The Annals of Statistics
2002, Vol. 30, No. 5, 1412–1440

PARAMETER PRIORS FOR DIRECTED ACYCLIC GRAPHICAL MODELS AND THE CHARACTERIZATION OF SEVERAL PROBABILITY DISTRIBUTIONS

BY DAN GEIGER AND DAVID HECKERMAN

Technion and Microsoft Research

We develop simple methods for constructing parameter priors for model choice among directed acyclic graphical (DAG) models. In particular, we introduce several assumptions that permit the construction of parameter priors for a large number of DAG models from a small set of assessments. We then present a method for directly computing the marginal likelihood of every DAG model given a random sample with no missing observations. We apply this methodology to Gaussian DAG models which consist of a recursive set of linear regression models. We show that the only parameter prior for complete Gaussian DAG models that satisfies our 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.

1. Introduction. Directed acyclic graphical (DAG) models have an increasing number of applications in statistics [Cowell, Dawid, Lauritzen and Spiegelhalter (1999)] as well as in decision analysis and artificial intelligence [Howard and Matheson (1981), Heckerman, Mamdani and Wellman (1995), 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 $\mathbf{X}$ is a directed graph with no directed cycles (i.e., a directed acyclic graph) having for every variable X_i in $\mathbf{X}$ 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 $\mathbf{X}$ given by $p(\mathbf{x}) = \prod_{i=1}^n p(x_i|\mathbf{pa}_i^m)$, where $\mathbf{x} = (x_1, \ldots, x_n)$ is a value for $\mathbf{X}$ (an n-tuple) and x_i is a value for X_i and where $\mathbf{pa}_i^m$ is the value for $\mathbf{Pa}_i^m$ as in $\mathbf{x}$. 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

Received November 1998; revised February 2002.

AMS 2000 subject classifications. Primary 62E10, 60E05; secondary 62A15, 62C10, 39B99.

Key words and phrases. Bayesian network, directed acyclic graphical model, Dirichlet distribution, Gaussian DAG model, learning, linear regression model, normal distribution, Wishart distribution.

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 latter 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 according to the same DAG. A DAG model is *complete* if it has no missing arcs. Note that any two complete DAG models for $\mathbf{X}$ encode the same assertions of conditional independence—namely, none. Also note that a complete DAG determines a unique ordering of the variables in which X_i precedes X_j if and only if $X_i \to X_j$ is an arc in this DAG.

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 are 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 $\mathbf{X}$ 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

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

where $\theta_1, \ldots, \theta_n$ 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 distributions 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 $d = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, 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

$$p(m^h|d) = cp(m^h)\, p(d|m^h) = cp(m^h) \int p(d|\theta_m, m^h)\, p(\theta_m|m^h)\, d\theta_m, \tag{2}$$

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, poses a serious computational challenge, because the number of DAG models grows faster than exponentially in n. Methods for searching through the space of model structures are discussed, for example, by Cooper and Herskovits

(1992), Heckerman, Geiger and Chickering (1995) 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 (1991) and Heckerman, Geiger and Chickering (1995) discuss methods for specifying the priors $p(m^h)$ via a small number of direct assessments.

Herein, we 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. Our method is based on a set of assumptions the most notable of which is the assumption that complete DAG models represent the same set of distributions, which implies that data cannot distinguish between two complete DAG models. Multivariate Gaussian, multinomial and multivariate t-distributions satisfy this assumption. Another assumption is *likelihood and prior modularity*, which says that the local distribution for x_i and its parameter priors depends only on the parents of x_i but not on the entire description of the structure. These assumptions, together with *global parameter independence*, introduced by Spiegelhalter and Lauritzen (1990), are the heart of the proposed methodology.

The methodology described herein for setting priors to DAG models and consequently calculating their marginal likelihoods is an extension of the results by Dawid and Lauritzen (1993) for decomposable graphical models. For decomposable graphical models, which form a set of models that can be regarded both as DAG models as well as undirected graphical models, the two methodologies are identical. Our specification of a formal set of assumptions followed by a technical derivation of this methodology provides an easy access to examine the validity of the approach and devise alternatives when needed.

The contributions of this paper are as follows: a methodology for specifying parameter priors for many DAG structures using a few direct assessments (Section 2); a formula that computes the marginal likelihood for every DAG model (Section 3); a specialization of this formula to an efficient computation for Gaussian DAG models (Section 4); and an analysis of complete Gaussian DAG models which shows that the only parameter prior that satisfies our assumptions is the normal-Wishart distribution (Section 5). 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 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 7, 9 and 10, respectively, proven in Section 5. It should be noted that a single principle, global parameter independence, is used to characterize three different distributions.

A similar characterization for the bivariate Wishart, bivariate normal, and bivariate normal-Wishart distributions has recently been obtained under the assumption that the pdf is strictly positive, and assuming also some additional independence constraints—termed standard local parameter independence [Geiger and Heckerman (1998)]. Another related result is the characterization of the Dirichlet distribution via global and local parameter independence [Geiger and Heckerman (1997), Járai (1998)].

2. The construction of parameter priors. In this section, we present 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 $\mathbf{X}$. 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, namely, that for every θ_{m1} there exists θ_{m2} such that $p(\mathbf{x}|\theta_{m1}, m_1^h) = p(\mathbf{x}|\theta_{m2}, m_2^h)$ and vice versa.*

Two examples where this assumption holds are quite common. One happens when $p(\mathbf{x}|\theta_{m1}, m_1^h)$ and $p(\mathbf{x}|\theta_{m2}, m_2^h)$ are multivariate normal distributions and the other happens when $\mathbf{X}$ consists of variables with finite domains and $p(\mathbf{x}|\theta_{m1}, m_1^h)$ and $p(\mathbf{x}|\theta_{m2}, m_2^h)$ are unrestricted discrete distributions. In these two cases, all the local distributions have the same functional form in every ordering of the variables. If the joint distribution for $\mathbf{X}$ is a multivariate t-distribution, then too, all local conditional distributions have the same functional form [e.g., DeGroot (1970)]. However, unlike the unrestricted discrete and multivariate normal distributions, for t-distributions, the parameters of the local distributions are dependent which violates Assumption 5 discussed below.

We now provide an example where this assumption 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 $X_1 \to X_2$, $X_1 \to X_3$, and $X_2 \to X_3$. Suppose further, that the local distributions $\mathcal{F}_{s1}$ of model m_1 are restricted to the logit

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

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 $X_1 \rightarrow X_2$, $X_1 \rightarrow X_3$, and $X_3 \rightarrow X_2$. Assumption 1 asserts that the families of local distributions for m_1 and m_2 are such that the set of joint distributions for $\mathbf{X}$ 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 logit distributions, then 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).

Note that whenever two DAG models represent the same set of probability distributions for $\mathbf{X}$, they must also specify the same set of independence assumptions. The example with the logit distributions highlights that the converse does not hold because every complete DAG represents the same independence assumptions, namely none, and yet complete DAG models can represent different sets of probability distributions.

Our definition of m^h, that the true joint pdf of a set of variables $\mathbf{X}$ is perfectly represented by m, and Assumption 1, which says that two complete models represent the same set of joint pdfs for $\mathbf{X}$, implies 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. Obviously, in the example with the logit distributions, the two models can be distinguished by data because they do not represent the same set of joint distributions.

