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

Peter J. F. Lucas,[1,2,*] Arjen Hommersom[1,†]
[1] *Institute for Computing and Information Sciences, Radboud University Nijmegen, 6525 EC Nijmegen, The Netherlands*
[2] *Leiden Institute of Advanced Computer Science, Leiden University, 2300 RA Leiden, The Netherlands*

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 modeled 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 generalized 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 in detail for the first time in this paper. Furthermore, the new theory is used as a basis to develop a relational theory of probabilistic interactions. It is also illustrated how this new theory can be used in connection with special probability distributions. © 2014 Wiley Periodicals, Inc.

1. INTRODUCTION

During the past two decades, probabilistic graphical models, and in particular Bayesian networks, have become popular methods for building applications involving uncertainty in many domains such as biology,[1,2] medicine,[3–5] and engineering.[6,7] Bayesian networks can be developed manually, e.g., by acquiring relevant knowledge from experts in a domain, and learnt from data, whereas mixing manual design and learning is also possible.[8] As a Bayesian network consists of a graph representation and an associated probability distribution, it is common to split up the task of developing a Bayesian network for a problem into two steps: (1) determining the graph or structure and (2) assessing the parameters of the probability

[*] Author to whom all correspondence should be addressed; e-mail: peterl@cs.ru.nl.
[†] e-mail: arjenh@cs.ru.nl.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 30, 209–235 (2015)

View this article online at wileyonlinelibrary.com. • DOI 10.1002/int.21698

distribution once the structure is known. In particular, this second step, estimation of the associated probabilistic parameters, is often challenging.

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, on the one hand, and $C_1, \ldots, C_n$, on the other hand, as the interactions between *causes* C_k and an *effect* E. Although not all Bayesian networks are causal networks,[9] causal knowledge is crucial in designing and interpreting Bayesian networks in particular problem domains. This insight has driven much of the early work on Bayesian networks, as is reflected in the seminal work by Pearl[11] 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.[10–13] It should be noted that in causal independence models the causes are not necessarily independent, but rather that the causes act independently on E. For this reason, the theory of causal independence is also called "intercausal independence" or "independence of causal interaction".[14] For historical reasons, we will use the term "causal independence."

So far, the theory of causal independence was restricted to modeling 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 modeled as disjunctive[10,11,15]—and the noisy-AND.[16] Interactions among continuous cause variables are usually modeled using standard statistical techniques, such as logistic regression and probit regression, typically by using iterative numerical methods that estimate the weight vector maximizing the likelihood of the data given the model.[17] Thus, these regression models resist manual construction based on a good 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 for the problem at hand, is seen as one of the key benefits of the technique. Finally, it is not possible to combine regression models with (discrete) causal-independence models.

In this paper, a new framework of causal-independence modeling is proposed. It builds upon the link we discovered between the theory of causal independence and the convolution theorem of probability theory. More in particular, the paper offers the following contributions:

- a new generalization of the theory of causal independence by offering support for the representation of discrete, continuous, and mixtures of discrete and continuous probabilistic interactions;
- a relational algebra that supports the modeling of relational interactions in a meaningful way.

We illustrate how the theory can be used for a number of different special probability distributions. By means of examples, we show how it can be deployed to solve actual problems. The developed theory can also be used in the context of probabilistic logic.[18] However, in this paper the focus is on exploitation in the context of Bayesian networks.

The structure of the paper is as follows. In Section 2, we provide a motivation why a relational theory of continuous and discrete probability distributions is needed. Section 3 offers the necessary background on Bayesian networks, causal independence, convolution, and the relationship between causal independence and convolution. Then, in Section 4, we prove that causal independence is equivalent to convolution, which allows us to generalize causal independence to the continuous case. This acts as a basis for the introduction of a relational language that provides extra expressive power for modeling interactions, which we introduce in Section 5. We then show how to use this language with different types of probability distributions in Section 6. In Section 7, we compare the research ideas explored in this paper to other work. Finally, in Section 8, we summarize what has been achieved and some future directions of research are mentioned.

