Modelling the Interactions between Discrete and Continuous Causal Factors in Bayesian Networks

Peter J.F. Lucas and Arjen Hommersom
Radboud University Nijmegen,
Institute for Computing and Information Sciences, The Netherlands
Email: {peterl,arjenh}@cs.ru.nl

Abstract

The theory of causal independence is frequently used to facilitate the assessment of the probabilistic parameters of discrete probability distributions of complex Bayesian networks. Although it is possible to include continuous parameters in Bayesian networks as well, such parameters could not, so far, be modelled by means of causal independence theory, as a theory of continuous causal independence was not available. In this paper, such a theory is developed and generalised such that it allows merging continuous with discrete parameters based on the characteristics of the problem at hand. This new theory is based on the discovered relationship between the theory of causal independence and convolution in probability theory, discussed for the first time in this paper. It is also illustrated how this new theory can be used in connection with special probability distributions.

1 Introduction

One of the major challenges in building Bayesian networks is to estimate the associated probabilistic parameters. As these parameters of a Bayesian network have the form of conditional probability distributions $P(E \mid C_1, \ldots, C_n)$, it has been beneficial to look upon the interaction between the associated random variables $E, C_1, \ldots, C_n$ as the interactions between *causes* C_k and an *effect* E. This insight has driven much of the early work (Pearl, 1988), and is still one of the main principles used to construct Bayesian networks for actual problems.

Causal principles have also been exploited in situations where the number of causes n becomes large, as the number of parameters needed to assess a family of conditional probability distributions for a variable E grows exponentially with the number of its causes. The theory of causal independence is frequently used in such situations, basically to decompose a probability table in terms of a small number of causal factors (Henrion, 1989; Pearl, 1988; Heckerman and Breese, 1996). However, so far this theory was restricted to the modelling of *discrete* probability distributions, where in particular three types of interaction are in frequent use: the noisy-OR and the noisy-MAX—in both cases, the interaction among variables is being modelled as disjunctive (Díez, 1993; Henrion, 1989; Pearl, 1988)—and the noisy-AND. Interactions among *continuous* cause variables are usually modelled by statistical techniques such as logistic regression and probit regression, typically by using iterative numerical methods that estimate the weight parameters by maximising the likelihood of the data given the model (Bishop, 2006). Clearly, these regression models resist manual construction based on a solid understanding of a problem domain; the fact that Bayesian networks can be constructed using a mixture of background knowledge and data, depending on the availability of knowledge and data of the problem at hand, is seen as one of the key benefits of the technique. Moreover, it is not possible to combine regression models with discrete causal independence models.

In this paper, a new framework of causal independence modelling is proposed. It builds upon the link we discovered between the theory of causal independence and the convolution theorem of probability theory. The framework is developed by generalising this theorem into an algebra that supports the modelling of interactions, whether discrete, continuous, or both, in a meaningful way.

Pp. 185–193 in *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM-2010)*, edited by Petri Myllymäki, Teemu Roos and Tommi Jaakkola. HIIT Publications 2010–2.

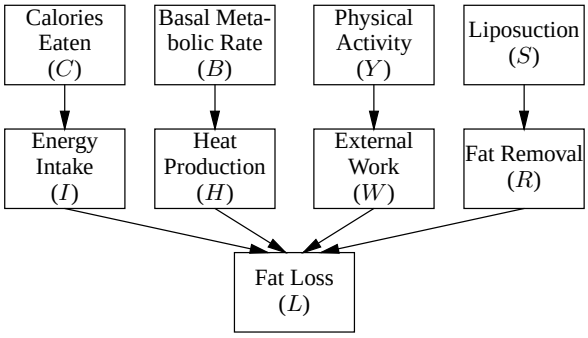

Figure 1: Causal factors that affect fat loss in humans.

2 Motivating Example

In biomedical modelling one often has to deal with a mixture of discrete and continuous causes that give rise to an effect. For example, the amount of *fat storage* in the human body is determined by the *energy balance*, i.e., the balance between energy intake and expenditure. A decrease in fat storage usually occurs whenever the energy intake is smaller than the energy expenditure. The energy expenditure is determined by the internal heat produced, which is mainly the basal metabolic rate (BMR), plus external work estimated by physical activity. Besides altering the energy balance, the storage can be decreased by means of *liposuction*. The energy variables are naturally represented as continuous variables, whereas 'Liposuction' is discrete.

