EM Algorithm for Symmetric Causal Independence Models

Rasa Jurgelenaite and Tom Heskes

Institute for Computing and Information Sciences, Radboud University Nijmegen,
Toernooiveld 1, 6525 ED Nijmegen, The Netherlands
{rasa, tomh}@cs.ru.nl

Abstract. Causal independence modelling is a well-known method both for reducing the size of probability tables and for explaining the underlying mechanisms in Bayesian networks. In this paper, we present the EM algorithm to learn the parameters in causal independence models based on the symmetric Boolean function. The developed algorithm enables us to assess the practical usefulness of the symmetric causal independence models, which has not been done previously. We evaluate the classification performance of the symmetric causal independence models learned with the presented EM algorithm. The results show the competitive performance of these models in comparison to noisy OR and noisy AND models as well as other state-of-the-art classifiers.

1 Introduction

Bayesian networks [1] are well-established as a sound formalism for representing and reasoning with probabilistic knowledge. However, because the number of conditional probabilities for the node grows exponentially with the number of its parents, it is usually unreliable if not infeasible to specify the conditional probabilities for the node that has a large number number of parents. The task of assessing conditional probability distributions becomes even more complex if the model has to integrate expert knowledge. While learning algorithms can be forced to take into account an expert's view, for the best possible results the experts must be willing to reconsider their ideas in light of the model's 'discovered' structure. This requires a clear understanding of the model by the domain expert. *Causal independence models* [2], [3], [4] can both limit the number of conditional probabilities to be assessed and provide the ability for models to be understood by domain experts in the field. The main idea of causal independence models is that causes influence a given common effect through intermediate variables and interaction function.

Causal independence assumptions are often used in practical Bayesian network models [5], [6]. However, most researchers restrict themselves to using only the logical OR and logical AND operators to define the interaction among causes. The resulting probabilistic submodels are called *noisy OR* and *noisy AND*; their underlying assumption is that the presence of either at least one cause or all

causes at the same time give rise to the effect. Several authors proposed to expand the space of interaction functions by other symmetric Boolean functions: the idea was already mentioned but not developed further in [7], analysis of the qualitative patterns was presented in [8], and assessment of conditional probabilities was studied in [9].

Even though for some real-world problems the intermediate variables are observable (see [10]), in many problems these variables are latent. Therefore, conditional probability distributions depend on unknown parameters which must be estimated from data, using *maximum likelihood* (ML) or *maximum a posteriori* (MAP). One of the most widespread techniques for finding ML or MAP estimates is the *expectation-maximization* (EM) algorithm. Meek and Heckerman [7] provided a general scheme how to use the EM algorithm to compute the maximum likelihood estimates of the parameters in causal independence models assumed that each local distribution function is collection of multinomial distributions. Vomlel [11] described the application of the EM algorithm to learn the parameters in the noisy OR model. However, the proposed schemes of the EM algorithm are specific to a given causal independence model, and hence not directly applicable to the general case of parameter learning in causal independence models.

Learning the parameters in causal independence models with a symmetric Boolean function as an interaction function (further referred to as the *symmetric causal independence models*) is the main topic of this paper. We develop an EM algorithm to learn the parameters in symmetric causal independence models. The presented algorithm enables us to assess the practical usefulness of this expanded class of causal independence models, which has not been done by other authors. The evaluation is done by using the symmetric causal independence models learned with the developed EM algorithm as classifiers. Experimental results show the competitive classification performance of these models in comparison with the noisy OR classifier as well as other widely-used classifiers.

The remainder of this paper is organised as follows. In the following section, we review Bayesian networks and discuss the semantics of symmetric causal independence models. In Section 3, we first describe the general scheme of the EM algorithm and then develop the EM algorithm for finding the parameters in symmetric causal independence models. Section 4 presents the experimental results, and conclusions are drawn in Section 5.