ASSUMPTION 2 (Regularity). *For every two complete DAG models m_1 and m_2 for $\mathbf{X}$ there exists a one-to-one mapping $k_{1,2}$ 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 $\theta_{m2} = k_{1,2}(\theta_{m1})$. The Jacobian $|\partial\theta_{m1}/\partial\theta_{m2}|$ exists and is nonzero for all values of Θ_{m1}.*

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

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

For example, suppose $x = \{x_1, x_2\}$ has a nonsingular bivariate normal pdf $f(\mathbf{x}) = N(\mathbf{x}|\mu, W)$ where μ is the vector of means and $W = (w_{ij})$ is the inverse of a positive definite covariance matrix. If we write $f(\mathbf{x}) = f_{x_1}(x_1) f_{x_2|x_1}(x_2|x_1)$

where $f_{x_1}(x_1) = N(x_1|e_1, 1/v_1)$ and $f_{x_2|x_1}(x_2|x_1) = N(x_2|e_{2|1} + b_{12}x_1, 1/v_{2|1})$, then the following well-known relationships are satisfied:

$$w_{11} = \frac{1}{v_1} + \frac{b_{12}^2}{v_{2|1}}, \quad w_{12} = -\frac{b_{12}}{v_{2|1}}, \quad w_{22} = \frac{1}{v_{2|1}}, \tag{4}$$
$$e_1 = \mu_1, \quad e_{2|1} = \mu_2 - b_{12}\mu_1.$$

Note that the transformation between $\{\mu, W\}$ and $\{e_1, v_1, e_{2|1}, v_{2|1}, b_{12}\}$ is one-to-one and onto as long as W is the inverse of a covariance matrix and the conditional variances $v_1, v_{2|1}$ are positive. The Jacobian of this transformation is given by,

$$\left| \frac{\partial w_{11}, w_{12}, w_{22}, \mu_1, \mu_2}{\partial v_1, v_{1|2}, b_{12}, e_1, e_{2|1}} \right| = v_1^{-2} v_{2|1}^{-3}. \tag{5}$$

Symmetric equations hold when $f(\mathbf{x})$ is written as $f_{x_2}(x_2) f_{x_1|x_2}(x_1|x_2)$ and so there is a one-to-one and onto mapping between $\{e_1, v_1, e_{2|1}, v_{2|1}, b_{12}\}$ and $\{e_2, v_2, e_{1|2}, v_{1|2}, b_{21}\}$. Note that the parameters μ, W for the joint space are instrumental for decomposing the needed mapping into a composition of two mappings.

ASSUMPTION 3 (Likelihood modularity). *For every two DAG models m_1 and m_2 for $\mathbf{X}$ 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 $\mathbf{X}$ 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 $\mathbf{X}$, $p(\theta_m|m^h) = \prod_{i=1}^n p(\theta_i|m^h)$.*

The likelihood and prior modularity assumptions have been used implicitly in the work of, for example, Cooper and Herskovits (1992), Spiegelhalter, Dawid, Lauritzen and Cowell (1993) and Buntine (1994). Heckerman, Geiger and Chickering (1995) 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."

Note that the first three assumptions concern the distribution of $\mathbf{X}$ whereas the last two assumptions concern the distribution of the parameters. Obviously, when the parameters $\theta_1, \ldots, \theta_n$ are not variation independent for every complete DAG model for $\mathbf{X}$, the assumption of global parameter independence is inconsistent with the model and cannot be true. Hence, Assumption 5 excludes, for example, the possibility that two local distributions share a common parameter. On the other

hand, even when the parameters are variation independent, it is possible to specify a prior distribution for θ that violates global parameter independence. Cowell, Dawid, Lauritzen and Spiegelhalter [(1999), pages 191 and 192] highlight this point.

The assumptions we have made lead to the following significant consequence: 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 4, 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 $\mathbf{X}$ from the prior of another complete model for $\mathbf{X}$. In so doing, we are merely performing coordinate transformations between parameters for different variable orderings in the factorization of the joint likelihood (3). Thus by specifying the 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 $\mathbf{X}$ from the prior $p(\theta_{mc}|m_c^h)$ for some complete model m_c. Due to global parameter independence we have $p(\theta_m|m^h) = \prod_{j=1}^n p(\theta_j|m^h)$ 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 $m_{c(i)}$ such that $\mathbf{Pa}_i^m = \mathbf{Pa}_i^{m_{c(i)}}$. As we have shown, the prior $p(\theta_{mc(i)}|m_{c(i)}^h)$ is obtained from $p(\theta_{mc}|m_c^h)$ for every pair of complete DAG models. Due to global parameter independence $p(\theta_{mc(i)}|m_{c(i)}^h)$ is a product one term of which is $p(\theta_i|m_{c(i)}^h)$. Finally, due to prior modularity $p(\theta_i|m^h)$ is equal to $p(\theta_i|m_{c(i)}^h)$.

This methodology of constructing priors is described by Heckerman, Geiger and Chickering (1995) for discrete DAG models and in Section 4 for Gaussian DAG models. Our method is equivalent to the method of compatible priors devised for decomposable graphical models [Dawid and Lauritzen (1993)]. Our arguments, via a set of assumptions, can be regarded as an axiomatic justification for compatible priors, and as an extension of this method to general DAG models and to any probability distributions that satisfy Assumptions 1–5. We are currently unaware, however, of additional probability distributions that satisfy these five assumptions.

The following theorem summarizes the general construction which was formulated to cover both cases: the discrete and the Gaussian.

THEOREM 1. *Given Assumptions* 1–5, *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 (2) and because 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 efficiently.

3. 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,

$$p(\theta_m|d, m^h) = \prod_{i=1}^{n} p(\theta_i|d, m^h). \tag{6}$$

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

$$p(\theta_i|d, m_1^h) = p(\theta_i|d, m_2^h). \tag{7}$$

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 $d = \{\mathbf{x}_1 = \{x_{11}, x_{21}, x_{31}\},\ \mathbf{x}_2 = \{x_{12}, x_{22}, x_{32}\}\}$, then we have $d^{\mathbf{Y}} = \{\{x_{11}, x_{21}\}, \{x_{12}, x_{22}\}\}$. 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

$$p(\mathbf{y}|d, m_c^h) = p(\mathbf{y}|d^{\mathbf{Y}}, m_c^h). \tag{8}$$

Given these observations, we can compute the marginal likelihood as follows, yielding an important component for searching DAG models via a Bayesian methodology.

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* 1–5 *imply*

$$p(d|m^h) = \prod_{i=1}^{n} \frac{p(d^{\mathbf{Pa}_i \cup \{X_i\}}|m_c^h)}{p(d^{\mathbf{Pa}_i}|m_c^h)}. \tag{9}$$

PROOF. From the rules of probability, we have

$$p(d|m^h) = \prod_{l=1}^{m} \int p(\mathbf{x}_l|\theta_m, m^h)\, p(\theta_m|d_l, m^h)\, d\theta_m, \tag{10}$$

where $d_l = \{\mathbf{x}_1, \ldots, \mathbf{x}_{l-1}\}$. Using (1) and (6) 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 lth data point.

Using likelihood modularity and (7), we get

$$p(d|m^h) = \prod_{l=1}^{m} \int \prod_{i=1}^{n} p\big(x_{il}|\mathbf{pa}_{il}, \theta_i, m^h_{c(i)}\big)\, p\big(\theta_i|d_l, m^h_{c(i)}\big)\, d\theta_m, \tag{11}$$

where $s_{c(i)}$ 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

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

Using (8), we obtain

$$\begin{aligned} p(d|m^h) &= \prod_{l=1}^{m} \prod_{i=1}^{n} \frac{p(x_{il}, \mathbf{pa}_{il}|d_l, m^h_{c(i)})}{p(\mathbf{pa}_{il}|d_l, m^h_{c(i)})} \\ &= \prod_{l=1}^{m} \prod_{i=1}^{n} \frac{p(x_{il}, \mathbf{pa}_{il}|d_l^{\mathbf{Pa}_i \cup \{X_i\}}, m^h_{c(i)})}{p(\mathbf{pa}_{il}|d_l^{\mathbf{Pa}_i}, m^h_{c(i)})} \qquad (12) \\ &= \prod_{i=1}^{n} \frac{p(d^{\mathbf{Pa}_i \cup \{X_i\}}|m^h_{c(i)})}{p(d^{\mathbf{Pa}_i}|m^h_{c(i)})}. \end{aligned}$$

By the likelihood modularity, complete model equivalence and regularity assumptions, we have that $p(d|m^h_{c(i)}) = p(d|m^h_c)$, $i = 1, \ldots, n$. Consequently, for any subset $\mathbf{Y}$ of $\mathbf{X}$, we obtain $p(d^{\mathbf{Y}}|m^h_{c(i)}) = p(d^{\mathbf{Y}}|m^h_c)$ by summing over the variables in $\mathbf{X} \setminus \mathbf{Y}$. Consequently, using (12), we get (9). □

An equivalent approach for computing the marginal likelihood (9) for decomposable discrete and Gaussian DAG models has been developed by Dawid and Lauritzen (1993) using compatible priors.

