Mach Learn (2008) 71: 133–153
DOI 10.1007/s10994-007-5041-7

Learning symmetric causal independence models

Rasa Jurgelenaite · Tom Heskes

Received: 25 October 2006 / Revised: 3 September 2007 / Accepted: 14 November 2007 /
Published online: 29 January 2008

Abstract Causal independence modelling is a well-known method for reducing the size of probability tables, simplifying the probabilistic inference and explaining the underlying mechanisms in Bayesian networks. Recently, a generalization of the widely-used noisy OR and noisy AND models, causal independence models based on symmetric Boolean functions, was proposed. In this paper, we study the problem of learning the parameters in these models, further referred to as symmetric causal independence models. We present a computationally efficient EM algorithm to learn parameters in symmetric causal independence models, where the computational scheme of the Poisson binomial distribution is used to compute the conditional probabilities in the E-step. We study computational complexity and convergence of the developed algorithm. The presented EM algorithm allows us to assess the practical usefulness of symmetric causal independence models. In the assessment, the models are applied to a classification task; they perform competitively with state-of-the-art classifiers.

Keywords Bayesian networks · Causal independence · EM algorithm

1 Introduction

Bayesian networks (Pearl 1988) are well-established as a sound formalism for representing and reasoning with probabilistic knowledge. However, because the number of conditional probabilities for each node grows exponentially with the number of its parents, it is usually

Editor: Kevin Murphy

This research, carried out in the TimeBayes project, was supported by the Netherlands Organization for Scientific Research (NWO) under project number FN4556 and a Vici grant (639.023.604) awarded to the second author.

R. Jurgelenaite (✉) · T. Heskes
Institute for Computing and Information Sciences, Radboud University Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands
e-mail: rasa@cs.ru.nl

T. Heskes
e-mail: tomh@cs.ru.nl

unreliable if not infeasible to specify the conditional probabilities for nodes that have a large number of parents. Furthermore, for large, richly-connected Bayesian networks, probabilistic inference is difficult or even intractable. Causal independence modelling (Díez 1993; Heckerman and Breese 1994; Srinivas 1993; Zhang and Poole 1996) can greatly reduce the number of conditional probabilities to be assessed or elicited from experts and simplify the probabilistic inference. Another desirable property of causal independence models is their ability to explain the underlying relationships among cause and effect variables.

Causal independence assumptions are often used in practical Bayesian network models (Kappen and Neijt 2002; Shwe et al. 1991). 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 gives rise to the effect. Several authors proposed to expand the space of functions that define interaction among causes by other symmetric Boolean functions: the idea was mentioned by Meek and Heckerman (1997); an analysis of the qualitative patterns was presented by Lucas (2005); the problem of probabilistic inference was studied by Jurgelenaite et al. (2006). The generalization preserves efficiency and understandability of the noisy-OR and noisy-AND models, while at the same time allowing more flexibility in modelling the interaction among causes.

Even though in some real-world problems the intermediate variables in causal independence models are observable (see Visscher et al. 2005), in many problems these variables are latent. In such problems, the conditional probability distribution of the effect given the causes depends on unknown parameters, which have to 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. A direct application of the EM algorithm to learn parameters in symmetric causal independence models is not tractable for models with many causes as the algorithm is exponential in the number of causes. At least two variants of the EM algorithm that make use of specific properties of causal independence models can be found in literature. Meek and Heckerman (1997) provided a general scheme to use the EM algorithm to compute the maximum likelihood estimate of the parameters in causal interaction models, which are a generalization of noisy-max and noisy-additive models that allows a more flexible model structure. Vomlel (2006) 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 too abstract or too specific to be directly applied 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, and study computational complexity and convergence of the algorithm. The EM algorithm to learn the parameters in symmetric causal independence models enables us to assess the practical usefulness of these extended models. In the assessment, we use symmetric causal independence models as classifiers. Experimental results show the competitive performance of symmetric causal independence models compared to the noisy OR model 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 Sect. 3, we first describe the general scheme of the EM algorithm and then develop the EM algorithm for parameter learning in symmetric causal independence models. The maxima

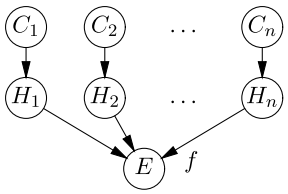

Fig. 1 Causal independence model

of the log-likelihood function for the symmetric causal independence models are examined in Sect. 4. Finally, Sect. 5 presents the experimental results, and conclusions are drawn in Sect. 6. The Appendix contains a number of properties that enable a reduction of computational complexity of the E-step of the EM algorithm.

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 the arcs $\mathbf{A}(G)$ represent the conditional 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 a variable). 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 use v to denote the realization of the random variable V, 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

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 Fig. 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 the way in which the

hidden effects H_i and, indirectly, also the causes C_i interact to yield the final effect E. The function f is defined in such a way that when a relationship, as modelled by the function f, between $H_1, \ldots, H_n$ 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 (Heckerman and Breese 1994; Pearl 1988). In terms of probability theory, this implies that it holds that $\Pr(h_i^+ \mid c_i^-) = 0$; as a consequence, it holds that $\Pr(h_i^- \mid c_i^-) = 1$. We make the same assumption in this paper.

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 (Henrion 1989). 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 effect E given the causes $C_1, \ldots, C_n$ is obtained from the causal parameters $\Pr(H_i \mid C_i)$ as follows (Zhang and Poole 1996):

$$\Pr(e \mid c_1, \ldots, c_n) = \sum_{f(h_1,\ldots,h_n)=e} \prod_{i=1}^{n} \Pr(h_i \mid c_i). \tag{1}$$

In this paper, we assume that the function f in (1) is a Boolean function. However, there are 2^{2^n} different n-ary Boolean functions (Enderton 1972; Wegener 1987); 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* (Wegener 1987) and the number reduces to 2^{n+1}.

Symmetric Boolean functions have an attractive property: they can be decomposed in terms of the exact Boolean functions, where the *exact* function ϵ_l checks whether there are exactly l trues among arguments, i.e. $\epsilon_l(h_1, \ldots, h_n) = \top$ if $\sum_{i=1}^{n} h_i = l$. A symmetric Boolean function can be decomposed in terms of the exact functions ϵ_i:

$$f(h_1, \ldots, h_n) = \bigvee_{i=0}^{n} \epsilon_i(h_1, \ldots, h_n) \wedge \gamma_i, \tag{2}$$

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

Another useful symmetric Boolean function is the *threshold* function τ_k, which checks whether there are at least k trues among the arguments, i.e. $\tau_k(h_1, \ldots, h_n) = \top$ if $\sum_{i=1}^{n} h_i \geq k$. To express it in the Boolean constants, we have: $\gamma_0 = \cdots = \gamma_{k-1} = \bot$ and $\gamma_k = \cdots = \gamma_n = \top$. The commonly used OR and AND functions are the extremes of a spectrum of the threshold functions: the OR function is a threshold function τ_k with $k = 1$ and the AND function is a threshold function τ_k with $k = n$.

