Learning Gaussian Networks

Dan Geiger
geiger02@gmail.com

David Heckerman
heckerma@hotmail.com

July 1994, Revised May 2021

Abstract

We describe scoring metrics for learning Bayesian networks from a combination of user knowledge and statistical data. Previous work has concentrated on metrics for domains containing only discrete variables, under the assumption that data represents a multinomial sample. In this paper, we extend this work, developing scoring metrics for domains containing only continuous variables under the assumption that continuous data is sampled from a multivariate normal distribution. Our work extends traditional statistical approaches for identifying vanishing regression coefficients in that we identify two important assumptions, called *event equivalence* and *parameter modularity*, that when combined allow the construction of prior distributions for multivariate normal parameters from a single *prior Bayesian network* specified by a user.

Corrections to the original text in red are taken from the 2021 update of J. Kuipers, G. Moffa, and D. Heckerman, Addendum on the scoring of Gaussian directed acyclic graphical models. *Annals of Statistics* 42, 1689-1691, Aug 2014 (arXiv:1402.6863). Other updates to the original are in blue.

1 Introduction

Several researchers have examined methods for learning Bayesian networks from data, including Cooper and Herskovits (1991,1992), Buntine (1991), Spiegelhalter et al. (1993), and Heckerman et al. (1994) (herein referred to as CH, Buntine, SDLC, and HGC, respectively). These methods all have the same basic components: a scoring metric and a search procedure. The metric computes a score that is proportional to the posterior probability of a network structure, given data and a user's prior knowledge. The search procedure generates networks for evaluation by the scoring metric. These methods use the two components to identify a network or set of networks with high relative posterior probabilities, and these networks are then used to predict future events.

Previous work has concentrated on domains containing only discrete variables, under the assumption that data is sampled from a multivariate discrete distribution. In this paper, we develop metrics for domains containing only continuous variables, under the assumption that continuous data is sampled from a multivariate normal (Gaussian) distribution. Previously, when working with continuous variables, the standard solution had been to transform each such variable x_i to a discrete one by splitting its domain into several mutually exclusive and exhaustive regions. Our metrics eliminate the need for this transformation. In addition, our metrics have the advantage that they use the low polynomial dimentionality of the parameter space of a mulitivariate normal distribution, whereas their discrete counterparts often require a parameter space that is exponential in the number of domain variables.

Our work can be viewed as an extension of traditional statistical approaches for identifying vanishing regression coefficients, such as those described in DeGroot (1970, Chapter 11). In particular, we translate two assumptions that we identified in HGC for domains containing only discrete variables, called parameter modularity and event equivalence, to domains containing continuous variables. The assumption of *parameter modularity,* addresses the relationship among prior distributions of parameters for different Bayesian-network structures. The property of *event equivalence* says that two Bayesian-network structures that represent the same set of independence assertions should correspond to the same event and thus receive the same score. We show that, when combined, these assumptions allow the construction of reasonable prior distributions for multivariate normal parameters from a single *prior Bayesian network* specified by a user.

Our identification of event equivalence arises from a subtle distinction between two types of Bayesian networks. The first type, called *belief networks*, represents only assertions of conditional independence and dependence. The second type, called *causal networks*, represents assertions of cause and effect as well as assertions of independence and dependence. In this paper, we argue that metrics for belief networks should satisfy event equivalence, whereas metrics for causal networks need not.

Our score-equivalent metrics for belief networks are similar to the metrics described by Dawid and Lauritzen (1993), except that our metrics score directed networks, whereas their metrics score undirected networks. In this paper, we concentrate on directed models rather than on undirected models, because we believe that users find the former easier to build and interpret.

We note that much of the mathematics involved in our derivations is borrowed from DeGroot's book, "Optimal Statistical Decisions," (1970).

2 Gaussian Belief Networks

Throughout this discussion, we consider a domain $\vec{x}$ of n continuous variables $x_1, \ldots, x_n$. We use $\rho(\vec{x}|\xi)$ to denote the joint probability density function (pdf) over $\vec{x}$ of a person with background knowledge ξ. We use $p(e|\xi)$ to denote the probability of a discrete event e.

A belief network for $\vec{x}$ represents a joint pdf over $\vec{x}$ by encoding assertions of conditional independence as well as a collection of pdfs. From the chain rule of probability, we know

$$\rho(x_1, \ldots, x_n|\xi) = \prod_{i=1}^{n} \rho(x_i|x_1, \ldots, x_{i-1}, \xi) \tag{1}$$

For each variable x_i, let $\Pi_i \subseteq \{x_1, \ldots, x_{i-1}\}$ be a set of variables that renders x_i and $\{x_1, \ldots, x_{i-1}\}$ conditionally independent. That is,

$$\rho(x_i|x_1, \ldots, x_{i-1}, \xi) = \rho(x_i|\Pi_i, \xi) \tag{2}$$

A belief network is a pair (B_S, B_P), where B_S is a belief-network structure that encodes the assertions of conditional independence in Equation 2, and B_P is a set of pdfs corresponding to that structure. In particular, B_S is a directed acyclic graph such that (1) each variable in U corresponds to a node in B_S, and (2) the parents of the node corresponding to x_i are the nodes corresponding to the variables in Π_i. (In the remainder of this paper, we use x_i to refer to both the variable and its corresponding node in a graph.) Associated with node x_i in B_S are the pdfs $\rho(x_i|\Pi_i, \xi)$. B_P is the union of these pdfs. Combining Equations 1 and 2, we see that any belief network for $\vec{x}$ uniquely determines a joint pdf for $\vec{x}$. That is,

$$\rho(x_1, \ldots, x_n|\xi) = \prod_{i=1}^{n} \rho(x_i|\Pi_i, \xi)$$

A *minimal belief network* is a belief network where Equation 2 is violated if any arc is removed. Thus, a minimal belief network represents both assertions of independence and assertions of dependence.