The causal model is presented in Figure 1 and the conditional probability distributions of fat loss are represented by: $P(L \mid C, B, Y, S)$. Somehow this distribution must be determined by the interaction between the intermediate causal variables concerned, expressed by $A \equiv (I \leq (H + W))$ (energy intake is less than or equal to heat production plus external work), with A standing for an appropriate energy balance. Furthermore, the binary (Boolean) effect variable fat loss L is defined as $L \equiv (A \vee R)$ (fat loss L is due to a change in the energy balance A or fat removal R). The techniques developed in this paper will allow one to exploit such information in building a Bayesian network.

3 Preliminaries

This section provides a review of the basics underlying the research of this paper.

3.1 Probability theory and Bayesian networks

In this paper we are concerned with both discrete and continuous probability distributions P, defined in terms functions f, called the probability mass function for the discrete case and density function for the continuous case. Associated with a mass and density function, respectively, are distribution functions, denoted by F. Random variables are denoted by upper case, e.g., X, I etc. Instead of $X = x$ we will frequently write simply x. This is also the notation used to vary over values in summation and integration and to indicate that a binary variable X has the value 'true'. The value 'false' of a binary variable X is denoted by $\bar{x}$. Finally, free variables are denoted by uppercase, e.g., X.

A *Bayesian network* is a concise representation of a joint probability distribution on a set of random variables (Pearl, 1988). It consists of an acyclic directed graph $G = (\mathcal{V}, \mathcal{A})$, where each node $V \in \mathcal{V}$ corresponds to a random variable and $\mathcal{A} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of arcs. The absence of arcs in the graph G models independences between the represented variables. In this paper, we give an arc $V \rightarrow V'$ a causal reading: the arc's direction marks V' as the *effect* of the *cause* V. In the following, causes will often be denoted by C_i and their associated effect variable by E.

Associated with the qualitative part of a Bayesian network are numerical parameters from the encoded probability distribution. With each variable V in the graph is associated a set of *conditional probability distributions* $P(V \mid \pi(V))$, describing the joint influence of values for the parents $\pi(V)$ of V on the probabilities of the variable V's values. These sets of probabilities constitute the quantitative part of the network. A Bayesian network represents a joint probability distribution of its variables and thus provides for computing any probability of interest.

3.2 Causal modelling

One popular way to specify interactions among statistical variables in a compact fashion is offered by the notion of *causal independence* (Heckerman and Breese, 1996). The global structure of a causal-independence model is shown in Figure 2; it expresses the idea that causes $C = (C_1, \ldots, C_n)$ influence a given common effect E through interme-

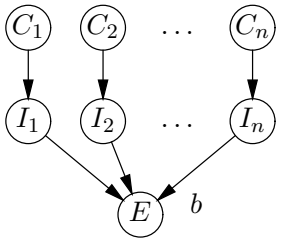

Figure 2: Causal independence model.

diate variables $I = (I_1, \ldots, I_n)$ and a Boolean, or Boolean-valued, function b, called the *interaction function*. The influence of each cause C_k on the common effect E is independent of each other cause C_j, $j \neq k$. The function b represents in which way the intermediate effects I_k, and indirectly also the causes C_k, interact to yield the final effect E. Hence, this function b is defined in such way that when a relationship, as modelled by the function b, between I_k, $k = 1, \ldots, n$, and $E = 1$ (true) is satisfied, then it holds that $b(I_1, \ldots, I_n) = 1$, denoted by $b(I_1, \ldots, I_n) = e$.

The conditional probability of the occurrence of the effect E given the causes $C_1, \ldots, C_n$, can be obtained from the conditional probabilities $P(I_k \mid C_k)$ as follows:

$$P_b(e \mid C_1, \ldots, C_n) = \sum_{b(i_1,\ldots,i_n)=e} \prod_{k=1}^{n} P(i_k \mid C_k) \quad (1)$$

Formula (1) is practically speaking not very useful, because the size of the specification of the function b is exponential in the number of its arguments. The resulting probability distribution is therefore in general computationally intractable, both in terms of space and time requirements. An important subclass of causal independence models, however, is formed by models in which the deterministic function b can be defined in terms of separate binary functions g_k, also denoted by $g_k(I_k, I_{k+1})$. Such causal independence models have been called *decomposable* causal independence models (Heckerman and Breese, 1996); these models are of significant practical importance. Often, all functions $g_k(I_k, I_{k+1})$ are identical for each k; a function $g_k(I_k, I_{k+1})$ may therefore be simply denoted by $g(I, I')$. Typical examples of decomposable causal independence models are the noisy-OR (Díez, 1993; Henrion, 1989; Pearl, 1988; Srinivas, 1993) and noisy-MAX (Díez, 1993; Heckerman and Breese, 1996; Srinivas, 1993) models, where the function g represents a logical OR and a MAX function, respectively.

