Learning Bayesian Networks: A Unification for Discrete and Gaussian Domains

David Heckerman and Dan Geiger*

Microsoft Research, Bldg 9S/1
Redmond 98052-6399, WA
heckerma@microsoft.com,dang@cs.technion.ac.il

Abstract

We examine Bayesian methods for learning Bayesian networks from a combination of prior knowledge and statistical data. In particular, we unify the approaches we presented at last year's conference for discrete and Gaussian domains. We derive a general Bayesian scoring metric, appropriate for both domains. We then use this metric in combination with well-known statistical facts about the Dirichlet and normal–Wishart distributions to derive our metrics for discrete and Gaussian domains.

1 Introduction

At last year's conference, we presented approaches for learning Bayesian networks from a combination of prior knowledge and statistical data. These approaches were presented in two papers: one addressing domains containing only discrete variables (Heckerman et al., 1994), and the other addressing domains containing continuous variables related by an unknown multivariate-Gaussian distribution (Geiger and Heckerman, 1994). Unfortunately, these presentations were substantially different, making the parallels between the two methods difficult to appreciate. In this paper, we unify the two approaches. In particular, we abstract our previous assumptions of likelihood equivalence, parameter modularity, and parameter independence such that they are appropriate for discrete and Gaussian domains (as well as other domains). Using these assumptions, we derive a domain-independent Bayesian scoring metric. We then use this general metric in combination with well-known statistical facts about the Dirichlet and normal–Wishart distributions to derive our metrics for discrete and Gaussian domains. In addition, we provide simple proofs that these assumptions are consistent for both domains.

Throughout this discussion, we consider a domain U of n variables $x_1, \ldots, x_n$. Each variable may be discrete—having a finite or countable number of states—or continuous. We use lower-case letters to refer to variables and upper-case letters to refer to sets of variables. We write $x_i = k$ to denote that variable x_i is in state k. When we observe the state for every variable in set X, we call this set of observations a *state* of X; and we write $X = k_X$ as a shorthand for the observations $x_i = k_i, x_i \in X$. The *joint space* of U is the set of all states of U. We use $p(X = k_X | Y = k_Y, \xi)$ to denote the generalized probability density that $X = k_X$ given $Y = k_Y$ for a person with current state of information ξ [DeGroot, 1970, p. 19]. We use $p(X|Y, \xi)$ to denote the generalized probability density function (gpdf) for X, given all possible observations of Y. The *joint gpdf* over U is the gpdf for U.

We use B_s to denote the structure of a Bayesian network, and Π_i to denote the parents of x_i in a given network. We assume the reader is familiar with Bayesian networks for the case where all variables in U are discrete. Here, we describe a Bayesian-network representation for continuous variables. In particular, consider the special case where all the variables in U are continuous and the joint probability density function for U is a multivariate (nonsingular) normal distribution. In this case, to be in line with more standard notation, we use $\vec{x}$ to denote the set of variables U. We have

$$\begin{aligned} p(\vec{x}|\xi) &= n(\vec{\mu}, \Sigma^{-1}) \\ &\equiv (2\pi)^{-n/2} |\Sigma|^{-1/2} e^{-1/2(\vec{x}-\vec{\mu})' \Sigma^{-1} (\vec{x}-\vec{\mu})} \end{aligned} \quad (1)$$

where $\vec{\mu}$ is an n-dimensional mean vector, and $\Sigma = (\sigma_{ij})$ is an $n \times n$ covariance matrix, which must be both symmetric and positive definite. Both $\vec{\mu}$ and Σ are implicitly functions of ξ. We shall find it convenient to refer to the *precision matrix* $W = \Sigma^{-1}$, whose elements are denoted by w_{ij}.

*Author's primary affiliation: Computer Science Department, Technion, Haifa 32000, Israel.

This joint density function can be written as a product of conditional density functions each being a normal distribution. Namely,

$$p(\vec{x}|\xi) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}, \xi) \tag{2}$$

$$p(x_i|x_1, \ldots, x_{i-1}, \xi) = n(\mu_i + \sum_{j=1}^{i-1} b_{ji}(x_j - \mu_j), 1/v_i) \tag{3}$$

where μ_i is the unconditional mean of x_i (i.e., the ith component of $\vec{\mu}$), v_i is the conditional variance of x_i given values for $x_1, \ldots, x_{i-1}$, and b_{ji} is a linear coefficient reflecting the strength of the relationship between x_j and x_i (e.g., DeGroot, p. 55).

Thus, we may interpret a multivariate-normal distribution as a Bayesian network, where there is no arc from x_j to x_i whenever $b_{ji} = 0$, $j < i$. Conversely, from a Bayesian network with conditional distributions satisfying Equation 3, we may construct a multivariate-normal distribution. We call this special form of a Bayesian network a *Gaussian network.* The name is adopted from Shachter and Kenley (1989) who first described Gaussian influence diagrams. We note that, in practice, it is typically easier to assess a Gaussian network than it is to assess directly a symmetric positive-definite precision matrix.

The transformations between $\vec{v} = \{v_1, \ldots, v_n\}$ and $B \equiv \{b_{ji} \mid j < i\}$ of a given Gaussian network G and the precision matrix W of the normal distribution represented by G are well known. In this paper, we need only the transformation from W to $\{\vec{v}, B\}$. We use the following recursive form given by Shachter and Kenley (1989). Let $W(i)$ denote the $i \times i$ upper left submatrix of W, $\vec{b}_i$ denote the column vector $(b_{1i}, \ldots, b_{i-1,i})$, and $\vec{b}_i'$ denote the transposition of $\vec{b}_i$. Then, for $i > 1$, we have

$$W(i+1) = \begin{pmatrix} W(i) + \frac{\vec{b}_{i+1}\vec{b}_{i+1}'}{v_{i+1}} & -\frac{\vec{b}_{i+1}}{v_{i+1}} \\ -\frac{\vec{b}_{i+1}'}{v_{i+1}} & \frac{1}{v_{i+1}} \end{pmatrix} \tag{4}$$

and $W(1) = \frac{1}{v_1}$.

Although Equation 3 is useful for the assessment of a Gaussian network, we shall sometimes find it convenient to write

$$p(x_i|x_1, \ldots, x_{i-1}, \xi) = n(\vec{m}_i + \sum_{j=1}^{i-1} b_{ji}x_j, 1/v_i) \tag{5}$$

where $m_i, i = 1, \ldots, n$ is defined by

$$m_i = \mu_i - \sum_{j=1}^{i-1} b_{ji}\mu_j \tag{6}$$

Note that m_i is the mean of x_i when all of x_i's parents are equal to zero.