Let us suppose that the joint probability density function for $\vec{x}$ is a multivariate (nonsingular) normal distribution. In this case, we write

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

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

This distribution can be written as a product of conditional distributions each being an independent normal distribution. Namely,

$$\rho(\vec{x}|\xi) = \prod_{i=1}^{n} \rho(x_i|x_1, \ldots, x_{i-1}, \xi) \tag{3}$$

$$\rho(x_i|x_1, \ldots, x_{i-1}, \xi) = n(m_i + \sum_{j=1}^{i-1} b_{ij}(x_j - m_j), 1/v_i) \tag{4}$$

where m_i is the unconditional mean of x_i, v_i is the conditional variance of x_i given values for $x_1, \ldots, x_{i-1}$, and b_{ij} is a linear coefficient reflecting the strength of the relationship between x_i and x_j (e.g., DeGroot, p.55).[1] Thus, we may interpret a multivariate normal distribution as a belief network, where $b_{ij} = 0$ $(j < i)$ implies that x_j is not a parent of x_i. We call this special form of a belief network a Gaussian belief network. The name is adopted from Shachter and Kenley (1989) who first described Gaussian influence diagrams.

More formally, a *Gaussian belief network* is a pair (B_S, B_P), where (1) B_S is a belief-network structure containing nodes $x_1, \ldots, x_n$ and no arc from x_j to x_i whenever $b_{ij} = 0, j < i$, (2) B_P is the collection of parameters $\vec{m} = (m_1, \ldots, m_n)$, $\vec{v} = \{v_1, \ldots, v_n\}$, and $\{b_{ij} \mid j < i\}$, and (3) the joint distribution over $\vec{x}$ is determined by Equations 3 and 4. Due to special properties of nonsingular normal distributions, a *minimal Gaussian belief network* is one were there is an arc from x_j to x_i if and only if $b_{ij} \neq 0$.

Given a multivariate normal density, we can generate a Gaussian belief network, and vice versa. The unconditional means $\vec{m}$ are the same in both representations. Shachter and Kenley (1989) describe the general transformation from $\vec{v}$ and $\{b_{ij} \mid i < j\}$ of a given Gaussian belief network G to the precision matrix W of the normal distribution represented by G. They use the following recursive formula in which $W(i)$ denotes the $i \times i$ upper left submatrix of W, $\vec{b}_i$ denotes the column vector $(b_{1,i}, \ldots, b_{i-1,i})$ and $\vec{b}'_i$ denotes the transposed vector $\vec{b}_i$ (i.e., the line vector $(b_{1,i}, \ldots, b_{i-1,i})$):

$$W(i+1) = \begin{pmatrix} W(i) + \frac{\vec{b}_{i+1}\vec{b}'_{i+1}}{v_{i+1}} & -\frac{\vec{b}_{i+1}}{v_{i+1}} \\ -\frac{\vec{b}'_{i+1}}{v_{i+1}} & \frac{1}{v_{i+1}} \end{pmatrix} \tag{5}$$

for $i > 0$, and $W(1) = \frac{1}{v_1}$. Equation 5 plays a key role in this paper.

For example, suppose $x_1 = n(m_1, 1/v_1), x_2 = n(m_2, 1/v_2)$, and $x_3 = n(m_3 + b_{13}(x_1 - m_1) + b_{23}(x_2 - m_2), 1/v_3)$. The belief-network structure defined by these equations is shown in Figure 1. The precision matrix is given by

$$W = \begin{pmatrix} \frac{1}{v_1} + \frac{b_{13}^2}{v_3} & \frac{b_{13}b_{23}}{v_3} & -\frac{b_{13}}{v_3} \\ \frac{b_{13}b_{23}}{v_3} & \frac{1}{v_2} + \frac{b_{23}^2}{v_3} & -\frac{b_{23}}{v_3} \\ -\frac{b_{13}}{v_3} & -\frac{b_{23}}{v_3} & \frac{1}{v_3} \end{pmatrix} \tag{6}$$

[1] The coefficients b_{ij} can be thought of as regression coefficients or expressed in terms of Yule's (1907) partial regression coefficient β.

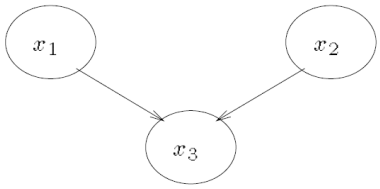

Figure 1: A belief-network structure for three variables.

Table 1: An complete database for the domain associated with the network shown in Figure 1.

	Variable values for each case		
Case	x_1	x_2	x_3
1	-0.78	-1.55	0.11
2	0.18	-3.04	-2.35
3	1.87	1.04	0.48
4	-0.42	0.27	-0.68
5	1.23	1.52	0.31
6	0.51	-0.22	-0.60
7	0.44	-0.18	0.13
8	0.57	-1.82	-2.76
9	0.64	0.47	0.74
10	1.05	0.15	0.20
11	0.43	2.13	0.63
12	0.16	-0.94	-1.96
13	1.64	1.25	1.03
14	-0.52	-2.18	-2.31
15	-0.37	-1.30	-0.70
16	1.35	0.87	0.23
17	1.44	-0.83	-1.61
18	-0.55	-1.33	-1.67
19	0.79	-0.62	-2.00
20	0.53	-0.93	-2.92

The Gaussian-belief-network representation of a multivariate normal distribution is better suited to model elicitation and understanding than is the standard representation [Shachter and Kenley, 1989]. To assess a Gaussian belief network, the user needs to specify (1) the unconditional mean of each variable x_i (m_i), (2) the relative importance of each parent x_j in determining the values of its child x_i (b_{ij}), and (3) a conditional variance for x_i given that its parents are fixed (v_i). Equation 5 then determines W. In contrast, when assessing a normal distribution directly, one needs to guarantee that the assessed covariance matrix is positive-definite—a task done by altering in some *ad hoc* manner the correlations stated by the user.