An important feature of (9), 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 $X_i \to X_j$ 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 $\mathbf{X}$ 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 developed by Chickering (1995) and independently by Andersson, Madigan and Perlman [(1997), Lemma 3.2] 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), Andersson, Madigan and Perlman (1997)]. *Two structures for $\mathbf{X}$ 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, Geiger and Chickering (1995).

Theorem 3 implies that if every pair of DAGs that differ by a single covered arc represents the same set of distributions, then every two independence equivalent DAGs represent the same set of distributions. Furthermore, a consequence of the next theorem is that Assumptions 1–5 imply that indeed every two independence equivalent DAGs represent the same set of distributions. Without these assumptions, two independence equivalent DAGs can represent different sets of distributions.

THEOREM 4. *Given Assumptions* 1–5, *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, (9) differs only in terms i and j. For both models the product of these terms is $p(d^{\pi \cup \{X_i, X_j\}} | m_c^h) / p(d^{\pi} | m_c^h)$. □

The conclusions of Theorem 2 and, consequently, of Theorem 4 are not justified when our assumptions are violated. In the example of the logit distributions, discussed in the previous subsection, which violates Assumption 1, 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 logit, there are certain joint distributions that can

be represented by one structure, but not the other, and so their marginal likelihoods will be different.

The implication of Theorem 4 is quite strong: all models in the same independence equivalence class are scored equivalently. This severely constrains possible parameter priors as shown in the next two sections. One possible approach to bypass our assumptions is to select one representative DAG model from each class of independence equivalent DAG models, assume global parameter independence only for these representatives, and evaluate the marginal likelihood only for these representatives. The search can then be conducted in the space of representative models as suggested in Spirtes and Meek (1995), Chickering (1996), and Madigan, Andersson, Perlman and Volinsky (1996). 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). Another approach is to modify the definition of m^h to allow independence equivalent DAG models to have different parameter priors. This alternative is needed when arcs have a causal interpretation. However, when choosing this alternative, the parameter prior for each model examined by a search procedure must be provided by a user as the search is being conducted, or a new mechanism to produce acceptable priors on-the-fly must be devised.

4. 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 (1), where each variable $X_i \in \mathbf{X}$ is continuous, and each local likelihood is the linear regression model

$$p(x_i|\mathbf{pa}_i^m, \theta_i, m^h) = N\left(x_i|m_i + \sum_{x_j \in \mathbf{pa}_i} b_{ji} x_j, 1/v_i\right), \tag{13}$$

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 is equivalent to $b_{ji} = 0$ in the DAG model. The local parameters are given by $\theta_i = (m_i, b_i, v_i)$, where b_i is the column vector $(b_{1i}, \ldots, b_{i-1,i})$ of regression coefficients. Furthermore, m_i is the conditional mean of X_i and v_i is the conditional variance of X_i.

For Gaussian DAG models, the joint likelihood $p(\mathbf{x}|\theta_m, m^h)$ obtained from (1) and (13) is an n-dimensional multivariate normal distribution with a mean vector μ and a 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

$$
\begin{aligned}
p(W|m_c^h) &= c(n, \alpha_w)|T|^{\alpha_w/2}|W|^{(\alpha_w - n - 1)/2} e^{-1/2\,\mathrm{tr}\{TW\}} \\
&\equiv \mathrm{Wishart}(W|\alpha_w, T)
\end{aligned} \tag{14}
$$

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

$$
c(n, \alpha_w) = \left[2^{\alpha_w n/2} \pi^{n(n-1)/4} \prod_{i=1}^{n} \Gamma\left(\frac{\alpha_w + 1 - i}{2} \right) \right]^{-1} \tag{15}
$$

[DeGroot (1970), page 57]. We provide interpretations for α_μ, α_w, ν and T later in this section. Note that in some expositions of the Wishart distribution, the inverse of T is used for the parameterization; T^{-1} is called the scale matrix [e.g., Press (1972), page 101].

This choice of a prior 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. Also define $W_{11.2} = W_{11} - W_{12}W_{22}^{-1}W_{12}'$ and recall that $((W^{-1})_{11})^{-1} = W_{11.2}$.

THEOREM 5. *If $f(\mu, W)$ is an n-dimensional normal-Wishart distribution, $n \geq 2$, with parameters ν, α_μ, α_w 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 $\{\mu_1, W_{11.2}\}$ is normal-Wishart with parameters ν_1, α_μ, T_{11}, and $\alpha_w - n + l$ 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 [(1972), pages 117–119]

carries out these computations for the Wishart distribution. Standard changes are needed to obtain the claim for the normal-Wishart distribution. A consequence of Theorem 5 is the following.

COROLLARY 6. *Let W be a $n \times n$ positive definite matrix of random variables. Let a, b, and c be three sets of indices of W. If $f(W_{ab.c}) = \text{Wishart}(W_{ab.c}|\alpha_1, T_1)$ and $f(W_{bc.a}) = \text{Wishart}(W_{bc.a}|\alpha_2, T_2)$, then $\alpha_1 - l_{ab} = \alpha_2 - l_{bc}$ where l_{ab} is the number of indices in the block a, b and l_{bc} is the number of indices in the block b, c.*

PROOF. The pdf for $W_{b.ac} = (W_{ab.c})_{b.a} = (W_{cb.a})_{b.c}$ is a Wishart distribution, and from the two alternative ways by which this pdf can be formed, using Theorem 5, it follows that $\alpha_1 - l_{ab} = \alpha_2 - l_{bc}$. □

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, $((W^{-1})_{11})^{-1} = W_{11} - W_{12} W_{22}^{-1} W_{12}' = W_{11.2}$ is the precision matrix associated with $X_1, \ldots, X_{n-1}$. Consequently, $\{m_n, b_n, v_n\}$ is independent of $\{\mu_1, W_{11.2}\}$. We now recursively repeat this argument with $\{\mu_1, W_{11.2}\}$ 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 choice of prior implies that the posterior $p(\mu, W|d, m_c^h)$ is also a normal-Wishart distribution [DeGroot (1970), page 178]. In particular, $p(\mu|W, d, m_c^h)$, where d is a sample of N complete cases, is multivariate normal with mean vector ν' given by

$$\nu' = \frac{\alpha_\mu \nu + N \overline{\mathbf{x}}_N}{\alpha_\mu + N} \tag{16}$$

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

$$R = T + S_N + \frac{\alpha_\mu N}{\alpha_\mu + N} (\nu - \overline{\mathbf{x}}_N)(\nu - \overline{\mathbf{x}}_N)', \tag{17}$$

where $S_N = \sum_{i=1}^{N} (\mathbf{x}_i - \overline{\mathbf{x}}_N)(\mathbf{x}_i - \overline{\mathbf{x}}_N)'$. From these equations, we see that α_μ and α_w can be thought of as effective sample sizes for μ and W, respectively.

