Causal Influence Coefficients: A Localised Maximum Entropy Approach to Bayesian Inference

Michael Ingleby
School of Computing and Mathematics
University of Huddersfield
HD1 3DH, UK
M.Ingleby@hud.ac.uk

Margaret West
M.M.West@hud.ac.uk

Abstract

We consider the problem of incomplete conditional probability tables in Bayesian nets, noting that marginal probabilities for an effect, given single cause are usually easy to elicit and can serve as constraints on the full CPT. A form of maximum entropy principle, local to an effect node is developed and contrasted with existing global methods. Exact maximum-entropy CPTs are computed and a conjecture about the exact solution for effects with a general number n of causes is examined.

1 Introduction

1.1 Background and Related Work

The use of directed acyclic graphs (DAGs) to represent both complex multivariate probability distributions and the dependency relations that reflect causality is very widespread. The recent review material in (Pearl 2000; Dawid 2001) gives the breadth and tenor of current practice. The usage encompasses Bayesian nets, and also nets enhanced with non-random decision or action nodes. Also, effect nodes with utilities attached are admitted, and the usual causal propagation procedures extended to derive distributions of utilities. With such enhancements, Bayesian nets are applicable whenever causal or diagnostic inference is needed in expert systems or other intelligent agents. The latter include dynamic decision nets designed to navigate uncertain environments.

A DAG, fully furnished with prior probability tables (PPTs) for all base nodes and conditional probability tables (CPTs) for all higher effect nodes, determines the joint probability density (JPD) of all its node variates. Within the rigorous framework of probability axioms, supported only by a few suppositions about independence and conditional independence, DAGs can *in principle* satisfy the need to reason about uncertainty of any kind of agent. The main barrier to *practical* use of DAGs is their heavy demand for frequency data with which to 'train' or 'populate' the tables needed for inference. The barrier is highest at effect nodes having a large number of direct causes. For example, a Boolean effect E with n Boolean causes $C_1, C_2, \ldots, C_n$ has a CPT with 2^n entries $P(E \mid (C_1 C_2, \ldots C_n), P(E \mid \neg C_1, C_2, \ldots, C_n)$, etc. If the causes and effects are represented on a scale of 5, as is common in human factors work, the table has 5^n entries. Obtaining frequency data sufficient to estimate such large numbers of CPT entries is costly, in some cases impossibly so.

In many environments, however, experts willingly assess the marginal influence of a single causal factor on an effect - by means of weights or influence coefficients justified by limited data. These are interpretable in a Bayesian framework as conditional probabilities, reinforcing the effect with weight $P(E \mid C_i)$ and inhibiting the effect with weight $P(E \mid \neg C_i)$. These marginals do not in general determine the full CPT for a multi-cause effect node, and hence do not provide a basis for the causal and diagnostic reasoning that make Bayesian nets such attractive processors of uncertain knowledge. We therefore suggest a maximal entropy (ME) principle that leads to a unique CPT by minimising a Shannon information measure. The principle amounts to minimising that part of the Shannon information in a CPT not already present in the marginal influence coefficients.It effectively views a node as a communication channel with inputs constrained by influence coefficients. This approach contrasts with others in the literature, notably

Dempster-Schaefer theory (Schafer 1979; Pearl 1988), and Huffman coding (e.g. (Heckermann 1996)). In Dempster-Schaefer theory, the idea of completing all the tables needed for strict Bayesian inference in a DAG is abandoned in favour of a focus on what can be proved from evidence. It operates with evidence sufficient only to partly complete the underlying JPD. In Huffman coding, on the other hand one minimises a global information-theoretic construct (the length of a binary code needed to represent training data and DAG connectivity) - but not, as we propose above, the local Shannon entropy at every effect node. Concerns for global information minimisation were also described in (Cheeseman 1983) and used for expert system building. The idea was essentially practical: complete the JPD with missing values by fixing these to maximise the Shannon entropy of the JPD.