In the case of continuous causal factors with a discrete effect variable, there are two main proposals for the conditional distribution of the discrete node (Bishop, 2006). Suppose we have a binary effect variable E and continuous parents $C_1, \ldots, C_n$. If E is modelled using a *logistic function*, then

$$P(e \mid C_1, \ldots, C_n) = \frac{\exp(b + w^T \varphi(C))}{1 + \exp(b + w^T \varphi(C))} \quad (2)$$

where $w^T = (w_1, \ldots, w_n)$ is a weight vector and $\varphi(C)$ a, possibly nonlinear, basis function applied to the causes C. The other option is to use the *probit regression model*, with

$$P(e \mid C_1, \ldots, C_n) = P(\Theta \leq (b + w^T \varphi(C))) \quad (3)$$

where $\Theta \sim N(0, 1)$. Although both types of model are flexible, it is very hard to come up with sensible weight vectors w and basis functions φ based only on available domain knowledge of the relations between causes.

3.3 The convolution theorem

A classical result from probability theory that is useful when studying sums of variables is the convolution theorem. The following well-known theorem (cf. (Grimmett and Stirzaker, 2001)) is central to the research reported in this paper.

Theorem 1. *Let f be a joint probability mass function of the random variables X and Y, such that $X + Y = z$. Then it holds that $P(X + Y = z) = f_{X+Y}(z) = \sum_x f(x, z - x)$.*

Proof. The (X, Y) space determined by $X + Y = z$ can be described as the union of disjoint sets (for each x): $\bigcup_x (\{X = x\} \cap \{Y = z - x\})$, from which the result follows. □

If X and Y are independent, then, in addition, the following corollary holds.

Corollary 1. *Let X and Y be two independent random variables, then it holds that*

$$\begin{aligned} P(X + Y = z) &= f_{X+Y}(z) \\ &= \sum_x f_X(x) f_Y(z - x) \quad (4) \end{aligned}$$

The probability mass function f_{X+Y} is in that case called the *convolution* of f_X and f_Y, and it is commonly denoted as $f_{X+Y} = f_X * f_Y$. The convolution theorem is very useful, as sums of random variables occur very frequently in probability theory and statistics. The convolution theorem can also be applied recursively, i.e.,

$$f_{X_1+\cdots+X_n} = f_{X_1} * \cdots * f_{X_n}$$

as follows from the recursive application of Equation (4):

$$P(X_1 + \cdots + X_n = z) = \sum_{y_{n-2}} \sum_{y_{n-3}} \cdots \sum_{y_1} \sum_{x_1} f_{X_1}(x_1) f_{X_2}(y_1 - x_1) \cdots f_{X_{n-1}}(y_{n-2} - y_{n-3}) f_{X_n}(z - y_{n-2}) \quad (5)$$

where we use the following equalities:

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_i &= Y_{i-1} + X_{i+1}, \ \ \forall i{:}\, 2 \leq i \leq n-2 \end{aligned}$$

Thus, $Y_{n-2} = X_1 + \cdots + X_{n-1}$, and $X_n = z - Y_{n-2}$. As addition is commutative and associative, any order in which the Y_i's are determined is valid.

The convolution theorem does not only hold for the addition of two random variables, but also for Boolean functions of random variables. However, in contrast to the field of real numbers where a value of a random variable X_n is uniquely determined by a real number z and y_{n-2} through $X_n = z - y_{n-2}$, in Boolean algebra values of Boolean variables only *constrain* the values of other Boolean variables. These constraints may yield a set of values, rather than a single value, which is still compatible with the convolution theorem. In the following, we use the notation $b(X, y) = z$ for such constraints, where the Boolean values y and z constrain X to particular values. For example, for $(X \vee y) = z$, where y, z stand for $Y = 1$ (Y has the value 'true') and $Z = 1$ (Z has the value 'true'), it holds that $X \in \{0, 1\}$.