In order to calculate the marginal likelihood of a Gaussian DAG model, we can work in the parametric space (μ, W). 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 $\alpha'_w = \alpha_w - n + l$, where $\mathbf{Y}$ is a subset of l coordinates. Thus, applying standard formulas pertaining to t-distributions [e.g., DeGroot (1970), pages 179–180], we obtain the terms in (9),

$$(18)\quad p(d^{\mathbf{Y}} | m_c^h) = (2\pi)^{-lN/2} \left(\frac{\alpha_\mu}{\alpha_\mu + N} \right)^{l/2} \frac{c(l, \alpha'_w)}{c(l, \alpha'_w + N)} |T_{\mathbf{Y}}|^{\alpha'_w/2} |R_{\mathbf{Y}}|^{-(\alpha'_w + N)/2},$$

where $R_{\mathbf{Y}} = ((R^{-1})_{\mathbf{YY}})^{-1}$ is the posterior parametric matrix restricted to the $\mathbf{Y}$ coordinates. Equations (9) and (18) together provide a way 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 (13)]. 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 Wishart$(1/v_n \mid \alpha_w + n - 1, T_{22} - T'_{12} T_{11}^{-1} T_{12})$ (i.e., a gamma distribution), and b_n has a multivariate normal pdf $N(b_n \mid T_{11}^{-1} T_{12}, T_{22}/v_n)$. Consequently, the degrees of freedom α_w and the parametric matrix T, which completely specify the Wishart prior distribution, are determined by the normal-gamma prior for the parameters of one regression model. Kadane, Dickey, Winkler, Smith and Peters (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 if these other priors have been elicited separately rather than computed. 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 effective sample size α_μ.

An alternative approach uses the observation that when $p(\mu, W | m_c^h)$ is normal-Wishart as we have described (with $\alpha_w > n + 1$), then $p(\mathbf{x} | m_c^h)$ is a multivariate t-distribution with γ degrees of freedom, location vector ν' and precision T', where

$$(19)\qquad \gamma = \alpha_w - n + 1, \qquad \nu' = \nu, \qquad T' = \frac{\alpha_\mu \gamma}{\alpha_\mu + 1} T^{-1}$$

[e.g., DeGroot (1970), page 180]. Thus, a person can assess the needed quantities by assessing α_μ, α_w, and a multivariate t-distribution for $\mathbf{X}$. Furthermore, rather than assess a multivariate t-distribution, which can be a difficult task, a person can—as an approximation—specify a multivariate-normal distribution having the same mean and covariance as the multivariate t-distribution. Note that the mean

and covariance of a multivariate t-distribution with γ degrees of freedom, location vector ν', and precision matrix T' is

$$\mathrm{E}(\mathbf{x}) = \nu', \qquad \mathrm{Cov}(\mathbf{x}) = \frac{\gamma}{\gamma - 2}\, T'^{-1} \tag{20}$$

[e.g., DeGroot (1970), pages 60, 61]. Finally, rather than assess a multivariate normal distribution directly, a person can assess a Gaussian DAG structure along with a value for each parameter. This method for constructing parameter priors for many DAG models has recently been applied to analysis of data in the domain of image compression [Thiesson, Meek, Chickering and Heckerman (1998)]. This method also provides a suitable Bayesian alternative for many of the examples discussed in Spirtes, Glymour and Scheines (2001).

5. 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 7). This theorem is phrased and proved 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 9 and 10).

We will use $\mathrm{tr}\{A + B\}$ to denote the sum of traces $\mathrm{tr}\{A\} + \mathrm{tr}\{B\}$ even when the dimensions of the square matrices A and B are different.

THEOREM 7. *Let W be an $n \times n$, $n \geq 3$, positive definite symmetric matrix of random variables and $f(W)$ be the 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 $W_{11.2} = W_{11} - W_{12}W_{22}^{-1}W_{12}'$ is independent of $\{W_{12}, W_{22}\}$ whenever $f(W)$ is a Wishart distribution is a well-known fact [Press (1972), pages 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 and since the transformations from $\{W_{11}, W_{12}, W_{22}\}$ to $\{W_{11.2}, W_{12}, W_{22}\}$ and to $\{W_{22.1}, W_{11}, W_{12}\}$ both have a Jacobian determinant of 1, we obtain the following functional equation:

$$\begin{aligned} f(W) &= f_1(W_{11} - W_{12}W_{22}^{-1}W_{12}')\, f_{2\|1}(W_{22}, W_{12}) \\ &= f_2(W_{22} - W_{12}'W_{11}^{-1}W_{12})\, f_{1\|2}(W_{11}, W_{12}), \end{aligned} \tag{21}$$

where a subscripted f denotes a pdf.

We divide the indices of W into two blocks, the first block (say, block 1) contains $n - 1$ indices and the second block (say, block 2) consists of one index. By the induction hypothesis, and since the independence conditions

on W also hold for $W_{11.2}$, we conclude that $W_{11.2}$ is distributed according to Wishart$(W_{11.2}|\ \alpha_1, T_1)$. Since this argument holds for every block of size $n-1$ of W, and since if a matrix V is distributed Wishart so is $V_{11.2}$ for any block of indices (Theorem 5), it follows that $W_{11.2}$ is distributed according to Wishart$(W_{11.2}|\ \alpha_1, T_1)$ also for blocks of size smaller than $n-1$.

Thus,

$$
\begin{aligned}
(22)\qquad & c_1|W_{11} - W_{12}W_{22}^{-1}W_{12}'|^{\beta_1} e^{\operatorname{tr}\{T_1(W_{11}-W_{12}W_{22}^{-1}W_{12}')\}} f_{2\|1}(W_{22}, W_{12}) \\
& \quad = c_2|W_{22} - W_{12}'W_{11}^{-1}W_{12}|^{\beta_2} e^{\operatorname{tr}\{T_2(W_{22}-W_{12}'W_{11}^{-1}W_{12})\}} f_{1\|2}(W_{11}, W_{12})
\end{aligned}
$$

where c_1 and c_2 are normalizing constants.

We now argue that $\beta_1 = \beta_2$. Divide the indices of W into three nonempty sets a, b, c such that block 1 consists of the indices in a, b and block 2 consists of the indices in c. The matrices $W_{ab.c}$ and $W_{bc.a}$ have a Wishart distribution, with, say, degrees of freedom α_1 and α_2, respectively, and so according to Corollary 6, $\alpha_1 - l_{ab} = \alpha_2 - l_{bc}$. Furthermore, $W_{c.ab} = (W_{bc.a})_{c.b}$ has a Wishart distribution with $\alpha_2 - l_{bc} + l_c$ degrees of freedom. Consequently, $\beta_1 = (\alpha_1 - l_{ab} - 1)/2$ is equal to $\beta_2 = (\alpha_2 - l_{bc} + l_c - l_c - 1)/2$. Let $\beta = \beta_1 = \beta_2$.

Define

$$
(23)\qquad F_{2\|1}(W_{22}, W_{12}) = c_1 f_{2\|1}(W_{22}, W_{12})/|W_{22}|^{\beta} e^{\operatorname{tr}\{T_2W_{22}+T_1(W_{12}W_{22}^{-1}W_{12}')\}},
$$

$$
(24)\qquad F_{1\|2}(W_{11}, W_{12}) = c_2 f_{1\|2}(W_{11}, W_{12})/|W_{11}|^{\beta} e^{\operatorname{tr}\{T_1W_{11}+T_2(W_{12}'W_{11}^{-1}W_{12})\}},
$$

substitute into (22), 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 (21), we obtain

$$
(25)\qquad f(W) = |W|^{\beta} e^{\operatorname{tr}\{T_1W_{11}+T_2W_{22}\}} H(W_{12})
$$

for some function H.

To show that $f(W)$ is Wishart we must find the form of H and show that it is proportional to $e^{2\operatorname{tr}\{T_{12}W_{12}\}}$ for some matrix T_{12}.