As an example, given the three-node network structure $x_1 \rightarrow x_3 \leftarrow x_2$, we have $b_{12} = 0$, $x_1 = n(m_1, 1/v_1), x_2 = n(m_2, 1/v_2)$, and $x_3 = n(m_3 + b_{13}(x_1 - m_1) + b_{23}(x_2 - m_2), 1/v_3)$. Also, the precision matrix corresponding to this network structure is given by

$$W = \begin{pmatrix} \frac{1}{v_1} + \frac{b_{13}^2}{v_3} & \frac{b_{13}b_{23}}{v_3} & -\frac{b_{13}}{v_3} \\ \frac{b_{13}b_{23}}{v_3} & \frac{1}{v_2} + \frac{b_{23}^2}{v_3} & -\frac{b_{23}}{v_3} \\ -\frac{b_{13}}{v_3} & -\frac{b_{23}}{v_3} & \frac{1}{v_3} \end{pmatrix} \tag{7}$$

Finally, it is important to note that two or more Bayesian-network structures for a given domain can be *equivalent* in the sense that the structures represent the same set of gpdfs for the domain (Verma and Pearl, 1990). For example, for the three variable domain $\{x, y, z\}$, each of the network structures $x \rightarrow y \rightarrow z$, $x \leftarrow y \rightarrow z$, and $x \leftarrow y \leftarrow z$ represents the gpdfs where x and z are conditionally independent of y, and are therefore equivalent. As another example, a *complete network structure* is one that has no missing edges. In a domain with n variables, there are $n!$ complete network structures. All complete network structures for a given domain represent the same set of gpdfs—namely, all possible gpdfs—and are therefore equivalent. In our proofs to follow, we require the following characterization of equivalent networks, proved by Chickering (in this proceedings).

Theorem 1 (Chickering, 1995) *Let B_{s1} and B_{s2} be two Bayesian-network structures, and $R_{B_{s1},B_{s2}}$ be the set of edges by which B_{s1} and B_{s2} differ in directionality. Then, B_{s1} and B_{s2} are equivalent if and only if there exists a sequence of $|R_{B_{s1},B_{s2}}|$ distinct arc reversals applied to B_{s1} with the following properties:*

1. *After each reversal, the resulting network structure contains no directed cycles and is equivalent to B_{s2}*
2. *After all reversals, the resulting network structure is identical to B_{s2}*
3. *If $x \rightarrow y$ is the next arc to be reversed in the current network structure, then x and y have the same parents in both network structures, with the exception that x is also a parent of y in B_{s1}*

2 A Bayesian Approach for Learning Bayesian Networks

Our Bayesian approach for learning Bayesian networks can be understood as follows. Suppose we have a domain of variables $\{x_1, \ldots, x_n\} = U$, and a set of cases

$\{C_1, \ldots, C_m\} = D$ where each case is a state of some or of all the variables in U. We sometimes refer to D as a database. We begin with the following *random-sample assumption*: the database is a random sample from some sample distribution with unknown parameters Θ_U, and this sample distribution satisfies the conditional-independence assertions of some network structure B_s for U. We define B_s^h to be the hypothesis that the sample distribution can be encoded in B_s.

Now, suppose that we wish to determine the gpdf $p(C|D,\xi)$—the generalized probability density function for a new case C, given the database and our current state of information ξ. Rather than reason about this distribution directly, we assume that the collection of hypotheses B_s^h corresponding to all network structures for U form a mutually exclusive and collectively exhaustive set[1] and compute

$$p(C|D,\xi) = \sum_{\text{all } B_s^h} p(C|D, B_s^h, \xi) \cdot p(B_s^h|D,\xi)$$

In practice, it is impossible to sum over all possible network structures. Consequently, we attempt to identify a small subset H of network-structure hypotheses that account for a large fraction of the posterior probability of the hypotheses. Rewriting the previous equation using the fact that $p(B_s^h|D,\xi) = p(D, B_s^h|\xi)/p(D|\xi)$, we obtain

$$p(C|D,\xi) \approx c \sum_{B_s^h \in H} p(C|D, B_s^h, \xi) \cdot p(D, B_s^h|\xi)$$

where c is the normalization constant $1/[\sum_{B_s^h \in H} p(D, B_s^h|\xi)]$. From this relation, we see that only the relative posterior probabilities $p(D, B_s^h|\xi)$ matter. Thus, we compute this relative posterior probability, or alternatively, a *Bayes' factor*—$p(B_s^h|D,\xi)/p(B_{s0}^h|D,\xi)$—where B_{s0} is some reference structure such as the empty graph. We call methods for computing these relative posterior probabilities Bayesian scoring metrics.

Extending the Bayesian analysis, we use Θ_{Bs} to denote the parameters of the sample distribution encoded in the network structure B_s given hypothesis B_s^h. That is, the parameters Θ_{Bs} determine the local gpdfs in B_p. From the rules of probability, we have

$$p(D, B_s^h|\xi) = p(B_s^h|\xi) \cdot \int p(\Theta_{Bs}|B_s^h,\xi)\; p(D|\Theta_{Bs}, B_s^h, \xi)\; d\Theta_{Bs} \tag{8}$$

The assessment of the network-structure priors $p(B_s^h|\xi)$ is treated elsewhere (e.g., Buntine, 1991, and Heckerman et al., 1995). In the following section, we introduce a set of assumptions that simplifies the assessment of the network-parameter priors $p(\Theta_{Bs}|B_s^h,\xi)$. In the remainder of this section, we show how to compute $p(D|\Theta_{Bs}, B_s^h, \xi)$.

A method for computing this term follows from our random-sample assumption. Namely, given hypothesis B_s^h, it follows that D can be separated into a set of random samples, where these random samples are determined by the structure of B_s. First, let us examine this decomposition when all the variables in U are discrete. Let $\theta_{X=k_X|Y=k_Y}$ denote the parameter corresponding to the probability $p(X = k_X|Y = k_Y, \xi)$, where X and Y are disjoint subsets of U. In addition, let x_{il} and Π_{il} denote the variable x_i and the parent set Π_i in the lth case, respectively; and let D_l denote the first $l-1$ cases in the database. Then, given B_s^h, we know that the observations of x_i in those cases where $\Pi_{il} = k_{\Pi_i}$ is a random sample with parameters $\Theta_{x_{il}|\Pi_{il}=k_{\Pi_i}}$. That is,

$$p(x_{il} = k_i|x_{1l} = k_1, \ldots, x_{(i-1)l} = k_{i-1}, D_l, \Theta_{Bs}, B_s^h, \xi) = \theta_{x_{il}=k_i|\Pi_{il}=k_{\Pi_i}} \tag{9}$$

where k_{Π_i} is the state of Π_{il} consistent with $\{x_{1l} = k_1, \ldots, x_{(i-1)l} = k_{i-1}\}$. Using Equation 9, we can compute $p(D|\Theta_{Bs}, B_s^h, \xi)$ for any database D and network structure B_s for discrete domain U.

Now consider a domain of continuous variables $\vec{x} = \{x_1, \ldots, x_n\}$, and suppose the database D is a random sample from a multivariate-normal distribution with parameters $\Theta_U = \{\vec{\mu}, W\}$. From our discussion in Section 1, it follows that, given hypothesis B_s^h, each variable x_i is a random sample from a normal distribution with mean $m_i + \sum_{x_j \in \Pi_i} b_{ji} x_j$ and variance v_i. Thus, with $\Theta_{Bs} = \{\vec{m}, B, \vec{v}\}$, we have

$$p(x_{il}|x_{1l}, \ldots, x_{(i-1)l}, D_l, \Theta_{Bs}, B_s^h, \xi) = n(m_i + \sum_{x_j \in \Pi_i} b_{ji} x_{jl}, 1/v_i) \tag{10}$$

Using Equation 10, we can compute $p(D|\Theta_{Bs}, B_s^h, \xi)$ for any D and B_s in a Gaussian domain.

The generalization of Equations 9 and 10 is straightforward, and we state it as our first formal assumption.