2.3 The Poisson binomial distribution

Using the property (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 are true as follows:

$$\Pr(e^+ \mid c_1, \ldots, c_n) = \sum_{\substack{0 \le l \le n \\ \gamma_l}} \sum_{\epsilon_l(h_1,\ldots,h_n)=\top} \prod_{i=1}^{n} \Pr(h_i \mid c_i). \tag{3}$$

As we will show next, the conditional probability of the occurrence of the effect E given the causes $C_1, \ldots, C_n$ in symmetric causal independence models is closely related to the Poisson binomial distribution from statistics.

Let l denote the number of successes in n independent trials, where p_i is the 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* (Edwards 1960; Le Cam 1960):

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

To put it in words, the Poisson binomial distribution is a discrete probability distribution of the number of successes in a sequence of n independent experiments, each of which yields success with a different probability. When all trials have the same probability of success p, $\mathrm{B}(l; \mathbf{p})$ reduces to the binomial distribution: $\mathrm{B}(l; p) = \binom{n}{l} p^l (1-p)^{n-l}$.

Let us define a vector of probabilistic parameters $\mathbf{p}(c_1, \ldots, c_n) = (p_1, \ldots, p_n)$ with $p_i = \Pr(h_i^+ \mid c_i)$. The relationship between the Poisson binomial probabilities and the conditional probability distribution given the Boolean exact function is as stated in the following proposition.

Proposition 1 *It holds that*:

$$\sum_{\epsilon_l(h_1,\ldots,h_n)=\top} \prod_{i=1}^{n} \Pr(h_i \mid c_i) = \mathrm{B}(l; \mathbf{p}(c_1, \ldots, c_n)). \tag{5}$$

Proof Note that in (3), the sum

$$\sum_{\epsilon_l(h_1,\ldots,h_n)=\top} \prod_{i=1}^{n} \Pr(h_i \mid c_i)$$

was defined as the probability that exactly l hidden variables are true. A hidden variable H_i can be seen as an independent trial whose probability of success is $\Pr(h_i^+ \mid c_i)$. Combining the definition of the vector of probabilistic parameters $\mathbf{p}(c_1, \ldots, c_n)$ and the definition of the Poisson binomial distribution, the result in the premise of this proposition is obtained. □

Now, we can establish the relationship between the conditional probability of the effect given the causes in a symmetric causal independence model and the Poisson binomial distribution.

Proposition 2 *It holds that*:

$$\Pr(e^+ \mid c_1, \ldots, c_n) = \sum_{i=0}^{n} \mathrm{B}(i; \mathbf{p}(c_1, \ldots, c_n)) \gamma_i.$$

Proof The proof follows from (3) and Proposition 1. □

The relationship described in Proposition 2 allows us to use the theory of the well-studied Poisson binomial distribution in the context of symmetric causal independence models. The properties of the Poisson binomial distribution will be of major importance in developing a computationally efficient EM algorithm for symmetric causal independence models.

3 EM algorithm

For a symmetric causal independence model to be complete, its structure, interaction function and parameters need to be determined. The structure of symmetric causal independence models is fixed, and the interaction function f is assumed to be known. To have a fully specified symmetric causal independence model, we need to estimate the unknown parameters in the model, i.e. the parameters of the conditional distributions of hidden variables $H_1, \ldots, H_n$. Given the assumption that absent causes do not contribute to the effect, the unknown parameters in the model are $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ where $\theta_i = \Pr(h_i^+ \mid c_i^+)$.

3.1 Maximum likelihood estimate and 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 focus on the problem of estimating the parameters when no additional information about the model is available, i.e. we do not have any prior knowledge about the parameters $\boldsymbol{\theta}$. To learn $\boldsymbol{\theta}$, we maximize the conditional log-likelihood

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

The value of $\boldsymbol{\theta}$ which maximizes the conditional log-likelihood is known as the *maximum likelihood estimate* for $\boldsymbol{\theta}$. Expectation-maximization (EM) algorithm (Dempster et al. 1977) 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.

EM algorithm can be explained and derived in several ways. We derive the EM algorithm on the basis of the explanation of EM in terms of lower-bound maximization (Neal and Hinton 1998). We start from the following simple identity:

$$\ln \Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) = \ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) - \ln \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}) \tag{6}$$

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 (6) does not depend on $\mathbf{H}$, so averaging over $\mathbf{H}$ yields

$$\begin{aligned} \ln \Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) = &\sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(old)}) \ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) \\ &- \sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(old)}) \ln \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}). \end{aligned} \tag{7}$$

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 (7) is guaranteed to increase the expected complete (conditional) log-likelihood.

Let us denote

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)}) = \sum_{j=1}^{N} \sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}). \tag{8}$$

The EM algorithm at each iteration maximizes the functional

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

3.2 Maximization step

To maximize $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)})$, we need to compute the partial derivatives of this functional with respect to each parameter, set them equal to zero and solve the system of equations. We start by transforming $\ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta})$ so that it becomes a sum of logarithms:

$$\begin{aligned} \ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) &= \ln \left(\Pr(e^j \mid \mathbf{h}) \prod_{i=1}^{n} \Pr(h_i \mid c_i^j, \theta_i) \right) \\ &= \ln \Pr(e^j \mid \mathbf{h}) + \sum_{i=1}^{n} \ln \Pr(h_i \mid c_i^j, \theta_i). \end{aligned} \tag{9}$$

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

$$\Pr(h_i \mid c_i^j, \theta_i) = c_i^j h_i \theta_i + c_i^j (1 - h_i)(1 - \theta_i) + (1 - c_i^j)(1 - h_i). \tag{10}$$

Combining (8–10), we obtain

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)}) = \sum_{j=1}^{N} \sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \left(\ln \Pr(e^j \mid \mathbf{h}) + \sum_{i=1}^{n} \ln \left(\theta_i c_i^j (2h_i - 1) + 1 - h_i \right) \right).$$

Then taking the partial derivatives with respect to each parameter θ_k and setting them equal to zero yields

$$\frac{\partial Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)})}{\partial \theta_k} = \sum_{j=1}^{N} \sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \frac{c_k^j (2h_k - 1)}{\theta_k c_k^j (2h_k - 1) + 1 - h_k} = 0. \tag{11}$$

Now let us define $\mathbf{h}_{\setminus k} = \{h_1, \ldots, h_{k-1}, h_{k+1}, \ldots, h_n\}$. Equation (11) can be simplified writing it as a sum over the states of the hidden variable H_k:

$$\sum_{1 \le j \le N} c_k^j \sum_{\mathbf{h}_{\setminus k}} \left(\Pr(\mathbf{h}_{\setminus k}, h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \frac{1}{\theta_k c_k^j} - \Pr(\mathbf{h}_{\setminus k}, h_k^- \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \frac{1}{1 - \theta_k c_k^j} \right) = 0. \tag{12}$$