Considering the three possible pairs of blocks formed with the sets of indices a, b and c, (25) can be rewritten as follows:

$$
\begin{aligned}
(26)\qquad f(W) &= |W|^{\beta_1} e^{\operatorname{tr}\{T_{aa}W_{aa}+T_{bb}W_{bb}+T_{cc}W_{cc}\}} e^{2\operatorname{tr}\{T_{ab}'W_{ab}+T_{ac}'W_{ac}+T_{bc}'W_{bc}\}} \\
&\quad \times H_1(W_{ac}, W_{bc}),
\end{aligned}
$$

$$
\begin{aligned}
(27)\qquad f(W) &= |W|^{\beta_2} e^{\operatorname{tr}\{S_{aa}W_{aa}+S_{bb}W_{bb}+S_{cc}W_{cc}\}} e^{2\operatorname{tr}\{S_{ab}'W_{ab}+S_{ac}'W_{ac}+S_{bc}'W_{bc}\}} \\
&\quad \times H_2(W_{ab}, W_{bc}),
\end{aligned}
$$

$$
\begin{aligned}
(28)\qquad f(W) &= |W|^{\beta_3} e^{\operatorname{tr}\{R_{aa}W_{aa}+R_{bb}W_{bb}+R_{cc}W_{cc}\}} e^{2\operatorname{tr}\{R_{ab}'W_{ab}+R_{ac}'W_{ac}+R_{bc}'W_{bc}\}} \\
&\quad \times H_3(W_{ab}, W_{ac}).
\end{aligned}
$$

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 (26) and (27) we obtain

$$\begin{aligned} &e^{2\,\mathrm{tr}\{(T'_{ac}-S'_{ac})W_{ac}\}} H_1(W_{ac}, W_{bc}) \\ &\quad = e^{2\,\mathrm{tr}\{(S'_{ab}-T'_{ab})W_{ab}+(S'_{bc}-T'_{bc})W_{bc}\}} H_2(W_{ab}, W_{bc}). \end{aligned} \tag{29}$$

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

$$H_1(W_{ac}, W_{bc}) = H_{12}(W_{bc}) e^{2\,\mathrm{tr}\{(S'_{ac}-T'_{ac})W_{ac}\}}$$

and by symmetric arguments, comparing (26) and (28),

$$H_1(W_{ac}, W_{bc}) = H_{13}(W_{ac}) e^{2\,\mathrm{tr}\{(R'_{bc}-T'_{bc})W_{bc}\}}.$$

Consequently, $H_{12}(W_{bc})$ is proportional to $e^{2\,\mathrm{tr}\{(R'_{bc}-T'_{bc})W_{bc}\}}$ and so, substituting into (25), $f(W)$ is found to be 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 Appendix we show that given the independence conditions for two coordinates, f must have the form

$$f(W) = c|W|^{\beta} e^{\mathrm{tr}\{TW\}} H(W_{12}), \tag{30}$$

where H is an arbitrary function, and that the marginal distributions of $W_{11.2}$ and $W_{22.1}$ are one-dimensional Wishart distributions.

We now treat the case $n = 3$ using these assertions about the case $n = 2$. Starting with (21), 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 (30), and that f_2 is a one-dimensional Wishart distribution. Proceeding parallel to (22)–(24), we obtain

$$H(a_{12} - b_1 b_2 / W_{22}) F_{2\|1}(W_{22}, W_{12}) = F_{1\|2}(W_{11}, W_{12}), \tag{31}$$

where (b_1, b_2) is the matrix W_{12}, a_{12} is the off-diagonal element of W_{11}, $a_{12} - b_1 b_2 / W_{22}$ 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 left-hand side depends on W_{11} only through a_{12}. Thus also the right-hand side depends on W_{11} only through a_{12}. Let b_1 and b_2 be fixed, $y = b_1 b_2 / W_{22}$ and $x = a_{12}$. Also let $F(t) = F_{2\|1}(b_1 b_2 / t, (b_1, b_2))$ and $G(a_{12}) = F_{1\|2}(W_{11}, (b_1, b_2))$. We can now rewrite (31) as $H(x - y) F(y) = G(x)$. Now set $z = x - y$, and obtain for every y and z,

$$H(z) F(y) = G(y + z), \tag{32}$$

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

Substituting this form of H into (30), we see that $W_{11.2}$ has a two-dimensional Wishart distribution. Recall that $W_{22.1}$ has a one-dimensional Wishart distribution. Consequently, we can apply the induction step starting from (22) 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 8 (Skitovich–Darmois). *Let $z_1, \ldots, z_k$ be independent random variables and α_i, β_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 by Kagan, Linnik and Rao (1973).

THEOREM 9. *Let W be an $n \times n$, $n \geq 2$, positive definite symmetric matrix 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:

$$f(\mu) = f_1(\mu_1) f_{2\|1}(\mu_2 + W_{22}^{-1} W_{12}' \mu_1) = f_2(\mu_2) f_{1\|2}(\mu_1 + W_{11}^{-1} W_{12} \mu_2), \tag{33}$$

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 nonempty 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 (33), 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 (1970), page 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 nonzero because W is positive definite and 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 nonsingular linear transformation and so $\{\mu_1, \mu_2\}$ also has a joint normal distribution $f(\mu) = N(\mu|\eta, A)$ where $A = (a_{ij})$ is a 2×2 precision matrix. Substituting this solution into (33) 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 $\gamma > 0$. □

The proofs of Theorems 7 and 9 can be combined to form the following characterization of the normal-Wishart distribution.

THEOREM 10. *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 the 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 to the partitioning μ_1, μ_2.*

PROOF SKETCH. 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:

$$\begin{aligned} f(\mu, W) &= f_1(\mu_1, W_{11} - W_{12}W_{22}^{-1}W_{12}')\, f_{2\|1}(\mu_2 + W_{22}^{-1}W_{12}'\mu_1, W_{22}, W_{12}) \\ &= f_2(\mu_2, W_{22} - W_{12}'W_{11}^{-1}W_{12})\, f_{1\|2}(\mu_1 + W_{11}^{-1}W_{12}\mu_2, W_{11}, W_{12}), \end{aligned} \tag{34}$$

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 (33) the solution of which is

$$
\begin{aligned}
f(\mu, W) &\propto N\big(\mu|\eta(W), \gamma(W)W\big) \\
&= N\big(\mu_2|\eta_2(W), \gamma(W)[W_{22} - W_{12}'W_{11}^{-1}W_{12}]\big) \\
&\quad \times N\big(\mu_1|\eta_1(W) + \eta_2(W)W_{11}^{-1}W_{12} - W_{11}^{-1}W_{12}\mu_2, \gamma(W)W_{11}\big),
\end{aligned} \tag{35}
$$

where both γ and $\eta = (\eta_1, \eta_2)$ potentially can be functions of W. To see that these quantities in fact do not depend on W, first note that the normal distributions for μ_2 and μ_1 in (35) must be proportional to the functions f_2 and $f_{1\|2}$ in (34), respectively. Comparing the form of f_2 with the normal distribution for μ_2, we see that $\gamma(W)$ and $\eta_2(W)$ can only depend on $W_{22} - W_{12}'W_{11}^{-1}W_{12}$. Comparing the form of $f_{1\|2}$ with the normal distribution for μ_1, we see that $\gamma(W)$ and $\eta_2(W)$ can only depend on $\{W_{11}, W_{12}\}$. Consequently, $\gamma(W)$ and $\eta_2(W)$ must be constant. Similarly, $\eta_1(W)$ must be a constant. Substituting these solutions into (34) and dividing by the common terms which are equal to $f(\mu|W)$ yields (21), the solution of which for f is a Wishart pdf. $\square$

Note that the conditions set on W in Theorem 10, 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.

6. 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.