Theorem 2. *Let f be a joint probability mass function of independent random, Boolean variables I and J and let b be a Boolean function defined on I and J, then it holds that*

$$P(b(I, J) = e) = \sum_i f_I(i) P(b(i, J) = e)$$

Proof. The (I, J) space defined by $b(I, J) = e$ can be decomposed as follows: $\bigcup_i \{I = i\} \cap \{J = j \mid b(i, j) = e\}$, where the expression $b(i, j) = e$ should be interpreted as a logical constraint on the Boolean values of the variable J. As in Theorem 1, the individual sets $\{I = i\} \cap \{J = j \mid b(i, j) = e\}$ are mutually exclusive. □

This theorem is illustrated by the following example.

Example 1. Consider the example given in Figure 1 as discussed in Section 2, and the Boolean relation $A \vee R \equiv L$, which expresses that fat loss L is due to changes in the energy balance A or fat removal R. By applying Theorem 2 the following results: $P(A \vee R = l) = \sum_a f_A(a) P(a \vee R = l) = f_A(a)\left(f_R(r) + f_R(\bar{r})\right) + f_A(\bar{a}) f_R(r) = f_A(a) f_R(r) + f_A(a) f_R(\bar{r}) + f_A(\bar{a}) f_R(r)$, where the term $(f_R(r) + f_R(\bar{r}))$ results from the logical constraint that $a \vee R = l$, i.e., $R \in \{0, 1\}$. Note that this is exactly the same result as for the noisy-OR model with the causal variables C marginalised out:

$$P_\vee(l) = \sum_{a \vee r = l} f_A(a) f_R(r) = P(A \vee R = l)$$

4 Convolution-based Causal Independence

In this section, we start to systematically explore the relationship between the convolution theorem of probability theory and the theory of causal independence.

4.1 General idea

The idea now is that we can use any Boolean-valued function, as long as the function is decomposable, to model causal interaction using the convolution theorem. A discrete causal independence model can also be written as follows:

$$P_b(e \mid C) = P(b(I_1, \ldots, I_n) = e \mid C)$$

where the right hand side can be determined as follows:

$$P(b(I_1, \ldots, I_n) = e \mid C) = \sum_{j_{n-2}} \sum_{j_{n-3}} \cdots \sum_{j_1} \sum_{i_1} f_{I_1}(i_1 \mid C_1) \cdot P_{I_2}(b_1(i_1, I_2) = j_1 \mid C_2) \cdots P_{I_n}(b_{n-1}(j_{n-1}, I_n) = e \mid C_n) \quad (6)$$

and the Boolean random variables J_k are defined in terms of I_l's dependent on the constraints imposed by the Boolean operators b_k. This can be proven by an inductive argument over all the cause variables. If we use a single operator $\odot$ that is commutative and associative, then the order of evaluation does not matter, and we can ignore parentheses: $b(I_1, \ldots, I_n) = I_1 \odot \cdots \odot I_n$ (Zhang and Poole, 1996; Lucas, 2005). However, if the single operator used to define the Boolean function b is neither commutative nor associative, then the order in which the Boolean expression is evaluated matters, and one should use parentheses.

The principles discussed above carry over to the continuous case. The convolution theorem for continuous variables X, Y, and Z, with $Z = X + Y$, has the following form:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\,\mathrm{d}x$$

where f_{X+Y}, f_X, and f_Y are probability density functions, and the variables X and Y are assumed to be independent. In the context of the theory of causal independence, we use convolution to compute the conditional probability density function $f_b(e \mid C)$, in a way very similar to the discrete case, where b is the causal interaction function.

4.2 A language for modelling interactions

To carry over the ideas of causal independence from the discrete case, we consider various operators for continuous variables. This will build up a rich language for modelling causal independence.