It can be shown that (12) is solved by

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

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 conditional probability of $\mathbf{H}$ given the 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:

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \frac{\Pr(h_k^+ \mid c_k^j, \theta_k^{(z)})}{\Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}^{(z)})} \sum_{\mathbf{h}_{\setminus k}} \Pr(e^j \mid \mathbf{h}_{\setminus k}, h_k^+) \prod_{\substack{1\le i\le n \\ i\ne k}} \Pr(h_i \mid c_i^j, \theta_i^{(z)}). \tag{14}$$

Computing the probability $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ from (14) in a straightforward way requires summing over 2^n terms, which is computationally expensive. In Chen et al. (1994), Howard (1972) it was shown that using recursive methods the Poisson binomial distribution can be computed in a quadratic number of operations with respect to n. Therefore, expressing (14) in terms of the Poisson binomial probabilities is an obvious way to reduce the computational cost of the algorithm.

Let us define $\hat{\boldsymbol{\theta}}^{(z)}_{(k=1)} = (\hat{\theta}_1^{(z)}, \ldots, \hat{\theta}_n^{(z)})$ where

$$\hat{\theta}_k^{(z)} = 1 \quad \text{and} \quad \hat{\theta}_i^{(z)} = \theta_i^{(z)}, \quad \forall_{i\ne k}.$$

Using the defined vector $\hat{\boldsymbol{\theta}}^{(z)}_{(k=1)}$ and $\Pr(h_k^+ \mid c_k^j, \theta_k^{(z)}) = c_k^j \theta_k^{(z)}$, (14) takes the form

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \frac{c_k^j \theta_k^{(z)} \Pr(e^j \mid \mathbf{c}^j, \hat{\boldsymbol{\theta}}^{(z)}_{(k=1)})}{\Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}. \tag{15}$$

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

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

From the following property of the Poisson binomial distribution (Darroch 1964):

$$\mathrm{B}(i;\mathbf{p}) = \mathrm{B}(i;\mathbf{p}_{\setminus k})(1-p_k) + \mathrm{B}(i-1;\mathbf{p}_{\setminus k})p_k \tag{16}$$

it follows that

$$\mathrm{B}(i;\hat{\mathbf{p}}^{(z,j)}_{(k=1)}) = \mathrm{B}(i-1;\mathbf{p}^{(z,j)}_{\setminus k}). \tag{17}$$

Using the identity (17) and Proposition 2, the left hand side of (15) can be expressed in terms of the Poisson binomial probabilities as follows:

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \begin{cases} \dfrac{p_k^{(z,j)} \sum_{i=0}^{n-1} \mathrm{B}(i;\mathbf{p}^{(z,j)}_{\setminus k})\gamma_{i+1}}{\sum_{i=0}^{n} \mathrm{B}(i;\mathbf{p}^{(z,j)})\gamma_i} & \text{if } e^j = 1, \\ \dfrac{p_k^{(z,j)}(1 - \sum_{i=0}^{n-1} \mathrm{B}(i;\mathbf{p}^{(z,j)}_{\setminus k})\gamma_{i+1})}{1 - \sum_{i=0}^{n} \mathrm{B}(i;\mathbf{p}^{(z,j)})\gamma_i} & \text{if } e^j = 0. \end{cases} \tag{18}$$

Summarizing, the $(z+1)$th iteration of 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

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

Subsequently, the probability $\mathrm{P}(h_k^+ \mid \mathbf{c}^j, e^j, \boldsymbol{\theta}^{(z)})$ is computed from (18) for each hidden variable $H_k, k = 1, \ldots, n$.

Maximization step: Update the parameter estimates using (13).

The expectation and maximization steps are repeated until a convergence criterion is satisfied.

3.4 Computational complexity of the expectation step

The computation of the conditional probabilities in the E-step is a probabilistic inference task; therefore, it can be performed by an inference algorithm. One way to perform an efficient inference in causal independence models is first to transform the models using parent divorcing (Olesen et al. 1989) or temporal transformation (Heckerman 1993) techniques and then to apply an exact inference algorithm. Zagorecki et al. (2006) used the latter technique to transform the probabilistic causal independence models, the models where the combination function does not need to be deterministic, to make inference efficient and applied the standard EM algorithm to learn the parameters. Another way to perform efficient inference in causal independence models is the VE_1 algorithm (Zhang and Poole 1996), which factorizes the conditional probabilities into a combination of smaller factors to obtain a finer-grain factorization of the joint probability and makes use of this factorization. Using one of the transformation techniques or the VE_1 algorithm, the conditional probability $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ can be computed in quadratic number of operations with respect to n; however, there was no investigation how these methods can be used for parameter learning in causal independence models.

The Poisson binomial distribution and, consequently, the conditional probability $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ can be computed in quadratic time with respect to n using recursive methods, and estimated in linear time with respect to n using approximation and bounding techniques (see Jurgelenaite et al. 2006, for a review of these techniques). A naive computation of the exact probabilities $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ in the expectation step requires $O(n^3)$ operations for every data instance $\mathbf{x}^j = (e^j, \mathbf{c}^j)$ at every iteration of the algorithm. For a problem with a large number of causes n, this cubic complexity can become a computational bottleneck. Using the theory of the Poisson binomial distribution, however, the computational complexity can be reduced to $O(n_*^2)$ operations, where n_* is the number of probabilistic parameters $p_i^{(z,j)}, i = 1, \ldots, n$ such that $0 < p_i^{(z,j)} < 1$. See the Appendix for details.

4 Analysis of the maxima of the log-likelihood function

Generally, there is no guarantee that the EM algorithm will converge to a global maximum of the log-likelihood. In this section, we investigate the maxima of the conditional log-likelihood function for symmetric causal independence models.

4.1 Noisy OR and noisy AND models

In this section, we show that the conditional log-likelihood for the noisy OR and the noisy AND models has only global maxima. Since the conditional log-likelihood function for these models is not necessarily concave, we use a monotonic transformation to prove the absence of the stationary points other than global maxima.

First, we establish a connection between the maxima of the log-likelihood function and the maxima of the corresponding composite function.

Proposition 3 (Global optimality condition for concave functions (Boyd and Vandenberghe 2004)) *Suppose* $h(\mathbf{q}) : \mathbf{Q} \to \Re$ *is concave and differentiable on* $\mathbf{Q}$. *Then,* $\mathbf{q}^* \in \mathbf{Q}$ *is a global maximum if and only if*

$$\nabla h(\mathbf{q}^*) = \left(\frac{\partial h(\mathbf{q}^*)}{\partial q_1}, \ldots, \frac{\partial h(\mathbf{q}^*)}{\partial q_n} \right)^T = \mathbf{0}.$$

Further, we consider the function

$$CLL(\boldsymbol{\theta}) = h(\mathbf{q}(\boldsymbol{\theta})).$$

Let $CLL(\boldsymbol{\theta})$ and $h(\mathbf{q}(\boldsymbol{\theta}))$ be twice differentiable functions, and let $\mathbf{q}(\boldsymbol{\theta})$ be a differentiable, injective function where $\boldsymbol{\theta}(\mathbf{q})$ is its inverse. Then, the following relationship between the stationary points of the functions CLL and h holds.

