Toward a Generalized Bayesian Network

Dawn E. Holmes

Department of Statistics and Applied Probability, South Hall,
University of California, Santa Barbara,
CA 93106, USA.

Abstract. The author's past work in this area has shown that the probability of a state of a Bayesian network, found using the standard Bayesian techniques, could be equated to the Maximum Entropy solution and that this result enabled us to find minimally prejudiced estimates of missing information in Bayesian networks. In this paper we show that in the class of Bayesian networks known as Bayesian trees, we are able to determine missing constraint values optimally using only the maximum entropy formalism. Bayesian networks that are specified entirely within the maximum entropy formalism, whether or not information is missing, are called generalized Bayesian networks. It is expected that further work will fully generalize this result.

Keywords: Bayesian networks, maximum entropy, *d*-separation.
PACS: 02.50.Cw, 89.70.+c , 05.70.–a, 65.40.Gr

INTRODUCTION

One of the major drawbacks of using Bayesian networks is that complete information, in the form of marginal and conditional probabilities must be specified before the usual updating algorithms are applied. Holmes [1] has shown that when all or some of this information is missing, it is possible to determine unbiased estimates using maximum entropy. The techniques thus developed depend on the property that the probability of a state of a fully-specified Bayesian network, found using the standard Bayesian techniques, can be equated to the maximum entropy solution. A fully-constrained Bayesian network is clearly a special case, both theoretically and practically, and a general theory has yet to be provided. As a first step toward a general theory a *generalized* Bayesian network is defined as one in which some, all or none of the essential information is missing. It is then shown that missing information can be estimated using the maximum entropy formalism (MaxEnt) alone, thus divorcing these results from their dependence on Bayesian techniques.

The techniques required for the current problem are substantially different to those used previously in that, although we still use the method of undetermined multipliers, we no longer equate the joint probability distributions given by the Bayesian and maximum entropy models in order to determine the Lagrange multipliers. Two preliminary results are described here. Firstly, we extend the 2-valued work of Holmes [2] and of Markham and Rhodes [3] by developing an iterative algorithm for updating probabilities in a multivalued multiway tree, Secondly, we use the Lagrange multiplier technique to find the probability of an arbitrary state in a Bayesian tree using only MaxEnt. We begin by defining a Bayesian network.

BAYESIAN NETWORKS

A Bayesian network is essentially a system of constraints; those constraints being determined by *d*-separation. Formally, a Bayesian network is defined as follows. Let:

(i) **V** be a finite set of vertices

(ii) **B** be a set of directed edges between vertices with no feedback loops. The vertices together with the directed edges form a directed acyclic graph $\mathbf{G} = \langle \mathbf{V},\mathbf{B} \rangle$

(iii) a set of events be depicted by the vertices of **G** and hence also represented by V, each event having a finite set of mutually exclusive outcomes

(iv) E_i be a variable which can take any of the outcomes e_i^j of the event i, $j = 1...n_i$

(v) **P** be a probability distribution over the combinations of events, i.e. **P** consists of all possible $P\left(\bigcap_{i\in\mathbf{V}} E_i\right)$.

Let C be the following set of constraints:

(2i) the elements of **P** sum to unity.

(2ii) for each event *i* with a set of parents M_i there are associated conditional probabilities $P\left(E_i \mid \bigcap_{j\in M_i} E_j\right)$ for each possible outcome that can be assigned to E_i and E_j.

(2iii) those independence relationships implied by *d*-separation in the directed acyclic graph.

Then $\mathbf{N} = \langle \mathbf{G},\mathbf{P},\mathbf{C} \rangle$ is a causal network if **P** satisfies **C**.

In a Bayesian network the property of *d*-separation identifies all the constraints as independencies and dependencies. In classical Bayesian network theory a prior distribution must be specified in order to apply the updating algorithms developed, for example, by Pearl [4] or Lauritzen and Spiegalhalter [5]. By working with the same set of constraints as those implied by *d*-separation, the MaxEnt formalism provides a means of determining the prior distribution when information is missing. The author has previously shown that the MaxEnt model with complete information is identical to the Bayesian model and has used this property to estimate the optimal prior distribution when information is missing. We now show that the MaxEnt model is not dependent on the Bayesian model for a class of Bayesian networks.