2 Symmetric Boolean Functions for Modelling Causal Independence

2.1 Bayesian Networks

A *Bayesian network* $\mathcal{B} = (G, \Pr)$ represents a factorised joint probability distribution on a set of random variables $\mathbf{V}$. It consists of two parts: (1) a qualitative part, represented as an acyclic directed graph (ADG) $G = (\mathbf{V}(G), \mathbf{A}(G))$, where there is a 1–1 correspondence between the vertices $\mathbf{V}(G)$ and the random variables in $\mathbf{V}$, and arcs $\mathbf{A}(G)$ 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 variable $V \in \mathbf{V}$ given the parents $\pi(V)$ of the corresponding vertex (interpreted as variables). The joint probability distribution Pr is factorised according to the structure of the graph, as follows:

$$\Pr(\mathbf{V}) = \prod_{V \in \mathbf{V}} \Pr(V \mid \pi(V)) \; .$$

Each variable $V \in \mathbf{V}$ has a finite set of mutually exclusive states. In this paper, we assume all variables to be binary; as an abbreviation, we will often use v^+ to denote $V = \top$ (true) and v^- to denote $V = \bot$ (false). We interpret $\top$ as 1 and $\bot$ as 0 in an arithmetic context. An expression such as

$$\sum_{\psi(H_1,\ldots,H_n)=\top} g(H_1, \ldots, H_n)$$

stands for summing $g(H_1, \ldots, H_n)$ over all possible values of the variables H_k for which the constraint $\psi(H_1, \ldots, H_n) = \top$ holds.

2.2 Semantics of Symmetric Causal Independence Models

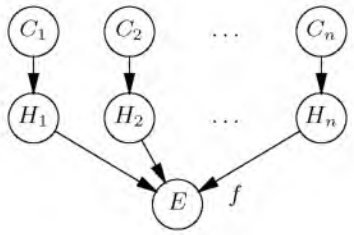

Fig. 1. Causal independence model.

Causal independence (also known as independence of causal influence) 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 hidden variables $H_1, \ldots, H_n$ and a deterministic function f, called the *interaction function*. The impact of each cause C_i on the common effect E is independent of each other cause $C_j, j \neq i$. The hidden variable H_i is considered to be a contribution of the cause variable C_i to the common effect E. The function f represents in which way the hidden effects H_i, and indirectly also the causes C_i, 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 $H_i, i = 1, \ldots, n$, and $E = \top$ is satisfied, then it holds that $f(H_1, \ldots, H_n) = \top$. It is assumed that $\Pr(e^+ \mid H_1, \ldots, H_n) = 1$ if $f(H_1, \ldots, H_n) = \top$, and $\Pr(e^+ \mid H_1, \ldots, H_n) = 0$ if $f(H_1, \ldots, H_n) = \bot$.

A causal independence model is defined in terms of the causal parameters $\Pr(H_i \mid C_i)$, for $i = 1, \ldots, n$ and the function $f(H_1, \ldots, H_n)$. Most papers on causal independence models assume that absent causes do not contribute to the effect [1]. In terms of probability theory this implies that it holds that $\Pr(h_i^+ \mid c_i^-) = 0$; as a consequence, it holds that $\Pr(h_i^- \mid c_i^-) = 1$. In this paper we make the same assumption.

In situations in which the model does not capture all possible causes, it is useful to introduce a *leaky cause* which summarizes the unidentified causes contributing to the effect and is assumed to be always present [12]. We model this leak term by adding an additional input $C_{n+1} = 1$ to the data; in an arithmetic context the leaky cause is treated in the same way as identified causes.

The conditional probability of the occurrence of the effect E given the causes $C_1, \ldots, C_n$, i.e., $\Pr(e^+ \mid C_1, \ldots, C_n)$, can be obtained from the causal parameters $\Pr(H_l \mid C_l)$ as follows [4]:

$$\Pr(e^+ \mid C_1, \ldots, C_n) = \sum_{f(H_1,\ldots,H_n)=\top} \prod_{i=1}^{n} \Pr(H_i \mid C_i) \; . \tag{1}$$

In this paper we assume that the function f in Equation (1) is a Boolean function. However, there are 2^{2^n} different n-ary Boolean functions [13], [14]; thus, the potential number of causal interaction models is huge. However, if we assume that the order of the cause variables does not matter, the Boolean functions become *symmetric* [14] and the number reduces to 2^{n+1}.