Lemma 1 *Suppose,* $\boldsymbol{\theta}^*$ *is a stationary point of* $CLL(\boldsymbol{\theta})$. *Then, there is a corresponding point* $\mathbf{q}(\boldsymbol{\theta}^*)$ *which is a stationary point of* $h(\mathbf{q}(\boldsymbol{\theta}))$.

Proof Since the function $\mathbf{q}(\boldsymbol{\theta})$ is differentiable and injective, its Jacobian matrix $\frac{\partial(q_1,\ldots,q_n)}{\partial(\theta_1,\ldots,\theta_n)}$ is positive definite. Therefore, from the chain rule it follows that if $\nabla CLL(\boldsymbol{\theta}^*) = \mathbf{0}$, then $\nabla h(\mathbf{q}(\boldsymbol{\theta}^*)) = \mathbf{0}$. □

Proposition 4 *If $h(\mathbf{q}(\boldsymbol{\theta}))$ is concave and $\boldsymbol{\theta}^*$ is a stationary point of $CLL(\boldsymbol{\theta})$, then $\boldsymbol{\theta}^*$ is a global maximum.*

Proof If $\boldsymbol{\theta}^*$ is a stationary point, then from Lemma 1 it follows that $\mathbf{q}(\boldsymbol{\theta}^*)$ is also stationary. From the global optimality condition for concave functions, the stationary point $\mathbf{q}(\boldsymbol{\theta}^*)$ is a maximum of $h(\mathbf{q}(\boldsymbol{\theta}))$; thus, from the definition of global maximum, we get that for all $\boldsymbol{\theta}$

$$CLL(\boldsymbol{\theta}) = h(\mathbf{q}(\boldsymbol{\theta})) \leq h(\mathbf{q}(\boldsymbol{\theta}^*)) = CLL(\boldsymbol{\theta}^*). \qquad \square$$

Given Proposition 4, the absence of local optima can be proven by introducing such a monotonic transformation $\mathbf{q}(\boldsymbol{\theta})$ that the composite function $h(\mathbf{q}(\boldsymbol{\theta}))$ would be concave. It is well known that the log-likelihood function for logistic regression is concave, i.e. has no local optima. We will use transformations that allow us to write the log-likelihood for the noisy OR and the noisy AND models in a similar form as that of the logistic regression model.

The conditional probability of the effect in the noisy OR model can be written as:

$$\Pr(e^+ \mid \mathbf{c}, \boldsymbol{\theta}) = 1 - \prod_{i=1}^{n} \Pr(h_i^- \mid c_i) = 1 - \prod_{i=1}^{n} (1-\theta_i)^{c_i} = 1 - \exp\left(\sum_{i=1}^{n} \ln(1-\theta_i) c_i\right).$$

Let us choose a monotonic transformation $q_i = -\ln(1-\theta_i)$, $i = 1, \ldots, n$. Then, the conditional probability of the effect in the noisy OR model equals

$$\Pr(e^+ \mid \mathbf{c}, \mathbf{q}) = 1 - e^{-\mathbf{q}^T \mathbf{c}}.$$

Let us define $z^j = \mathbf{q}^T \mathbf{c}^j$ and $f(z^j) = \Pr(e^+ \mid \mathbf{c}^j, \mathbf{q})$, then the function h reads

$$h(\mathbf{q}) = \sum_{j=1}^{N} e^j \ln f(z^j) + (1 - e^j) \ln(1 - f(z^j)). \tag{19}$$

Since $f'(z^j) = 1 - f(z^j)$, the first derivative of h is

$$\frac{\partial h(\mathbf{q})}{\partial \mathbf{q}} = \sum_{j=1}^{N} \frac{f'(z^j)(e^j - f(z^j))}{f(z^j)(1 - f(z^j))} \mathbf{c}^j = \sum_{j=1}^{N} \frac{e^j - f(z^j)}{f(z^j)} \mathbf{c}^j.$$

To prove that the function h is concave, we need to prove that its Hessian matrix is negative semidefinite. The Hessian matrix of h reads

$$\frac{\partial^2 h(\mathbf{q})}{\partial \mathbf{q} \partial \mathbf{q}^T} = -\sum_{j=1}^{N} \frac{1 - f(z^j)}{f(z^j)^2} e^j \mathbf{c}^j \mathbf{c}^{jT} \leq 0.$$

As the Hessian matrix of h is negative semidefinite, the function h is concave. Therefore, from Proposition 4 it follows that every stationary point of the log-likelihood function for the noisy OR model is a global maximum.

The conditional probability of the effect in the noisy AND model can be written as:

$$\Pr(e^+ \mid \mathbf{c}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \Pr(h_i^+ \mid c_i) = \prod_{i=1}^{n} \theta_i^{c_i} = \exp\left(\sum_{i=1}^{n} \ln \theta_i c_i\right).$$

Let us choose a monotonic transformation $q_i = \ln \theta_i, i = 1, \ldots, n$. Then the conditional probability of the effect in the noisy AND model equals

$$\Pr(e^+ \mid \mathbf{c}, \mathbf{q}) = e^{\mathbf{q}^T \mathbf{c}}.$$

Let us define $z^j = \mathbf{q}^T \mathbf{c}^j$ and $f(z^j) = \Pr(e^+ \mid \mathbf{c}^j, \mathbf{q})$. The function h is the same as for the noisy OR model in (19). Combined with $f'(z^j) = f(z^j)$, it yields the first derivative of h

$$\frac{\partial h(\mathbf{q})}{\partial \mathbf{q}} = \sum_{j=1}^{N} \frac{f'(z^j)(e^j - f(z^j))}{f(z^j)(1 - f(z^j))} \mathbf{c}^j = \sum_{j=1}^{N} \frac{e^j - f(z^j)}{1 - f(z^j)} \mathbf{c}^j$$

and Hessian matrix

$$\frac{\partial^2 h(\mathbf{q})}{\partial \mathbf{q} \partial \mathbf{q}^T} = -\sum_{j=1}^{N} \frac{f(z^j)}{(1 - f(z^j))^2} (1 - e^j) \mathbf{c}^j \mathbf{c}^{jT} \leq 0.$$

Hence, the function h is concave, and the log-likelihood for the noisy AND model has no other stationary points than global maxima.