4.2.1 Boolean-valued continuous operators

Moving to the continuous case, first let I be a set of independent continuous causal random variables with associated probability density $f(I \mid C)$. Consider the Boolean-valued decomposable functions b, i.e., functions $b : I \to \{0, 1\}$, such that constraints on some variables $I' \subset I$ imposed by b are measurable sets of values for I'. We now wish to use the theory of causal independence in order to decompose the probability mass $f_b(e \mid C)$. If $I = \{J, K\}$ are continuous intermediate variables and $C = \{C_J, C_K\}$ the relevant causal variables, then:

$$\begin{aligned} f_b(e \mid C) &= P(b(J,K) = e \mid C) \\ &= \iint_{b(j,k)=e} f_{JK}(j,k \mid C)\,\mathrm{d}k\,\mathrm{d}j \\ &= \int_{-\infty}^{\infty} f_J(j \mid C_J) \int_{b(j,k)=e} f_K(k \mid C_K)\,\mathrm{d}k\,\mathrm{d}j \quad (7) \\ &= \int_{-\infty}^{\infty} f_J(j \mid C_J) P(b(j,K) = e \mid C_K)\mathrm{d}j \quad (8) \end{aligned}$$

The constraint $b(j, K) = e$ determines a subspace of the real numbers for variable K over which the density function f_K is integrated.

For a general n-ary Boolean-valued function b of continuous variables, we can apply this equation recursively, which gives:

$$\begin{aligned} f_b(e \mid C) = P(b(I_1, I_2, \ldots, I_n) = e \mid C) = \\ \int_{-\infty}^{\infty} f_{I_1}(i_1 \mid C_1) \int_{b(i_1,i_2,\ldots,i_n)=e} f_{I_2}(i_2 \mid C_2) \cdots \\ \cdot \int_{b(i_1,\ldots,i_n)=e} f_{I_n}(i_n \mid C_n)\,\mathrm{d}i_n \cdots \mathrm{d}i_1 \quad (9) \end{aligned}$$

If b is defined on both discrete and continuous variables, then this yields a mix of sums and integrals by repeated application of Theorem 2 and Eq. (8).

Analogously to the convolution notation, we define an operator $\textcircled{b}$ for denoting this decomposition for any Boolean function such that:

$$\textcircled{b}\,(f_{I_1}^{C_1}, \ldots, f_{I_n}^{C_n})(e) = f^{C}_{b(I_1,\ldots,I_n)}(e) = f_b(e \mid C)$$

where the superscripts C_1 and C_2 represent conditioning on the corresponding variables. This allows us to deal with complex combinations of such operators in a compact fashion.

If b is binary, we use an infix notation; e.g., $\textcircled{\vee}$ denotes the decomposition of two densities f_J and f_K using a logical OR. Returning to the fat loss problem (denoted by the variable L with l standing for $L = 1$) of Example 1, we have:

$$(f_A \mathbin{\textcircled{\vee}} f_R)(l) = \sum_a f_A(a) P((a \vee R) = l)$$

which is again the noisy-OR operator.

In the following section, a language that supports Boolean combinations of relations is developed.

4.2.2 Relational operators

The relational operators are treated similarly to convolutions and Boolean operators by viewing a

relation and a value of a random variable as a constraint on the other variables. First, basic operators to build up our language are basic relational operators, such as $=, \leq, >$. Consider $\leq$:

$$P_{\leq}(e \mid C) = P((I_1 \leq I_2) = e \mid C) = \iint_{(i_1 \leq i_2)=e} f(i_1, i_2 \mid C)\,\mathrm{d}i_1\,\mathrm{d}i_2 \quad (10)$$

If I_1 and I_2 independent, then the following equality results:

$$\begin{aligned} P_{\leq}(e \mid C) &= \int_{-\infty}^{\infty} f_{I_1}(i_1 \mid C_1) \cdot P((i_1 \leq I_2) = e \mid C_2)\,\mathrm{d}i_1 \\ &= \int_{-\infty}^{\infty} f_{I_1}(i_1 \mid C_1) \cdot \int_{i_i}^{\infty} f_{I_2}(i_2 \mid C_2)\,\mathrm{d}i_2\,\mathrm{d}i_1 \end{aligned}$$

A similar expression can be derived for $>$, while $P((I_1 = I_2) = e \mid C) = 0$ as $P((I_2 = i_1 \mid C_2) = 0$ for continuous variables I_1 and I_2. This expression implies that, in case I_1 and I_2 are independent, the relation can be decomposed. As a result, we can use the notation as introduced earlier to obtain operators $\circledR$:

$$\begin{aligned} (f_{I_1}^{C_1} \circledR f_{I_2}^{C_2})(e) &= f_R(e \mid C) \\ &= P(R(I_1, I_2) = e \mid C) \end{aligned}$$

where R is one of the basic relational operators.