Assumption 1 (Random Sample) *Let $D = \{C_1, \ldots, C_m\}$ be a database, and B_s be a network structure for U determined by variable ordering $(x_1, \ldots, x_n)$. Let $\Theta(x_i, \Pi_i)$ denote the parameters of the network associated with variable x_i. Then, for all variables $x_i \in U$,*

$$p(x_{il}|x_{1l}, \ldots, x_{(i-1)l}, D_l, \Theta_{Bs}, B_s^h, \xi) = f(\Theta(x_i, \Pi_i), x_{il}, \Pi_{il}) \tag{11}$$

[1] We comment on this assumption in the following section.

where f is some function of the parameters $\Theta(x_i, \Pi_i)$ and the database entries x_{il} and Π_{il}.

In the discrete case, we have $\Theta(x_i, \Pi_i) = \Theta_{x_i|\Pi_i}$, and $f(\Theta(x_i, \Pi_i), x_{il}, \Pi_{il}) = \Theta_{x_{il}|\Pi_{il}}$. In the Gaussian case, we have $\Theta(x_i, \Pi_i) = \{m_i, v_i, \vec{b}_i\}$, and $f(\Theta(x_i, \Pi_i), x_{il}, \Pi_{il}) = n(m_i + \sum_{x_j \in \Pi_i} b_{ji} x_{jl}, 1/v_i)$.

3 Informative Priors

In this section, we derive a general approach for assessing the network-parameter priors $p(\Theta_{Bs}|B_s^h, \xi)$. Our derivation is based on four assumptions that are abstracted from our previous work.

Assumption 2 (Likelihood Equivalence) *Given two network structures B_{s1} and B_{s2} such that $p(B_{s1}^h|\xi) > 0$ and $p(B_{s2}^h|\xi) > 0$, if B_{s1} and B_{s2} are equivalent, then $p(\Theta_U|B_{s1}^h, \xi) = p(\Theta_U|B_{s2}^h, \xi)$.*

Informally, the assumption states that the observation of a database does not help to discriminate equivalent network structures. We note that an equivalent way to state likelihood equivalence is that $p(D|B_{s1}^h, \xi) = p(D|B_{s2}^h, \xi)$ for all databases D, whenever B_{s1} and B_{s2} are equivalent.[2]

The motivation for this assumption is different for acausal Bayesian networks—Bayesian networks that represent only assertions of conditional independence—and causal Bayesian networks. For acausal networks, likelihood equivalence is not an assumption, but rather a consequence of our definition of B_s^h. In particular, recall that the hypothesis B_s^h is true iff the parameters Θ_U satisfy the conditional independence assertions of B_s. Therefore, by definition of network-structure equivalence, if B_{s1} and B_{s2} are equivalent, then $B_{s1}^h = B_{s2}^h$.[3] For example, in the domain $\{x_1, x_2, x_3\}$, the equivalent network structures $x_1 \rightarrow x_2 \rightarrow x_3$ and $x_1 \leftarrow x_2 \leftarrow x_3$ both correspond to the assertion $\theta_{x_1,x_3|x_2} = \theta_{x_1|x_2}\theta_{x_3|x_2}$. Consequently, $B_{x_1 \rightarrow x_2 \rightarrow x_3}^h = B_{x_1 \leftarrow x_2 \leftarrow x_3}^h$. This property, which we call *hypothesis equivalence*, implies likelihood equivalence. We note that, given hypothesis equivalence, we should score equivalence classes of network structures—not individual network structures—when learning acausal Bayesian networks.

For causal Bayesian networks, we must modify the definition of B_s^h to include the assertion that each non-root node in B_s is a direct causal effect of its parents. Consequently, the property of hypothesis equivalence is contradicted by the new definition. Nonetheless, we have found that the assumption of likelihood equivalence is reasonable for learning causal networks in many domains. (For a detailed discussion of this point, see Heckerman in this proceedings.)

The next assumption was adopted implicitly in our previous work.

Assumption 3 (Structure Possibility) *Given a domain U, $p(B_{sc}^h|\xi) > 0$ for all complete network structures B_{sc}.*

As we shall see, the assumption allows us to make good use of the property of likelihood equivalence. Although it is an assumption of convenience, we have found it to be reasonable for many real-world network-learning problems.

The remaining two assumptions are abstractions of assumptions made either explicitly or implicitly by all researchers who have considered Bayesian-network learning (e.g., Cooper and Herskovits, 1991, 1992; Buntine, 1991; Spiegelhalter et al., 1993). These assumptions are made mostly for computational convenience, although they are reasonable for many domains.

Assumption 4 (Global Parameter Independence) *For all network structures B_s,*

$$p(\Theta_{Bs}|B_s^h, \xi) = \prod_{i=1}^{n} p(\Theta(x_i, \Pi_i)|B_s^h, \xi)$$

Assumption 4 says that the parameters associated with each variable in a network structure are independent. This assumption was first introduced under the name of global independence by Spiegelhalter and Lauritzen (1990).

Assumption 5 (Parameter Modularity) *Given two network structures B_{s1} and B_{s2} such that $p(B_{s1}^h|\xi) > 0$ and $p(B_{s2}^h|\xi) > 0$, if x_i has the same parents in B_{s1} and B_{s2}, then*

$$p(\Theta(x_i, \Pi_i)|B_{s1}^h, \xi) = p(\Theta(x_i, \Pi_i)|B_{s2}^h, \xi)$$

[2] We assume this equivalence is well known, although we have not found a proof in the literature.

[3] We note that there is a flaw with our definition of B_s^h for acausal Bayesian networks. In particular, the definition implies that hypotheses associated with different network-structure equivalence classes will not be mutually exclusive. For example, in the two-binary-variable domain, the hypotheses B_{xy}^h and $B_{x \rightarrow y}^h$ (corresponding to the empty network structure, and the network structure $x \rightarrow y$, respectively) both include the possibility $\theta_{xy} = \theta_x \theta_y$. This flaw is potentially troublesome, because mutual exclusivity is important for our Bayesian interpretation of network learning (Equation 2). Nonetheless, because the densities $p(\Theta_{Bs}|B_s^h, \xi)$ must be integrable and hence bounded, the overlap of hypotheses will be of measure zero, and we may use Equation 2 without modification. For example, in our two-binary-variable domain, given the hypothesis $B_{x \rightarrow y}^h$, the probability that B_{xy}^h is true (i.e., $\theta_y = \theta_{y|x}$) has measure zero.

For example, in our two-binary-variable domain, x has the same parents (none) in the network structure $x \rightarrow y$ and the structure contains no arc. Consequently, the probability density for $\Theta(x, \emptyset)$ would be the same for both of these structures. We call this property parameter modularity, because it says that the densities for parameters $\Theta(x_i, \Pi_i)$ depend only on the structure of the network that is local to variable x_i—namely, on the parents of x_i.