4.2 General case

The EM algorithm is guaranteed to converge to the local maxima or saddle points. Thus, we can only be sure that the global maximum, i.e. a point $\boldsymbol{\theta}^*$ such that $CLL(\boldsymbol{\theta}^*) \geq CLL(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$, will be found if the log-likelihood has neither saddle points nor local maxima. However, the log-likelihood function for a causal independence model with any symmetric Boolean function does not always fulfill this requirement as is shown in the following counterexample.

Example 1 Let us assume a data set $\mathbf{D} = \{(1, 1, 1, 1), (1, 0, 1, 0)\}$ and the interaction function ϵ_1, i.e. $\gamma_1 = 1$ and $\gamma_0 = \gamma_2 = \gamma_3 = 0$. To learn the unknown parameters in the model describing this interaction, we have to maximize the conditional log-likelihood function

$$\begin{aligned} CLL(\boldsymbol{\theta}) = {} & \ln[\theta_1(1 - \theta_2)(1 - \theta_3) + (1 - \theta_1)\theta_2(1 - \theta_3) + (1 - \theta_1)(1 - \theta_2)\theta_3] \\ & + \ln[1 - \theta_1(1 - \theta_3) - (1 - \theta_1)\theta_3]. \end{aligned}$$

Depending on the choice for initial parameter settings $\boldsymbol{\theta}^{(0)}$, the EM algorithm converges to one of the maxima:

$$CLL(\boldsymbol{\theta})_{\max} = \begin{cases} 0 & \text{at } \boldsymbol{\theta} = (0, 1, 0), \\ -1.386 & \text{at } \boldsymbol{\theta} \in \{(\theta_1, 0, \frac{1}{2}), (\frac{1}{2}, 0, \theta_3)\}. \end{cases}$$

Obviously, only the point $\boldsymbol{\theta} = (0, 1, 0)$ is a global maximum of the log-likelihood function while the other obtained points are local maxima.

The existence of local maxima can also be shown graphically. Let us take the two points $\boldsymbol{\theta}' = (0, 1, 0)$ and $\boldsymbol{\theta}'' = (\frac{4}{5}, 0, \frac{1}{2})$, then the resulting log-likelihood given the convex combinations of $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$:

$$CLL(\lambda \boldsymbol{\theta}' + (1 - \lambda)\boldsymbol{\theta}''), \quad \lambda \in [0, 1]$$

is convex as shown in Fig. 2.

The discussed counterexample proves that in the general case the EM algorithm for symmetric causal independence models does not necessarily converge to a global maximum.

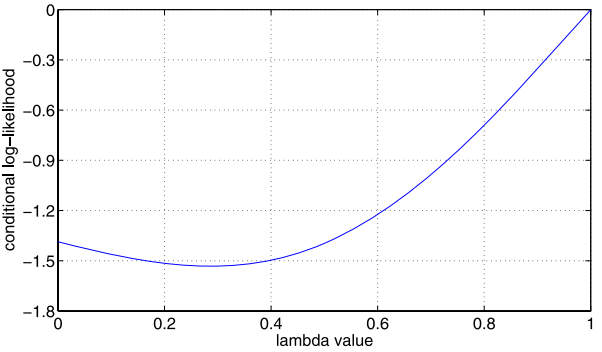

Fig. 2 The log-likelihood $CLL(\lambda\boldsymbol{\theta}' + (1-\lambda)\boldsymbol{\theta}'')$ from Example 1 as a function of λ

5 Experimental results

The EM algorithm developed in this paper allows us to assess the practical significance of symmetric causal independence models. To do so, we evaluated symmetric causal independence models on the basis of their classification performance.

For our experiments, we chose to use two data sets that are different in their causal interpretation and size. The non-Hodgkin lymphoma data set consists of the factors that influence the result of the treatment, and, for this reason, the models learned from this data set can be argued to follow the causal interpretation. The second data set consisting of Reuters news stories does not follow the causal interpretation. This data set is important because of its size; experiments on the Reuters data collection allowed us to test the EM algorithm on large symmetric causal independence models where the number of cause variables for some document classes is in the hundreds.

5.1 Evaluation scheme

We modelled 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. However, the models of document classes in Reuters data set had tens or even hundreds of causes, making learning the models with all threshold functions computationally expensive. Therefore, for this data collection, we only learned the models with the threshold functions τ_2, τ_3, τ_4, the closest threshold functions to the OR function, which was shown to perform well on this data collection (Vomlel 2006). To give a feeling to what extent classification performance of symmetric causal independence models is influenced by the choice of an interaction function, we report all results obtained.

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 *fn (false negatives)* 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 used *accuracy*, which is a measure of correctly classified cases,

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

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

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

and *area under the ROC curve*, which is estimated by a generalization of the Mann-Whitney U statistic (Bamber 1975):

$$AUC = \frac{\sum_{(\mathbf{c}^i, e^{i+}) \in \mathbf{D}_{test}} \sum_{(\mathbf{c}^j, e^{j-}) \in \mathbf{D}_{test}} (\mathbf{I}_{\{\Pr(e^+|\mathbf{c}^i) > \Pr(e^+|\mathbf{c}^j)\}} + \frac{1}{2}\mathbf{I}_{\{\Pr(e^+|\mathbf{c}^i) = \Pr(e^+|\mathbf{c}^j)\}})}{(tp + fn)(tn + fp)},$$

where $\mathbf{I}_{\{\cdot\}}$ denotes the indicator variable whose value is one if its argument is true and zero otherwise.

To measure the significance of any difference between the classification performance of symmetric causal independence models and other classifiers, we used the exact version of McNemar's test (Salzberg 1997). Let n be the number of cases for which the two classifiers produce different output, and let s be the number of cases where the output of the classifier with higher accuracy was correct, while the output of the other classifier was wrong. Under the null hypothesis that the two classifiers perform equally well, we computed the two-sided p-value

$$p = 2\sum_{i=s}^{n} \frac{n!}{i!(n-i)!}(0.5)^n.$$

5.2 Non-Hodgkin lymphoma data set

This data set contains data from the patients with gastric non-Hodgkin lymphoma (NHL) collected by clinical experts from the Netherlands Cancer Institute (see Lucas et al. 1998, for a thorough description of the disease and collected data).

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 (Bast et al. 2000). We learned the symmetric 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, denotes the endoscopically verified result of the treatment, six to eight weeks after treatment; the positive state of this variable, complete remission, defines a situation in which all clinical signs of disease disappear with the treatment. The following pretreatment prognostic factors are available: (1) age; (2) general health status; (3) bulky disease; (4) histological classification; (5) stage of the cancer; (6) clinical signs (hemorrhage, perforation, obstruction due to the disease). The prognostic factors correspond to the cause variables in the model.