A GENERALIZED BAYESIAN NETWORK WITH MAXIMUM ENTROPY

Consider the knowledge domain represented by a set, *K*, of multivalued events a_i. Associated with each event is a variable E_v. The general state *S* of the causal tree is the conjunction $\bigcap_{v\in \mathbf{V}} E_v$. A particular state is obtained by assigning some e_v^j to each E_v. It is assumed that the probability of a state is non-zero. The number of states N_S in the tree is given by:

$$N_S = \prod_{i\in \mathbf{V}} n_i$$

where n_i is the number of values possessed by the *i*th event. States are numbered from $1,...,N_S$ and denoted by $S_i : i = 1,...,N_S$, and the probability of a state is denoted by $P(S_i)$. To determine a minimally prejudiced probability distribution **P**, using the maximum entropy formalism, we maximize

$$H = -\sum_{i=1}^{N_s} P(S_i) \ln P(S_i) \tag{1.1}$$

in accordance with the constraints implied by *d*-separation. These constraints are given in the form of marginal or conditional probabilities that represent the current state of knowledge of the domain.

Let a sufficient set of constraints be denoted by **C**, where each constraint $C_j \in \mathbf{C}$. Each constraint is assigned a unique Lagrange multiplier λ_j, where *j* represents the subscripts corresponding to the events on the associated edge. For the edge $\langle a_1, b_1 \rangle$, the Lagrange multipliers are $\lambda(b_1^1, a_1^1)$, $\lambda(b_1^1, a_1^2),...,\lambda(b_1^m, a_1^p)$ where event a_1 has *p* outcomes and event b_1 has *m* outcomes. Without loss of generality we consider the constraints arising from a typical edge $\langle a_1, b_1 \rangle$ thus:

$$P(e_b^j \mid e_a^i) = \beta(b_j, a_i) \qquad i = 1,...,N_S; \quad j = 1,...,m \tag{1.2}$$

Since **P** is a probability distribution we also require the normalization constraint:

$$\sum_{i=1}^{N_s} P(S_i) = 1 \tag{1.3}$$

The Lagrange multiplier λ_0 is associated with the sum to unity. Applying the theory of Lagrange multipliers transforms the problem into that of maximizing:

$$F = H - \sum_{all\ j} \lambda_j C_j \tag{1.4}$$

By partially differentiating (1.4) with respect to $P(S_i)$ and λ_j, we see that the contribution to the expression for a maximum from H is given by:

$$-\left(1 + \ln P(S_i)\right) \quad i = 1,\ldots,N_S \tag{1.5}$$

Similarly, the contribution made by each causal constraint and the sum to unity to the expression for a maximum is given by

$$-\sum_{\substack{C_j \in C \\ i=1,\ldots,N_s}} \lambda_j \frac{\partial C_j}{\partial P(S_i)} = 0 \tag{1.6}$$

resulting in a combined expression:

$$-\left(1 + \ln P(S_i)\right) - \sum_{\substack{C_j \in C \\ i=1,\ldots,N_s}} \lambda_j \frac{\partial C_j}{\partial P(S_i)} = 0 \tag{1.7}$$

and hence

$$P(S_i) = e^{-1} \prod_{\substack{C_j \in C \\ i=1,\ldots,N_s}} \exp\left(\left(-\lambda_j\right) \frac{\partial C_j}{\partial P(S_i)} \right) \tag{1.8}$$

In order to further consider the probability of a state, as given in (1.8), we first need to transform the given constraints into expressions containing the sums of probabilities of states. These causal constraints given in (1.2) are thus expressed in the form:

$$\left(1 - \beta\left(b_1^j, a_1^i\right)\right) \sum_{x \in X} P\left(S_x\right) - \beta\left(b_1^j, a_1^i\right) \sum_{y \in Y} P\left(S_y\right) = 0 \tag{1.9}$$

where $X = \left\{ x \mid \sum_x P\left(S_x\right) = P\left(e_{a_1}^i e_{b_1}^j\right) \right\}$ and $Y = \left\{ y \mid \sum_y P\left(S_y\right) = \sum_{\substack{k=1 \\ k \neq j}}^{k=m} P\left(e_{a_1}^i e_{b_1}^k\right) \right\}$

This defines a family of constraint equations for the arbitrary edge $\langle a_1, b_1 \rangle$. The root node is a special case of equations (1.2) since the information is given in the form of marginal probabilities and hence they need not be considered separately.