1.2 Preliminary Definitions

We illustrate our local ME approach by working out in detail the maximum entropy CPTs for 3-cause and 4-cause nodes. The former arises in simple hardware systems with one failure mode arising from joint action of three independent causal factors. For nodes of arity n, we need a ladder representation based on ternary nodes. The general idea and notation is as follows: Consider Figure 1.2 which shows a factor

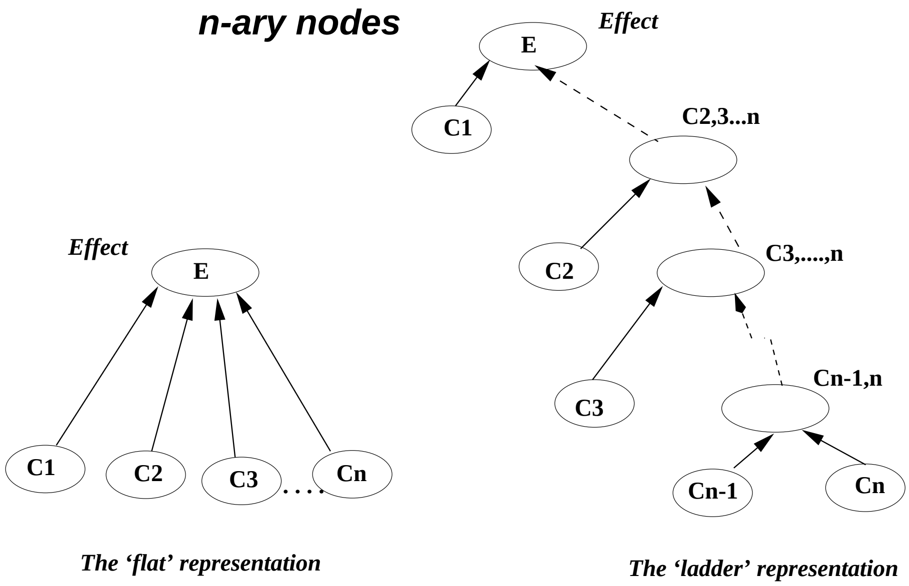

Figure 1: n causes

node E of a CPT with parents $C_1, C_2, C_3, \ldots, C_n$ where variables $C_1, C_2, \ldots, C_n$ take boolean values $c_1, c_2, \ldots, c_n$, where $c_i \in \{0, 1\}$. *Two* representations of the node are shown, the 'flat' representation and the 'ladder'.The ladder graph shows the decomposition of an n-ary node as a ladder of successive ternary nodes involving joint causes, and allows us to use ternary node solutions to build up the CPT for the n-ary case.

The associated probability distribution functions (pdfs) are denoted:

$$p(c_1, c_2, \ldots, c_n) \equiv P((C_1 = c_1) \wedge (C_2 = c_2) \wedge \ldots \wedge (C_n = c_n))$$
$$p(E \mid c_1, c_2, \ldots, c_n) \equiv P(E \mid (C_1 = c_1) \wedge (C_2 = c_2) \wedge \ldots \wedge (C_n = c_n))$$
$$p(c_1, c_2, \ldots, c_n \mid E) \equiv P((C_1 = c_1) \wedge (C_2 = c_2) \wedge \ldots \wedge (C_n = c_n) \mid E)$$
$$p(c_1, c_2, \ldots, c_n; \; E) \equiv P((C_1 = c_1) \wedge (C_2 = c_2) \wedge \ldots \wedge (C_n = c_n) \wedge E)$$

From Bayes' rule:

$$p(c_1, c_2, \ldots, c_n; \; E) = p(c_1, c_2, \ldots, c_n) p(E \mid c_1, c_2, \ldots, c_n)$$

2 Nodes of Arity 3

We develop the maximum entropy calculation for nodes of small arity, commencing with arity 3. Consider the probability distribution for nodes of arity 3 defined by $u_i = p(c_1^i, c_2^i, c_3^i;\ \ E),\ c^i \in \{0,1\},\ i = 1 \ldots 8$. Variables u_i can be determined from the following linear equations:

$$\begin{aligned}
&u_1 + u_2 + u_3 + u_4 = P(C_1 \wedge E) = P(C_1)P(E \mid C_1) = v_1 \\
&u_5 + u_6 + u_7 + u_8 = P(\neg\, C_1 \wedge E) = v_2 \\
&u_1 + u_2 + u_5 + u_6 = P(C_2 \wedge E) = P(C_2)P(E \mid C_2) = v_3 \\
&u_3 + u_4 + u_7 + u_8 = P(\neg\, C_2 \wedge E) = v_4 \\
&u_1 + u_3 + u_5 + u_7 = P(C_3 \wedge E) = v_5 \\
&u_2 + u_4 + u_6 + u_8 = P(\neg\, C_3 \wedge E) = v_6
\end{aligned}$$

