Parameter Estimation in Large Causal Independence Models

Rasa Jurgelenaite Peter Lucas

Department of Information and Knowledge Systems
Radboud University Nijmegen, The Netherlands
Email: {rasa,peter}@cs.ru.nl

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

Bayesian networks offer an appealing language with associated set of tools for building models of domains with inherent uncertainty. However, a significant bottle-neck in constructing Bayesian networks, whether done manually or by learning from data, is the size of their underlying probability tables. The theory of causal independence has been put forward as systematic way to cope with this situation; it allows decomposing a probability distribution in terms of Boolean interactions among local parameters.

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.

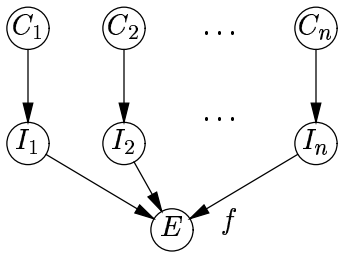

Figure 1: Causal independence model.

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 $\mathbf{V}$. It consists of two parts: (1) a qualitative part, represented as an acyclic directed graph $G = (\mathbf{V}(G), \mathbf{A}(G))$, whose vertices $\mathbf{V}(G)$ correspond to the random variables in $\mathbf{V}$, and arcs $\mathbf{A}(G) \subseteq \mathbf{V}(G) \times \mathbf{V}(G)$ represent the conditional (in)dependencies between the variables; (2) a quantitative part Pr consisting of local probability distributions $\Pr(V \mid \pi(V))$, for each vertex $V \in \mathbf{V}(G)$ given its parents $\pi(V)$. In this paper, we assume all variables to be binary; as an abbreviation, we will often use v to denote $V = \top$ (true) and $\bar{v}$ to denote $V = \bot$ (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 impact of each cause C_k on the common effect E is independent of each other cause $C_j, j \neq k$. The function f represents in which way the intermediate effects I_k, and indirectly also the causes C_k, interact to yield the final effect E. Hence, the function f is defined in such a way that when a relationship, as modelled by the function f, between $I_k, k = 1, \ldots, n$, and $E = \top$ is satisfied, then it holds that $e = f(I_1, \ldots, I_n)$. It is assumed that $\Pr(e \mid I_1, \ldots, I_n) = 1$ if $f(I_1, \ldots, I_n) = e$, and $\Pr(e \mid I_1, \ldots, I_n) = 0$ if $f(I_1, \ldots, I_n) = \bar{e}$.

The conditional probability of the occurrence of the effect E given the

causes $C_1, \ldots, C_n$ can be computed as follows [3]:

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

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 the blood, which can be modelled by means of an exclusive OR ($\otimes$) (See Figure 2).

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, 4]. 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 [4].

There are 8 symmetric binary Boolean functions, of which 6 suitable as a basis for defining n-ary 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. From now one, the following notation is adopted: $\vee$ (OR), $\wedge$ (AND), $\otimes$ (exclusive OR), $\leftrightarrow$ (bi-implication).

Recall that the function $f_\vee(I_1, \ldots, I_n)$ yields the value *true* if there is at least one variable I_j with the value *true*. Therefore, the probability

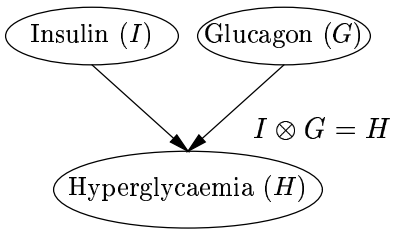

Figure 2: Causal interaction, described as an exclusive OR, denoted by $\otimes$.

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.

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

distribution for the OR causal independence model is defined as follows:

$$\Pr_{\vee}(e \mid C_1, \ldots, C_n) \;=\; 1 - \prod_{k=1}^{n} \Pr(\bar{\imath}_k \mid C_k) \tag{2}$$

For the AND causal independence model we obtain:

$$\Pr_{\wedge}(e \mid C_1, \ldots, C_n) \;=\; \prod_{k=1}^{n} \Pr(i_k \mid C_k) \tag{3}$$