3 A Metric for Gaussian Belief Networks

We are interested in computing a score for a Gaussian belief-network structure, given a set of cases $D = \{\vec{x}_1, \ldots, \vec{x}_m\}$. Each *case* $\vec{x}_i$ is the observation of one or more variables in $\vec{x}$. We sometimes refer to D as a *database*. Table 1 is an example of a database for the three-node domain of the Gaussian belief network shown in Figure 1.

Our scoring metrics are based on five assumptions, the first of which is the following:

Assumption 1 *The database D is a random sample from a multivariate normal distribution with unknown means $\vec{m}$ and unknown precision matrix W.*

Because every Gaussian belief network is equivalent to a multivariate normal distribution, Assumption 1 is equivalent to stating that the database D is a random sample from a Gaussian belief network with unknown parameters, $\vec{v}, B = \{b_{ij} | j < i\}, \vec{m}$.

A Bayesian measure of the goodness of a network structure is its posterior probability given a database:

$$p(B_S|D,\xi) = c\ p(B_S|\xi)\ \rho(D|B_S,\xi)$$

where $c = 1/\rho(D|\xi) = 1/\sum_{B_S} p(B_S|\xi)\ \rho(D|B_S,\xi)$ is a normalization constant. For even small domains, however, there are too many network structures to sum over in order to determine the constant. Therefore we use $p(B_S|\xi)\ \rho(D|B_S,\xi) = \rho(D,B_S|\xi)$ as our score.

Also problematic is our use of the term B_S as an argument of a probability. In particular, B_S is a belief-network structure, not an event. Thus, we need a definition of an event B_S^e that corresponds to structure B_S (the superscript "e" stands for *e*vent). A natural definition for this event is that B_S^e holds true iff the database is a random sample from a *minimal* Gaussian belief network with structure B_S —that is, iff for all $j < i$, $b_{ij} \neq 0$ if and only if there is an arc from x_j to x_i in B_S. For example the event B_S^e corresponding to the Gaussian belief network of Figure 1, is the event $\{b_{12} = 0, b_{13} \neq 0, b_{23} \neq 0\}$.

This definition has the following desirable property. When two belief-network structures represent the same assertions of conditional independence, we say that they are *isomorphic*. For example, in the three variable domain $\{x_1, x_2, x_3\}$, the network structures $x \rightarrow x_2 \rightarrow x_3$ and $x_1 \leftarrow x_2 \rightarrow x_3$ represent the same assertion: x_1 and x_3 are independent given x_2. Given the definition of B_S^e, it can be shown that events B_{S1}^e and B_{S2}^e are equivalent if and only if the structures B_{S1} and B_{S2} are isomorphic. That is, the relation of isomorphism induces an equivalence class on the set of events B_S^e. We call this property *event equivalence*.

There is a problem with the definition, however. In particular, events corresponding to some non-isomorphic network structures are not mutually exclusive. For example, in the four-variable domain $\{x_1, x_2, x_3, x_4\}$, consider the structures $x_1 \Rightarrow B \Leftarrow x_4$ and $x_1 \Rightarrow B \Rightarrow x_4$, where B is the subnetwork structure $x_2 \rightarrow x_3$, and $x \Rightarrow B$ means that there is an arc from x to both variables in B. The events corresponding to these structures both include the situation where x_1 and x_4 are marginally independent. Arbitrary overlaps between events can make scores difficult to interpret and use. For example, the prediction of future events by averaging over multiple models cannot be justified. In our case, however, we can repair the definition of B_S^e so as to make non-equivalent events mutually exclusive, without affecting our mathematical results or the intuitive understanding of events by the user. In particular, all overlaps will be of measure zero with respect to the events that create the overlap. Thus, given a set of overlapping events, we simply exclude the intersection from all but one of the events. We note that this revised definition retains the property of event equivalence.

Proposition 1 (Event Equivalence) *Belief-network structures B_{S1} and B_{S2} are isomorphic if and only if $B_{S1}^e = B_{S2}^e$.*

Because the score for network structure B_S is $\rho(D, B_S^e|\xi)$, an immediate consequence of the property of event equivalence is score equivalence.

Proposition 2 (Score Equivalence) *The scores of two isomorphic belief-network structures must be equal.*

Given the property of event equivalence, we technically should score each belief-network-structure equivalence class, rather than each belief-network structure. Nonetheless, users find it intuitive to work with (i.e., construct and interpret) belief networks. Consequently, we continue our presentation in terms of belief networks, keeping Proposition 2 in mind.

3.1 Complete Gaussian Belief Networks

We first derive $\rho(D, B_S^e|\xi)$, assuming B_S is the structure of a complete Gaussian belief network. A *complete Gaussian belief network* is one with no missing edges. Applying the property of event equivalence, we know that the event associated with any complete belief network is the same; and we use $B_{S_C}^e$ to denote this event.

To motivate the derivation, consider the following expansion of $\rho(D|B^e_{S_C}, \xi)$:

$$\rho(D|B^e_{S_C}, \xi) = \prod_{l=1}^{m} \rho(C_l|C_1, \ldots, C_{l-1}, B^e_{S_C}, \xi) =$$

$$\prod_{l=1}^{m} \int \rho(C_l|\vec{m}, W, B^e_{S_C}, \xi)\ \rho(\vec{m}, W|C_1, \ldots, C_{l-1}, B^e_{S_C}, \xi)\ d\vec{m}\ dW$$

Thus, we can derive the metric if we find a conjugate distribution for the parameters $\vec{m}$ and W such that the integral above has a closed form solution.

The next assumption leads to such a conjugate distribution. If all variables in a case are observed, we say that the case is *complete*. If all cases in a database are complete, we say that the database is *complete*.

Assumption 2 *All databases are complete.*[2]

Given this assumption, the following distribution is conjugate for multivariate-normal sampling.