This system of linear equations is under-determined, as is the case for higher arity nodes. Here, there are 6 equations in 8 unknowns.Inspection of these reveals the fact that row 1 plus row 2 equals row 3 plus row 4 which in turn equals row 5 plus row 6. We thus have 4 independent equations in 8 unknowns, and therefore expect that the solution will require 4 parameters $x_1 \ldots x_4$.

2.1 Solution of the linear equations

To solve in a systematic way, we write the equations in terms of an augmented matrix and use row manipulation to transform the augmented matrix to its echelon form. The most straightforward way is to utilise our knowledge about the dependencies: thus R3 becomes R3 -R2 + R4 - R1, a null row, and similarly for R6. The augmented matrix and its re-arranged form is

$$\begin{pmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & v_1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & v_2 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & v_3 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & v_4 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & v_5 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & v_6
\end{pmatrix}
\rightarrow
\begin{pmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & v_1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & v_6 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & v_4 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & v_2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

As expected, the matrix indicates 4 linearly independent equations in 8 unknowns, so the solution set involves parameters. We fix these by letting: $u_4 = x_1, u_6 = x_2, u_7 = x_3, u_8 = x_4$. Substituting these values we thus obtain solutions

$$\begin{aligned}
&u_1 = -v_4 - v_2 + v_5 + x_1 + x_2 + x_3 + 2x_4; \\
&u_2 = v_6 - x_1 - x_2 - x_4; \\
&u_3 = v_4 - x_1 - x_3 - x_4; \\
&u_5 = -x_2 - x_3 - x_4 + v_2
\end{aligned}$$

2.2 Maximum Entropy Calculation

The maximum entropy principle was originally developed by Jaynes (Jaynes 1957). The total entropy or *uncertainty* of the above pdf is $H(x_1, x_2, x_3, x_4)$ where $H(x_1, x_2, x_3, x_4) = -\Sigma_i u_i \log u_i$ Let

$$H_j = \frac{\partial}{\partial x_j} H(x_1, x_2, x_3, x_4) \qquad (j = 1 \ldots 4)$$

By maximising the uncertainty H we *minimise* the Shannon information and this requires $H_j = 0,\ j = 1 \ldots 4$. The simplest way of calculating H_j is to utilise the chain rule:

$$\partial(u_i \log u_i)/\partial x_j = (1 + \log u_i)\partial u_i/\partial x_j.$$

Setting $H_j = 0,\ j = 1 \ldots 4$ and simplifying yields:

$$u_3/u_1 = u_8/u_6; \qquad u_4/u_2 = u_8/u_6; \qquad u_4/u_2 = u_7/u_5$$

To solve we require a multiplier λ

$$u_4/u_2 = u_3/u_1 = u_8/u_6 = u_7/u_5 = \lambda$$

Substitution into the equations in u_i yields $\lambda = v_4/v_3$ and hence u_i in terms of v_i which determines the CPT. Substituting for $u_1 \ldots u_8$ and $v_1 \ldots v_6$ and we obtain the following solution:

$$P(C_1 \wedge C_2 \wedge C_3 \wedge E) = \frac{P(C_1 \wedge E)P(C_2 \wedge E)P(C_3 \wedge E)}{P(E)^2}$$
$$P(C_1 \wedge C_2 \wedge \neg C_3 \wedge E) = \frac{P(C_1 \wedge E)P(C_2 \wedge E)P(\neg C_3 \wedge E)}{P(E)^2}$$
$$\ldots$$
$$P(\neg C_1 \wedge \neg C_2 \wedge \neg C_3 \wedge E) = \frac{P(\neg C_1 \wedge E)P(\neg C_2 \wedge E)P(\neg C_3 \wedge E)}{P(E)^2}$$

We observe that the CPT we obtain is the CPT we would have obtained if we had assumed that $C_1 \wedge E, C_2 \wedge E, C_3 \wedge E$ were independent events. The (simpler) derivation for nodes of arity 2 and more complex derivation for nodes of arity 4 also yield a result consistent with independence of $C_i \wedge E$. The next section attempts to generalise, to nodes of arity n.