Given Assumptions 2 through 5, we can construct the priors $p(\Theta_{Bs}|B_s^h, \xi)$ for every network structure B_s in U from the single prior $p(\Theta_U|B_{sc}^h, \xi)$, where B_{sc} is any complete network structure for U. As an illustration of this construction, consider again our two-binary-variable domain. Given the prior density $p(\theta_{xy}, \theta_{x\bar{y}}, \theta_{\bar{x}y}|B_{x\rightarrow y}^h, \xi)$, we construct the priors $p(\Theta_{Bs}|B_s^h, \xi)$ for each of the three network structures in the domain. First, consider the network structure $x \rightarrow y$. The joint-space parameters and parameters for this structure are related as follows:

$$\theta_{xy} = \theta_x \theta_{y|x} \quad \theta_{\bar{x}y} = (1-\theta_x)(\theta_{y|\bar{x}}) \quad \theta_{x\bar{y}} = \theta_x(1-\theta_{y|x})$$

Thus, we may obtain $p(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}|B_{x\rightarrow y}^h, \xi)$ from the given density by changing variables:

$$p(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}|B_{x\rightarrow y}^h, \xi) = J_{x\rightarrow y} \cdot p(\theta_{xy}, \theta_{\bar{x}y}, \theta_{x\bar{y}}|B_{x\rightarrow y}^h, \xi) \tag{12}$$

where $J_{x\rightarrow y}$ is the Jacobian of the transformation

$$\begin{aligned} J_{x\rightarrow y} &= \begin{vmatrix} \partial\theta_{xy}/\partial\theta_x & \partial\theta_{\bar{x}y}/\partial\theta_x & \partial\theta_{x\bar{y}}/\partial\theta_x \\ \partial\theta_{xy}/\partial\theta_{y|x} & \partial\theta_{\bar{x}y}/\partial\theta_{y|x} & \partial\theta_{x\bar{y}}/\partial\theta_{y|x} \\ \partial\theta_{xy}/\partial\theta_{y|\bar{x}} & \partial\theta_{\bar{x}y}/\partial\theta_{y|\bar{x}} & \partial\theta_{x\bar{y}}/\partial\theta_{y|\bar{x}} \end{vmatrix} \\ &= \theta_x(1-\theta_x) \end{aligned} \tag{13}$$

The Jacobian $J_{B_{sc}}$ for the transformation from Θ_U to Θ_{Bsc} in an arbitrary discrete domain is given in Section 5.1.

Next, consider the network structure $x \leftarrow y$. By Assumption 3, the hypothesis $B_{x\leftarrow y}^h$ is also possible, and, by likelihood equivalence, we have $p(\theta_{xy}, \theta_{\bar{x}y}, \theta_{x\bar{y}}|B_{x\leftarrow y}^h, \xi) = p(\theta_{xy}, \theta_{\bar{x}y}, \theta_{x\bar{y}}|B_{x\rightarrow y}^h, \xi)$. Therefore, we can compute the density for the network structure $x \leftarrow y$ using the Jacobian $J_{x\leftarrow y} = \theta_y(1-\theta_y)$.

Finally, consider the empty network structure. Given the assumption of global parameter independence, we may obtain the densities $p(\theta_x|B_{xy}^h, \xi)$ and $p(\theta_y|B_{xy}^h, \xi)$ separately. To obtain the density for θ_x, we first extract $p(\theta_x|B_{x\rightarrow y}^h, \xi)$ from the density for the network structure $x \rightarrow y$. This extraction is straightforward, because, by global parameter independence, the parameters for $x \rightarrow y$ must be independent. Then, we use parameter modularity, which says that $p(\theta_x|B_{xy}^h, \xi) = p(\theta_x|B_{x\rightarrow y}^h, \xi)$. To obtain the density for θ_y, we extract $p(\theta_y|B_{x\leftarrow y}^h, \xi)$ from the density for the network structure $x \leftarrow y$, and again apply parameter modularity. The approach is summarized in Figure 1.

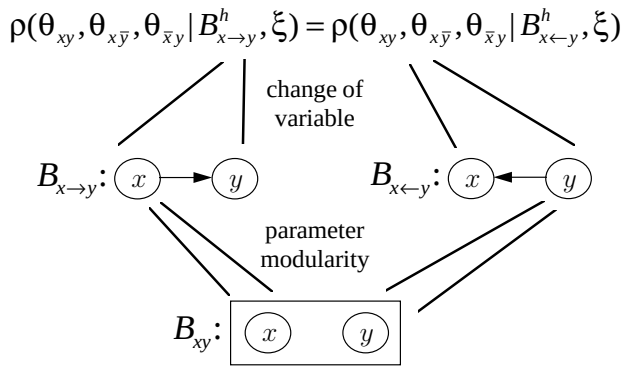

Figure 1: A computation of the parameter densities for the three network structures of the two-binary-variable domain $\{x, y\}$. The approach computes the densities from $p(\theta_{xy}, \theta_{x\bar{y}}, \theta_{\bar{x}y}|B_{x\rightarrow y}^h, \xi)$, using likelihood equivalence, global parameter independence, and parameter modularity.

In general, we have the following construction.

Theorem 2 *Given domain U and a probability density $p(\Theta_U|B_{sc}^h, \xi)$ where B_{sc} is some complete network structure for U, Assumptions 2 through 5 determine $p(\Theta_{Bs}|B_s^h, \xi)$ for any network structure B_s in U.*

We note that our construction assumes that Assumptions 2 through 5 are consistent. We demonstrate consistency in Section 7.

4 A General Metric for Complete Data

In this section, we derive a general metric from Assumptions 1 through 5 and the following additional assumption:

Assumption 6 (Complete Data) *The database is complete. That is, it contains no missing data.*

We make this assumption only as a computational convenience. The reader should recognize that random-sample assumption and the informative priors developed in Section 3 can be used in conjunction with well-known statistical techniques to score incomplete databases as well. Such techniques include filling in missing data based on the data that is present [Titterington, 1976, Spiegelhalter and Lauritzen, 1990], the EM algorithm [Dempster et al., 1977], and Gibbs sampling [Madigan and Raftery, 1994].

Given our assumptions, we obtain the following lemmas.[4]

Lemma 3 (Posterior Parameter Independence) *Given the random-sample assumption (Assumption 1), global parameter independence (Assumption 4), and the assumption of no missing data (Assumption 6), we have*

$$p(\Theta_{Bs}|D, B_s^h, \xi) = \prod_{i=1}^{n} p(\Theta(x_i, \Pi_i)|D, B_s^h, \xi)$$

for all network structures B_s ($p(B_s^h|\xi) > 0$) and databases D.

Lemma 4 (Posterior Parameter Modularity) *Given the random-sample assumption (Assumption 1), global parameter independence (Assumption 4), parameter modularity (Assumption 5), and the assumption of no missing data (Assumption 6), if x_i has the same parents in any two network structures B_{s1} and B_{s2} ($p(B_{s1}^h|\xi) > 0, p(B_{s2}^h|\xi) > 0$), then*

$$p(\Theta(x_i, \Pi_i)|D, B_{s1}^h, \xi) = p(\Theta(x_i, \Pi_i)|D, B_{s2}^h, \xi)$$

for all databases D.

In the following lemma and in subsequent discussions, we need the notion of a *database D restricted to $X \subseteq U$*—that is the projection of database D onto the subset X—denoted D^X. For example, given domain $U = \{x_1, x_2, x_3\}$ and database $D = \{C_1 = \{x_1 = 1, x_2 = 2, x_3 = 1\}, C_2 = \{x_1 = 2, x_2 = 2, x_3 = 1\}\}$, we have $D^{\{x_1, x_2\}} = \{C_1 = \{x_1 = 1, x_2 = 2\}, C_2 = \{x_1 = 2, x_2 = 2\}\}$.

