CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Radboud Repository

Radboud Repository
Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a preprint version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/60099

Please be advised that this information was generated on 2017-12-06 and may be subject to change.

Parameter Estimation in Large Causal Models

Rasa Jurgelenaite and **Peter Lucas**[1]

Abstract. The assessment of a probability distribution that is associated with a Bayesian network is a challenging task, even if its topology is sparse. Special probability distributions, based on the notion of causal independence, have therefore been proposed, as these allow defining a probability distribution in terms of Boolean combinations of local distributions. In Bayesian networks which need to model a large number of interactions among causal mechanisms even this approach becomes infeasible. We investigate the use of equivalence classes of binomial distributions as a means to define such very large Bayesian networks.

1 INTRODUCTION

As a consequence of the success of using Bayesian networks in solving realistic problems, increasingly complicated situations are being tackled. We are in particular interested in the modelling of biomedical knowledge, for example in fields such as genetics and immunology; in these fields hundreds to thousands of interactions between variables may need to be captured in a probabilistic model. Clearly, such models cannot be constructed and handled without exploiting (potentially hypothetical) knowledge about underlying causal mechanisms and associated simplifying assumptions.

The aim of the present work was to develop a theory that allows defining interactions between a huge number of causal factors.

2 PRELIMINARIES

2.1 Bayesian networks and causal modelling

A *Bayesian network* $\mathcal{B} = (G, \Pr)$ represents a factorised joint probability distribution on a set of variables V. It consists of an acyclic directed graph G, and a joint probability distribution $\Pr$ defined in terms of local probability distributions $\Pr(V_i \mid \pi(V_i))$, for each node $V_i \in V(G)$ given its parents $\pi(V_i)$. In this paper, we assume all variables to be binary; as an abbreviation, we will often use v_i to denote $V_i = \top$ (true) and $\bar{v}_i$ to denote $V_i = \bot$ (false). Bayesian networks are often seen as attractive tools because of the ease with which cause-effect relationships can be modelled.

2.2 Probabilistic representation of interactions

Causal independence [3] is a popular way to specify interactions among cause variables. The global structure of a causal independence model is shown in Figure 1; it expresses the idea that causes $C_1, \ldots, C_n$ influence a given common effect E through intermediate variables $I_1, \ldots, I_n$ and a deterministic function f, called the *interaction function*. The conditional probability of the occurrence of the effect E given the causes $C_1, \ldots, C_n$ can be computed as follows [3]:

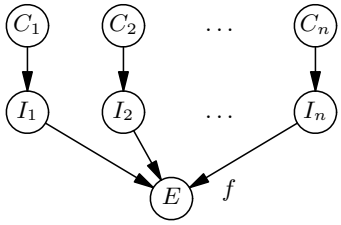

Figure 1. Causal independence model.

$$\Pr(e \mid C_1, \ldots, C_n) = \sum_{f(I_1,\ldots,I_n)=e} \prod_{1 \leq k \leq n} \Pr(I_k \mid C_k) \qquad (1)$$

Absent causes do not contribute to the effect, i.e. $\Pr(i_k \mid \bar{c}_k) = 0$. As an example, consider the interaction between insulin and glucagon, two important hormones involved in the regulation of glucose levels in blood; their effect on glucose levels in blood can be modelled by means of an exclusive OR ($\otimes$).

2.3 Symmetric causal independence models

The function f in equation (1) is actually a Boolean function. However, there are 2^{2^n} different n-ary Boolean functions [2]. Consequently, the potential number of causal interaction models is huge. However, in the case of causal independence it is usually assumed that the function f is decomposable to identical, binary functions. In addition, it is attractive to assume that the order of the cause variables does not matter; thus, it makes sense to restrict causal independence models to symmetric Boolean functions, where the order of arguments is irrelevant.

There are 8 symmetric binary Boolean functions, of which 6 suitable as a basis for defining Boolean functions, as these are all commutative and associative [3]. Logical truth and falsity are constants, and act as the global extremes in a partial order among Boolean functions. As such they give rise to trivial causal independence models. The remaining four causal independence models are defined in terms of the logical OR, AND, XOR and bi-implication. We use $*$ to denote a commutative, associative binary operator. Table 1 gives the truth tables for the n-nary Boolean functions of interest.