An important symmetric Boolean function is the *exact* Boolean function ϵ_l, which has function value true, i.e. $\epsilon_l(H_1, \ldots, H_n) = \top$, if $\sum_{i=1}^{n} \nu(H_i) = l$ with $\nu(H_i)$ equal to 1, if H_i is equal to true and 0 otherwise. A symmetric Boolean function can be decomposed in terms of the exact functions ϵ_l as [14]:

$$f(H_1, \ldots, H_n) = \bigvee_{i=0}^{n} \epsilon_i(H_1, \ldots, H_n) \wedge \gamma_i \tag{2}$$

where γ_i are Boolean constants depending only on the function f. For example, for the Boolean function defined in terms of the OR operator we have $\gamma_0 = \bot$ and $\gamma_1 = \ldots = \gamma_n = \top$.

Another useful symmetric Boolean function is the *threshold* function τ_k, which simply checks whether there are at least k trues among the arguments, i.e. $\tau_k(H_1, \ldots, H_n) = \top$, if $\sum_{j=1}^{n} \nu(H_j) \geq k$ with $\nu(H_j)$ equal to 1, if H_j is equal to true and 0 otherwise. To express it in the Boolean constants we have: $\gamma_0 = \cdots = \gamma_{k-1} = \bot$ and $\gamma_k = \cdots = \gamma_n = \top$. Causal independence model based on the Boolean threshold function further will be referred to as the *noisy threshold models*.

2.3 The Poisson Binomial Distribution

Using the property of Equation (2) of the symmetric Boolean functions, the conditional probability of the occurrence of the effect E given the causes $C_1, \ldots, C_n$

can be decomposed in terms of probabilities that exactly l hidden variables $H_1, \ldots, H_n$ are true, as follows:

$$\Pr(e^+ \mid C_1, \ldots, C_n) = \sum_{\substack{0 \leq l \leq n \\ \gamma_l}} \sum_{\epsilon_l(H_1,\ldots,H_n)} \prod_{i=1}^{n} \Pr(H_i \mid C_i) \ . \tag{3}$$

Let l denote the number of successes in n independent trials, where p_i is a probability of success in the ith trial, $i = 1, \ldots, n$; let $\mathbf{p} = (p_1, \ldots, p_n)$, then $\mathrm{B}(l; \mathbf{p})$ denotes the *Poisson binomial distribution* [15]:

$$\mathrm{B}(l; \mathbf{p}) = \left\{ \prod_{i=1}^{n} (1 - p_i) \right\} \sum_{1 \leq j_1 < \ldots < j_l \leq n} \prod_{z=1}^{l} \frac{p_{j_z}}{1 - p_{j_z}} \ . \tag{4}$$

Let us define a vector of probabilistic parameters $\mathbf{p}(C_1, \ldots, C_n) = (p_1, \ldots, p_n)$ with $p_i = \Pr(h_i^+ \mid C_i)$. Then the connection between the Poisson binomial distribution and the class of symmetric causal independence models is as follows.

Proposition 1. *It holds that:*

$$\Pr(e^+ \mid C_1, \ldots, C_n) = \sum_{i=0}^{n} \mathrm{B}(i; \mathbf{p}(C_1, \ldots, C_n)) \gamma_i \ .$$

3 EM Algorithm

In this section, we first describe the general scheme of the EM algorithm. Then we develop the EM algorithm that finds the unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ of a symmetric causal independence model where $\theta_i = \Pr(h_i^+ \mid c_i^+)$.

3.1 Basic EM

Let $\mathbf{D} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ be a data set of independent and identically distributed settings of the observed variables in a symmetric causal independence model, where

$$\mathbf{x}^j = (\mathbf{c}^j, e^j) = (c_1^j, \ldots, c_n^j, e^j) \ .$$

We assume that no additional information about the model is available. Therefore, to learn the parameters of the model we maximize the conditional log-likelihood

$$CLL(\boldsymbol{\theta}) = \ln(CL(\boldsymbol{\theta})) = \sum_{j=1}^{N} \ln \Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) \ .$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ are unknown parameters of the model.