3 Extension to nodes of arity n

Consider $C_1, C_2, \ldots, C_n$ which take boolean values $c_1, c_2, \ldots, c_n$. There are $2n$ equations in $M = 2^n$ unknowns, $u_1, u_2, \ldots, u_M$. Of these, $n+1$ are independent. To solve these equations requires $L = 2^n - (n+1)$ arbitrary parameters $x_1 \ldots x_L$ and to solve in a systematic way, we again use row manipulation to transform the augmented matrix to its echelon form. An attempt can be made to obtain a solution, but the solution requires many assumptions to be made, including symmetry. The transformation of the matrix to its echelon form and subsequent calculation of maximum entropy leads us to deduce a result requiring multipliers $\lambda_0, \lambda_1, \lambda_3, \ldots$ where for example:
$u_1/u_2 = u_3/u_4 \ldots = \lambda_0$; $\quad u_1/u_3 = u_5/u_7 \ldots = \lambda_1$; $\quad u_1/u_5 = u_9/u_{13} \ldots = \lambda_2$.
We can evaluate some of the multipliers; for example the first and second linear constraints are:

$$(1 + \lambda_0)(u_2 + \ldots + u_{M/2}) = v_1; \qquad (1 + \lambda_0)(u_{2+M/2} + \ldots + u_M) = v_2$$

and we have $(1 + \lambda_0)(u_2 + \ldots + u_M) = v_1 + v_2$. The last constraint is: $u_2 + u_4 + \ldots + u_M = v_{2n}$ so that, assuming the u factor is non-zero: $(1 + \lambda_0) = (v_1 + v_2)/v_{2n}$.

Proceeding in a similar manner to the 3-node case, we conjecture that for all values of $c_i \in \{0, 1\}$:

$$P((C_1 = c_1) \wedge (C_2 = c_2) \wedge \ldots \wedge (C_n = c_n) \wedge E) = \frac{P((C_1 = c_1) \wedge E)P((C_2 = c_2) \wedge E) \ldots P((C_n = c_n) \wedge E)}{P(E)^{n-1}}$$

The conjecture is consistent with our results for small n, that the calculation of maximum entropy for a system with a factor node E with parents $C_1, C_2, C_3, \ldots, C_n$ yields the result obtainable assuming independence of $C_1 \wedge E, C_2 \wedge E, C_3 \wedge E, \ldots, C_n \wedge E$. We have not succeeded in proving the conjecture using only probability axioms and the conditional independencies that follow from causal links. The following inductive argument is inspired by the ladder decomposition, and may be called the 'ladder method'.

3.1 An Inductive Approach

To reason by induction, we assume that the maximum entropy principle is associative in the following sense: to extend from a 2-node to a 3-node one can replace C_2 by $C_{2,3}$ for the '2 node':

$$P(C_1 \wedge C_{2,3} \wedge E) = \frac{P(C_1 \wedge E)P(C_{2,3} \wedge E)}{P(E)}; \; P(C_1 \wedge \neg C_{2,3} \wedge E) = \frac{P(C_1 \wedge E)P(\neg C_{2,3} \wedge E)}{P(E)};$$
$$P(\neg C_1 \wedge C_{2,3} \wedge E) = \frac{P(\neg C_1 \wedge E)P(C_{2,3} \wedge E)}{P(E)}; \; P(\neg C_1 \wedge \neg C_{2,3} \wedge E) = \frac{P(\neg C_1 \wedge E)P(\neg C_{2,3} \wedge E)}{P(E)}$$

Since $C_{2,3} = C_2 \wedge C_3$, we can rewrite as follows, utilising the initial definitions of u_i:

$$P(C_1 \wedge C_2 \wedge C_3 \wedge E) = \frac{P(C_1 \wedge E)P(C_2 \wedge C_3 \wedge E)}{P(E)}; \; P(C_1 \wedge \neg (C_2 \wedge C_3) \wedge E) = \frac{P(C_1 \wedge E)P(\neg (C_2 \wedge C_3) \wedge E)}{P(E)}$$
$$P(\neg C_1 \wedge C_2 \wedge C_3 \wedge E) = \frac{P(\neg C_1 \wedge E)P(C_2 \wedge C_3 \wedge E)}{P(E)}; \; P(\neg C_1 \wedge \neg (C_2 \wedge C_3) \wedge E) = \frac{P(\neg C_1 \wedge E)P(\neg (C_2 \wedge C_3) \wedge E)}{P(E)}$$