2. MOTIVATING EXAMPLE

In biomedical modeling, 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 less 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 $P(L \mid C, B, Y, S)$. To accurately estimate this distribution, we may consider the underlying physiological mechanism that the causes induce, expressed by the intermediate causal variables I, H, W, and R. In addition, there are deterministic interactions between the intermediate causal variables, for example, we may want to model that

$$A \equiv (I \leq (H + W))$$

(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). While existing methods may be geared toward representing the same distribution, for example by adding an additional node A and using other representational tricks, none of the existing methods provide support to represent this knowledge directly and to reason about its properties. The techniques developed in this paper will allow one to exploit such information in building a Bayesian network in a general way.

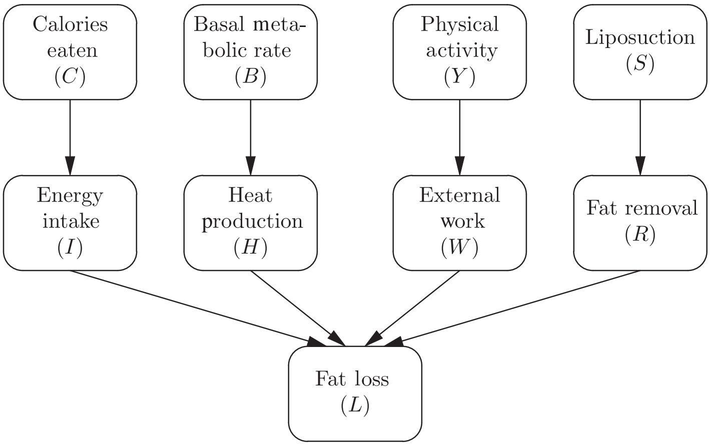

Figure 1. Causal factors that affect fat loss in humans.

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 of functions f, called a probability mass function for the discrete case and density function for the continuous case. When we use Boolean expressions to define probability distributions, we also use P. We sometimes use the notation f_g to indicate that a function g is associated with a probability function. Associated with a mass and density function, respectively, are distribution functions, denoted by F, and defined in terms of mass and density functions as usual.[19] 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 $\bar{x}$. We will denote sets of random variables by $\mathbf{X}$, i.e., $\mathbf{X} = \{X_1, \ldots, X_n\}, n \geq 1$ or $\mathbf{X} = \varnothing$. Finally, free variables are denoted by uppercase, e.g., X or $\mathbf{X}$. From the context, it will become apparent which meaning is intended.

A *Bayesian network* is a concise representation of a joint probability distribution on a set of random variables.[11] It consists of an acyclic directed graph $G = (\mathbf{V}, \mathbf{A})$, where each node $V \in \mathbf{V}$ corresponds to a random variable and $\mathbf{A} \subseteq \mathbf{V} \times \mathbf{V}$ is a set of arcs. The absence of arcs in the graph G models independences between the represented variables. Informally speaking, we take an arc $V \rightarrow V'$ between the nodes V and V' to represent an influential relationship between the associated variables of V and V'. If this arc is given a causal reading, the arc's direction marks

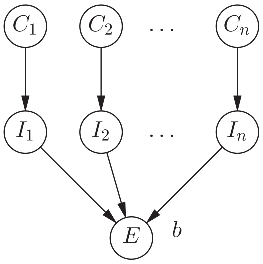

Figure 2. Causal-independence model.

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. The distinction between cause and effect is only meant to support Bayesian network modeling in collaboration with domain experts.

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 with 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 Modeling

One popular way to specify interactions among statistical variables in a compact fashion is offered by the notion of *causal independence.*[12] The global structure of a causal-independence model is shown in Figure 2; it expresses the idea that causes $\mathbf{C} = \{C_1, \ldots, C_n\}$ influence a given common effect E through intermediate variables $\mathbf{I} = \{I_1, \ldots, I_n\}$ 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 what way the intermediate effects I_k, and indirectly also the causes C_k, interact to yield the final effect E. Thus, the function b is defined such that when a relationship between I_k, $k = 1, \ldots, n$, and $E = 1$ (*true*) is satisfied, then it holds that $b(I_1, \ldots, I_n) = 1$ (*true*), denoted for ease of understanding by $b(I_1, \ldots, I_n) = e$.