Based on medical literature, we converted the data to binary form and defined each variable such that a false state corresponds to a risk factor that accounts for impaired complete remission rate. The resulting model is shown in Fig. 3; the name of the variable indicates a true state of the cause or effect. 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, the other 30 patients failed to achieve complete remission. As the data set is small, we could use a leave-one-out cross-validation scheme both to test the performance of the model and to avoid data overfitting. Classification performance measures for

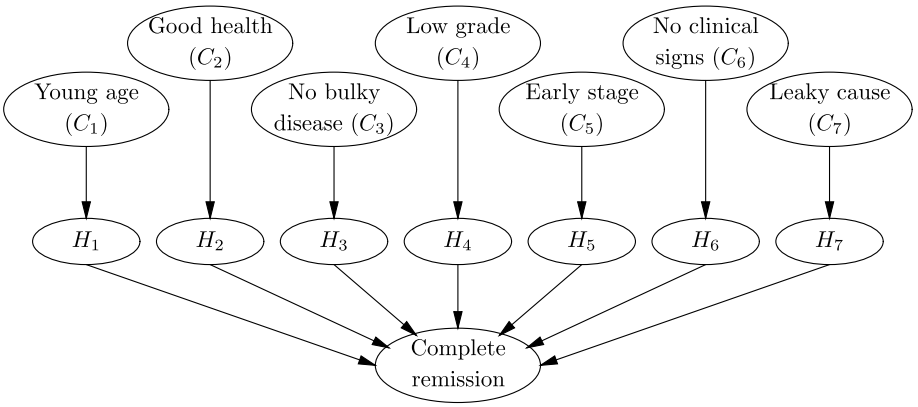

Fig. 3 Causal independence model modelling complete remission following the treatment of non-Hodgkin lymphoma. The variable 'Young age' represents a patient younger than 60 years, the variable 'Early stage' stands for the first clinical stage of NHL, and the variable 'No clinical presentation' represents a patient who has no hemorrhage, no perforation and no obstruction

Table 1 Classification performance measures for symmetric causal independence models (SCIMs) with the interaction function τ_k, $k = 1, \ldots, 7$ for the non-Hodgkin lymphoma data set

Classifier	Accuracy (%)	F-measure	*AUC*
SCIM, τ_1 (noisy OR)	75.2	0.854	0.785
SCIM, τ_2	83.2	0.896	0.832
SCIM, τ_3	82.4	0.891	0.834
SCIM, τ_4	78.4	0.857	0.775
SCIM, τ_5	71.2	0.795	0.733
SCIM, τ_6	56.8	0.625	0.661
SCIM, τ_7 (noisy AND)	36.8	0.288	0.584

symmetric causal independence models with the interaction function τ_k, $k = 1, \ldots, 7$ are reported in Table 1. The results indicate that the interaction between the pretreatment variables and the outcome of the treatment is best modelled by the interaction function τ_2. Note that the symmetric causal independence model with the function τ_2 outperforms the noisy OR model, while the noisy AND model is a poor choice to model the given problem.

To compare symmetric causal independence models with other classification algorithms, we evaluated the classification performance of a few widely-used classifiers. The experiments were performed using the Weka system with its default settings (Witten and Frank 2005). The results reported in Table 2 show that the classification performance of the symmetric causal independence model with the function τ_2 is very similar to that of naive Bayes, logistic regression and multilayer perceptron. However, this symmetric causal independence model outperformed the noisy OR model, decision tree and support vector machine with two-sided p-values of 0.0063, 0.0042 and 0.0923, respectively.

Table 2 Classification performance measures for different classifiers for the non-Hodgkin lymphoma data set

Classifier	Accuracy (%)	F-measure	*AUC*
SCIM, τ_2	83.2	0.896	0.832
Naive Bayes	84.0	0.899	0.821
Logistic regression	82.4	0.885	0.823
Multilayer perceptron	82.4	0.885	0.780
Decision tree (C4.5)	73.6	0.832	0.708
Support vector machine	77.6	0.861	0.625

5.3 Reuters data set

For the second part of our experiments, we use the Reuters-21578 text categorization collection containing the Reuters news stories preprocessed by Karčiauskas (2002). The comparison of the noisy OR model and a few widely-used classifiers on this data collection was reported in Vomlel (2006), and the results showed the competitive performance of the noisy OR model. Therefore, we aim to show that extended symmetric causal independence models perform competitively with the noisy OR model.

The data set has been split into training (7769 documents) and test (3018 documents) sets. For every one of the ten document classes, the most informative features were selected using the expected information gain as a feature selection criteria. Each document class was classified separately against all other classes. We chose to use the same threshold for the expected information gain as in Vomlel (2006). The number of selected features varied from 23 for the *corn* document class to 307 for the *earn* document class. Classification performance measures for symmetric causal independence models with the interaction function τ_k, $k = 1, \ldots, 4$ are given in Tables 3–5. Even though the threshold to select the relevant features was tuned for the noisy OR model, for five document classes a symmetric causal independence model with another interaction function than the OR function provides better results. For the documents classes *earn* and *trade* the difference in performance is significant with double-sided p-values of 0.0016 (SCIM, τ_2) and 0.0154 (SCIM, τ_4), respectively.

6 Discussion

In this paper, we have developed a computationally efficient EM algorithm to learn parameters in symmetric causal independence models, where the computational scheme of the Poisson binomial distribution was used for the computation of the conditional probabilities in the expectation step. We also investigated the maxima of the log-likelihood function for symmetric causal independence models and showed that the log-likelihood for the noisy OR and the noisy AND models has only global maxima. The presented algorithm allowed us to evaluate the utility of the extended 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. Competitive performance of the extended symmetric causal independence models present them as a potentially useful additional tool to the set of classifiers.

Even though we described symmetric causal independence models as models constructed on the basis of different behavioural patterns among causes and effects, this description

Table 3 Accuracy of symmetric causal independence models (SCIMs) with the interaction function τ_k, $k = 1, \ldots, 4$ for the Reuters data set; N_{Class} is the number of documents in the corresponding class. The highest accuracy obtained for each document class is shown in bold

Class	N_{Class}	Noisy OR	SCIM, τ_2	SCIM, τ_3	SCIM, τ_4
Earn	1087	96.3	**97.2**	**97.2**	96.8
Acq	719	93.1	**93.2**	**93.2**	93.0
Crude	189	**98.1**	**98.1**	97.6	97.7
Money-fx	179	95.8	95.8	95.9	**96.0**
Grain	149	**99.2**	99.0	98.2	97.9
Interest	131	96.5	**96.8**	96.7	96.7
Trade	117	96.6	97.0	**97.3**	**97.3**
Ship	89	**98.9**	98.8	98.7	98.6
Wheat	71	**99.5**	99.2	98.8	98.5
Corn	56	**99.7**	99.4	99.1	98.8

Table 4 F-measure of symmetric causal independence models with the interaction function τ_k, $k = 1, \ldots, 4$ for the Reuters data set; N_{Class} is the number of documents in the corresponding class. The highest F-measure obtained for each document class is shown in bold

Class	N_{Class}	Noisy OR	SCIM, τ_2	SCIM, τ_3	SCIM, τ_4
Earn	1087	95.0	**96.1**	**96.1**	95.6
Acq	719	**85.3**	84.3	84.5	83.8
Crude	189	84.5	**85.7**	80.7	81.0
Money-fx	179	60.9	62.1	62.6	**62.7**
Grain	149	**92.7**	89.9	80.7	77.2
Interest	131	40.2	**55.0**	53.3	54.0
Trade	117	51.0	61.2	**63.7**	**63.7**
Ship	89	**79.5**	77.7	74.5	71.5
Wheat	71	**90.3**	81.8	71.4	66.2
Corn	56	**91.8**	83.6	72.5	61.5

should not limit the use of the framework. When causal interpretation cannot be applied, symmetric causal independence models can be used as merely a technique to reduce the number of parameters and to simplify the inference problem in Bayesian networks.