The expectation-maximization (EM) algorithm [16] is a general method to find the maximum likelihood estimate of the parameters in probabilistic models, where the data is incomplete or the model has hidden variables.

We start from the following simple identity:

$$\ln \Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) = \ln \Pr(\mathbf{H}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) - \ln \Pr(\mathbf{H} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}) \tag{5}$$

and take expectations of both sides, treating $\mathbf{H}$ as a random variable with the distribution $\Pr(\mathbf{H} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(old)})$, where $\boldsymbol{\theta}^{(old)}$ is the current (old) guess. The left hand side of Equation (5) does not depend on $\mathbf{H}$, so averaging over $\mathbf{H}$ yields

$$\begin{aligned} \ln \Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) = & \sum_{\mathbf{H}} \Pr(\mathbf{H} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(old)}) \ln \Pr(\mathbf{H}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) \\ & - \sum_{\mathbf{H}} \Pr(\mathbf{H} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(old)}) \ln \Pr(\mathbf{H} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}) \ . \end{aligned} \tag{6}$$

The key result for the EM algorithm is that the last term in the above equation is maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(old)}$, thus any increase of the first term on the right side of Equation (6) is guaranteed to increase the expected complete (conditional) log-likelihood.

Let us denote

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)}) = \sum_{j=1}^{N} \sum_{\mathbf{H}} \Pr(\mathbf{H} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \ln \Pr(\mathbf{H}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) \ . \tag{7}$$

The EM algorithm at each iteration maximizes this functional:

$$\boldsymbol{\theta}^{(z+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)}) \ .$$

In the next subsection, we find the values of the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ that maximize the function $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)})$ for the symmetric causal independence model.

3.2 Maximization Step

We start by transforming $\ln \Pr(\mathbf{H}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta})$ so that it becomes a sum of logarithms:

$$\ln \Pr(\mathbf{H}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) = \ln \Pr(e^j \mid \mathbf{H}) + \sum_{i=1}^{n} \ln \Pr(H_i \mid c_i^j, \theta_i) \ . \tag{8}$$

The conditional probability $\Pr(H_i \mid c_i^j, \theta_i)$ can be written in the form

$$\Pr(H_i \mid c_i^j, \theta_i) = c_i^j H_i \theta_i + c_i^j (1 - H_i)(1 - \theta_i) + (1 - c_i^j)(1 - H_i) \ . \tag{9}$$

Combining (7), (8) and (9), we obtain

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)}) = & \sum_{j=1}^{N} \sum_{\mathbf{H}} \Pr(\mathbf{H} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \cdot \\ & \left(\ln \Pr(e^j \mid \mathbf{H}) + \sum_{i=1}^{n} \ln \left(\theta_i c_i^j (2H_i - 1) + 1 - H_i \right) \right) \ . \end{aligned}$$

We can maximize this result by computing the partial derivatives of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)})$ with respect to $\theta_k : k = 1, \ldots, n$ and setting them to zero:

$$\frac{\partial Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)})}{\partial \theta_k} = \sum_{j=1}^{N} \sum_{\mathbf{H}} \Pr(\mathbf{H} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \frac{c_k^j (2H_k - 1)}{\theta_k c_k^j (2H_k - 1) + 1 - H_k} = 0 \ . \quad (10)$$

Now let us define $\mathbf{H}_{\setminus k} = \{H_1, \ldots, H_{k-1}, H_{k+1}, \ldots, H_n\}$. Then Equation (10) can be simplified writing it as a sum over the states of the hidden variable H_k:

$$\sum_{1 \leq j \leq N} c_k^j \sum_{\mathbf{H}_{\setminus k}} \left(\frac{\Pr(\mathbf{H}_{\setminus k}, h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}{\theta_k} - \frac{\Pr(\mathbf{H}_{\setminus k}, h_k^- \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}{1 - \theta_k} \right) = 0 \ .$$

It can be shown that Equation (10) is solved by

$$\theta_k = \frac{\sum_{1 \leq j \leq N} c_k^j \Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}{\sum_{1 \leq j \leq N} c_k^j} \ . \quad (11)$$

It is easy to check whether this extremum is a maximum by computing the second partial derivatives of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)})$ with respect to $\theta_k, k = 1, \ldots, n$. The matrix formed from these second partial derivatives is negative semidefinite, and hence this stationary point is indeed always a maximum of the function $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)})$.