Substituting (1.8) into (1.9) gives:

$$\left(1-\beta\left(b_1^j,a_1^i\right)\right)\sum_{x\in X}\prod_{C_j\in C}\exp\left(\left(-\lambda_{b_1^j,a_1^i}\right)\frac{\partial C_j}{\partial P(S_x)}\right)-\beta\left(b_1^j,a_1^i\right)\sum_{y\in Y}\prod_{C_j\in C}\exp\left(\left(-\lambda_{b_1^j,a_1^i}\right)\frac{\partial C_j}{\partial P(S_y)}\right)=0 \tag{1.10}$$

Now consider the probability of the state with event a_1 instantiated with its *i*th outcome and event b_1 with its *j*th outcome, denoted by $P\left(S_{b_1^j,a_1^i}\right)$. We see that when $x\in X$, $P\left(S_{b_1^j,a_1^i}\right)$ contains the expression:

$$\exp\left(\left(-\lambda\left(b_1^j,a_1^i\right)\right)\left(1-\beta\left(b_1^j,a_1^i\right)\right)\right)$$

Similarly, when $y\in Y$, $P\left(S_{b_1^j,a_1^i}\right)$ contains the terms:

$$\exp\left(-\lambda\left(b_1^j,a_1^i\right)\right)\prod_{k=1}^{k=m-1}\exp\left(\left(-\lambda\left(b_1^k,a_1^i\right)\right)\left(-\beta\left(b_1^k,a_1^i\right)\right)\right)$$

Hence $P\left(S_{b_1^j,a_1^i}\right)$ contains the terms

$$\exp\left(\left(-\lambda\left(b_1^j,a_1^i\right)\right)\left(1-\beta\left(b_1^j,a_1^i\right)\right)\right)\prod_{\substack{k=1\\k\neq j}}^{k=m-1}\exp\left(\left(-\lambda\left(b_1^k,a_1^i\right)\right)\left(-\beta\left(b_1^k,a_1^i\right)\right)\right)$$

arising from the edge $\left\langle a_1,b_1\right\rangle$. Re-arranging gives

$$\exp\left(\left(-\lambda\left(b_1^j,a_1^i\right)\right)\right)\prod_{k=1}^{k=m-1}\exp\left(-\lambda\left(b_1^k,a_1^i\right)\left(-\beta\left(b_1^k,a_1^i\right)\right)\right)$$

Since this constraint is typical we see that for all states belonging to $X\in x$:

$$\exp\left(\left(-\lambda\left(b_1^j,a_1^i\right)\right)\left(1-\beta\left(b_1^j,a_1^i\right)\right)\right) \tag{1.11}$$

and for all states belonging to $Y\in y$:

$$\exp\left(-\lambda\left(b_1^j,a_1^i\right)\right)\prod_{k=1}^{k=m-1}\exp\left(\left(-\lambda\left(b_1^k,a_1^i\right)\right)\left(-\beta\left(b_1^k,a_1^i\right)\right)\right) \tag{1.12}$$

From equations (1.11) and (1.12) we see that (1.10) becomes:

$$\exp\left(-\lambda\left(b_1^j,a_1^i\right)\right)\left(1-\beta\left(b_1^j,a_1^i\right)\right)\sum_{x\in X}\prod_{C_j\in\mathbf{C}-C_{\left(b_1^j,a_1^i\right)}}\exp\left(\left(-\lambda_{b_1^j,a_1^i}\right)\frac{\partial C_j}{\partial P(S_x)}\right)-$$

$$\beta\left(b_1^j,a_1^i\right)\sum_{y\in Y}\prod_{C_j\in\mathbf{C}-C_{\left(b_1^j,a_1^i\right)}}\exp\left(\left(-\lambda_{b_1^j,a_1^i}\right)\frac{\partial C_j}{\partial P(S_y)}\right)=0$$