Lemma 5 *Let X be a subset of U, and B_{sc} ($p(B_{sc}^h|\xi) > 0$) be a complete network structure for any ordering where the variables in X come first. Given the random-sample assumption (Assumption 1), global parameter independence (Assumption 4), and the assumption of no missing data (Assumption 6),*

$$p(X|D, B_{sc}^h, \xi) = p(X|D^X, B_{sc}^h, \xi)$$

for all databases D.

Readers familiar with the concept of d-separation will recognize that Lemmas 3 and 5 can be readily obtained from graphical manipulations applied to the Bayesian-network representation of the random-sample assumption and the assumption of global parameter independence.

We can now derive the general metric.

[4]The proofs are simple and are omitted.

Theorem 6 *Given a domain U, let B_s be any network structure for U and B_{sc} be a some complete network structure for U. Then, given Assumptions 2 through 6,*

$$p(D, B_s^h|\xi) = p(B_s^h|\xi) \cdot \prod_{i=1}^{n} \frac{p(D^{\Pi_i, x_i}|B_{sc}^h, \xi)}{p(D^{\Pi_i}|B_{sc}^h, \xi)} \qquad (14)$$

for any database D.

Proof: From the rules of probability, we obtain

$$p(D|B_s^h, \xi) = \prod_{l=1}^{m} \int p(\Theta_{Bs}|D_l, B_s^h, \xi) \cdot \prod_{i=1}^{n} p(x_{il}|x_{1l}, \ldots, x_{(i-1)l}, D_l, \Theta_{Bs}, B_s^h, \xi)\ d\Theta_{Bs}$$

For every x_i with parents Π_i in B_s, let B_{SC,Π_i,x_i} be a complete network structure with variable ordering Π_i, x_i followed by the remaining variables. By Assumption 3, $p(B_{SC,\Pi_i,x_i}|\xi) > 0$. Using Assumption 1 and Lemmas 3 and 4, we get

$$p(D|B_s^h, \xi) = \prod_{l=1}^{m} \int \prod_{i=1}^{n} p(\Theta(x_i, \Pi_i)|D_l, B_{SC,\Pi_i,x_i}, \xi) \cdot p(x_{il}|\Pi_{1l}, D_l, \Theta(x_i, \Pi_i), B_{SC,\Pi_i,x_i}, \xi)\ d\Theta_{Bs}$$

Decomposing the integral over Θ_{Bs} into integrals over the individual parameter sets $\Theta(x_i, \Pi_i)$, and performing the integrations, we have

$$p(D|B_s^h, \xi) = \prod_{l=1}^{m} \prod_{i=1}^{n} p(x_{il}|\Pi_{1l}, D_l, B_{SC,\Pi_i,x_i}, \xi)$$

Also, using Lemma 5, we obtain

$$\begin{aligned} p(D|B_s^h, \xi) &= \prod_{l=1}^{m} \prod_{i=1}^{n} \frac{p(x_{il}, \Pi_{1l}|D_l, B_{SC,\Pi_i,x_i}, \xi)}{p(\Pi_{1l}|D_l, B_{SC,\Pi_i,x_i}, \xi)} \\ &= \prod_{l=1}^{m} \prod_{i=1}^{n} \frac{p(x_{il}, \Pi_{1l}|D_l^{\Pi_i,x_i}, B_{SC,\Pi_i,x_i}, \xi)}{p(\Pi_{1l}|D_l^{\Pi_i}, B_{SC,\Pi_i,x_i}, \xi)} \\ &= \prod_{i=1}^{n} \frac{p(D^{\Pi_i,x_i}|B_{SC,\Pi_i,x_i}, \xi)}{p(D^{\Pi_i}|B_{SC,\Pi_i,x_i}, \xi)} \qquad (15) \end{aligned}$$

By likelihood equivalence, we have that $p(D|B_{SC,\Pi_i,x_i}, \xi) = p(D|B_{sc}^h, \xi)$. Consequently, for any subset X of U, we obtain $p(D^X|B_{SC,\Pi_i,x_i}, \xi) = p(D^X|B_{sc}^h, \xi)$ by summing over the variables in $D^{U \setminus X}$. Applying this result to Equation 15, we get Equation 14. □

We call Equation 14 the Be (*B*ayesian likelihood *e*quivlent) metric.

5 Special-Case Metrics

Our general metric is powerful, because it tells us that if we know how to compute $p(D^X|B^h_{sc},\xi)$ for any subset X of U under the assumption that the domain contains no structure (i.e., there are no independencies), then we can compute the probability of any database when there is structure. Therefore, the Be metric allows us to leverage much of the work in the statistics literature, as statisticians have long dealt with the former problem. In this section, we illustrate this claim by deriving likelihood-equivalent metrics for the discrete and Gaussian cases.

5.1 The BDe Metric

Suppose all variables in U are discrete. Recall that we use $\theta_{X=k_X|Y=k_Y}$ denote the multinomial parameter corresponding to probability $p(X=k_X|Y=k_Y,\xi)$. In addition, we use $\Theta_{X|Y}$ denote the collection of parameters $\theta_{X=k_X|Y=k_Y}$ for all states of sets X and Y. If Y is empty, we simply write Θ_X. Thus, for example, $\Theta_U = \Theta_{x_1,\ldots,x_n}$ represents the multinomial parameters of the joint space of U.

Let us assume that the parameter set Θ_U has a Dirichlet distribution when conditioned on a hypothesis corresponding to some complete network structure B_{sc}:

$$p(\Theta_{x_1,\ldots,x_n}|B^h_{sc}) = \prod_{x_1,\ldots,x_n} \theta^{N'_{B_{sc}}p(x_1,\ldots,x_n|B^h_{sc},\xi)-1}_{x_1,\ldots,x_n} \tag{16}$$

where $N'_{B_{sc}}$ is the equivalent sample size of the Dirichlet distribution associated with a complete network structure B_{sc}. DeGroot (1970, p. 50) shows that, for any subset X of U, Θ_X also has a Dirichlet distribution:

$$p(\Theta_X|B^h_{sc},\xi) = \prod_X \theta^{N'_{B_{sc}}p(X|B^h_{sc},\xi)-1}_X \tag{17}$$

Now, it is a well-known statistical result that, if a discrete variable x with r states has a Dirichlet distribution with exponents $N'_1-1,\ldots,N'_r-1$, then

$$p(D|\xi) = \frac{\Gamma(\sum_{k=1}^r N'_k)}{\Gamma(\sum_{k=1}^r N'_k + N_k)} \prod_{k=1}^r \frac{\Gamma(N'_k+N_k)}{\Gamma(N'_k)} \tag{18}$$

where D is a database for variable x and N_k is the number of times x takes on state k in D. Also, because U is discrete, any subset X of U can also be thought of as a single discrete variable with $\prod_{x_i\in X} r_i$ states. Therefore, Equations 17 and 18 allow us to compute each term in the Be metric (Equation 14). To express the resulting metric for a given network structure B_s, we use $q_i = \prod_{x_i\in\Pi_i} r_i$ to denote the number of states of Π_i in B_s, and $\Pi_i = j$ to denote that Π_i has assumed the jth state, $j = 1,\ldots,q_i$.