Theorem 3 (DeGroot, p.178) *Suppose that $\vec{x_1}, \ldots, \vec{x_l}$ is a random sample from a multivariate normal distribution with an unknown value of the mean vector $\vec{m}$ and an unknown value of the precision matrix W. Suppose that the prior joint distribution of $\vec{m}$ and W is the normal-Wishart distribution: the conditional distribution of $\vec{m}$ given W is $n(\vec{\mu}_0, \nu W)$ such that $\nu > 0$, and the marginal distribution of W is a Wishart distribution with $\alpha > n-1$ degrees of freedom and precision matrix T_0, denoted by $w(\alpha, T_0)$. Then the posterior joint distribution of $\vec{m}$ and W given $\vec{x_i}$, $i = 1, \ldots, l$, is as follows: The conditional distribution of $\vec{m}$ given W is a multivariate normal distribution with mean vector $\vec{\mu}_l$ and a precision matrix $(\nu + l)W$, where*

$$\overline{X}_l = \frac{1}{l}\sum_{i=1}^{l} \vec{x}_i, \quad \vec{\mu}_l = \frac{\nu\vec{\mu}_0 + l\overline{X}_l}{\nu + l}. \tag{7}$$

and the marginal of W is $w(\alpha + l, T_l)$, where S_l and T_l are given by

$$S_l = \sum_{i=1}^{l} (\vec{x}_i - \overline{X}_l)(\vec{x}_i - \overline{X}_l)' \tag{8}$$

and

$$T_l = T_0 + S_l + \frac{\nu l}{\nu + l}(\vec{\mu}_0 - \overline{X}_l)(\vec{\mu}_0 - \overline{X}_l)' \tag{9}$$

In this theorem, $\overline{X}_l$ and S_l are the sample mean and scatter matrix of the database, respectively. Also, an n dimensional Wishart distribution with α degrees of freedom and matrix T_0 is given by

$$\rho(W|\xi) = w(\alpha, T_0) \equiv c(n, \alpha)|T_0|^{\alpha/2}|W|^{(\alpha-n-1)/2}e^{-1/2tr\{T_0W\}} \tag{10}$$

where $tr\{T_0W\}$ is the sum of the diagonal elements of T_0W and

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

The parameters ν, α, $\vec{\mu}_0$, and T_0 are implicit functions of the user's background knowledge ξ. The quantities ν and α can be thought of as the effective sample sizes of the normal and Wishart components of the prior, respectively.

Summarizing our discussion so far, we make the following assumption:

[2]SDLC present a survey of approximation methods for handling missing data in the context of discrete variables. Some of these methods in modified form can be applied to Gaussian networks.

Assumption 3 *The prior distribution $\rho(\vec{m}, W|B^e_{S_C}, \xi)$ is a normal-Wishart distribution as given in Theorem 3.*

From Equation 5, this assumption fixes the distribution $\rho(\vec{m}, \vec{v}, B|B^e_{S_C}, \xi)$. Nonetheless, we shall sometimes find it easier to specify the prior density in the space of W, rather then in the space of parameters describing a Gaussian belief network.

If $\rho(\vec{x}|\vec{m}, W, B^e_{S_C}, \xi) = n(\vec{m}, W)$ and if $\rho(\vec{m}, W|B^e_{S_C}, \xi)$ is a normal-Wishart distribution as specified by Theorem 3, then $\rho(\vec{x}|B^e_{S_C}, \xi)$, defined by

$$\rho(\vec{x}|B^e_{S_C}, \xi) = \int \rho(\vec{x}|\vec{m}, W, B^e_{S_C}, \xi)\ \rho(\vec{m}, W, B^e_{S_C}, \xi)\ d\vec{m}\ dW$$

is an n dimensional multivariate t distribution with $\gamma = \alpha - n + 1$ degrees of freedom, location vector $\vec{\mu}_0$, and a precision matrix $T'_0 = \frac{\nu\gamma}{\nu+1}T_0^{-1}$. This result can be derived by first integrating over $\vec{m}$ using Equation 6 on p.178 of DeGroot with sample size equal to one, and then integrating over W following an approach similar to that on pp.179–180 of DeGroot. Also, using Equation 3 on p.180 of DeGroot, the t distribution $\rho(\vec{x}|B^e_{S_C}, \xi)$ can be written in a less traditional form as follows:

$$\rho(\vec{x} \mid B^e_{S_C}, \xi) = (2\pi)^{-n/2}\left(\frac{\nu}{\nu+1}\right)^{n/2}\frac{c(n,\alpha)}{c(n,\alpha+1)}|T_0|^{\alpha/2}|T_1|^{-(\alpha+1)/2} \tag{11}$$

where T_1 is defined by Equation 9 with $l = 1$.

Combining these facts with Theorem 3, we know that $\rho(C_l|C_1, \ldots, C_{l-1}, B^e_{S_C}, \xi)$ is a multivariate t distribution with parameters $\nu + l - 1$, $\alpha + l - 1$, $\vec{\mu}_{l-1}$, and T_{l-1}. Consequently, we obtain

$$\begin{aligned}
\rho(D \mid B^e_{S_C}, \xi) &= \prod_{l=1}^{m}\rho(C_l \mid C_1, \ldots, C_{l-1}, B^e_{S_C}, \xi) \\
&= \prod_{l=1}^{m}\left((2\pi)^{-n/2}\left(\frac{\nu+l-1}{\nu+l}\right)^{n/2}\frac{c(n,\alpha+l-1)}{c(n,\alpha+l)}\frac{|T_{l-1}|^{\frac{\alpha+l-1}{2}}}{|T_l|^{\frac{\alpha+l}{2}}}\right) \\
&= (2\pi)^{-nm/2}\left(\frac{\nu}{\nu+m}\right)^{n/2}\frac{c(n,\alpha)}{c(n,\alpha+m)}|T_0|^{\frac{\alpha}{2}}|T_m|^{-\frac{\alpha+m}{2}}
\end{aligned} \tag{12}$$

Multiplying Equation 12 by the prior probability $p(B^e_{S_C}|\xi)$ yields a metric for scoring $B^e_{S_C}$.