In the next subsection, we derive the expectation step which corresponds to computing the conditional probabilities $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ for all $k = 1, \ldots, n$, $j = 1, \ldots, N$ where $c_k^j = 1$.

3.3 Expectation Step

Using Bayes rule, we can write the probability of $\mathbf{H}$ given a data sample $\mathbf{x}^j$ and the parameters $\boldsymbol{\theta}^{(z)}$ as follows:

$$\Pr(\mathbf{H} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \frac{\Pr(e^j \mid \mathbf{H}) \prod_{i=1}^{n} \Pr(H_i \mid c_i^j, \boldsymbol{\theta}^{(z)})}{\Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}^{(z)})} \ .$$

By marginalizing $\mathbf{H}_{\setminus k}$ out we obtain the conditional probability of the hidden variable H_k being true:

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \frac{\Pr(h_k^+ \mid c_k^j, \theta_k^{(z)})}{\Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}^{(z)})} \cdot \sum_{\mathbf{H}_{\setminus k}} \Pr(e^j \mid \mathbf{H}_{\setminus k}, h_k^+) \prod_{\substack{1 \leq i \leq n \\ i \neq k}} \Pr(H_i \mid c_i^j, \theta_i^{(z)}) \ . \quad (12)$$

Let us define $\hat{\boldsymbol{\theta}}_{(k=1)}^{(z)} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$ where $\hat{\theta}_k^{(z)} = 1$ and $\hat{\theta}_i^{(z)} = \theta_i^{(z)}, \ \forall_{i \neq k}$. Using the defined vector $\hat{\boldsymbol{\theta}}_{(k=1)}^{(z)}$ and $\Pr(h_k^+ \mid c_k^j, \theta_k^{(z)}) = c_k^j \theta_k^{(z)}$, Equation (12) takes the

form

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \frac{c_k^j \theta_k^{(z)} \Pr(e^j \mid \mathbf{c}^j, \hat{\boldsymbol{\theta}}_{(k=1)}^{(z)})}{\Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}^{(z)})} \ . \tag{13}$$

Now we can express the obtained result in terms of the Poisson binomial probabilities. First, let us define

$$\begin{aligned} &\mathbf{p}^{(z,j)} = (p_1^{(z,j)}, \ldots, p_n^{(z,j)}) \quad \text{where} \quad p_i^{(z,j)} = \theta_i^{(z)} c_i^j \ , \\ &\hat{\mathbf{p}}_{(k=1)}^{(z,j)} = (\hat{p}_1^{(z,j)}, \ldots, \hat{p}_n^{(z,j)}) \quad \text{where} \quad \hat{p}_k = 1 \text{ and } \hat{p}_i^{(z,j)} = \theta_i^{(z)} c_i^j, \forall_{i \neq k} \ . \end{aligned}$$

From the following property of the Poisson binomial distribution [17]:

$$\mathrm{B}(i; \mathbf{p}) = \mathrm{B}(i; \mathbf{p}_{\setminus k})(1 - p_k) + \mathrm{B}(i-1; \mathbf{p}_{\setminus k}) p_k \tag{14}$$

it follows that

$$\mathrm{B}(i; \hat{\mathbf{p}}_{(k=1)}^{(z,j)}) = \mathrm{B}(i-1; \mathbf{p}_{\setminus k}^{(z,j)}) \ .$$

Using the last identity and Proposition 1 the left hand side of (13) can be expressed in terms of the Poisson binomial probabilities as follows:

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \begin{cases} \dfrac{p_k^{(z,j)} \sum_{i=0}^{n-1} \mathrm{B}\left(i; \mathbf{p}_{\setminus k}^{(z,j)}\right) \gamma_{i+1}}{\sum_{i=0}^{n} \mathrm{B}\left(i; \mathbf{p}^{(z,j)}\right) \gamma_i} & \text{if } e^j = 1 \ , \\ \dfrac{p_k^{(z,j)} \left(1 - \sum_{i=0}^{n-1} \mathrm{B}\left(i; \mathbf{p}_{\setminus k}^{(z,j)}\right) \gamma_{i+1}\right)}{1 - \sum_{i=0}^{n} \mathrm{B}\left(i; \mathbf{p}^{(z,j)}\right) \gamma_i} & \text{if } e^j = 0 \ . \end{cases} \tag{15}$$

Summarizing, the EM algorithm for symmetric causal independence models is given by:

Expectation step: For every instance $\mathbf{x}^j = (\mathbf{c}^j, e^j)$ with $j = 1, \ldots, N$, we form

$$\mathbf{p}^{(z,j)} = (p_1^{(z,j)}, \ldots, p_n^{(z,j)}) \quad \text{where} \quad p_i^{(z,j)} = \theta_i^{(z)} c_i^j \ .$$

Subsequently, the probability $\mathrm{P}(h_k^+ \mid \mathbf{c}^j, e^j, \boldsymbol{\theta}^{(z)})$ is computed from (15) for all hidden variables H_k with $k = 1, \ldots, n$.

Maximization step: Update the parameter estimates for all $k = 1, \ldots, n$ using Equation (11).

4 Experimental Results

The introduced EM algorithm enables us to evaluate the practical significance of the symmetric causal independence models. As it is difficult to provide an interpretation of the learned parameters, we evaluate the learned symmetric causal independence models based on their classification performance.

4.1 Evaluation Scheme

Since we do not have an efficient algorithm to perform a search in the space of symmetric Boolean functions, we chose to model the interaction among cause and effect variables by means of Boolean threshold functions, which seem to be the most probable interaction functions for the given domains.

Given the model parameters $\boldsymbol{\theta}$, the testing data $\mathbf{D}_{test}$ and the classification threshold $\frac{1}{2}$, the classifications and misclassifications for both classes are computed. Let *tp (true positives)* stand for the number of data samples $(\mathbf{c}^j, e^{j+}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) \geq \frac{1}{2}$ and *fp (false positives)* stand for the number of data samples $(\mathbf{c}^j, e^{j+}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) < \frac{1}{2}$. Likewise, *tn (true negatives)* is the number of data samples $(\mathbf{c}^j, e^{j-}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) < \frac{1}{2}$ and *fp (false positives)* is the number of data samples $(\mathbf{c}^j, e^{j-}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) \geq \frac{1}{2}$. To evaluate the classification performance we use *accuracy*, which is a measure of correctly classified cases,

$$\eta = \frac{tp + tn}{tp + tn + fn + fp} ,$$

and *F-measure*, which combines *precision* $\pi = \frac{tp}{tp+fp}$ and *recall* $\rho = \frac{tp}{tp+fn}$,

$$F = \frac{2\pi\rho}{\pi + \rho} .$$

4.2 Non-Hodgkin Lymphoma Data Set

For our experiments we use a database with data from the patients with gastric non-Hodgkin lymphoma (NHL) collected by the clinical experts from the Netherlands Cancer Institute (NKI). The data set consists of the factors that influence the result of treatment, and hence the learned models can be argued to follow the causal interpretation. We will cover only the basic facts; a thorough description of the disease and collected data can be found in [18].

Gastric non-Hodgkin lymphoma is a type of cancer of the lymphatic system, the disease-fighting network spread throughout the body, which originates in the stomach. Response to treatment is one of the most important prognostic indicators of a long-term disease-free survival, particularly in patients with aggressive NHL [19]. We learn a causal independence model that models the interaction between the early outcome of the treatment and the pretreatment prognostic factors. The early outcome of the treatment, i.e. the effect in the model, stands for endoscopically verified result of the treatment, six to eight weeks after treatment with complete remission defining a situation in which all clinical signs of disease disappear with the treatment. The following pretreatment information, i.e. the causes in the model, is available: (1) age; (2) general health status; (3) bulky disease; (4) histological classification; (5) stage of the cancer; (6) clinical signs (hemmorhage, perforation, obstruction) due to the disease.