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:

$$\begin{aligned}
&\Pr_{\otimes}(e \mid C_1, ..., C_n) = \sum_{I_1 \otimes \cdots \otimes I_n} \prod_{k=1}^{n} \Pr(I_k \mid C_k) \\
&= \Pr(\bar{\imath}_1 \mid C_1) \cdots \Pr(\bar{\imath}_n \mid C_n) \cdot \\
&\sum_{\substack{1 \leq k \leq n \\ odd(k)}} \sum_{j_1=1}^{n-k+1} \cdots \sum_{j_t=j_{t-1}+1}^{n-k+t} \cdots \sum_{j_k=j_{k-1}+1}^{n} \frac{\Pr(i_{j_1} \mid C_{j_1})}{\Pr(\bar{\imath}_{j_1} \mid C_{j_1})} \cdots \frac{\Pr(i_{j_k} \mid C_{j_k})}{\Pr(\bar{\imath}_{j_k} \mid C_{j_k})}
\end{aligned} \tag{4}$$

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:

$$\begin{aligned}
&\Pr_{\leftrightarrow}(e \mid C_1, \ldots, C_n) = \sum_{I_1 \leftrightarrow \cdots \leftrightarrow I_n} \prod_{k=1}^{n} \Pr(I_k \mid C_k) \\
&= \Pr(i_1 \mid C_1) \cdots \Pr(i_n \mid C_n) \cdot \\
&\left(1 + \sum_{\substack{1 \leq k \leq n \\ even(k)}} \sum_{j_1=1}^{n-k+1} \cdots \sum_{j_t=j_{t-1}+1}^{n-k+t} \cdots \sum_{j_k=j_{k-1}+1}^{n} \frac{\Pr(\bar{\imath}_{j_1} | C_{j_1})}{\Pr(i_{j_1} | C_{j_1})} \cdots \frac{\Pr(\bar{\imath}_{j_k} | C_{j_k})}{\Pr(i_{j_k} | C_{j_k})} \right)
\end{aligned} \tag{5}$$

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 distributions. 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 $[2(t-1)\varepsilon, 2t\varepsilon)$. The number of equivalence classes is equal to $r = \frac{1}{2\varepsilon}$. Further, we assume that all intermediate variables from the same equivalence class have the same probability of success $\Pr(i_t \mid C_t) = (2t-1)\varepsilon$ and apply the concepts of the binomial distribution to estimate the probability distribution of the t-th equivalence class $\sum_{I_{m_t} * \cdots * I_{m_t+n_t-1}} \prod_{k=m_t}^{m_t+n_t-1} \Pr(I_k|C_k)$, where $C_{m_t}, \ldots, C_{m_t+n_t-1}$ are the cause variables that belong to the t-th equivalence class, m_t and n_t respectively denote the index of the first variable and the number of variables in the equivalence class.

To determine the probability distribution of the effect variable E based on the probability distributions of contributing equivalence classes, exactly the same combining functions are employed as when combining single probability distributions $\Pr(I_k \mid C_k)$ associated with cause variables C_k.

4 Analysis of probabilistic behaviour

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

Section 3 mentioned a scheme to combine the effects of the individual equivalence classes. Here it is therefore permitted to restrict the mathematical analysis to one equivalence class of binomial distributions as the

analysis for the other equivalence classes is identical.

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. $p = \Pr(i_t \mid C_t)$.

In our treatment we combine various causal independence models based on similarity in behaviour. For example, the OR and AND causal independence models possess similar behaviours, which in most cases appear to be each other opposites. Analogous remarks can be made for the two other types of causal independence models. The following propositions show that OR and AND causal independence models yield monotonic behaviour for any probability p with the exception of the bounds $p \in \{0, 1\}$. The proofs are omitted because of lack of space.

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 $p \in (0, 1)$ then $S_n^* \in [p, 1)$ for $* = \vee$, and $S_n^* \in (0, p]$ for $* = \wedge$; otherwise, if $p \in \{0, 1\}$ then $S_n^* = p$ for both $* = \vee$ and $* = \wedge$.*

Proposition 2 *If $p \in (0, 1)$ then a sequence $\langle S_n^* \rangle$ is strictly monotonically increasing for $* = \vee$, and strictly monotonically decreasing for $* = \wedge$.*

It appears that the sequences converge to one of their bounds. As we try to understand the behaviour of large causal independence models, the rate of convergence is clearly also relevant. The first derivative of $F(S_n^*) = S_{n+1}^*$ can serve as a basis for this. If $* = \vee$ then $F(S_n^*) = 1 - (1 - p)S_n^*$; thus the larger value of p, the faster the sequence converges to 1. If $* = \wedge$ then $F(S_n^*) = pS_n^*$; thus the smaller value of p, the faster the sequence converges to 0. Figure 3(left) illustrates the results for the OR; the plot for the AND shows similar, however decreasing, behaviour.