In terms of probability theory, the notion of causal independence can be formalized for the occurrence of effect E as follows: By standard probability theory,

$$P(E \mid C_1, \ldots, C_n) = \sum_{i_1, \ldots, i_n} P(E \mid i_1, \ldots, i_n, C_1, \ldots, C_n) P(i_1, \ldots, i_n \mid C_1, \ldots, C_n) \tag{1}$$

meaning that the causes $\mathbf{C} = \{C_1, \ldots, C_n\}$ influence the common effect E through the intermediate effects $I_1, \ldots, I_n$. The deterministic probability distribution $P(E \mid I_1, \ldots, I_n)$ corresponds to the Boolean function b, where $b(I_1, \ldots, I_n) = e$ if $P(e \mid I_1, \ldots, I_n) = 1$; otherwise, $b(I_1, \ldots, I_n) = \bar{e}$ if $P(e \mid I_1, \ldots, I_n) = 0$. To ease the notation in the remainder of this paper, we will from now on always consider the case where $E = 1$ denoted by e; the case for $\bar{e}$ can be derived from the positive case. Note that the effect variable E is conditionally independent of $C_1, \ldots, C_n$ given the intermediate variables $I_1, \ldots, I_n$, and that each variable I_k is only dependent on its associated variable C_k; hence, it holds that

$$P(e \mid I_1, \ldots, I_n, C_1, \ldots, C_n) = P(e \mid I_1, \ldots, I_n)$$

and

$$P(I_1, \ldots, I_n \mid C_1, \ldots, C_n) = \prod_{k=1}^{n} P(I_k \mid C_k)$$

Formula (1) can now be simplified to

$$P_b(e \mid \mathbf{C}) = \sum_{b(i_1,\ldots,i_n)=e} \prod_{k=1}^{n} P(i_k \mid C_k) \tag{2}$$

Formula (2) 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 for large values of n, 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;[12] 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 $g(I, I')$, and the function b is obtained by recursive application of g. Typical examples of decomposable causal-independence models are the noisy-OR[11,12] and noisy-MAX[10,15] models, where the function g represents a logical OR and a MAX function, respectively. Decomposable functions are normally defined in terms of *algebraic expressions,*[20] where in particular unary and binary operators are used—Boolean operators for Boolean expressions.[21]

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.[17] Suppose we have a binary effect variable E and continuous parents $C_1, \ldots, C_n$. If E is modeled using a *logistic function*, then

$$P(e \mid C_1, \ldots, C_n) = \frac{\exp(w_0 + w^T \varphi(\mathbf{C}))}{1 + \exp(w_0 + w^T \varphi(\mathbf{C}))} \tag{3}$$

where w_0 is called the intercept, $w^T = (w_1, \ldots, w_n)$ is a weight vector ,and $\varphi(\mathbf{C})$ is a, possibly nonlinear, basis function applied to the causes $\mathbf{C}$. The other option is to use the *probit model*, with

$$P(e \mid C_1, \ldots, C_n) = P(\Theta \leq (w_0 + w^T \varphi(\mathbf{C}))) \tag{4}$$

where $\Theta \sim N(0, 1)$, i.e., Θ is distributed following a standard Gaussian distribution with mean 0 and variance 1. The logistic model has certain advantages compared to the probit model,[22] e.g., it can easily be generalized to multivalued discrete variables. 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. Instead, these models are typically used in regression analysis, where the parameters are estimated from data.

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. Causal Independence as Convolution

A classical result from probability theory that is useful when studying sums of variables is the convolution theorem. The following theorem[19] is central to the research reported in this paper, and is used for deriving the convolution theorem as a special case (see below).

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): $\bigcup_x (\{X = x\} \cap \{Y = z - x\})$, and the sets $\{X = x\} \cap \{Y = z - x\}$ are mutually disjoint for each x. Thus,

$$\begin{aligned} P(X + Y = z) &= P\left(\bigcup_x (\{X = x\} \cap \{Y = z - x\})\right) \\ &= \sum_x P(X = x, Y = z - x) \\ &= \sum_x f(x, z - x), \end{aligned}$$

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*

$$\begin{aligned} P(X+Y=z) &= f_{X+Y}(z) \\ &= \sum_x P(X=x)P(Y=z-x) \\ &= \sum_x f_X(x) f_Y(z-x) \\ &= \sum_y f_X(z-y) f_Y(y) \end{aligned} \tag{5}$$