and hence

$$\exp\left(-\lambda\left(b_1^j,a_1^i\right)\right)=\frac{\beta\left(b_1^j,a_1^i\right)\sum_{y\in Y}\prod_{C_j\in\mathbf{C}-C_{\left(b_1^j,a_1^i\right)}}\exp\left(\left(-\lambda_{b_1^j,a_1^i}\right)\frac{\partial C_j}{\partial P(S_y)}\right)}{\left(1-\beta\left(b_1^j,a_1^i\right)\right)\sum_{x\in X}\prod_{C_j\in\mathbf{C}-C_{\left(b_1^j,a_1^i\right)}}\exp\left(\left(-\lambda_{b_1^j,a_1^i}\right)\frac{\partial C_j}{\partial P(S_x)}\right)} \tag{1.13}$$

This expression enables us to update Lagrange multipliers using an iterative algorithm. However, as we show in the next section, we can solve for the Lagrange multipliers algebraically, thus producing a solution identical to that given in earlier papers, using techniques outside of the MaxEnt formalism. See for example, Holmes [6]

SOLVING FOR THE LAGRANGE MULTIPLIERS: EXAMPLE

For the purposes of illustration we consider a three valued causal binary tree with three nodes A, B and C. Let

$$E_a=\left\{e_a^1e_a^2e_a^3\right\},\ E_b=\left\{e_b^1e_b^2e_b^3\right\},\ E_c=\left\{e_c^1e_c^2e_c^3\right\}$$

denote the outcomes of events a, b and c respectively, which are mutually exclusive and collectively exhaustive. The required information, given by conditional probabilities associated with each outcome, is as follows:

$$\begin{aligned}&\sum_{i=0}^{26}P(S_i)=1\quad\text{(constraint 0)}\\&P(e_a^i)=\alpha(a_i);\ i=1,2;\ \text{(constraints 1 and 2)}\\&P(e_b^j\,|\,e_a^i)=\beta(b_ja_i),\ P(e_b^j\,|\,e_a^i)=\beta(b_ja_i);\ \ i=1,2,3;\ j=1,2;\ \text{(constraints 3 - 8)}\\&P(e_c^j\,|\,e_a^i)=\beta(c_ja_i),\ P(e_c^j\,|\,e_a^i)=\beta(c_ja_i);\ \ i=1,2,3;\ j=1,2;\ \text{(constraints 9 - 14)}\end{aligned} \tag{1.14}$$

This system can be in any of 27 states, labeled 0-26, as follows:

$S_0:\ e_{\mathrm{a}}^1e_b^1e_c^1$	$S_1:\ e_{\mathrm{a}}^1e_b^1e_c^2$	$S_2:\ e_{\mathrm{a}}^1e_b^1e_c^3$	$S_3:\ e_{\mathrm{a}}^1e_b^2e_c^1$	$S_4:\ e_{\mathrm{a}}^1e_b^2e_c^2$	$S_5:\ e_{\mathrm{a}}^1e_b^2e_c^3$
$S_6:\ e_{\mathrm{a}}^1e_b^3e_c^1$	$S_7:\ e_{\mathrm{a}}^1e_b^3e_c^2$	$S_8:\ e_{\mathrm{a}}^1e_b^3e_c^3$			

The remaining states are similarly defined but with $E_a=e_a^2$ for states 9 – 17 and $E_a=e_a^3$ for states 18 – 26. Each constraint in (1.14) can be expressed in terms of state probabilities, as in (1.9). For example, constraint 3 gives:

$$\left(1-\beta(b_1a_1)\right)\sum_{i=0,1,2}P(S_i)-\beta(b_1a_1)\left(\sum_{i=3,4,5}P(S_i)+\sum_{i=6,7,8}P(S_i)\right)=0 \tag{1.15}$$