Subsequently, we look at the extension of this language with convolutions of the interaction between variables and constants. A constant k can be described by a uniform probability distribution with density function

$$f_J(j) = \begin{cases} 1/\delta & \text{if } j \in (k - \delta/2, k + \delta/2] \\ 0 & \text{otherwise} \end{cases}$$

for $\delta \in \mathbb{R}^+$ very small, then

$$\begin{aligned} P((I \leq J) = e) &= (f_I \circledS f_k)(e) \\ &= \int_{-\infty}^{k} f_I(i)\,\mathrm{d}i = P(I \leq k) \end{aligned}$$

as one would expect. For convenience, we have written f_k for this density function f_J and will do so in the following.

For modelling the interaction between convolutions of variables, let I a set of continuous random variables and K a set of constants. Then, a *sum-relation* is a Boolean-valued function b such that

$$b(I) = R(\sum_{k=1}^{n} V_k, \sum_{l=1}^{m} W_l)$$

where $V \subseteq I \cup K$, $W \subseteq I \cup K$, and R is a relational operator.

If V and W do not overlap in variables except for the constants, the sums of V and W are independent. In that case, the relation can be decomposed by Eq. (9). So we have the following proposition.

Proposition 1. *The causal independence model of a sum-relation $R(\sum_{k=1}^{n} V_k, \sum_{l=1}^{m} W_l)$ with continuous interaction variables I can be written as:*

$$\begin{aligned} &P(R(\sum_{k=1}^{n} V_k, \sum_{l=1}^{m} W_l) = e) \\ &\quad = (f_{V_1+\cdots+V_n} \circledR f_{W_1+\cdots+W_m})(e) \end{aligned}$$

if $V \cap W \cap I = \varnothing$.

Example 2. Recall the example in Figure 1 as discussed in Section 2. The causal independence model of the energy balance A can be written as:

$$\begin{aligned} &P((I \leq H + W) = a \mid C, B, Y) \\ &= (f_I^C \circledS f_{H+W}^{\{B,Y\}})(a) = (f_I^C \circledS (f_H^B * f_W^Y))(a) \end{aligned}$$

where $*$ is the convolution operator.

This approach could be extended easily to other operators, such as subtraction, but we refrain from this because of space limitations.

4.2.3 Boolean combinations of relations

Sum-relations can now be combined using Boolean functions in a uniform manner. Let I_c be a set of continuous causal random variables, I_d a set of discrete causal random variables, and $I = I_c \cup I_d$. A *Boolean combination bc* is a Boolean-valued function defined on I as follows:

$$bc(I) = b(R_1(V_1), \ldots, R_n(V_n), I_d)$$

where b is a Boolean function and $R_1, \ldots, R_n$ a set of sum-relations.

If the continuous variables in the Boolean combinations of relations are partitioned, Eq. (6) can be applied to obtain the following proposition.

Proposition 2. *The causal independence model of a Boolean combination of sum-relations $b(R_1(V_1), \ldots, R_2(V_n))$, can be written as:*

$$P(b(R_1(V_1), R_2(V_2)) = e \mid C) = (f^{C_1}_{R_1(V_1)} \textcircled{b} f^{C_2}_{R_2(V_2)})(e)$$

if $V_1 \cap V_1 = \varnothing$.

Example 3. Again, consider the example in Figure 1 as discussed in Section 2. We are now in the position to decompose the full causal independence function representing fat loss L.

$$\begin{aligned} &P((I \leq H + W) \vee R) = l \mid C, B, Y, S) \\ &= P((R \vee (I \leq H + W)) = l \mid C, B, Y, S) \\ &= f^{\{C,B,Y,S\}}_{R \vee (I \leq H+W)}(l) \\ &= (f^S_R \textcircled{\vee} f_{\mathrm{L} \leq \mathrm{H+W}})(l) \\ &= (f^S_R \textcircled{\vee} (f^C_I \textcircled{\leq} (f^B_H * f^Y_W)))(l) \end{aligned}$$

5 Special Probability Distributions

In this section, the theory developed in the previous sections is illustrated by actually choosing special probability distributions to model problems.