The probability mass function f_{X+Y} defined by the sum of random variables X and Y is called the *convolution* of f_X and f_Y, and it is commonly denoted as

$$f_{X+Y} = f_X * f_Y.$$

This 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.,

$$f_{X_1+\cdots+X_n} = f_{X_1} * \cdots * f_{X_n} \tag{6}$$

as follows from the recursive application of Equation 5:

$$\begin{aligned} P(X_1+\cdots+X_n=z) = \sum_{y_{n-2}} \sum_{y_{n-3}} \cdots \sum_{y_1} \sum_{x_1} f_{X_1}(x_1) f_{X_2}(y_1-x_1) \cdots \\ f_{X_{n-1}}(y_{n-2}-y_{n-3}) f_{X_n}(z-y_{n-2}) \end{aligned} \tag{7}$$

where we use the following equalities:

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= Y_1 + X_3 \\ \vdots & \quad \vdots \\ Y_{n-3} &= Y_{n-4} + X_{n-2} \\ Y_{n-2} &= Y_{n-3} + X_{n-1} \end{aligned}$$

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_is are determined is valid.

From now on, we will use the term convolution in its general sense describing virtually any operation on random variables, not just addition. This is how it is used by Williamson[23] where other arithmetical operations on random variables, i.e., subtraction, multiplication, and division, are discussed. In particular, it turns out that 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 proof is almost identical to that of Theorem 1. 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 result 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:

$$\begin{aligned} P(A \vee R = l) &= \sum_a f_A(a) P(a \vee R = l) \\ &= f_A(a)(f_R(r) + f_R(\bar{r})) + f_A(\bar{a}) f_R(r) \\ &= f_A(a) f_R(r) + f_A(a) f_R(\bar{r}) + f_A(\bar{a}) f_R(r), \end{aligned}$$

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 marginalized out:

$$\begin{aligned} P_\vee(l) &= \sum_{a \vee r = l} f_A(a) f_R(r) \\ &= f_A(a) f_R(r) + f_A(a) f_R(\bar{r}) + f_A(\bar{a}) f_R(r) \\ &= P(A \vee R = l). \end{aligned}$$

4.2. Generalizing Causal Independence

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. Then, the hypothesis is that a discrete causal independence model can also be written as follows:

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

where the right-hand side can be determined as follows:

$$P(b(I_1, \ldots, I_n) = e \mid \mathbf{C}) = \sum_{j_{n-2}} \sum_{j_{n-3}} \cdots \sum_{j_1} \sum_{i_1} f_{I_1}(i_1 \mid C_1)\, P_{I_2}(b_1(i_1, I_2) = j_1 \mid C_2) \\ \cdots P_{I_n}(b_{n-1}(j_{n-1}, I_n) = e \mid C_n) \tag{8}$$

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. Equation 8 can be proven by an inductive argument over all the cause variables using Theorem 2. In particular, if we use a single operator $b_k = \odot$ that is commutative and associative, then the order of evaluation does not matter and we can ignore parentheses: $b(I_1, \ldots, I_n) = I_1 \odot \cdots \odot I_n$.[13,50] However, if the single operator used to define the Boolean function b is not commutative or associative, then the order in which the Boolean expression is evaluated matters, and one should use parentheses.

THEOREM 3. *Let $P_b(e \mid \mathbf{C})$ be defined as a causal-independence model in terms of the Boolean function b on Boolean random variables $I_1, \ldots, I_n$, then it holds that*

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

Proof. The proof is by induction on n, the number of cause variables. The essence of the proof is that a probability distribution $P_{I_k}(b_l(j_l, I_k) = j_p \mid C_k)$ is always equal to $c_1 P_{I_k}(I_k = 1 \mid C_k) + c_2 P_{I_k}(I_k = 0 \mid C_k)$, with $c_1, c_2 \in \{0, 1\}$, i.e., the Boolean constraint function b_l determines whether or not $P_{I_k}(I_k = 1 \mid C_k)$, or $P_{I_k}(I_k = 0 \mid C_k)$, or both, or none of these are entered into the equation. This is exactly what $b(I_1, \ldots, I_n)$ does in the formula for causal independence; in fact, it holds that $c_1 = b(I_1, \ldots, i_k, \ldots, I_n)$ for certain values of $\{I_1, \ldots, I_n\} \backslash \{I_k\}$ and likewise for c_2. □

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. As for discrete probability distributions, convolution can be used recursively. Thus, in the context of the theory of causal independence, we use convolution to compute the conditional probability density function $f_b(e \mid \mathbf{C})$, in a way very similar to the discrete case, where b is the causal interaction function.