Consider the parameter prior for $\{m_n, b_n, v_n\}$ when the prior for $\{\mu, W\}$ is a normal-Wishart as specified by (14) and (15). By a change of variables, we get

$$
\begin{aligned}
&f_n(m_n, b_n, v_n) \\
&\quad = \text{Wishart}(1/v_n \mid \alpha + n - 1, T_{22} - T_{12}'T_{11}^{-1}T_{12}) \\
&\qquad \times N(b_n \mid T_{11}^{-1}T_{12}, T_{22}/v_n)N(m_n \mid \nu_n, \alpha_\mu/v_n),
\end{aligned}
$$

where the first block (T_{11}) corresponds to $X_1, \ldots, X_{n-1}$ and the second one-dimensional block (T_{22}) 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, $m_n^* = (m_n - \nu_n)/(\alpha_\mu/v_n)^{1/2}$ 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 11. *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 10, 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 Wishart$(W|\alpha, T)H(w_{12})$, where H is some real function and w_{12} is the off-diagonal element of W, satisfies global parameter independence (as shown in the Appendix) 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.

7. Discussion. The formula for the marginal likelihood applies whenever Assumptions 1–5 are satisfied, not only for Gaussian DAG models. Another important special case is when all variables in $\mathbf{X}$ are discrete and all local distributions are multinomial. This case has been treated in Heckerman and Geiger (1995) and Geiger and Heckerman (1997) under the additional assumption of local parameter independence which was introduced by Spiegelhalter and Lauritzen (1990). Our generalized derivation herein dispenses with this assumption and unifies the derivation in the discrete case with the derivation needed for Gaussian DAG models.

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. The difficulty with this approach is that the marginal likelihood no longer has closed form and therefore approximate methods such as MCMC are usually employed to compute the marginal likelihood. Also

the elicitation of hierarchical priors is often difficult. Other alternative approaches have been discussed at the end of Section 3.

We conclude with a technical comment. Equation (21), which encodes global parameter independence for an unknown covariance matrix, is an interesting example of a *matrix functional equation*. The domain of each unknown function is a nonsingular matrix and the range is R. A well-known functional equation of this sort is the equation

$$f(XY) = f(X)f(Y), \tag{36}$$

where X and Y are nonsingular matrices. The general solution of this equation is $f(X) = |X|^\alpha$ or $f(X) = |X|^\alpha \operatorname{sgn}(|X|)$ [e.g., Aczél (1966)]. When the domain of f is the set of positive definite matrices, the solution is simply $f(X) = |X|^\alpha$.

We note that the solution of (36) is obtained for matrices over arbitrary fields. Only algebraic manipulations are used in its proof. It seems reasonable to believe and interesting to investigate whether a solution to (21) can be obtained via purely algebraic manipulations. The proof technique that we have employed, however, especially for the base case of the induction, uses the fact that the matrices are over the real numbers.

APPENDIX

We now characterize the pdfs of an unknown 2×2 precision matrix that satisfy global parameter independence. This result has been obtained in Geiger and Heckerman (1998) under additional regularity conditions.

THEOREM 12. *Let W be a 2×2 positive definite symmetric matrix with random entries w_{11}, w_{12} and w_{22} and let $f(W)$ be the pdf of W. Then, $f(W) = |W|^\beta e^{\operatorname{tr}\{TW\}} H(w_{12})$ where H is a real function if and only if $w_{11} - w_{12}^2/w_{22}$ is independent of $\{w_{12}, w_{22}\}$ and $w_{22} - w_{12}^2/w_{11}$ is independent of $\{w_{12}, w_{11}\}$.*

PROOF. That $w_{11} - w_{12}^2/w_{22}$ is independent of $\{w_{12}, w_{22}\}$ whenever $f(W)$ is a Wishart distribution [e.g., when $H(x) = \text{constant}$] is a well-known fact [Press (1972), pages 117–119]. Consequently, this claim holds for any real function H. We prove the other direction by solving the functional equation, which is implied by the given independence assumptions,

$$\begin{aligned} f(W) &= f_1(w_{11} - w_{12}^2/w_{22}) f_{2\|1}(w_{22}, w_{12}) \\ &= f_2(w_{22} - w_{12}^2/w_{11}) f_{1\|2}(w_{11}, w_{12}), \end{aligned} \tag{37}$$

where a subscripted f denotes a pdf. To solve this functional equation, namely to find all pdfs that satisfy it, we use techniques described in Aczél (1966) and results from Járai (1986, 1998).

Let w_{12} be a value such that the integral of $f_{2\|1}(x, w_{12})$ over the domain of x is not identically zero. Such a value for w_{12} exists because $f_{2\|1}(x, w_{12})$ integrates to 1 over its domain. Without loss of generality, suppose this value of w_{12} is 1, lest we can modify the scale using the transformations $w_{11} \leftarrow w_{12}w_{11}$ and $w_{22} \leftarrow w_{12}w_{22}$. We rewrite (37) as

$$f_1(w_{11} - 1/w_{22}) f_{2\|1}(w_{22}, 1) = f_2(w_{22} - 1/w_{11}) f_{1\|2}(w_{11}, 1). \tag{38}$$

We claim that all density functions satisfying (38) must be positive everywhere and smooth. This is shown in Lemmas 14 and 16 at the end of the proof. Consequently, we can take the logarithm of (38) and then take derivatives. First, we take the logarithm and rename the functions. We get

$$g_1(w_{11} - 1/w_{22}) + g_{2\|1}(w_{22}) = g_2(w_{22} - 1/w_{11}) + g_{1\|2}(w_{11}), \tag{39}$$

where $g_1(x) = \ln f_1(x)$, $g_{2\|1}(x) = \ln f_{2\|1}(x, 1)$, and where g_2 and $g_{1\|2}$ are defined analogously.

We take a mixed second derivative with respect to w_{11} and w_{22} of (39). We get

$$g_1''(w_{11} - 1/w_{22})/w_{22}^2 = g_2''(w_{22} - 1/w_{11})/w_{11}^2. \tag{40}$$

By substituting $w_{11} = w_{22}$ we obtain $g_1'' = g_2''$. We denote this function by h and so,

$$w_{11}^2 h(w_{11} - 1/w_{22}) = w_{22}^2 h(w_{22} - 1/w_{11}). \tag{41}$$

It is easy to show, using this functional equation for h, that if h is zero at some point then h must be identically zero; if h is positive at one point then h is positive everywhere, and if h is negative at one point then h is negative everywhere. We now take a derivative wrt w_{11} and a derivative with respect to w_{22},

$$2w_{11}h(w_{11} - 1/w_{22}) + w_{11}^2 h'(w_{11} - 1/w_{22}) = \{w_{22}/w_{11}\}^2 h'(w_{22} - 1/w_{11}),$$
$$2w_{22}h(w_{22} - 1/w_{11}) + w_{22}^2 h'(w_{22} - 1/w_{11}) = \{w_{11}/w_{22}\}^2 h'(w_{11} - 1/w_{22}).$$

From these equations, and using (41) we get

$$2(w_{22} + 1/w_{11})h(w_{22} - 1/w_{11}) = -(w_{22}^2 - 1/w_{11}^2)h'(w_{22} - 1/w_{11}).$$

Consequently,

$$h'(x)/h(x) = -2/x,$$

where $x = w_{22} - 1/w_{11}$. This equation holds for every $x \in R^+$. Assuming h is positive everywhere, we have $(\ln h(x))' = -2/x$ and so $\ln h(x) = \ln x^{-2} + c'$ where c' is a constant. If h is negative everywhere, we have $(\ln -h(x))' = -2/x$ and so $\ln(-h(x)) = \ln x^{-2} + c'$. Consequently, whether h is positive everywhere, negative everywhere, or identically zero, it has the form $h(x) = c/x^2$ where c is a constant. Recall that $h = (\ln f_1)''$. Hence, $f_1(x) = c_1 x^{-c} e^{c_2 x}$ and similarly $f_2(x) = c_1' x^{-c} e^{c_2' x}$ (i.e., one-dimensional Wishart distributions with the same

degrees of freedom). We conclude the proof by substituting f_1 and f_2 into (37) and proceeding as in (22)–(25). $\square$

The next lemma shows that every positive everywhere pdf that satisfies (38) must be smooth. Our lemma is an immediate consequence of Járai's theorem which we now state.