In (1.14), sets *X* and *Y* as defined in (1.9), contain states 3,4,5 and 6,7,8 respectively. Using the equation for probability of a state given by (1.6) together with (1.15) enables us to find the values of all the Lagrange multipliers. Expanding (1.15) and simplifying gives an expression for $\exp(-\lambda_3)$ in terms of known information, together with certain unknown Lagrange multipliers thus:

$$\exp(-\lambda_3)=\left(\frac{\beta(b_1a_1)}{1-\beta(b_1a_1)}\right)\times \left(\frac{1+\exp(-\lambda_4)+\exp(-\lambda_5)+\exp(-\lambda_6)+\exp(-\lambda_4)\exp(-\lambda_5)+\exp(-\lambda_4)\exp(-\lambda_6)}{1+\exp(-\lambda_5)+\exp(-\lambda_6)}\right) \tag{1.16}$$

Following the same procedure but with

$$\left(1-\beta(b_2a_1)\right)\sum_{i=3,4,5}P(S_i)-\beta(b_2a_1)\left(\sum_{i=0,1,2}P(S_i)+\sum_{i=6,7,8}P(S_i)\right)=0 \tag{1.17}$$

leads to

$$\exp(-\lambda_4)=\left(\frac{\beta(b_2a_1)}{1-\beta(b_2a_1)}\right)\times \left(\frac{1+\exp(-\lambda_3)+\exp(-\lambda_5)+\exp(-\lambda_6)+\exp(-\lambda_3)\exp(-\lambda_5)+\exp(-\lambda_3)\exp(-\lambda_6)}{1+\exp(-\lambda_5)+\exp(-\lambda_6)}\right) \tag{1.18}$$

Using equations (1.16) and (1.17) we find, by factorization and substitution, that:

$$\exp(-\lambda_3)=\left(\frac{\beta(b_1a_1)}{1-\beta(b_1a_1)}\right)\left(1+\frac{\beta(b_2a_1)}{1-\beta(b_2a_1)}\left(1+\exp(-\lambda_3)\right)\right) \tag{1.19}$$

hence

$$\exp(-\lambda_3)=\frac{\beta(b_1a_1)}{\beta(b_1a_1)\beta(b_2a_1)+\left(1-\beta(b_2a_1)\right)\left(1-\beta(b_1a_1)\right)} \tag{1.20}$$

and so

$$\exp(-\lambda_3) = \frac{\beta(b_1 a_1)}{\beta(b_3 a_1)} \tag{1.21}$$

The remaining Lagrange multipliers are found similarly, and so the probability of each state can be determined.

REMARKS

For the class of Bayesian networks discussed here, the non-linear independence constraints implied by *d*-separation are preserved by the maximum entropy formalism and do not need to be explicitly stated.

Having shown how to find the Lagrange multipliers and thus the probability of each state, methods previously developed by Holmes and Rhodes [1] can be used to determine missing information since these depend only on the maximum entropy formalism. We have seen in this paper how to derive expressions for estimating missing information in tree-like Bayesian networks without equating the maximum entropy and Bayesian models.

The next step in the current project will be to develop the theory required to deal with the non-linear constraints inherent in singly connected networks without recourse to methods outside of the maximum entropy formalism.

REFERENCES

1. Holmes D.E. and Rhodes P.C [1998] 'Reasoning with Incomplete Information in a Multivalued Multiway Causal Tree Using the Maximum Entropy Formalism'. *International Journal of Intelligent Systems.*Vol.13, No 9, pp 841-859.
2. Holmes D.E. [2004] 'Maximizing Entropy for Inference in a Class of Multiply Connected Networks'. The 24th Conference on Maximum Entropy and Bayesian methods, American Institute of Physics.
3. Markham M.J. and Rhodes P.C [1999] 'Maximizing Entropy to deduce an Initial Probability Distribution for a Causal Network'. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems.Vol.7*, No 1, pp 63-80.
4. Pearl J. [1988] *Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference.* Morgan Kaufmann Publishers.
5..Lauritzen S.L. and Spiegelhalter D.J. [1988] 'Local Computations with Probabilities on Graphical Structures and their Applications to Expert Systems'. *J.Royal Statist.Soc.* **B50**, No.2. pp 154-227
6. Holmes D.E. [1999] 'Efficient Estimation of Missing Information in Multivalued Singly Connected Networks Using Maximum Entropy'. In *Maximum Entropy and Bayesian Methods* pp 289-300. W.von der Linden, V.Dose, R.Fischer and R.Preuss. (Eds.) Kluwer Academic, Dordrecht, Netherlands.