n$. We call this property the *standard local independence* for Gaussian DAG models.

This observation leads to the following corollary of our characterization theorems.

Corollary 10 *If global parameter independence holds for every complete Gaussian DAG model over $X_1, \ldots, X_n$ ($n \geq 3$), then standard local parameter independence also holds for every complete Gaussian DAG model over $X_1, \ldots, X_n$.*

This corollary follows from the fact that global parameter independence implies that, due to Theorem 9, the parameter prior is a normal-Wishart, and for this prior, we have shown that standard local parameter independence must hold.

It is interesting to note that when $n = 2$, there are distributions that satisfy global parameter independence

but do not satisfy standard local parameter independence. In particular, a prior for a 2 × 2 positive definite matrix W which has the form $W(W|\alpha, T)H(w_{12})$, where H is some real function and w_{12} is the off-diagonal element of W, satisfies global parameter independence but need not satisfy standard local parameter independence. Furthermore, if standard local parameter independence is assumed, then $H(w_{12})$ must be proportional to $e^{aw_{12}}$, which means that, for $n = 2$, the only pdf for W that satisfies global and standard local parameter independence is the bivariate Wishart distribution. In contrast, for $n > 2$, global parameter independence alone implies a Wishart prior.

5   Discussion

The formula for the marginal likelihood applies whenever Assumptions 1 through 5 are satisfied, not only for Gaussian DAG models. Another important special case is when all variables in **X** are discrete and all local distributions are multinomial. This case has been treated in (Heckerman et al. (1995; Geiger and Heckerman, 1997) under the additional assumption of local parameter independence. Our generalized derivation herein dispenses this assumption and unifies the derivation in the discrete case with the derivation needed for Gaussian DAG models.

Furthermore, our proof also suggests that the only parameter prior for complete discrete DAG models with $n \geq 3$ variables that satisfies Assumptions 1 through 5 is the Dirichlet distribution. The added assumption of local parameter independence, which is essential for the characterization of the Dirichlet distribution when $n = 2$ (Geiger and Heckerman, 1997), seems to be redundant when $n \geq 3$, just as it is redundant for the characterization of the normal-Wishart distribution.

Our characterization means that the assumption of global parameter independence when combined with the definition of m^h, the assumption of complete model equivalence, and the regularity assumption, may be too restrictive. One common remedy for this problem is to use a hierarchical prior $p(\theta|\eta)p(\eta)$ with hyperparameters η. When such a prior is used for Gaussian DAG models our results show that for every value of η for which global parameter independence holds, $p(\theta|\eta)$ must be a normal-Wishart distribution. Another possible approach is to select one representative DAG model from each class of equivalent DAG models, assume global parametr independence only for these representatives, and evaluate the marginal likelihood only for these representatives. The difficulty with this approach is that when projecting a prior from a complete DAG model to a DAG model with missing edges, one needs to perform additional high dimensional integrations, before using the parameter modularity property (see Section 2). The assumption of global parameter independence for all complete DAGs rather than one, removes the need for this additional integration. A final approach is to modify the definition of m^h to allow equivalent DAG models to have different parameter priors.

Acknowledgments

References

Aczél , J. (1966). *Lectures on Functional Equations and Their Applications*. Academic Press, New York.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley and Sons, New York.

Buntine, W. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225.

Chickering, D. (1995). A transformational characterization of equivalent Bayesian networks structures. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence*, Montreal, QU, pages 87–98. Morgan Kaufmann.

Cooper, G. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

Dawid, P. and Lauritzen, S. (1993). Hyper Markov laws in statistical analysis of decomposable graphical models. *Annals of Statistics*, 21:1272–1317.

DeGroot, M. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Friedman, N., and Goldszmidt, M. (1997). Sequential update of Bayesian networks structure. In *Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence*, Providence, RI, pages 165–174. Morgan Kaufmann.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence*, Seattle, WA, pages 235–243. Morgan Kaufmann.

Geiger, D. and Heckerman, D. (1997). A characterization of the Dirichlet distribution through global and local parameter independence. *Annals of Statistics*, 25(3):1344–1369.

Geiger, D. and Heckerman, D. (1998). A characterization of the bivariate Wishart distribution. *Probability and Mathematical Statistics*, 18(1):197–243, pages 119–131.

Heckerman, D. and Geiger, D. (1995). Learning Bayesian networks: A unification for discrete and Gaussian domains. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence*, Montreal, QU, pages 274–284. Morgan Kaufmann. See also Technical Report TR-95-16, Microsoft, Redmond, WA, February 1995.

Heckerman, D., Geiger, D., and Chickering, D. (1995a). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20, pages 197–243.

Heckerman, D., Mamdani, A., and Wellman, M. (1995b). Real-world applications of Bayesian networks. *Communications of the ACM*, 38.

Howard, R. and Matheson, J. (1981). Influence diagrams. In Howard, R. and Matheson, J., editors, *Readings on the Principles and Applications of Decision Analysis*, volume II, pages 721–762. Strategic Decisions Group, Menlo Park, CA.

Járai , A. (1998). On regular solutions of functional equations. *Aequationes Mathematicae*, 30, pages 21–54.

Járai , A. (1998). Regularity property of the functional equation of the Dirichlet distribution. *Aequationes Mathematicae*, 56, pages 37–46.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters, S.C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal American Statistical Association*, 75, pages 845–854.

Kagan, A.M., Linnik, Y.V., and Rao C.R. (1973). *Characterization problems in Mathematical Statistics*. Wiley & Sons.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Press S. (1971). *Applied Multivariate Analysis*, Holt, Rinehart, and Winston, New York.

Shachter, R. and Kenley, C. (1989). Gaussian influence diagrams. *Management Science*, 35:527–550.

Spiegelhalter, D., Dawid, A., Lauritzen, S., and Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8:219–282.

Spiegelhalter, D. and Lauritzen, S. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.

Thiesson B., Meek C., Chickering D., and D. Heckerman (1998). Computationally efficient methods for selecting among mixtures of graphical models. In *Proceedings of the Sixth Valencia International Meeting on Bayesian Statistics*, Alcossebre, Spain, pages 223-240,

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence*, Boston, MA, pages 220–227. Morgan Kaufmann.