Based on the medical literature we converted the data to binary form and chose the state of every variable that corresponds to the presence of the cause/effect.

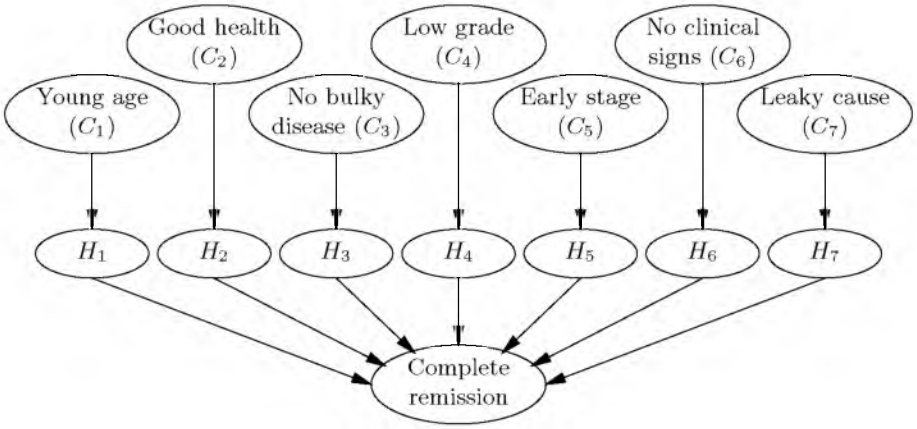

Fig. 2. Causal independence model modelling complete remission following treatment of non-Hodgkin lymphoma. The variable 'Young age' represents a patient younger than 60 years, the variable 'Early stage' stands for the first clinical stage of NHL, and the variable 'No clinical signs' represents a patient who has no hemorrhage, no perforation and no obstruction.

The resulting model is shown in Figure 2 where the name of the variable indicates its positive state. To learn the parameters of the model we used 125 patient cases with no missing data. 95 of the patients had complete remission six to eight weeks after the treatment and for the other 30 patients the disease did not disappear. As the data set is small, a leave-one-out cross-validation scheme was employed both to evaluate the performance of the model and to avoid data overfitting. Classification performance measures for symmetric causal independence models with the interaction function τ_k, $k = 1, \ldots, 7$ are listed in Table 1. The results show that the interaction between the pretreatment variables and the outcome of the treatment is best modelled by the interaction function τ_2. Note that noisy threshold model with the threshold $k = 2$ outperforms the noisy OR model, while the noisy AND model is a poor choice to model the given problem.

Table 1. Classification performance measures for noisy threshold models with the threshold $k = 1, \ldots, 7$ for Non-Hodgkin Lymphoma data set.

Causal independence model	Accuracy (%)	F-measure
noisy OR	75.2	0.854
noisy threshold $k = 2$	83.2	0.896
noisy threshold $k = 3$	82.4	0.891
noisy threshold $k = 4$	78.4	0.857
noisy threshold $k = 5$	71.2	0.795
noisy threshold $k = 6$	56.8	0.625
noisy AND	36.8	0.288

In order to see how well the causal independence models classify compared with other classification algorithms, we evaluated the classification performance of a few widely-used classifiers on NHL data set. The experiments were performed using the Weka system [20]. The results reported in Table 2 show that noisy threshold model provides very similar results to those of naive Bayes, logistic regression and multilayer perceptron and outperforms decision tree and support vector machine classifiers.

Table 2. Classification performance measures for different classifiers for Non-Hodgkin Lymphoma data set. Weka's default parameter settings were used.

Classifier	Accuracy (%)	F-measure
noisy threshold $k = 2$	83.2	0.896
naive Bayes	84.0	0.899
logistic regression	82.4	0.885
multilayer perceptron	82.4	0.885
decision tree (C4.5)	73.6	0.832
support vector machine	77.6	0.861

5 Discussion

In this paper, we developed the EM algorithm to learn the parameters in symmetric causal independence models and studied its computational complexity and convergence. The presented algorithm enabled us to evaluate the utility of symmetric causal independence models. The reported experimental results indicate that it is unnecessary to restrict causal independence models to only two interaction functions, logical OR and logical AND. Additionally, competitive performance of symmetric causal independence models present them as a potentially useful additional tool to the set of classifiers.