THEOREM 13 [Járai (1986, 1998)]. *Let X_i be an open subset of R^{r_i} $(i = 1, 2, \ldots, n)$, T be an open subset of R^s, Y be an open subset of R^k, Z_i be an open subset of R^{m_i} $(i = 1, 2, \ldots, n)$, D be an open subset of $T \times Y$ and let Z be a Euclidean space. Consider the functions $f : T \to Z$, $g_i : D \to X_i$, $f_i : X_i \to Z_i$, $h_i : D \times Z_i \to Z$ $(i = 1, 2, \ldots, n)$. Suppose that $0 \le p \le \infty$ and*:

(i) *for each $(t, y) \in D$,*

$$f(t) = \sum_{i=1}^{n} h_i\big(t, y, f_i\big(g_i(t, y)\big)\big);$$

(ii) *h_i is $p + 1$ times continuously differentiable $(1 \le i \le n)$;*

(iii) *g_i is $p+2$ times continuously differentiable and for each $t \in T$ there exists a $y \in Y$ such that $(t, y) \in D$ and $\frac{\partial g_i}{\partial y}(t, y)$ has rank r_i $(1 \le i \le n)$.*

Then:

(iv) *if f_i $(i = 1, 2, \ldots, n)$ is Lebesgue measurable and* (ii), (iii) *are satisfied with $p = 0$ then f is continuous on T*;

(v) *if f_i $(i = 1, 2, \ldots, n)$ is continuous and* (ii), (iii) *are satisfied with $p = 0$ then f is continuously differentiable on T*;

(vi) *if f_i $(i = 1, 2, \ldots, n)$ is p times continuously differentiable and* (ii), (iii) *are satisfied then f is $p + 1$ times continuously differentiable on T.*

This theorem is stated in Járai (1998) and its proof is based on Theorems 3.3, 5.2 and 7.2 of Járai (1986). A simple corollary of Járai 's theorem is the following.

LEMMA 14. *All Lebesgue measurable real functions $l_1, l_2, l_{1\|2}$ and $l_{2\|1}$ defined on R^+ which satisfy*

$$l_1(y - 1/t) + l_{2\|1}(t) = l_2(t - 1/y) + l_{1\|2}(y) \tag{42}$$

for every $y, t > 0$ such that $yt > 1$, are p times continuously differentiable where p is arbitrarily large.

PROOF. The proof follows closely the lines of reasoning that Járai (1998) applied to another functional equation.

Using statement (iv) of Theorem 13 we show that $l_{2\|1}$ is continuous. To match Járai's theorem notation we define $f = l_{2\|1}$, $f_1 = -l_1$, $f_2 = l_2$, $f_3 = l_{1\|2}$,

$h_i(t, y, w) = w$ for $i = 1, 2, 3$, $g_1(t, y) = (y - 1/t)$, $g_2(t, y) = (t - 1/y)$ and $g_3(t, y) = y$. The only nonobvious condition to check is that for each $t \in R^+$ there exists a $y \in R^+$ such that $ty > 1$ and $\frac{\partial g_i}{\partial y}(t, y)$ has rank r_i, $1 \le i \le n$. But here the rank is 1 and so we just need to observe that there exists a y such that $\frac{\partial g_i}{\partial y}(t, y)$ is not zero.

To show that l_1 is continuous, rewrite (42) as

$$l_1(t) + l_{2\|1}(y) = l_2\left(\frac{ty^2}{ty+1}\right) + l_{1\|2}(t + 1/y), \tag{43}$$

where $t, y > 0$. Now define $f = l_1$, $f_1 = -l_{2\|1}$, $f_2 = l_2$, $f_3 = l_{1\|2}$, $h_i(t, y, w) = w$ for $i = 1, 2, 3$, $g_1(t, y) = y$, $g_2(t, y) = \frac{ty^2}{ty+1}$ and $g_3(t, y) = t + 1/y$. Observe that the conditions of Járai's theorem hold and so $f = l_1$ is continuous. By the symmetry of the equation, l_2 and $l_{1\|2}$ are also continuous on R^+.

Now we can apply statement (v) of Járai's theorem. We obtain, in the same way as above, that all four functions are continuously differentiable. Finally, applying statement (vi) of Járai's theorem in the same way, we get that all four functions are twice continuously differentiable. Repeating this process shows that all four functions are p times continuously differentiable for every $p > 0$. □

The next theorem and lemma show that every pdf that satisfies (38) must be positive everywhere and so taking the logarithm of this equation, as we have done, is legitimate. We denote by λ^s the s-dimensional Lebesgue measure and by λ the one-dimensional Lebesgue measure.

THEOREM 15 [Járai (1995, 1998)]. *Let $X_1, \ldots, X_n$ be orthogonal subspaces of R^r of dimensions $r_1, \ldots, r_n$, respectively. Suppose that $r_i \ge 1$ $(1 \le i \le n)$ and $\sum_{i=1}^n r_i = r$. Let U be an open subset of R^r and $F : U \to R^m$ be a continuously differentiable function. For each $x \in U$, let N_x denote the nullspace of $F'(x)$. Let p_i denote the orthogonal projection of X onto X_i. Suppose that* $\dim N_x = r - m$ *and $p_i(N_x) = X_i$ $(i = 1, \ldots, n)$ for all $x \in U$. Let A_i be a subset of X_i $(i = 1, \ldots, n)$. If $A_1 \times A_2 \times \cdots \times A_n \subset U$ and A_i is λ^{r_i} measurable with $\lambda^{r_i}(A_i) > 0$ $(1 \le i \le n)$, then $F(A_1 \times A_2 \times \cdots \times A_n)$ contains a nonempty open set.*

Recall that if $X_1, \ldots, X_n$ are the standard orthogonal axes of R^n, then $p_i(X_1, \ldots, X_n) = X_i$, and $P_i(N_x) = \{x | (X_1, \ldots, X_{i-1}, x, X_{i+1}, \ldots, X_n) \in N_x\}$.

LEMMA 16. *Let f, g, h, k be nonnegative real functions that are Lebesgue integrable with integral $c > 0$. If these functions satisfy*

$$f(s - 1/t)g(t) = h(t - 1/s)k(s) \tag{44}$$

for every $s, t > 0$ such that $st > 1$, then they are everywhere positive.

PROOF. The proof follows closely the lines of reasoning that Járai (1998) applied to another functional equation.

Let $\{f = 0\}$ denote the set of points in the domain of f for which f is zero and let $\{f \neq 0\}$ denote the complementary set of all points in the domain for which f is not zero, namely, the set of points for which f is positive. Similar notation is used for the functions g, h and k. The idea of the proof is to show that the set $\{f = 0\}$ and the set $\{f \neq 0\}$ are both open and therefore, since the domain of f is connected, one of these sets must be empty. The set $\{f \neq 0\}$ cannot be empty because f is nonnegative and integrates to a positive constant and so $\{f = 0\}$ must be empty as claimed by the theorem. Similar arguments show that g, h and k are also positive everywhere.

The proof proceeds in three steps. First, we use Theorem 15 to establish that the set $\{g \neq 0\}$ contains a nonempty open set (i.e., it contains an inner point). Then we show that every point in $\{f \neq 0\}$ is an inner point and so $\{f \neq 0\}$ is open. Finally, we show that every point in $\{f = 0\}$ is an inner point and so $\{f = 0\}$ is open as well. Similar arguments work for g, h and k.