4.3. Decomposition with Continuous Variables

Moving to the continuous case, consider the Boolean-valued decomposable functions b, i.e., functions $b : \mathbf{I} \to \{0, 1\}$, such that constraints on some variables $\mathbf{I}' \subseteq \mathbf{I}$ imposed by b are measurable sets of values for $\mathbf{I}'$. We now wish to use the theory of causal independence to decompose the probability mass of e given $\mathbf{C}$, i.e., $f_b(e \mid \mathbf{C})$.

First observe that Theorem 3 also holds for the continuous case. The representation of e is thus fully determined by $P(b(I_1, \ldots, I_n) = e \mid \mathbf{C})$. The decomposition of this distribution now follows a similar pattern as the discrete case.

THEOREM 4. *Let f be a joint probability density function of independent random, continuous intermediate variables J and K and the related continuous variables $\mathbf{C} = \{C_J, C_K\}$, and let b be a Boolean function, then it holds that*

$$P(b(J, K) = e \mid \mathbf{C}) = \int_{-\infty}^{\infty} f_J(j \,|\, C_J) P(b(j, K) = e \,|\, C_K)\,\mathrm{d}j$$

Proof.

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

Note that 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

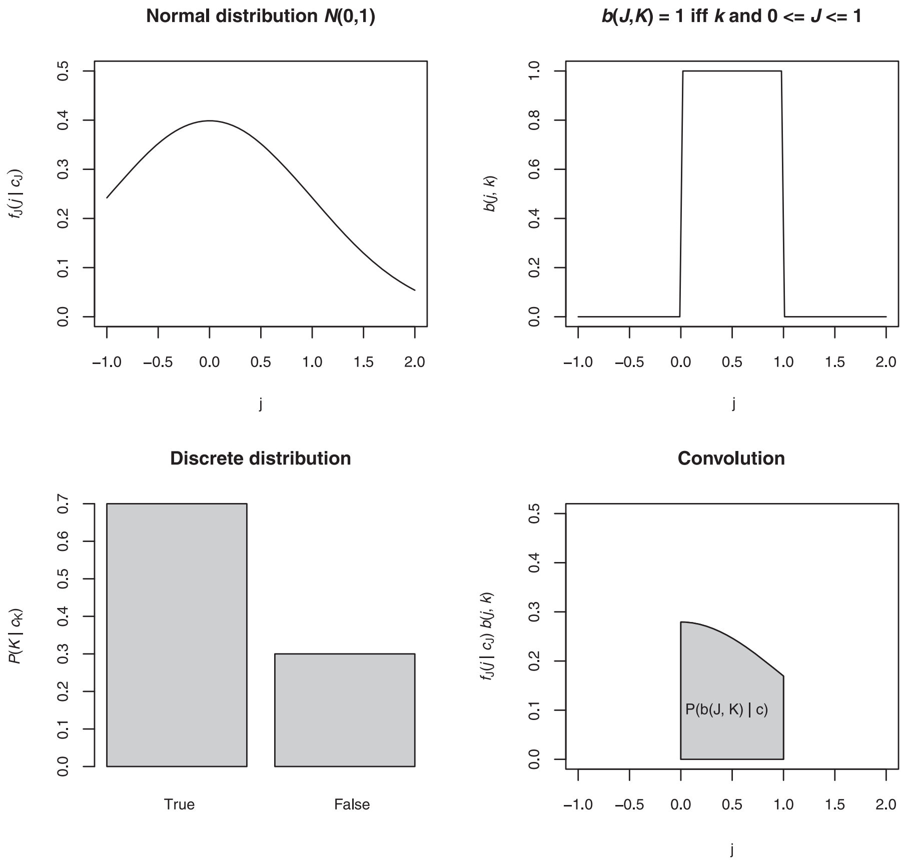

Figure 3. Illustration of a generalized convolution with continuous and discrete variables.

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

If b is defined on both discrete and continuous variables, then this yields a mix of sums and integrals by repeated application of Theorems 2 and 4.