The study of the properties of the causal independence models with XOR and bi-implication interactions revealed surprisingly complicated behaviours. In addition to the expected bounds of 0 and 1, the sequences has an additional bound at $\frac{1}{2}$.

Proposition 3 *Let $\langle S_n^* \rangle$ be a sequence as defined above. For each member S_n^* of the sequence it holds that:*

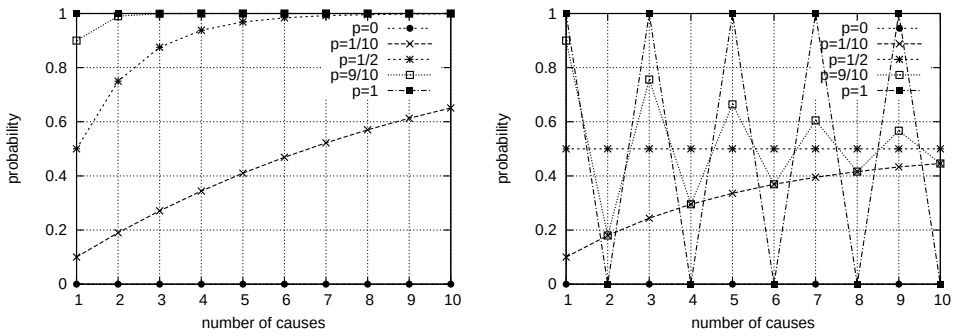

Figure 3: Patterns of causal independence model. Left: OR; right: XOR.

- *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 4 *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, i.e. when $p \in (\frac{1}{2}, 1]$, $* = \otimes$, and $p \in [0, \frac{1}{2})$, $* = \leftrightarrow$. Let the sequence $\langle S_n^* \rangle$ be divided into two sequences: $S_1^*, S_3^*, \ldots$, denoted by $\langle S_{odd(n)}^* \rangle$, and $S_2^*, S_4^*, \ldots$, denoted by $\langle S_{even(n)}^* \rangle$. We have the following proposition:

Proposition 5 *Let* $\langle S_{odd(n)}^* \rangle$ *and* $\langle S_{even(n)}^* \rangle$ *be sequences as defined above. For each member of the sequences it holds that:*

- *if* $* = \otimes$ *and* $p \in (\frac{1}{2}, 1]$ *then* $S_{odd(n)}^* \in (\frac{1}{2}, p]$, $S_{even(n)}^* \in [2p(1-p), \frac{1}{2})$;
- *if* $* = \leftrightarrow$ *and* $p \in [0, \frac{1}{2})$ *then* $S_{odd(n)}^* \in [p, \frac{1}{2})$, $S_{even(n)}^* \in (\frac{1}{2}, p^2 + (1-p)^2]$.

Proposition 6 *Let* $\langle S_{odd(n)}^* \rangle$ *and* $\langle S_{even(n)}^* \rangle$ *be sequences as defined above. Then it holds that:*

- *if* $p \in (\frac{1}{2}, 1]$ *and* $* = \otimes$ $\langle S^*_{odd(n)} \rangle$ *is strictly monotonically decreasing, and* $\langle S^*_{even(n)} \rangle$ *is strictly monotonically increasing;*

- *if* $p \in [0, \frac{1}{2})$ *and* $* = \leftrightarrow$ $\langle S^*_{odd(n)} \rangle$ *is strictly monotonically increasing, and* $\langle S^*_{even(n)} \rangle$ *is strictly monotonically decreasing.*

From the propositions above we conclude that despite their complicated behaviours, 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 3(right) 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. Quite naturally, the theory of causal independence served as a starting point for such networks. As was argued, even if resorting to this theory it quickly becomes infeasible to assess probability distributions for 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.

[4] I. Wegener, *The complexity of Boolean Functions.* John Wiley, New York, 1987.