If we use the results for arity 2 (assuming we can) then

$$u_1 = \frac{P(C_1 \wedge E)P(C_2 \wedge C_3 \wedge E)}{P(E)} = \frac{P(C_1 \wedge E)}{P(E)} \frac{P(C_2 \wedge E)P(C_3 \wedge E)}{P(E)}$$

To obtain the rest of the pdf, use: $P(C_1 \wedge \neg\, C_2 \wedge C_3 \wedge E) + P(C_1 \wedge C_2 \wedge C_3 \wedge E) = P(C_1 \wedge C_3 \wedge E)$. Thus $u_1 = P(C_1 \wedge C_3 \wedge E) - u_3$ (from definitions) and so

$$u_3 = P(C_1 \wedge C_3 \wedge E) - \frac{P(C_1 \wedge E)P(C_2 \wedge E)P(C_3 \wedge E)}{P(E)P(E)} = \frac{P(C_1 \wedge E)P(C_3 \wedge E)P(\neg\, C_2 \wedge E)}{P(E)^2}$$

(as before). The expressions for $u_2, u_4, u_5, u_6, u_7, u_8$ can also be determined and also yield the same result. Thus, assuming that maximum entropy is associative we can proceed straightaway to induction. The rule is

"Assuming that the maximum entropy criteria for a factor node E of a CPT with parents $C_1, C_2, C_3, \ldots, C_n$ for $n \geq 2$ yields, for each value of C_i:

$$P(C_1 \wedge C_2 \wedge \ldots \wedge C_n \wedge E) = \frac{P(C_1 \wedge E)P(C_2 \wedge E)\ldots P(C_n \wedge E)}{P(E)^{n-1}}$$

then for a CPT with parents $C_1, C_2, \ldots, C_{n+1}$,

$$P(C_1 \wedge C_2 \wedge \ldots \wedge C_{n+1} \wedge E) = \frac{P(C_1 \wedge E)P(C_2 \wedge E)\ldots P(C_{n+1} \wedge E)}{P(E)^{n}}$$

where the base case is $n = 2$."

From this type of associativity assumption, the result follows easily using:

$$P(C_1 \wedge \ldots \wedge C_i \wedge \ldots \wedge C_{n+1} \wedge E) \ + \ P(C_1 \wedge \ldots \wedge \neg\, C_i \wedge \ldots \wedge C_{n+1} \wedge E) = \\ P(C_1 \wedge \ldots \wedge C_{i-1} \wedge C_{i+1} \wedge \ldots C_{n+1} \wedge E)$$

4 Conclusions and Further Work

Firstly, we need to relate the associativity assumption to the more usual ideas about conditional independence in causal nets. At present the connection has not been established in a general way. Secondly, and more specifically, we need to compare the ladder method with exact solutions of low arity nodes. At present we only have explicit solutions up to arity 4. Thirdly, we need to make further comparisons with other popular belief-maintenance methods - such as the relatively rigorous Schaefer-Dempster theory of evidence and the various semi-empirical ways of expressing extended influence in terms of only a few marginal influence coefficients.

References

Cheeseman, P. (1983). A method of computing generalised bayesian probability values for expert systems. In *IJCAI- 83 Proceedings*, pp. 198 – 202.

Dawid, A. P. (2001, June). Influence diagrams for causal modelling and inference. Technical Report 221, Department of Statistical Science, UCL, UK.

Heckermann, D. (1996). A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, WA, USA.

Jaynes, E. T. (1957, May). Information theory and statistical mechanics. *Physical Review 106*(4), 620–630.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. CA, USA: Morgan Kauffman.

Pearl, J. (2000). *Causality*. Cambridge University Press, UK.

Schafer, G. (1979). *A Mathematical Theory of Evidence*. Princeton University Press, USA.