The current study has examined the problem of parameter learning in symmetric causal independence models, but 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.

The EM algorithm presented in this paper learns parameters in models where both cause and effect variables are assumed to be binary. However, causal independence models do not have to be limited to binary variables. Researchers proposed several schemes to generalize the noisy OR model to multivalued variables (Díez 1993; Henrion 1989; Srinivas 1993). Extension of the framework of symmetric causal independence models to handle multivalued variables and adjustment of the proposed EM algorithm to this generalization is another research problem of interest.

Table 5 Area under the ROC curve of symmetric causal independence models with the interaction function τ_k, $k = 1, \ldots, 4$ for the Reuters data set; N_{Class} is the number of documents in the corresponding class. The highest *AUC* obtained for each document class is shown in bold

Class	N_{Class}	Noisy OR	SCIM, τ_2	SCIM, τ_3	SCIM, τ_4
Earn	1087	0.995	**0.996**	0.995	0.992
Acq	719	**0.972**	**0.972**	0.957	0.928
Crude	189	0.994	**0.995**	0.994	0.983
Money-fx	179	0.971	**0.973**	0.957	0.915
Grain	149	**0.999**	0.997	0.985	0.952
Interest	131	0.961	**0.963**	0.949	0.915
Trade	117	0.979	**0.985**	0.982	0.973
Ship	89	0.979	**0.986**	0.938	0.842
Wheat	71	**0.997**	0.995	0.994	0.956
Corn	56	**0.998**	0.996	0.982	0.939

Acknowledgements The authors are grateful to Henk Boot and Babs Taal for providing the non-Hodgkin's lymphoma data, Gytis Karčiauskas for the preprocessed Reuters data and Jiří Vomlel for sharing his code and insights. We would also like to thank the anonymous reviewers for their constructive comments.

Appendix

Even if we use recursive methods to compute the Poisson binomial distribution, the computation of the probabilities $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$, $k = 1, \ldots, n$ for a data sample $\mathbf{x}^j$ requires $O(n^3)$ operations. This appendix explains how the theory of the Poisson binomial distribution can be applied to further reduce the computational complexity of the expectation step of the EM algorithm for symmetric causal independence models.

A.1 Reducing the size of the input

The obvious way to reduce the size of the input is to remove those probabilistic parameters from $\mathbf{p}$ that equal zero or one and adapt the Boolean constants accordingly. Let us define three new vectors:

$$\mathbf{p}_{\setminus(0)} = (p_i; i = 1 \ldots n,\, p_i \neq 0),$$
$$\mathbf{p}_{\setminus(1)} = (p_i; i = 1 \ldots n,\, p_i \neq 1),$$
$$\mathbf{p}_{\setminus(0,1)} = (p_i; i = 1 \ldots n,\, p_i \neq 0,\, p_i \neq 1).$$

Using (16) iteratively, we obtain the relationships $\mathrm{B}(i; \mathbf{p}) = \mathrm{B}(i - n_1; \mathbf{p}_{\setminus(1)})$ and $\mathrm{B}(i; \mathbf{p}) = \mathrm{B}(i; \mathbf{p}_{\setminus(0)})$, where n_1 is the number of elements of $\mathbf{p}$ equal to 1. Combining these two results yields:

$$\mathrm{B}(i; \mathbf{p}) = \mathrm{B}(i - n_1; \mathbf{p}_{\setminus(0,1)}).$$

Since $\mathrm{B}(i-n_1;\mathbf{p}_{\setminus(0,1)})=0$ for all $i<n_1$ and $i>n-n_0$ (with n_0 being the number of zero elements of $\mathbf{p}$), we can write the conditional probability of the effect as:

$$\Pr(e^+\mid \mathbf{c}^j,\boldsymbol{\theta}^{(z)})=\sum_{i=0}^{n}\mathrm{B}(i-n_1;\mathbf{p}_{\setminus(0,1)}^{(z,j)})\gamma_i=\sum_{i=0}^{n-n_0-n_1}\mathrm{B}(i;\mathbf{p}_{\setminus(0,1)}^{(z,j)})\gamma_{i+n_1}. \tag{20}$$

Equation (20) allows the reduction of the size of the input from n to $n-n_0-n_1$. Note that, given the assumption $\Pr(h_i^+\mid c_i^-)=0$, n_0 is equal to or larger than the number of 'inactive' causes in a given data sample $\mathbf{x}^j$.

A.2 Specific cases

From (20) and the Poisson binomial identity (16), we derived a few special cases when we do not need to compute the Poisson binomial distribution in order to compute the probability $\Pr(h_k^+\mid e^j,\mathbf{c}^j,\boldsymbol{\theta}^{(z)})$:

$$\Pr(h_k^+\mid e^j,\mathbf{c}^j,\boldsymbol{\theta}^{(z)})=\begin{cases}p_k^{(z,j)} & \text{if } p_k^{(z,j)}\in\{0,1\},\\ \frac{p_k^{(z,j)}}{1-p_k^{(z,j)}} & \text{if } \max(u\mid\gamma_u=e^j,u=0,\ldots,n-n_0)=0,\\ 1 & \text{if } \min(u\mid\gamma_u=e^j,u=0,\ldots,n-n_0)=n-n_0.\end{cases} \tag{21}$$

A.3 Number of operations

At the expectation step of the $(z+1)$th iteration of the EM algorithm for a given data sample $\mathbf{x}^j$, we need to compute the Poisson binomial distributions

$$\mathrm{B}(0;\mathbf{p}^{(z,j)}),\ldots,\mathrm{B}(n;\mathbf{p}^{(z,j)}),\quad\text{and}$$

$$\mathrm{B}(0;\mathbf{p}_{\setminus k}^{(z,j)}),\ldots,\mathrm{B}(n-1;\mathbf{p}_{\setminus k}^{(z,j)}).$$

Given (21), we need to compute only those distributions $\mathrm{B}(0;\mathbf{p}_{\setminus k}^{(z,j)}),\ldots,\mathrm{B}(n-1;\mathbf{p}_{\setminus k}^{(z,j)})$ for which $p_k^{(z,j)}\notin\{0,1\}$. Using the recursive method to compute the Poisson binomial distribution and (20), computation of each Poisson binomial distribution requires $(n-n_0-n_1)^2$ operations. The data sample $\mathbf{x}^j$ requires at most $(n-n_0-n_1)^3$ operations. Thus, the computational complexity of the EM algorithm was reduced but it remained cubic.