Table 1. The truth tables for some n-ary symmetric Boolean functions; $k = \sum_{j=1}^{n} \nu(I_j)$, with $\nu(I_j) = 1$ if I_j is equal to true and 0 otherwise.

$I_1 \vee \cdots \vee I_n$	$I_1 \wedge \cdots \wedge I_n$	$I_1 \otimes \cdots \otimes I_n$	$I_1 \leftrightarrow \cdots \leftrightarrow I_n$
$k \geq 1$	$k = n$	*odd*(k)	*even*$(n-k)$

[1] Department of Information and Knowledge Systems, Radboud University Nijmegen, Nijmegen, The Netherlands, email: {rasa, peterl}@cs.kun.nl

Due to space limitations, we only consider XOR and bi-implication in this paper. The function $f_{\otimes}(I_1, \ldots, I_n)$ yields the value *true* if there are an odd number of variables I_j with the value *true*. Therefore, in order to determine the probability of the effect variable E, $\Pr(e \mid C_1, \ldots, C_n)$, the probabilities for all cause variable combinations with an odd number of present causes have to be added. We have:

$$\Pr_{\otimes}(e|C_1, ..., C_n) = \sum_{I_1 \otimes \cdots \otimes I_n} \prod_{k=1}^{n} \Pr(I_k|C_k)$$
$$= \Pr(\bar{\imath}_1|C_1) \cdots \Pr(\bar{\imath}_n|C_n) \cdot$$
$$\sum_{\substack{1 \leq k \leq n \\ odd(k)}} \sum_{j_1=j_0+1}^{n-k+1} \cdots \sum_{j_t=j_{t-1}+1}^{n-k+t} \frac{\Pr(i_{j_1}|C_{j_1})}{\Pr(\bar{\imath}_{j_1}|C_{j_1})} \cdots \frac{\Pr(i_{j_t}|C_{j_t})}{\Pr(\bar{\imath}_{j_t}|C_{j_t})} \quad (2)$$

where $t = 1, \ldots, k$ and $j_0 = 0$.

The function value $f_{\leftrightarrow}(I_1, \ldots, I_n)$ is *true* if there are an even number of variables I_j with the value *false*. Thus, to determine $\Pr(e \mid C_1, \ldots, C_n)$ the probabilities for all cause variable combinations with an even number of absent causes have to be added:

$$\Pr_{\leftrightarrow}(e \mid C_1, \ldots, C_n) = \sum_{I_1 \leftrightarrow \cdots \leftrightarrow I_n} \prod_{k=1}^{n} \Pr(I_k \mid C_k)$$
$$= \Pr(i_1 \mid C_1) \cdots \Pr(i_n \mid C_n) \cdot$$
$$\left(1 + \sum_{\substack{1 \leq k \leq n \\ even(k)}} \sum_{j_1=j_0+1}^{n-k+1} \cdots \sum_{j_t=j_{t-1}+1}^{n-k+t} \frac{\Pr(\bar{\imath}_{j_1}|C_{j_1})}{\Pr(i_{j_1}|C_{j_1})} \cdots \frac{\Pr(\bar{\imath}_{j_t}|C_{j_t})}{\Pr(i_{j_t}|C_{j_t})} \right) \quad (3)$$

where $t = 1, \ldots, k$ and $j_0 = 0$.

3 EQUIVALENCE CLASSES OF BINOMIAL DISTRIBUTIONS

The larger the number of causal mechanisms n becomes, the more likely that the parameters $\Pr(I_k \mid C_k)$ of a causal independence model become arbitrarily close to each other. Hence, one way to simplify the estimation of the probability distribution is to group parameters in particular equivalence classes.

The binomial distribution is one of the most commonly used discrete probability distribution. Cause variables can be treated as trials of an experiment satisfying the requirements of a binomial distribution, as the number of cause variables n is known in advance, all cause variables have two states, are independent, and the probability of occurrence of each cause is the same.

We organise the intermediate variables $I_1, \ldots, I_n$ and their associated variables $C_1, \ldots, C_n$ by their influence on the common effect E, in accordance to the increasing order of the associated probabilistic parameters $\Pr(I_k \mid C_k)$. Next, we choose a small $\varepsilon \in \mathbb{R}^+$, which determines how much the probabilities may vary inside an equivalence class. An intermediate variable I_k belongs to the t-th equivalence class if its probability of success $\Pr(i_k \mid C_k)$ falls into the interval $[2(t-1)\varepsilon, 2t\varepsilon)$; we also assume that $\Pr(i_t \mid C_t) = (2t-1)\varepsilon$.