3.2 General Gaussian Belief Networks

We now consider an arbitrary Gaussian belief network B_S. To form a prior distribution for the parameters of B_S, we make two additional assumptions:

Assumption 4 (Parameter Independence) *For every Gaussian belief network B_S, $\rho(\vec{v}, B|B^e_S, \xi) = \prod_{i=1}^{n}\rho(v_i, \vec{b_i}|B^e_S, \xi)$.*

We note that this assumption is consistent with Assumption 3, because if $\rho(W|B^e_{S_C}, \xi)$ is a Wishart distribution, then $\rho(\vec{v}, B|B^e_{S_C}, \xi)$, obtained from $\rho(W|B^e_{S_C}, \xi)$ by using Equation 5 and the Jacobian $\partial W/\partial\vec{v}B$ of this transformation, is equal to $\prod_{i=1}^{n}\rho(v_i, \vec{b_i}|B^e_{S_C}, \xi)$. The derivation of this claim is given in the Appendix (Theorem 7).

Assumption 5 (Parameter Modularity) *If x_i has the same parents in two Gaussian belief networks B_{S1} and B_{S2}, then $\rho(v_i, \vec{b_i}|B^e_{S1}, \xi) = \rho(v_i, \vec{b_i}|B^e_{S2}, \xi)$.*

Assumption 4 has been made in discrete contexts by many researchers (e.g., CH, Buntine, SDLC, and HGC). Assumption 5 has also been made by these same researchers, but HGC were the first researchers to make the assumption explicit and to emphasize its importance for generating prior distributions. Parameter modularity plays a similar important role in the current development. In particular, this assumption, in conjunction with the property of event equivalence and our previous assumptions allows us to determine the joint prior distribution of the parameters $\vec{m}, \vec{v}, B$ associated with any Gaussian network B_S from the joint density $\rho(\vec{m}, W|B^e_{S_C})$.

To see this fact, first note that, by the definition of the event B^e_S, $\rho(\vec{m}|\vec{v}, B, B^e_S, \xi) = \rho(\vec{m}|\vec{v}, B, B^e_{S_C}, \xi)$. The latter distribution is determined by $\rho(\vec{m}|W, B^e_{S_C}, \xi)$, which is given. Second, from Assumption 4, we obtain $\rho(\vec{v}, B|B^e_S, \xi)$ by determining $\rho(v_i, \vec{b_i}|B^e_S, \xi)$ for each i. By Assumption 5, however, $\rho(v_i, \vec{b_i}|B^e_S, \xi)$ is equal to $\rho(v_i, \vec{b_i}|B^e_{S'_C}, \xi)$ for any complete network structure $B_{S'_C}$ where the parents of x_i are the same as are those in B_S. By event equivalence and Assumption 4, we obtain $\rho(v_i, \vec{b_i}|B^e_{S'_C}, \xi)$ from the given density $\rho(W|B^e_{S_C}, \xi)$.

From Assumptions 1 through 5, we derive $\rho(D|B^e_S, \xi)$. To do so, we need the following theorem whose proof is provided in the Appendix. [Note: a derivation from weaker assumptions is given in D. Geiger and D. Heckerman, Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions, *The Annals of Statistics*, 30: 1412-1440, Oct 2002.]

Theorem 4 *If $\rho(\vec{x}|\vec{m}, W, D, \xi)$ is a multivariate normal distribution, and $\rho(\vec{m}|W, D, B^e_S, \xi)$ is a multivariate normal distribution with a precision matrix νW, $\nu > 0$, then $\rho(x_i|x_1, \ldots, x_{i-1}, \vec{v}, B, D, B^e_S, \xi) = \rho(x_i|\Pi_i, v_i, \vec{b_i}, D^{x_i\Pi_i}, B^e_{S'}, \xi)$, where $B_{S'}$ is any network where x_i has the same parents as in B_S, and $D^{x_i\Pi_i}$ is the database D restricted to the variables in $\{x_i\} \cup \Pi_i$. In particular, this claim holds for any complete Gaussian belief network $B_{S_C} = B_{S'}$ in which Π_i and x_i appear before any other variables, and Π_i appears before x_i.*

Let $D_l = \{C_1, \ldots, C_{l-1}\}$ and C_l be an instance of $x_1, \ldots, x_n$. In the following derivation, we use x_i and Π_i to represent the instance of x_i and Π_i in the lth case. Theorem 4 yields,

$$\begin{aligned}\rho(D|\vec{v}, B, B^e_S, \xi) &= \prod_{l=1}^{m}\prod_{i=1}^{n} \rho(x_i|x_1, \ldots, x_{i-1}, \vec{v}, B, D_l, B^e_S, \xi) \\ &= \prod_{l=1}^{m}\prod_{i=1}^{n} \frac{\rho(x_i, \Pi_i|v_i, \vec{b_i}, D_l^{x_i\Pi_i}, B^e_S, \xi)}{\rho(\Pi_i|v_i, \vec{b_i}, D_l^{x_i\Pi_i}, B^e_S, \xi)}\end{aligned}$$

and

$$\rho(\Pi_i|v_i, \vec{b_i}, D_l^{x_i\Pi_i}, B^e_S, \xi) = \rho(\Pi_i|v_i, \vec{b_i}, D_l^{\Pi_i}, B^e_S, \xi)$$

By combining these equations, we obtain the following *likelihood separability property:*

$$\rho(D|\vec{v}, B, B^e_S, \xi) = \prod_{i=1}^{n} \frac{\rho(D^{x_i\Pi_i}|v_i, \vec{b_i}, B^e_S, \xi)}{\rho(D^{\Pi_i}|v_i, \vec{b_i}, B^e_S, \xi)} \tag{13}$$