The number of operations can be reduced from cubic to quadratic by the use of (16). To do so, we first need to compute the distribution $\mathrm{B}(0;\mathbf{p}^{(z,j)}),\ldots,\mathrm{B}(n;\mathbf{p}^{(z,j)})$, and then iteratively calculate $\mathrm{B}(0;\mathbf{p}_{\setminus k}^{(z,j)}),\ldots,\mathrm{B}(n-1;\mathbf{p}_{\setminus k}^{(z,j)})$ either from

$$\mathrm{B}(i;\mathbf{p}_{\setminus k}^{(z,j)})=\frac{\mathrm{B}(i;\mathbf{p}^{(z,j)})-\mathrm{B}(i-1;\mathbf{p}_{\setminus k}^{(z,j)})p_k^{(z,j)}}{1-p_k^{(z,j)}} \tag{22}$$

starting with $\mathrm{B}(-1;\mathbf{p}_{\setminus k}^{(z,j)})=0$, or from

$$\mathrm{B}(i-1;\mathbf{p}_{\setminus k}^{(z,j)})=\frac{\mathrm{B}(i;\mathbf{p}^{(z,j)})-\mathrm{B}(i;\mathbf{p}_{\setminus k}^{(z,j)})(1-p_k^{(z,j)})}{p_k^{(z,j)}} \tag{23}$$

starting with $\mathrm{B}(n;\mathbf{p}_{\setminus k}^{(z,j)})=0$.

This scheme works because (16) reduces to $\mathrm{B}(0;\mathbf{p}^{(z,j)}) = \mathrm{B}(0;\mathbf{p}^{(z,j)}_{\setminus k})(1-p_k^{(z,j)})$ and $\mathrm{B}(n;\mathbf{p}^{(z,j)}) = \mathrm{B}(n-1;\mathbf{p}^{(z,j)}_{\setminus k})p_k^{(z,j)}$ when $i=0$ and $i=n$, respectively. Note that we know $p_k^{(z,j)}$ and we have previously computed $\mathrm{B}(0;\mathbf{p}^{(z,j)}),\ldots,\mathrm{B}(n;\mathbf{p}^{(z,j)})$. To reduce the effect of roundoff error, which may occur using the iterative method, we advise to use the bottom-up (22) approach when $p_k^{(z,j)} < \frac{1}{2}$ and the top-down (23) approach when $p_k^{(z,j)} > \frac{1}{2}$.

Using the results presented in this appendix, the probability $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ can be computed in at most $O((n-n_0-n_1)^2)$ operations.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387–415.

Bast, R. C., Kufe, D. W., Pollock, R. E., Weichselbaum, R. R., Holland, J. F., & Frei, E. (2000). *Cancer medicine—5 review*. Ontario: B.C. Decker.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Chen, X. H., Dempster, A. P., & Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, *81*, 457–469.

Darroch, J. (1964). On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, *35*, 1317–1321.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*, 1–38.

Díez, F. J. (1993). Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In *Proceedings of the ninth conference on uncertainty in artificial intelligence* (pp. 99–105).

Edwards, A. W. P. (1960). The meaning of binomial distribution. *Nature*, *186*, 1074.

Enderton, H. B. (1972). *A mathematical introduction to logic*. San Diego: Academic.

Heckerman, D. (1993). Causal independence for knowledge acquisition and inference. In *Proceedings of the ninth conference on uncertainty in artificial intelligence* (pp. 122–127).

Heckerman, D., & Breese, J. S. (1994). A new look at causal independence. In *Proceedings of the tenth conference on uncertainty in artificial intelligence* (pp. 286–292).

Henrion, M. (1989). Some practical issues in constructing belief networks. *Uncertainty in Artificial Intelligence*, *3*, 161–173.

Howard, S. (1972). Discussion on Professor Cox's paper. *Journal of the Royal Statistical Society, Series B*, *34*, 210–211.

Jurgelenaite, R., Lucas, P. J. F., & Heskes, T. (2006). *Noisy threshold functions for modelling causal independence in Bayesian networks* (Technical report ICIS-R06014). Radboud University Nijmegen.

Kappen, H. J., & Neijt, J. P. (2002). *Promedas, a probabilistic decision support system for medical diagnosis* (Technical report). SNN—UMCU.

Karčiauskas, G. (2002). *Text categorization using hierarchical Bayesian networks classifiers*. Master thesis, Aalborg University.

Le Cam, L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, *10*, 1181–1197.

Lucas, P. J. F. (2005). Bayesian network modelling through qualitative patterns. *Artificial Intelligence*, *163*, 233–263.

Lucas, P. J. F., Boot, H., & Taal, B. (1998). Computer-based decision support in management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine*, *37*, 206–219.

Meek, C., & Heckerman, D. (1997). Structure and parameter learning for causal independence and causal interaction models. In *Proceedings of the thirteenth conference on uncertainty in artificial intelligence* (pp. 366–375).

Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in graphical models*. Cambridge: MIT Press.

Olesen, K. G., Kjærulff, U., Jensen, F., Falck, B., Andreassen, S., & Andersen, S. K. (1989). A Munin network for the median nerve—a case study on loops. *Applied Artificial Intelligence*, *3*, 384–403.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Mateo: Morgan Kaufman.

Salzberg, S. L. (1997). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, *1*, 317–327.

Shwe, M. A., Middleton, B., Heckerman, D. E., Henrion, M., Horvitz, E. J., Lehmann, H. P., & Cooper, G. F. (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I—The probabilistic model and inference algorithms. *Methods of Information in Medicine*, *30*, 241–255.

Srinivas, S. (1993). A generalization of the noisy-or model. In *Proceedings of the ninth conference on uncertainty in artificial intelligence* (pp. 208–215).

Wegener, I. (1987). *The complexity of Boolean functions*. New York: Wiley.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

Visscher, S., Lucas, P. J. F., Bonten, M., & Schurink, K. (2005). 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* (pp. 161–172).

Vomlel, J. (2006). Noisy-or classifier. *International Journal of Intelligent Systems*, *21*, 381–398.

Zagorecki, A., Voortman, M., & Druzdzel, M. J. (2006). Decomposing local probability distributions in Bayesian networks for improved inference and parameter learning. In *Proceedings of the nineteenth international Florida artificial intelligence research society conference* (pp. 860–865).

Zhang, N. L., & Poole, D. (1996). Exploiting causal independence in Bayesian networks inference. *Journal of Artificial Intelligence Research*, *5*, 301–328.