4 ANALYSIS OF PROBABILISTIC BEHAVIOUR

In this section, we study the properties of the causal independence models introduced above.

Let $S_1^*, S_2^*, \ldots$ be a sequence, abbreviated to $\langle S_n^* \rangle$; throughout this section, a member S_n^* of this sequence represents a sum of products of probability distribution in an equivalence class of binomial distributions, i.e.: $S_n^* = \sum_{I_1 * \cdots * I_n} \prod_{t=1}^{n} \Pr(I_t \mid C_t)$. We assume the probability $\Pr(i_t \mid C_t)$ to be constant, i.e. $p = \Pr(i_t \mid C_t)$.

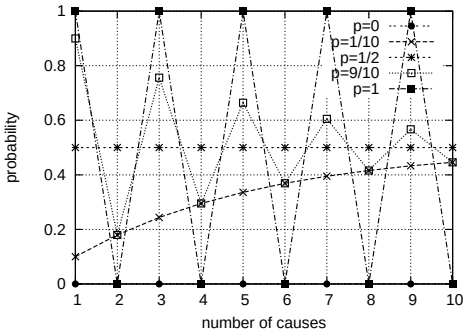

Figure 2. Patterns of the XOR causal independence model.

Due to lack of space, only the situation for the XOR and bi-implication causal independence models are considered here. In addition to the expected bounds of 0 and 1, the sequences have an additional bound at $\frac{1}{2}$.

Proposition 1 *Let $\langle S_n^* \rangle$ be a sequence as defined above. For each member S_n^* of the sequence it holds that:*

- *if $p \in [0, \frac{1}{2})$ then $S_n^* \in [p, \frac{1}{2})$ for $* = \otimes$, and $S_n^* \in [p, \frac{1}{2}) \cup (\frac{1}{2}, p^2 + (1-p)^2]$ for $* = \leftrightarrow$;*
- *otherwise, if $p \in (\frac{1}{2}, 1]$ then $S_n^* \in [2p(1-p), \frac{1}{2}) \cup (\frac{1}{2}, p]$ for $* = \otimes$, and $S_n^* \in (\frac{1}{2}, p]$ for $* = \leftrightarrow$.*

Proposition 2 *A sequence $\langle S_n^* \rangle$ is*

- *strictly monotonically increasing if $p \in (0, \frac{1}{2})$ and $* = \otimes$,*
- *strictly monotonically decreasing if $p \in (\frac{1}{2}, 1)$ and $* = \leftrightarrow$,*
- *constant $S_n^* = p$ if $p \in \{0, \frac{1}{2}\}$ and $* = \otimes$, $p \in \{\frac{1}{2}, 1\}$ and $* = \leftrightarrow$,*
- *non monotonic if $p \in (\frac{1}{2}, 1]$ and $* = \otimes$, $p \in [0, \frac{1}{2})$ and $* = \leftrightarrow$.*

The propositions above yield insight into the behaviour of the sequences but leave questions about non-monotonic behaviour unanswered. We have proved (not shown here) that the sequences converge to $\frac{1}{2}$. As $F'(S_n^*) = |1 - 2p|$ for $* \in \{\otimes, \leftrightarrow\}$ the rate of convergence depends on the value of p; the closer the value of p is to $\frac{1}{2}$, the faster the sequence converges to $\frac{1}{2}$. Figure 2 illustrates this behaviour; the plot for the bi-implication is similar.

5 DISCUSSION

In this paper, we addressed the problem of parameter estimation in very large Bayesian networks. Our solution was to group local probability distributions into equivalence classes using probability intervals, and to use a suitably defined probability distribution as a basis for assessment. As far as we know, this is the first paper offering a systematic analysis of the global probabilistic patterns that occur in large Bayesian networks based on the theory of causal independence.

REFERENCES

[1] F.J.Díez, *Parameter adjustment in Bayes networks. The generalized noisy OR-gate*. UAI'93, pp. 99-105, 1993.

[2] H.B. Enderton, *A Mathematical Introduction to Logic*. Academic Press, San Diego, 1972.

[3] P.J.F. Lucas, *Bayesian network modelling by qualitative patterns*. Proc ECAI-2002, pp. 690-694, 2002.