Theorem 7 (BDe Metric) *Given domain U, and network structure B_s and database D for U, let N_{ijk} denote the number of times that $x_i = k$ and $\Pi_i = j$ in the database D; and let $N_{ij} = \sum_{k=1}^{r_i}$ denote the number of times that $\Pi_i = j$ in a database D. Then, if $p(\Theta_U|B^h_{sc},\xi)$ is Dirichlet with equivalent sample size N' for some complete network structure B_{sc}, and if Assumptions 2 through 6 hold, then*

$$\begin{aligned} p(D,B^h_s|\xi) &= p(B^h_s|\xi)\cdot\prod_{i=1}^n\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})} \\ &\quad\cdot\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})} \end{aligned} \tag{19}$$

where

$$\begin{aligned} N'_{ijk} &= N'\cdot p(x_i=k,\Pi_i=j|B^h_{sc},\xi) \\ N'_{ij} &= \sum_{k=1}^{r_i} N'_{ijk} = N'\cdot p(\Pi_i=j|B^h_{sc},\xi) \end{aligned} \tag{20}$$

Equations 19 and 20 are the BDe (*B*ayesian *D*irichlet likelihood *e*quivalent) metric, originally derived in Heckerman et al. (1994).

The assumption that $p(\Theta_U|B^h_{sc},\xi)$ is Dirichlet is not as arbitrary as it may seem at first glance. In discrete domains, we can assume not only that the parameters corresponding to each variable are independent, but that the parameters corresponding to each state of every variable's parents are independent. Spiegelhalter and Lauritzen (1990) call this added assumption *local independence*. Geiger and Heckerman (in this proceedings) show that likelihood equivalence, structure possibility, global and local parameter independence, and the assumption that $p(\Theta_U|B^h_{sc},\xi)$ is positive imply that $p(\Theta_U|B^h_{sc},\xi)$ must be Dirichlet.

5.2 The BGe Metric

Suppose that all variables in $U = \vec{x}$ are continuous, and that the database is a random sample from a multivariate-normal distribution. Let us assume that the parameter set $\{\vec{\mu},W\}$ has a normal–Wishart distribution when conditioned on B^h_{sc} for some complete network structure B_{sc}. Namely, assume that $p(\vec{\mu}|W,B^h_{sc},\xi)$ is a multivariate-normal distribution with mean $\vec{\mu}_0$ and precision matrix $N'_\mu W$ ($N'_{\vec{\mu}} > 0$); and that $p(W|B^h_{sc},\xi)$ is a Wishart distribution with N'_T degrees of freedom ($N'_T > n-1$) and positive-definite precision matrix T_0. That is,

$$p(W|B^h_{sc},\xi) = c\ |W|^{(N'_T-n-1)/2} e^{-1/2\,tr\{T_0W\}} \tag{21}$$

where c is a normalization constant [DeGroot, 1970, p. 57].

It is well known that the normal–Wishart distribution is a conjugate family for multivariate-normal sampling (e.g., DeGroot, 1970, p. 178). Given a database $D = \{\vec{x_1}, \ldots, \vec{x_m}\}$, let $\overline{\vec{x}}_m$ and S_m denote its sample mean and variance, respectively. Then, given the normal–Wishart prior we have described, the posterior density $p(\vec{\mu}, W|D, B^h_{sc}, \xi)$ is also a normal–Wishart distribution. In particular, $p(\vec{\mu}|W, D, B^h_{sc}, \xi)$ is multivariate normal with mean vector $\vec{\mu}_m$ given by

$$\vec{\mu}_m = \frac{N'_{\vec{\mu}}\vec{\mu}_0 + m\overline{\vec{x}}_m}{N'_{\vec{\mu}} + m} \tag{22}$$

and precision matrix $(N'_{\vec{\mu}} + m)W$; and $p(W|D, B^h_{sc}, \xi)$ is a Wishart distribution with $N'_T + m$ degrees of freedom and precision matrix T_m given by

$$T_m = T_0 + S_m + \frac{N'_{\vec{\mu}} m}{N'_{\vec{\mu}} + m}(\vec{\mu}_0 - \overline{\vec{x}}_m)(\vec{\mu}_0 - \overline{\vec{x}}_m)' \tag{23}$$

From these equations, we see that $N'_{\vec{\mu}}$ and N'_T can be thought of as equivalent sample sizes for the mean μ_0 and the precision matrix T_0, respectively.

Given domain $U = \{x_1, \ldots, x_n\}$, subset X of U, and vector $\vec{y} = (y_1, \ldots, y_n)$, let $\vec{y}^X$ denote the vector formed by the components y_i of $\vec{y}$ such that $x_i \in X$. Similarly, given matrix M, let M^X denote the submatrix of M containing elements m_{ij} such that $x_i, x_j \in X$. It is well known that if D is a random sample from an n-dimensional multivariate-normal distribution whose parameters $\{\vec{\mu}, W\}$ have a normal–Wishart distribution with constants $\vec{\mu}_0$, $N'_{\vec{\mu}}$, T_0 and N'_T, then D^X is a random sample from an $|X|$-dimensional multivariate distribution with parameters $\{\vec{\mu}^X, W^X\}$, and these parameters have normal–Wishart distribution with constants $\vec{\mu}^X_0$, $N'_{\vec{\mu}}$, T^X_0 and N'_T. Furthermore, the formula for $p(D|B^h_{sc}, \xi)$ given the normal–Wishart prior is known (e.g., the probability may be obtained by integrating the left-hand-side of Equation 8, DeGroot, 1970, p. 179, over the parameters). Consequently, the evaluation of $p(D^X|B^h_{sc}, \xi)$ in Equation 14 is straightforward.

Theorem 8 (BGe Metric) *Given domain $\vec{x} = \{x_1, \ldots, x_n\}$, assume $p(\vec{\mu}, W|B^h_{sc}, \xi)$ is an n-dimensional normal–Wishart distribution with constants $\vec{\mu}_0, N'_{\vec{\mu}}, T_0$, and N'_T. Given a database $D = \{C_1, \ldots, C_m\}$ and a subset X of $\vec{x}$ with l elements, Assumptions 2 through 6 imply the Be metric, where each term is given by*

$$p(D^X|B^h_{sc}, \xi) = \pi^{-lm/2}\left(\frac{N'_{\vec{\mu}}}{N'_{\vec{\mu}} + m}\right)^{l/2} \cdot \frac{c(l, N'_T + m)}{c(l, N'_T)} |T^X_0|^{\frac{N'_T}{2}} |T^X_m|^{-\frac{N'_T + m}{2}} \tag{24}$$

where

$$c(l, N'_T) = \prod_{i=1}^{l} \Gamma\left(\frac{N'_T + 1 - i}{2}\right) \tag{25}$$

and T_m is the precision matrix of the posterior normal–Wishart distribution given by Equation 23.

The Be metric in combination with Equation 24 defines the BGe (*B*ayesian *G*aussian likelihood *e*quivalent) metric, originally derived in Geiger and Heckerman (1994). We note that assumptions similar to those used to show the inevitability of the Dirichlet distribution for discrete domains imply that the normal-Wishart assumption is inevitable for Gaussian domains (see Geiger and Heckerman in this proceedings).

The BDe and BGe metrics may be combined to score domains containing both discrete variables and continuous variables. Namely, let $U = U_d \cup U_c$ where all variables in U_d and U_c are discrete and continuous, respectively. Suppose that the observations of U_d in the database are a random sample from a multivariate-discrete distribution, and the observations of the U_c given each state of U_d are a random sample from a multivariate-normal distribution. Finally, suppose that Θ_{U_d} has a Dirichlet distribution, and that $\Theta_{U_c|U_d=k}$ has a normal–Wishart distribution for every state k of U_d. Then, we can apply the Be metric to any network structure B_s where the variables in U_d precede the variables in U_c, using Equation 18 to evaluate terms for discrete variables, and Equations 24 and 25 to evaluate terms for continuous variables.