5.1 Bernoulli distribution

As an example of discrete distributions, we take the simplest one: the Bernoulli distribution. This distribution has a probability mass function f such that $f(0) = 1 - p$ and $f(1) = p$. Let $P(I_k \mid c_k)$ be Bernoulli distributions with parameters p_k where $k = \{1, 2\}$. Suppose the interaction between C_1 and C_2 is modelled by $\leq$, then the effect variable E also follows a Bernoulli distribution with parameter:

$$\begin{aligned} P_{\leq}(e \mid c_1, c_2) &= (f^{c_1}_{I_1} \textcircled{\leq} f^{c_1}_{I_2})(e) \\ &= \sum_{i_1} f_{I_1}(i_1 \mid c_1) P((i_1 \leq I_2) = e \mid c_2) \\ &= p_1 - p_1 p_2 + 1 \end{aligned}$$

By the same reasoning, we obtain the parameters of the resulting distribution when $\bar{c}_1$ or $\bar{c}_2$.

5.2 Exponential distribution

In order to model the time it takes for the effect to take place due to the associated cause, we use the exponential probability distribution with distribution function $F(t) = 1 - e^{-\lambda t}$, where $t \in \mathbb{R}^+_0$ is the time it takes before the effect occurs. The associated probability density function is $f(t) = F'(t) = \lambda e^{-\lambda t}$. Now, let I_1 and I_2 stand for two of such temporal random variables such that $I_1 \leq I_2$, meaning that intermediate effect I_1 does not occur later than I_2. The probability mass of E to occur is:

$$\begin{aligned} P_{\leq}(e \mid C) &= (f^{c_1}_{I_1} \textcircled{\leq} f^{c_1}_{I_2})(e) \\ &= \int_{-\infty}^{\infty} f_{I_1}(i_1 \mid c_1) P((i_1 \leq I_2) = e \mid c_2) \,\mathrm{d}i_1 \\ &= \int_{-\infty}^{\infty} f_{I_1}(i_1 \mid c_1) \int_0^{\infty} f_{I_2}(i_1 + \delta \mid c_2) \,\mathrm{d}\delta \,\mathrm{d}i_1 \\ &= \int_{-\infty}^{\infty} \lambda_1 e^{-\lambda_1 i_1} e^{-\lambda_2 i_1} \,\mathrm{d}i_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} \end{aligned}$$

where we use a delay $\delta \geq 0$. If $\lambda_1 = \lambda_2$, then $P_{I_1 \leq I_2}(e \mid C) = 1/2$.

5.3 Conditional Gaussian distribution

The most common hybrid distribution for Bayesian networks is the conditional Gaussian distribution (Lauritzen and Wermuth, 1989). We illustrate the theory for the case when a continuous interaction variable I has a continuous cause variable C. The distribution of I is given in this model by $f(i \mid C) = N(\alpha + \beta C, \sigma^2)$. Let I_1 and I_2 be two such random variables with causal variables C_1 and C_2. It is well-known that variable E with $f_{I_1 - I_2}(e \mid C)$ is distributed Gaussian with mean $\alpha_1 + \beta_1 C_1 - \alpha_2 - \beta_2 C_2$ and variance $\sigma_1^2 + \sigma_2^2$. Similarly, the convolution of two Gaussian variables is a Gaussian variable with the sums of means and variances. Because of space limitations, the derivations are omitted.

Here we illustrate the relational operator $\leq$. The probability $P_{\leq}(e \mid C)$ can be obtained by

$$\begin{aligned} P_{\leq}(e \mid C) &= f^{C_1}_{I_1} \textcircled{\leq} f^{C_2}_{I_2} \\ &= (f^{C_1}_{I_1} \textcircled{-} f^{C_2}_{I_2}) \textcircled{\leq} 0 = F_J(0) \\ &= \tfrac{1}{2}\left[1 + \mathrm{erf}\left(\frac{-(\alpha_1 + \beta_1 c_1 - \alpha_2 - \beta_2 c_2)}{\sqrt{2(\sigma_1^2 + \sigma_2^2)}}\right)\right] \\ &= P(\Theta \leq b + w_1 c_1 + w_2 c_2) \end{aligned}$$

where $b = \frac{\alpha_2 - \alpha_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$, $w_1 = \frac{-\beta_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$, $w_2 = \frac{\beta_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$, and $\Theta \sim N(0, 1)$, which is a probit regression model (cf. Section 3.2).