By Bayes rule, $\rho(\vec{v}, B|D, B^e_S, \xi)$ is proportional to $\rho(D|\vec{v}, B, B^e_S, \xi)\rho(\vec{v}, B|B^e_S, \xi)$. Thus, because $\rho(D|\vec{v}, B, B^e_S, \xi)$ factors as shown by Equation 13, and $\rho(\vec{v}, B|B^e_S, \xi)$ factors as given by Assumption 4, we obtain the following *posterior parameter independence* property:

$$\rho(\vec{v}, B|D, B^e_S, \xi) = \prod_{i=1}^{n} \rho(v_i, \vec{b_i}|D^{x_i\Pi_i}, B^e_S, \xi)$$

In a similar manner, whenever x_i has the same parents in two Gaussian belief networks B_S and $B_{S'}$, by using Equation 13 where B^e_S in the right hand side is replaced by $B^e_{S'}$ and using Assumption 5,

we obtain the *posterior parameter modularity* property:

$$\rho(v_i, \vec{b_i} | D^{x_i \Pi_i}, B_S^e, \xi) = \rho(v_i, \vec{b_i} | D^{x_i \Pi_i}, B_{S'}^e, \xi)$$

Now, we have

$$\rho(D|B_S^e, \xi) = \prod_{l=1}^{m} \rho(C_l|D_l, B_S^e, \xi), \tag{14}$$

$$\rho(C_l|D_l, B_S^e, \xi) = \prod_{i=1}^{n} \rho(x_i|x_1, \ldots, x_{i-1}, D_l, B_S^e, \xi)$$

$$\rho(x_i|x_1, \ldots, x_{i-1}, D_l, B_S^e, \xi) = \int \rho(x_i|x_1, \ldots, x_{i-1}, D_l, \vec{v}, B, B_S^e, \xi)\rho(\vec{v}, B \mid D_l, B_S^e, \xi) d\vec{v}B \tag{15}$$

By applying Theorem 4 to the first term of the right-hand-side of Equation 15, and posterior parameter independence and posterior parameter modularity to the second term, we obtain

$$\begin{aligned} \rho(x_i|x_1, \ldots, x_{i-1}, D_l, B_S^e, \xi) &= \int \rho(x_i|\Pi_i, v_i, \vec{b_i}, D_l^{x_i \Pi_i}, B_{S_C}^e, \xi)\ \rho(v_i, \vec{b_i}|D_l^{x_i \Pi_i}, B_{S_C}^e, \xi)\ dv_i \vec{b_i} \\ &= \rho(x_i|\Pi_i, D_l^{x_i \Pi_i}, B_{S_C}^e, \xi) \end{aligned}$$

Therefore,

$$\rho(C_l|D_l, B_S^e, \xi) = \prod_{i=1}^{n} \frac{\rho(x_i, \Pi_i|D_l^{x_i \Pi_i}, B_{S_C}^e, \xi)}{\rho(\Pi_i|D_l^{x_i \Pi_i}, B_{S_C}^e, \xi)} \tag{16}$$

Furthermore, because $\rho(\Pi_i|D_l^{x_i \Pi_i}, B_{S_C}^e, \xi)$ is a multivariate t distribution, we know that

$$\rho(\Pi_i|D_l^{x_i \Pi_i}, B_{S_C}^e, \xi) = \rho(\Pi_i|D_l^{\Pi_i}, B_{S_C}^e, \xi)$$

(DeGroot, p.60). Thus, combining Equations 14 and 16, we have

$$\rho(D|B_S^e, \xi) = \prod_{i=1}^{n} \frac{\rho(D^{x_i \Pi_i}|B_{S_C}^e, \xi)}{\rho(D^{\Pi_i}|B_{S_C}^e, \xi)} \tag{17}$$

where each term in 17 is of the form given in Equation 12. Multiplying Equation 17 by $p(B_S^e|\xi)$, we obtain a metric for an arbitrary Gaussian belief network B_S. (This development is incomplete, as it requires a recipe for deriving the parameters of the prior for subsets of the domain variables from the prior for all domain variables. The recipe implicit in an example given in the original version—deleted in this version—is incorrect. For a correction, see the 2021 update of D. Geiger and D. Heckerman, Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions, *The Annals of Statistics*, 30: 1412-1440, Oct 2002.) We call this metric BGe which stands for *B*ayesian metric for *G*aussian networks having score *e*quivalence.

3.3 Score Equivalence

In making the assumptions of parameter independence and parameter modularity, we have—in effect—specified the prior densities for the multinomial parameters in terms of the structure of a belief network. Consequently, there is the possibility that this specification violates the property of score equivalence. The following theorem, however, demonstrates that our specification implies score equivalence.

Theorem 5 (Score Equivalence) *If B_{S1} and B_{S2} are isomorphic belief-network structures, then $\rho(D|B_{S1}^e, \xi)$ and $\rho(D|B_{S2}^e, \xi)$ as computed by Equation 17 are equal.*

Proof: In Heckerman et al. (1994, Theorem 10), we show that a belief network structure can be transformed into an isomorphic structure by a series of arc reversals, such that, whenever an arc from x_i to x_j is reversed, $\Pi_i = \Pi_j \setminus \{x_i\}$. Thus, our claim follows if we can prove it for the case where B_{S1} and B_{S2} differ by a single arc reversal with this restriction.