We start by rewriting (44) in two symmetric ways. First as

$$f(y)g(z) = h(x(y,z))k(w(y,z)) \tag{45}$$

for all $y > 0$ and $z > 0$, where $x(y,z) = yz^2/(yz+1)$ and $w(y,z) = y + 1/z$; second as

$$f(y(x,w))g(z(x,w)) = h(x)k(w) \tag{46}$$

for all $x > 0$, and $w > 0$ where $y(x,w) = xw^2/(xw+1)$ and $z(x,w) = x + 1/w$.

Step 1. We show that $\{g \neq 0\}$ contains an inner point. Since both h and k integrate to a positive constant, there must exist two λ-measurable sets A_h in $\{h \neq 0\}$ and A_k in $\{k \neq 0\}$ such that $\lambda(A_h) > 0$ and $\lambda(A_k) > 0$. The image of these sets under $z(x,w) = x + 1/w$ contains an inner point z according to Theorem 15. This theorem is applicable because the nullspace of z' is $\{a(1/w^2, 1))|a > 0\}$ and its projection on either of the two coordinates is R^+. Due to (46), and because the right-hand side is not zero for any $x \in A_h$ and $w \in A_k$, each term on the left-hand side is also not zero. Consequently, their image under $z(x,w)$, which includes an inner point, belongs to $\{g \neq 0\}$.

Step 2. Let y be an arbitrary point in $\{f \neq 0\}$. We now show that y is an inner point and so $\{f \neq 0\}$ is open. Let z be an inner point in $\{g \neq 0\}$. It follows that the image of a sufficiently small open set containing z under $x(y,z) = yz^2/(yz+1)$ and the image under $w(y,z) = y + 1/z$ are open sets. These images belong to $\{h \neq 0\}$ and $\{k \neq 0\}$, respectively, because the left-hand side of (45) is positive. Now we fix x in the image and vary w in a small open neighborhood. Then y is varied in a small open neighborhood. Since the right-hand side of (45) is positive, the neighborhood of y belongs to $\{f \neq 0\}$ and so y is an inner point. Similar arguments show that $\{g \neq 0\}$ is open as well. By the symmetry of (44) the same claim holds for h and k.

Step 3. Let y be an arbitrary point in $\{f = 0\}$. We now show that y is an inner point and so $\{f = 0\}$ is open. Let z be an inner point in $\{g \neq 0\}$. It follows that the image of a sufficiently small open set containing z under $x(y, z) = yz^2/(yz + 1)$ and the image under $w(y, z) = y + 1/z$ are open sets. Since the left-hand side of (45) is zero, at least one term in the right-hand side must be zero. If x is in $\{h = 0\}$, then fix x. As we vary w in a small open neighborhood in the image, g remains positive due to continuity. Also y is varied in a small open neighborhood. Since the right-hand side of (45) is zero, the neighborhood of y belongs to $\{f = 0\}$ and so y is an inner point. The other case occurs when w is in $\{k = 0\}$, in which case we fix w and vary x in a small neighborhood. Similar arguments show that $\{g = 0\}$ is open as well. By the symmetry of (44), the same claim holds for h and k. □

Note that Lemmas 14 and 16 together imply that every pdf that solves (44) must be positive everywhere and smooth.

Acknowledgments. We thank Chris Meek for helping us shape the definition of DAG models and correcting earlier versions of this manuscript, Bo Thiesson for implementing the proposed scheme, and Jim Kajiya for his help in regard to the characterization theorems. We also thank János Aczél, Enrique Castillo, Clark Glymour, Antal Járai, Gérard Letac, Hélène, Peter Spirtes and the reviewers for their useful suggestions. The first four sections of this work with different emphasizes [including, e.g., details of the derivation of equation (18)] have been reported in Geiger and Heckerman (1994) and Heckerman and Geiger (1995). A short version of this work appeared in Geiger and Heckerman (1999).

REFERENCES

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

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541.

BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, New York.

BUNTINE, W. (1994). Operations for learning with graphical models. *J. Artificial Intelligence Research* **2** 159–225.

CHICKERING, D. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence*, Montreal 87–98. Morgan Kaufmann, San Francisco.

CHICKERING, D. (1996). Learning Bayesian networks from data. Ph.D. dissertation, Univ. California, Los Angeles.

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

COWELL, R., DAWID, A. P., LAURITZEN, S. and SPIEGELHALTER, D. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

DAWID, A. P. and LAURITZEN, S. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317.

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

FRIEDMAN, N. and GOLDSZMIDT, M. (1997). Sequential update of Bayesian network structures. In *Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence* 165–174. Morgan Kaufmann, Providence, RI.

GEIGER, D. and HECKERMAN, D. (1994). Learning Gaussian networks. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence* 235–243. Morgan Kaufmann, San Francisco.

GEIGER, D. and HECKERMAN, D. (1997). A characterization of the Dirichlet distribution through global and local parameter independence. *Ann. Statist.* **25** 1344–1369.

GEIGER, D. and HECKERMAN, D. (1998). A characterization of the bivariate Wishart distribution. *Probab. Math. Statist.* **18** 119–131.

GEIGER, D. and HECKERMAN, D. (1999). Parameter priors for directed graphical models and the characterization of several probability distributions. In *Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence* 216–225. Morgan Kaufmann, San Francisco.

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* 274–284. Morgan Kaufmann, San Francisco.

HECKERMAN, D., GEIGER, D. and CHICKERING, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20** 197–243.

HECKERMAN, D., MAMDANI, A. and WELLMAN, M. (1995). Real-world applications of Bayesian networks. *Comm. ACM* **38**.

HOWARD, R. and MATHESON, J. (1981). Influence diagrams. In *The Principles and Applications of Decision Analysis* **2** (R. Howard and J. Matheson, eds.) 721–762. Strategic Decisions Group, Menlo Park, CA.

JÁRAI, A. (1986). On regular solutions of functional equations. *Aequationes Math.* **30** 21–54.

JÁRAI, A. (1998). Regularity property of the functional equation of the Dirichlet distribution. *Aequationes Math.* **56** 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. *J. Amer. Statist. Assoc.* **75** 845–854.

KAGAN, A. M., LINNIK, Y. V. and RAO, C. R. (1973). *Characterization Problems in Mathematical Statistics*. Wiley, New York.

MADIGAN, D., ANDERSSON, S. A., PERLMAN, M. D. and VOLINSKY, C. T. (1996). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm. Statist. Theory Methods* **25** 2493–2519.

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

PRESS, J. S. (1972). *Applied Multivariate Analysis*. Holt, Rinehart and Winston, New York.

SHACHTER, R. and KENLEY, C. (1989). Gaussian influence diagrams. *Management Sci.* **35** 527–550.

SPIEGELHALTER, D., DAWID, A., LAURITZEN, S. and COWELL, R. (1993). Bayesian analysis in expert systems (with discussion). *Statist. Sci.* **8** 219–283.

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. (2001). *Causation, Prediction, and Search*. MIT Press.

SPIRTES, P. and MEEK, C. (1995). Learning Bayesian networks with discrete variables from data. In *Proceedings of First International Conference on Knowledge Discovery and Data Mining* 294–299. Morgan Kaufmann, San Francisco.

THIESSON, B., MEEK, C., CHICKERING, D. and HECKERMAN, D. (1998). Computationally efficient methods for selecting among mixtures of graphical models. In *Bayesian Statistics 6* (J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 631–656. Clarendon Press, Oxford.

VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence* 220–227. Morgan Kaufmann, San Francisco.

COMPUTER SCIENCE DEPARTMENT
TECHNION – ISRAEL INSTITUTE
OF TECHNOLOGY
BUILDING TAUB 616
HAIFA 32000
ISRAEL
E-MAIL: dang@cs.technion.ac.il

MICROSOFT RESEARCH
REDMOND, WASHINGTON 98052-6399
E-MAIL: heckerma@microsoft.com