Example 4. Consider the energy balance A as decomposed in Example 2. Suppose all causal and interaction variables are conditionally Gaussian. Suppose the balance is negative, i.e., a is true, then,

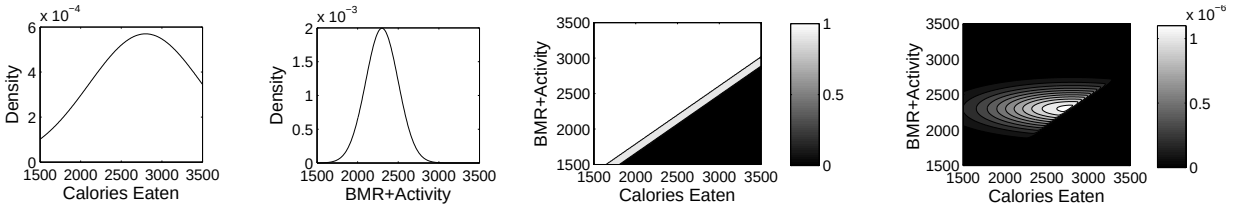

Figure 3: Example distributions, where, from left to right, the first figure shows the density of $C \sim N(2800, 700)$; the second figure shows the density of $B + Y \sim N(2300, 200)$; the third figure shows the probability distributions $P(A \mid C, B + Y)$ with $A \equiv I \leq (H + W)$ where $I \sim N(0.9 \cdot C, 200)$ and $H + W \sim N(1.1 \cdot (B + Y), 300)$; finally, the figure on the right shows the joint density of $\{A, C, B + Y\}$.

$(f_H^B * f_W^Y)(a)$ represents a distribution $N(\alpha_H + \alpha_W + \beta_H C_B + \beta_W C_Y, \sigma_H^2 + \sigma_W^2)$, i.e., the sum of the mean and variance. Using the above, it follows that the probability of a is:

$$P(a) = (f_I^C \circledS (f_H^B * f_W^Y))(a)$$

which is a probit regression model with $b = (\alpha_I - \alpha_H - \alpha_W)/\sigma'$, $w_C = \beta_I/\sigma'$, $w_B = -\beta_H/\sigma'$, and $w_Y = -\beta_W/\sigma'$, where $\sigma' = \sqrt{\sigma_I^2 + \sigma_H^2 + \sigma_W^2}$.

In Figure 3 a number of plots are given to illustrate this model for some realistic parameters. Note that the energy balance distributions depicted in the third figure are split up into 0 (too much intake), 1 (too much energy expenditure), and an uncertain band in the middle.

6 Conclusions

We presented a new algebraic framework for causal independence modelling of Bayesian networks that goes beyond what has been available so far. In contrast to other approaches, the framework supports the modelling of discrete as well as of continuous variables, either separately or mixed.

The design of the framework was inspired by the convolution theorem of probability theory, and it was shown that this theorem easily generalises to convolution with Boolean-valued functions. We also studied a number of important modelling operators. Contrary to regression models, we were thus able to model interactions between variables using knowledge at hand. Furthermore, the theory was illustrated by a number of typical probability distributions which one needs to use when actually building Bayesian network models for problems. Finally, although some of the results suggest that standard tools for solving the inference problem can be used, such as the probit model for the conditional Gaussian distribution, more research is required and such we intend to undertake in the near future.

References

C.M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

F.J. Díez. 1993. Parameter adjustment in Bayes networks: the generalized noisy OR-gate. In *UAI'93*, pages 99–105.

G. Grimmett and D. Stirzaker. 2001. *Probability and Random Processes*. Oxford University Press, Oxford.

D. Heckerman and J.S. Breese. 1996. Causal independence for probabilistic assessment and inference using Bayesian networks. *IEEE Transactions on Systems, Man and Cybernetics*, 26(6):826–831.

M. Henrion. 1989. Some practical issues in constructing belief networks. In J.F. Lemmer and L.N. Kanal, editors, *Uncertainty in Artificial Intelligence*, pages 161–173, Amsterdam. Elsevier.

S.L. Lauritzen and N. Wermuth. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17:31–57.

P.J.F. Lucas. 2005. Bayesian network modelling through qualitative patterns. *AI*, 163:233–263.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Palo Alto.

S. Srinivas. 1993. A generalization of the noisy-OR model. In *UAI'93*, pages 208–215.

N.L. Zhang and D. Poole. 1996. Exploiting causal independence in Bayesian network inference. *JAIR*, 5:301–328.