So, let B_{S1} and B_{S2} be two isomorphic 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 parents of x_i in B_{S1}. By the cited theorem, $R \cup \{x_i\}$ is the parents of x_j in B_{S1}, R is the parents of x_j in B_{S2}, and $R \cup \{x_j\}$ is the 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 17 that can differ are those involving x_i and x_j. For B_{S1}, these terms are

$$\frac{\rho(D^{x_i R}|B^e_{S_C},\xi)}{\rho(D^R|B^e_{S_C},\xi)} \frac{\rho(D^{x_i x_j R}|B^e_{S_C},\xi)}{\rho(D^{x_i R}|B^e_{S_C},\xi)} = \frac{\rho(D^{x_i x_j R}|B^e_{S_C},\xi)}{\rho(D^R|B^e_{S_C},\xi)}$$

whereas for B_{S2}, they are

$$\frac{\rho(D^{x_j R}|B^e_{S_C},\xi)}{\rho(D^R|B^e_{S_C},\xi)} \frac{\rho(D^{x_i x_j R}|B^e_{S_C},\xi)}{\rho(D^{x_j R}|B^e_{S_C},\xi)} = \frac{\rho(D^{x_i x_j R}|B^e_{S_C},\xi)}{\rho(D^R|B^e_{S_C},\xi)}$$

Thus, $\rho(D|B^e_{S1},\xi) = \rho(D|B^e_{S2},\xi)$. □

3.4 Encoding Prior Knowledge: The Prior Gaussian Belief Network

From the previous discussion, we see that there are three components of a user's prior knowledge that are relevant to learning Gaussian networks: (1) the prior probabilities $p(B^e_S|\xi)$, (2) the effective sample sizes α and ν, and (3) the parameters $\vec{\mu}_0$ and T_0. The assessment of the prior probabilities $p(B^e_S|\xi)$ is straightforward. Buntine and HGC, for example, describe methods that facilitate these assessments. In addition, a user can assess the effective sample sizes directly. In this section, we concentrate on the assessment of $\vec{\mu}_0$ and T_0.

Using (1) our previous observation that $p(\vec{x}|B^e_{S_C},\xi)$ is a multivariate t distribution, and (2) Equation 11 on p.61 of DeGroot with $\alpha > n+1$, we obtain

$$\mathrm{E}(\vec{x}|B^e_{S_C},\xi) = \vec{\mu}_0 \qquad \mathrm{Cov}(\vec{x}|B^e_{S_C},\xi) = \frac{\nu+1}{\nu} \frac{1}{\alpha-n-1} T_0 \tag{18}$$

Thus, a person can assess a Gaussian belief network for $\mathrm{E}(\vec{x}|B^e_{S_C},\xi)$ and $\mathrm{Cov}(\vec{x}|B^e_{S_C},\xi)$, and then compute $\vec{\mu}_0$ and T_0 using Equations 18. We call this belief network a *prior belief network*.

4 Metrics for Gaussian Causal Networks

People often have knowledge about the causal relationships among variables in addition to knowledge about conditional independence. Such causal knowledge is stronger than is conditional-independence knowledge, because it allows us to derive beliefs about a domain after we intervene. Causal networks, described—for example—by Spirtes et al. (1993), Pearl and Verma (1991), and Heckerman and Shachter (1994) represent such causal relationships among variables. In particular, a causal network for U is a belief network for U, wherein it is asserted that each nonroot node x is caused by its parents. The precise meaning of cause and effect is not important for our discussion. The interested reader should consult the previous references.

The event C^e_S is the same as that for a belief-network structure, except that we also include in the event the assertion that each nonroot node is caused by its parents. Thus, in contrast to the case for belief networks, it is not appropriate to require the properties of event equivalence or score equivalence. For example, consider a domain containing two variables x and y. Both the causal network C_{S1} where x points to y and the causal network C_{S2} where y points to x represent

the assertion that x and y are dependent. The network C_{S1}, however, in addition represents the assertion that x causes y, whereas the network C_{S2} represents the assertion that y causes x. Thus, the events C^e_{S1} are C^e_{S2} are not equal. Indeed, it is reasonable to assume that these events—and the events associated with any two different causal-network structures—are mutually exclusive.

In principle, then, a user may assign a (possibly different) prior distribution to the parameters $\vec{m}$, $\vec{v}$, and B to every complete Gaussian causal network, constrained only by the assumption of parameter modularity. The prior distributions for parameters of incomplete networks would then be determined by parameter modularity. We call this general metric BG, as it is a superset of the BGe metric. For practical reasons, however, the assessment process should be constrained. One alternative is to use the BGe metric. A more general alternative is to continue to use the prior network to compute $\vec{\mu}_0$ and T_0, but to allow effective sample size to vary for different variables and different parent sets of each variable. We call this metric the BGp metric, where "p" stands for *p*rior network.

5 Summary and Future Work

We have described metrics for learning belief networks and causal networks from a combination of user knowledge and statistical data for domains containing only continuous variables. An important contribution has been our elucidation of the property of event equivalence and the assumption of parameter modularity. We have shown that these properties, when combined, allow a statistician to compute a reasonable prior distribution for the parameters of any Gaussian belief network, given a single prior Gaussian belief network provided by a user.

A legitimate concern with our approach is that the multivariate model is too restrictive. In practice, when this model is inappropriate, statisticians will typically turn to a more general model where each continuous variable conditioned on its parents is assumed to be a mixture of multivariate normal distributions. In Geiger and Heckerman (1994), we derive metrics for domains containing both discrete and continuous variables, subject to the restriction that a domain can be decomposed into disjoint sets of continuous variables where each such set is conditioned by a set of discrete variables. We note that this work, when combined with approximation methods that handle missing data, provides a method for learning with multivariate mixtures.

In the discrete case, a complete network has one parameter for each instance of $\vec{x}$. Consequently, it is easy to overfit such a structure with data; and the metrics developed for discrete domains provide a means by which we can avoid such overfitting. In the continuous case, a complete network has only $n + n(n-1)/2$ parameters. Thus, it is possible that the errors introduced by our methods, arising from heuristic search in an exponential space to find one or a handful of structures with high scores outweigh the benefits associated with decreasing the degree of overfitting. We leave this concern for future experimentation.

Acknowledgments

We thank Wray Buntine and anonymous reviewers for useful suggestions.

References

[Cooper and Herskovits, 1991] Cooper, G. and Herskovits, E. (January, 1991). Technical Report SMI-91-1, Section of Medical Informatics, University of Pittsburgh.

[Cooper and Herskovits, 1992] Cooper, G. and Herskovits, E. (1992). *Machine Learning*, 9:309–347.

[Dawid and Lauritzen, 1993] Dawid, A. and Lauritzen, S. (1993). *Annals of Statistics*, 21:1272–1317.