To illustrate this, consider Figure 3, where a continuous variable J is combined with a binary variable K using a Boolean function b such that

$$
b(J, K) = \begin{cases} 0 \leq J \leq 1 & \text{if } k \ (K = 1) \\ 0 & \text{otherwise} \end{cases}
$$

In this case, by applying Theorem 4, we obtain $P(b(J, K) = e \mid \mathbf{c}) = \int_0^1 f_J(j \mid c_J)P(k \mid c_K)\,\mathrm{d}j = P(0 \leq J \leq 1 \mid c_J) \cdot P(k \mid c_K)$, as illustrated in Figure 3.

5. A RELATIONAL PROBABILISTIC LANGUAGE

We consider various operators for continuous variables, which will build up a rich language for modeling causal independence in terms of relations between continuous variables.

5.1. Boolean-Valued Continuous Operators

Analogously to the convolution notation (6), we define an operator $\textcircled{b}$ for denoting this decomposition for any Boolean function such that

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

where the superscripts $\mathbf{C}$ and $C_1, \ldots, C_n$ represent conditioning of the mass or density functions on the corresponding superscript 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:

$$\left(f_J^{C_J} \textcircled{b}\, f_K^{C_k}\right)(e) = f_{b(J,K)}^{C}(e) = f_{b(J,K)}(e \mid \mathbf{C})$$

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 $L = \mathit{true}$) of Example 1, we have

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

which is again the noisy-OR operator for discrete random variables.

In the following section, a language that supports Boolean combinations of relations is developed and studied. This language will be built up using well-known algebraic operators. Their algebraic properties in fact carry over to the convolution-based decomposition as shown by the following proposition.

PROPOSITION 1. *Given a set of interaction variables* $\mathbf{I}$ *with* $\{I_1, \ldots, I_n\} \subseteq \mathbf{I}$ *and* $\{I'_1, \ldots, I'_m\} \subseteq \mathbf{I}$*, and its associated causal variables* C_j *for each* I_j *and* C'_k *for each* I'_k*,* $1 \leq j \leq n, 1 \leq k \leq m$*, if* $b(I_1, \ldots, I_n)$ *is a Boolean expression that is equivalent to the Boolean expression* $b'(I'_1, \ldots, I'_m)$*, then*

$$\textcircled{b}\left(f_{I_1}^{C_1}, \ldots, f_{I_n}^{C_n}\right) = \textcircled{b'}\left(f_{I'_1}^{C'_1}, \ldots, f_{I'_m}^{C'_m}\right)$$

Proof. Take a value for the effect variable E. By definition, $\textcircled{\scriptsize b}\,(f_{I_1}^{C_1}, \ldots, f_{I_n}^{C_n})(e) = f_b(e \mid \mathbf{C})$. By Theorem 3, this is equal to $P(b(I_1, \ldots, I_n) = e \mid \mathbf{C})$. Owing to the equivalence of the Boolean expressions, we have $P(b'(I_1', \ldots, I_m') = e \mid \mathbf{C})$, which again implies $f_{b'}(e \mid \mathbf{C})$. □

While it is a trivial proof, the property is very useful in practice as it allows for algebraic manipulations of the convolution operator based on the properties of the Boolean function. For example, since the $\vee$ operator is commutative, i.e., it holds that $I_1 \vee I_2$ iff $I_2 \vee I_1$ for all I_1 and I_2, the proposition states that $f_{I_1}^{C_1} \textcircled{\scriptsize$\vee$}\, f_{I_2}^{C_2} = f_{I_2}^{C_2} \textcircled{\scriptsize$\vee$}\, f_{I_1}^{C_1}$. Also for continuous Boolean operators, this can be used, e.g., it directly follows that $\textcircled{\scriptsize$\geq 0$}\, f_J^{C_J} = \textcircled{\scriptsize$\not< 0$}\, f_J^{C_J}$. We study some of these operators in more detail next, in particular when we use this to build causal independence models.

5.2. 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 relational operators*, such as $=, <, \leq, >, \ldots$, to build up our relational language are studied. Consider $\leq$:

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

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

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

A similar expression can be derived for $>$, whereas $P((I_1 = I_2) = e \mid \mathbf{C}) = 0$ 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 $\textcircled{\scriptsize R}$:

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

where R is one of the basic relational operators.