6 Informative Priors from a Prior Network

Given our assumptions, $p(\Theta_U|B^h_{sc}, \xi)$ determines a Bayesian scoring metric. In this section, we discuss the assessment of this distribution.

For discrete domains, we can assess $p(\Theta_U|B^h_{sc}, \xi)$ by assessing (1) the joint probability distribution for the first cases to be seen in the database $p(U|B^h_s, \xi)$ and (2) the equivalent sample size N' for the domain. Methods for assessing N' are discussed in (e.g.) Heckerman et al. (1995). To assess $p(U|B^h_s, \xi)$, we can construct a Bayesian network for the first case to be seen. We call this Bayesian network a *prior network.* The unusual aspect of this assessment is the conditioning hypothesis B^h_{sc} (see Heckerman et al. [1995] for a discussion).

We can assess $p(\Theta_U|B^h_{sc}, \xi)$ in the Gaussian case using a prior network as well. In this case, however, we require two equivalent samples sizes ($N'_{\vec{\mu}} > 0$ and $N'_T > n - 1$). The details are discussed in last year's proceedings [Geiger and Heckerman, 1994]. Examples of the assessment of $p(\Theta_U|B^h_{sc}, \xi)$ for discrete

and Gaussian domains, and examples of the metrics that result from these assessments are also given in last year's proceedings.

7 Consistency of the Assumptions

The assumptions of likelihood equivalence, structure possibility, global parameter independence, and parameter modularity may not be consistent. In particular, the assumptions of global parameter independence and modularity are constraints on parameter densities among individual network structures, whereas likelihood equivalence is a constraint on parameter densities among network-structure equivalence classes. Furthermore, our choices $p(\Theta_U|B^h_{sc},\xi)$ is Dirichlet and $p(\vec{\mu},W|B^h_{sc},\xi)$ is normal-Wishart may not be consistent with the assumptions of likelihood equivalence and global parameter independence. In this section, we demonstrate consistency in each case.

7.1 Consistency of the Dirichlet Assumption

First, we show that the assumption $p(\Theta_U|B^h_{sc},\xi)$ is Dirichlet is consistent with the assumptions of likelihood equivalence and global parameter independence for complete network structures.

To see the potential for inconsistency, consider again our approach for constructing priors in the two-binary-variable domain. Suppose we choose the density

$$p(\theta_{xy},\theta_{\bar{x}y},\theta_{x\bar{y}}|B^h_{x\to y},\xi) = \frac{c}{(\theta_{xy}+\theta_{x\bar{y}})(1-(\theta_{xy}+\theta_{x\bar{y}}))} = \frac{c}{\theta_x(1-\theta_x)}$$

where c is a normalization constant. By Equations 12 and 13 we obtain

$$p(\theta_x,\theta_{y|x},\theta_{y|\bar{x}}|B^h_{x\to y},\xi) = c$$

for the network structure $x \to y$. This density satisfies the assumption of global (and local) parameter independence. Using likelihood equivalence, however, we have for the network structure $y \to x$

$$p(\theta_y,\theta_{x|y},\theta_{x|\bar{y}}|B^h_{x\leftarrow y},\xi) = \frac{c\cdot\theta_y(1-\theta_y)}{\theta_x(1-\theta_x)} = \frac{c\cdot\theta_y(1-\theta_y)}{(\theta_y\theta_{x|y}+(1-\theta_y)\theta_{x|\bar{y}})(1-(\theta_y\theta_{x|y}+(1-\theta_y)\theta_{x|\bar{y}}))}$$

This density satisfies neither global (nor local) parameter independence.

When $p(\Theta_U|B^h_{sc},\xi)$ is Dirichlet, however, likelihood equivalence implies global (and local) parameter independence for all complete network structures. This result is proved for the two-variable case in Dawid and Lauritzen (1993, Lemma 7.2) and for the general case in Heckerman et al. (1995, Theorem 3), which we summarize here.

Theorem 9 *Let B_{sc} be any complete network structure for domain $U = \{x_1,\ldots,x_n\}$. The Jacobian for the transformation from Θ_U to Θ_{Bsc} is*

$$J_{B_{sc}} = \prod_{i=1}^{n-1}\prod_{x_1,\ldots,x_i}[\theta_{x_i|x_1,\ldots,x_{i-1}}]^{[\prod_{j=i+1}^{n} r_j]-1} \quad (26)$$

Theorem 10 *Given a domain $U = \{x_1,\ldots,x_n\}$, if the parameters Θ_U have a Dirichlet distribution with parameters $N'_{x_1,\ldots,x_n}$—that is,*

$$p(\Theta_U|\xi) = c\cdot\prod_{x_1,\ldots,x_n}[\theta_{x_1,\ldots,x_n}]^{N'_{x_1,\ldots,x_n}-1} \quad (27)$$

then, for any complete network structure B_{sc} in U, the density $p(\Theta_{Bsc}|\xi)$ satisfies global and local parameter independence. In particular,

$$p(\Theta_{Bsc}|\xi) = c\cdot\prod_{i=1}^{n}\prod_{x_1,\ldots,x_i}[\theta_{x_i|x_1,\ldots,x_{i-1}}]^{N'_{x_i|x_1,\ldots,x_{i-1}}-1} \quad (28)$$

where

$$N'_{x_i|x_1,\ldots,x_{i-1}} = \sum_{x_{i+1},\ldots,x_n} N'(x_1,\ldots,x_n) \quad (29)$$

Proof: The result follows by multiplying the right-hand-side of Equation 27 by the Jacobian in Theorem 9, using the relation $\theta_{x_1,\ldots,x_n} = \prod_{i=1}^{n}\theta_{x_i|x_1,\ldots,x_{i-1}}$, and collecting powers of $\theta_{x_i|x_1,\ldots,x_{i-1}}$. □

It is interesting to note that each set of conditional parameters $\Theta_{x_i|x_1,\ldots,x_{i-1}}$ also has a Dirichlet distribution.

7.2 Consistency of the Normal–Wishart Assumption

Next, we show that the assumption $p(\vec{\mu},W|B^h_{sc},\xi)$ is normal–Wishart is consistent with the assumptions of likelihood equivalence and global parameter independence for complete network structures.