The current study has only examined the problem of learning conditional probabilities of hidden variables. The problem of learning an optimal interaction function has not been addressed. Efficient search in symmetric Boolean function space is a possible direction for future research.

Acknowledgments. This research, carried out in the TimeBayes project, was supported by the Netherlands Organization for Scientific Research (NWO) under project number FN4556. The authors are grateful to Henk Boot and Babs Taal for the provided non-Hodgkin's lymphoma data. We would also like to thank Jiří Vomlel for sharing his code and insights.

References

1. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kauffman Publishers (1988)
2. Díez, F.J.: Parameter Adjustment in Bayes Networks. The generalized noisy OR-gate. In: Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence (1993) 99–105
3. Heckerman D., Breese, J.S.: A New Look at Causal Independence. In: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (1994) 286–292
4. Zhang, N.L., Poole, D.: Exploiting Causal Independence in Bayesian Networks Inference. Journal of Artificial Intelligence Research, Vol. 5 (1996) 301–328
5. Kappen, H.J., Neijt, J.P.: Promedas, a Probabilistic Decision Support System for Medical Diagnosis. Technical report, SNN - UMCU (2002)
6. Shwe, M.A., Middleton, B., Heckerman, D.E., Henrion, M., Horvitz, E.J., Lehmann, H.P., Cooper, G.F.: Probabilistic Diagnosis using a Reformulation of the INTERNIST-1/QMR Knowledge Base, I – The Probabilistic Model and Inference Algorithms. Methods of Information in Medicine, Vol. 30 (1991) 241–255
7. Meek, C., Heckerman, D.: Structure and Parameter Learning for Causal Independence and Causal Interaction Models. In: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (1997) 366–375
8. Lucas, P.J.F.: Bayesian Network Modelling Through Qualitative Patterns. Artificial Intelligence, Vol. 163 (2005) 233–263
9. Jurgelenaite, R., Lucas, P.J.F., Heskes, T.: Noisy Threshold Functions for Modelling Causal Independence in Bayesian Networks. Technical report ICIS–R06014, Radboud University Nijmegen (2006)
10. Visscher, S., Lucas, P.J.F., Bonten, M., Schurink, K.: Improving the Therapeutic Performance of a Medical Bayesian Network using Noisy Threshold Models. In: Proceedings of ISBMDA 2005, the 6th International Symposium on Biological and Medical Data Analysis (2005) 161–172
11. Vomlel, J.: Noisy-or Classifier. International Journal of Intelligent Systems, Vol. 21 (2006) 381–398
12. Henrion, M.: Some Practical Issues in Constructing Belief Networks. Uncertainty in Artificial Intelligence, Vol. 3 (1989) 161–173
13. Enderton, H.B.: A Mathematical Introduction to Logic. Academic Press, San Diego (1972)
14. Wegener, I.: The Complexity of Boolean Functions. John Wiley & Sons, New York (1987)
15. Le Cam, L.: An Approximation Theorem for the Poisson Binomial Distribution. Pacific Journal of Mathematics, Vol. 10 (1960) 1181–1197
16. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Vol. 39 (1977) 1–38
17. Darroch, J.: On the Distribution of the Number of Successes in Independent Trials. The Annals of Mathematical Statistics, Vol. 35 (1964) 1317–1321
18. Lucas, P.J.F., Boot, H., Taal, B.: Computer-based Decision Support in Management of Primary Gastric non-Hodgkin Lymphoma. Methods of Information in Medicine, Vol. 37 (1998) 206–219
19. Bast, R.C., Kufe, D.W., Pollock, R.E., Weichselbaum, R.R., Holland, J.F., Frei, E.: Cancer Medicine - 5 Review. B C Decker Inc., Ontario (2000)
20. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)