[DeGroot, 1970] DeGroot, M. (1970). McGraw-Hill, New York.

[Geiger and Heckerman, 1994] Geiger, D. and Heckerman, D. (March, 1994). Technical Report MSR-TR-94-10, Microsoft.

[Heckerman et al., 1994] Heckerman, D., Geiger, D., and Chickering, D. (1994b). In this proceedings.

[Heckerman and Shachter, 1994] Heckerman, D. and Shachter, R. (1994). In this proceedings.

[Pearl and Verma, 1991] Pearl, J. and Verma, T. (1991). In Allen, J., Fikes, R., and Sandewall, E., editors, *Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452. Morgan Kaufmann, New York.

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

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

[Spirtes et al., 1993] Spirtes, P., Glymour, C., and Scheines, R. (1993). Springer-Verlag, New York.

[Yule, 1907] Yule, G. (1907). *Proceedings of the Royal Society of London,* Series A, 79:182–193.

Appendix

Theorem 6 *The Jacobian J for the change of variables from W to $\{\vec{v}, B\}$ is given by*

$$J = \partial W / \partial \vec{v} B = \prod_{i=1}^{n} v_i^{-(i+1)} \tag{19}$$

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

$$\begin{pmatrix} J(i-1) & 0 & 0 \\ 0 & -\frac{1}{v_i} I_{i-1,i-1} & 0 \\ 0 & 0 & -\frac{1}{v_i^2} \end{pmatrix} \tag{20}$$

where $I_{k,k}$ is the identity matrix of size $k \times k$. Thus, the absolute value of $J(i)$ is given by,

$$|J(i)| = \frac{1}{v_i^{i+1}} \cdot |J(i-1)| \tag{21}$$

which gives Equation 19. □

Theorem 7 *If $\rho(W|\xi)$ has an n-dimensional Wishart distribution, then*

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

Proof: By assumption, we have

$$\rho(W|\xi) = c\ |W|^{(\alpha-n-1)/2} e^{-1/2 tr\{T_0 W\}} \tag{22}$$

Thus, we must express Equation 22 in terms of $\{\vec{v}, B\}$, multiply by the Jacobian given by Theorem 6, and show that the resulting function factors as a function of i. From Equation 5, we get

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

so that the determinant in Equation 22 factors as a function of i. Also, Equation 5 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 22 factors as a function of i. □

Theorem 4 *If $\rho(\vec{x}|\vec{m}, W, D, B_S^e, \xi)$ is a multivariate normal distribution, and $\rho(\vec{m}|W, D, B_S^e, \xi)$ is a multivariate normal distribution with precision matrix νW, $\nu > 0$, then $\rho(x_i|x_1, \ldots, x_{i-1}, \vec{v}, B, D, B_S^e, \xi) = \rho(x_i|\Pi_i, v_i, \vec{b_i}, D^{x_i\Pi_i}, B_{S'}^e, \xi)$ where $B_{S'}$ is any network where x_i has the same parents as in B_S, and $D^{x_i\Pi_i}$ is the database D restricted to the variables in $\{x_i\} \cup \Pi_i$.*

Proof: Using

$$\rho(\vec{x}|W, D, B_S^e, \xi) = \int \rho(\vec{x}|\vec{m}, W, D, B_S^e, \xi)\ \rho(\vec{m}|W, D, B_S^e, \xi)\ d\vec{m}$$

and Assumptions 1 and 3, we obtain

$$\rho(\vec{x}|W, D, B_S^e, \xi) = c\ |W|^{1/2} \cdot e^{-\frac{1}{2}\frac{\nu}{\nu+1}\sum_{i,j=1}^{n}(x_i-\mu_{Di})(x_j-\mu_{Dj})w_{ij}} \tag{23}$$

where $\vec{\mu}_D$ is the posterior mean after seeing D, given by Equation 7 of Theorem 3.

The marginal distribution $\rho(x_1, \ldots, x_i|\xi)$ of a normal distribution $n(\vec{m}, W)$ is a normal distribution $n(\vec{m}_i, W_i)$, where $\vec{m}_i$ and W_i are the terms in $\vec{m}$ and W that correspond to $x_1, \ldots, x_i$. Thus, using $|W| = \prod_{i=1}^{n} v_i^{-1}$, Equation 23 becomes

$$\rho(x_1, \ldots, x_i|W, D, B_S^e, \xi) = c\ |W_i|^{1/2} \cdot e^{-\frac{1}{2}\frac{\nu}{\nu+1}\sum_{j,k=1}^{i}(x_j-\mu_{jD})(x_k-\mu_{kD})w_{jk}} \tag{24}$$

By expressing W in terms of $\vec{v}$ and B using Equation 5, we obtain

$$\frac{\rho(x_1, \ldots, x_i|\vec{v}, B, D, B_S^e, \xi)}{\rho(x_1, \ldots, x_{i-1}|\vec{v}, B, D, B_S^e, \xi)} = c \cdot v_i^{-1/2} \cdot e^{-\frac{1}{2}\frac{\nu}{\nu+1}A} \tag{25}$$

where

$$A = tr\left[(\vec{x} - \vec{\mu}_D)_i(\vec{x} - \vec{\mu}_D)_i' \begin{pmatrix} \frac{\vec{b}_i\vec{b}_i'}{v_i} & -\frac{\vec{b}_i}{v_i} \\ -\frac{\vec{b}_i'}{v_i} & v_i \end{pmatrix}\right] \tag{26}$$

where $(\vec{x} - \vec{\mu}_D)_i$ is the column vector of the i elements of $(\vec{x} - \vec{\mu}_D)$ that correspond to $x_1, \ldots, x_i$. Starting with any network $B_{S'}$, such that the parents of x_i are the same as in B_S, we obtain exactly Equations 25 and 26. Furthermore, because $\vec{\mu}_D$ depends only on $D^{x_i\Pi_i}$, the theorem is established. □