Theorem 11 *The Jacobian for the change of variables from W to $\{\vec{v},B\}$ is given by*

$$J_{\vec{v},B} = |\partial W/\partial\vec{v}B| = \prod_{i=1}^{n} v_i^{-(i+1)} \quad (30)$$

Proof: Let $J(i)$ denote the Jacobian for the first i variables in W. Then $J(i)$ has the following form:

$$J(i) = \begin{vmatrix} J(i-1) & 0 & 0 \\ 0 & -\frac{1}{v_i}I_{i-1,i-1} & 0 \\ 0 & 0 & -\frac{1}{v_i^2} \end{vmatrix} \quad (31)$$

where $I_{k,k}$ is the identity matrix of size $k \times k$. Thus, we have

$$J(i) = \frac{1}{v_i^{i+1}} \cdot J(i-1) \tag{32}$$

which gives Equation 30. □

Theorem 12 *The Jacobian for the change of variables from $\vec{\mu}$ to $\vec{m}$ is given by $J_{\vec{m}} = 1$.*

Proof: From Equation 6, $J_{\vec{m}}$ is the determinant of a triangular matrix whose diagonal elements are 1. □

Theorem 13 *If $\{\vec{\mu}, W\}$ has a normal–Wishart distribution given background information ξ, then*

$$p(\vec{m}, \vec{v}, B|\xi) = \prod_{i=1}^{n} p(m_i, v_i, \vec{b_i}|\xi)$$

Proof: To prove the theorem, we factor $p(\vec{m}|\vec{v}, B, \xi)$ and $p(\vec{v}, B|\xi)$ separately. By assumption, we know that $p(\vec{\mu}|W)$ is a multivariate-normal distribution with mean μ_0 and precision matrix $N'_{\vec{\mu}}W$. Transforming this result to conditional distributions for μ_i, we obtain

$$p(\mu_i|\mu_1, \ldots, \mu_{i-1}, \vec{v}, B, \xi) = \left(\frac{N'_{\vec{\mu}}}{2\pi v_i}\right)^{1/2} \cdot \exp\left\{\frac{\left(\mu_i - \mu_{0i} - \sum_{j=1}^{i-1} b_{ji}(\mu_j - \mu_{0j})\right)^2}{2v_i/N'_{\vec{\mu}}}\right\}$$

for $i = 1, \ldots, n$. Letting $m_{0i} = \mu_{0i} - \sum_{j=1}^{i-1} b_{ji}\mu_{0j}$ for each i, we get

$$p(\mu_i|\mu_1, \ldots, \mu_{i-1}, \vec{v}, B, \xi) = \left(\frac{N'_{\vec{\mu}}}{2\pi v_i}\right)^{1/2} \cdot \exp\left\{\frac{(m_i - m_{0i})^2}{2v_i/N'_{\vec{\mu}}}\right\}$$

Thus, collecting terms for each i and using the Jacobian $J_{\vec{m}} = 1$, we have

$$p(\vec{m}|\vec{v}, B, \xi) = \prod_{i=1}^{n} n(m_{0i}, N'_{\vec{\mu}}/v_i) \tag{33}$$

In addition, by assumption, we have

$$p(W|\xi) = c|W|^{(\alpha-n-1)/2} e^{-1/2\, tr\{T_0 W\}} \tag{34}$$

From Equation 4, we have

$$|W(i)| = \frac{1}{v_i}|W(i-1)| = \prod_{i=1}^{n} v_i^{-1}$$

so that the determinant in Equation 34 factors as a function of i. Also, Equation 4 implies (by induction) that each element w_{ij} in W is a sum of terms each being a function of $\vec{b_i}$ and v_i. Consequently, the exponent in Equation 34 factors as a function of i. Thus, given the Jacobian $J_{\vec{v},B}$, which also factors as a function of i, we obtain

$$p(\vec{v}, B|\xi) = \prod_{i=1}^{n} p(v_i, \vec{b_i}|\xi) \tag{35}$$

Equations 33 and 35 imply the theorem. □

7.3 Consistency of Likelihood Equivalence, Structure Possibility, Parameter Independence, and Parameter Modularity

As mentioned, the assumptions of likelihood equivalence, structure possibility, global parameter independence, and parameter modularity may not be consistent. To understand the potential for inconsistency, note that we obtained the Be metric (Equation 14) for all network structures using likelihood equivalence applied only to complete network structures in combination with the assumptions of structure possibility, global parameter independence, parameter modularity. Thus, it could be that the Be metric for incomplete network structures is not likelihood equivalent. Nonetheless, the following theorem shows that the Be metric is likelihood equivalent for all network structures—that is, given structure possibility, global parameter independence, and parameter modularity, likelihood equivalence for incomplete structures is implied by likelihood equivalence for complete network structures. Consequently, the assumptions are consistent.

Theorem 14 (Likelihood Equivalence) *If B_{s1} and B_{s2} are equivalent network structures for domain U, then, for all databases D, $p(D|B^h_{s1}, \xi) = p(D|B^h_{s2}, \xi)$, where each likelihood is computed by the Be metric (Equation 14).*

Proof: By Theorem 1, we know that a network structure can be transformed into an equivalent structure by a series of arc reversals. Thus, we can demonstrate likelihood equivalence in general if we can do so for the case where two equivalent structures differ by a single arc reversal. So, let B_{s1} and B_{s2} be two equivalent network structures that differ only in the direction of the arc between x_i and x_j (say $x_i \rightarrow x_j$ in B_{s1}). Let R be the set of parents of x_i in B_{s1}. By Theorem 1, we know that $R \cup \{x_i\}$ is the set of parents of x_j in B_{s1}, R is the set of parents of x_j in B_{s2}, and $R \cup \{x_j\}$ is the set of parents of x_i in B_{s2}. Because the two structures differ only in the reversal of a single arc, the only terms in the product of Equation 14 that can differ are those

involving x_i and x_j. For B_{s1}, these terms are

$$\frac{p(D^{x_i R}|B^h_{sc},\xi)}{p(D^R|B^h_{sc},\xi)}\frac{p(D^{x_i x_j R}|B^h_{sc},\xi)}{p(D^{x_i R}|B^h_{sc},\xi)} = \frac{p(D^{x_i x_j R}|B^h_{sc},\xi)}{p(D^R|B^h_{sc},\xi)}$$

whereas for B_{s2}, they are

$$\frac{p(D^{x_j R}|B^h_{sc},\xi)}{p(D^R|B^h_{sc},\xi)}\frac{p(D^{x_i x_j R}|B^h_{sc},\xi)}{p(D^{x_j R}|B^h_{sc},\xi)} = \frac{p(D^{x_i x_j R}|B^h_{sc},\xi)}{p(D^R|B^h_{sc},\xi)}$$

These terms are equal, and consequently, so are the likelihoods. □

Acknowledgments

We thank Peter Spirtes for identifying an error with Equation 24.

References

[Buntine, 1991] Buntine, W. (1991). Theory refinement on Bayesian networks. In *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence,* Los Angeles, CA, pages 52–60. Morgan Kaufmann.

[Chickering, 1995] Chickering, D. (1995). A transformational characterization of equivalent Bayesian network structures. In this proceedings.

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

[Cooper and Herskovits, 1991] Cooper, G. and Herskovits, E. (January, 1991). A Bayesian method for the induction of probabilistic networks from data. Technical Report SMI-91-1, Section on Medical Informatics, Stanford University.

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

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

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38.

[Geiger and Heckerman, 1994] 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 and Heckerman, 1995] Geiger, D. and Heckerman, D. (1995). A characterization of the Dirichlet distribution with application to learning Bayesian networks. In this proceedings.

[Heckerman, 1995] Heckerman, D. (1995). A Bayesian approach for learning causal networks. In this proceedings.

[Heckerman et al., 1994] Heckerman, D., Geiger, D., and Chickering, D. (1994). Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 293–301. Morgan Kaufmann.

[Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, to appear.

[Madigan and Raftery, 1994] Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89:1535–1546.

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

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

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

[Titterington, 1976] Titterington, D. (1976). Updating a diagnostic system using unconfirmed cases. *Applied Statistics*, 25:238–247.

[Verma and Pearl, 1990] 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.