Subsequently, we look at the extension of this language with convolutions of density functions of the interaction between variables and constants. For convenience, we ignore in the subsequent derivations that we are dealing with *conditional* density functions. A constant k can be described by a uniform probability distribution with a 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

$$P((I \leq J) = e) = (f_I \textcircled{\scriptsize$\leq$} f_k)(e) = \int_{-\infty}^{k} f_I(i)\,\mathrm{d}i = P(I \leq k)$$

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 modeling the probabilistic interaction between variables, let $\mathbf{I_c}$ be a set of continuous random variables and $\mathbf{K}$ be a set of constants. Then, a *linear-equations relation* is a Boolean-valued function b such that

$$b(\mathbf{I_c}) = R\left(\sum_{i=1}^{n} c_i X_i, \sum_{j=1}^{m} d_j Y_j\right)$$

such that $\mathbf{X} = \{X_1, \ldots, X_n\}$, $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$, $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{I_c} \cup \mathbf{K}$, c_i and d_j are real numbers, and R is a relational operator.

If the sets $\mathbf{X}$ and $\mathbf{Y}$ are disjoint with respect to variables in $\mathbf{I_c}$, whereas overlap for constants is allowed, the sums of $\mathbf{X}$ and $\mathbf{Y}$ are independent. In that case, the relation can be decomposed using Equation 9, yielding the following proposition.

PROPOSITION 2. *The causal-independence model of a linear-equations relation*

$$R\left(\sum_{i=1}^{n} c_i X_i, \sum_{j=1}^{m} d_j Y_j\right)$$

with continuous interaction variables $\mathbf{I_c}$ *can be written as*

$$\begin{aligned} P(b(\mathbf{I_c}) = e) &= P\left(R\left(\sum_{i=1}^{n} c_i X_i, \sum_{j=1}^{m} d_j Y_j\right) = e\right) \\ &= (f_{c_1 X_1 + \cdots + c_n X_n} \textcircled{\scriptsize$R$} f_{d_1 Y_1 + \cdots + d_m Y_m})(e) \end{aligned}$$

if $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{I_c}$ *and* $\mathbf{X} \cap \mathbf{Y} = \varnothing$.

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

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

where $*$ is the (sum) convolution operator.

5.3. Boolean Combinations of Relations

Linear-equations relations can now be combined in a uniform manner using Boolean functions, which allows us to complete the relational language. Let again $\mathbf{I_c}$ be a set of continuous causal random variables, $\mathbf{I_d}$ a set of discrete causal random variables, and $\mathbf{I} = \mathbf{I_c} \cup \mathbf{I_d}$. A *Boolean combination bc* is a Boolean-valued function defined on $\mathbf{I}$ as follows:

$$bc(\mathbf{I}) = b(R_1(\mathbf{X}^1, \mathbf{Y}^1), \ldots, R_r(\mathbf{X}^r, \mathbf{Y}^r), \mathbf{I_d})$$

where b is a Boolean function if all variables in $\mathbf{I_d}$ are Boolean (binary), and otherwise a Boolean-valued function, and $\{R_1, \ldots, R_r\}$ a set of r linear-equations relations given $\mathbf{I_c}$ and some set of constants $\mathbf{K}$.

If the set of continuous variables is partitioned into disjoint sets, then we have ensured that each of the relations in the Boolean combination is independent of each other. To see that independence is not generally the case for nondisjoint sets, consider, for example, the relations $I_1 \leq I_2$ and $I_2 \leq I_3$ and assume that I_2 is normally distributed with mean 1 and arbitrary variance, Furthermore, assume that $I_1 = 1$ and $I_3 = 0$, i.e., I_1 and I_3 are constant random variables. Then $P(I_1 \leq I_2) = P(I_2 \geq 1) = 1/2 \neq 0 = P(I_2 \geq 1 \mid I_2 \leq 0) = P(I_1 \leq I_2 \mid I_2 \leq I_3)$. It follows that the relation $I_1 \leq I_2$ is indeed dependent of the relation $I_2 \leq I_3$.

If the continuous variables in the Boolean combinations of relations are partitioned, Equation 8 can be applied to obtain the following proposition.

PROPOSITION 3. *The causal-independence model of a Boolean combination of linear-equations relations* $b(R_1(\mathbf{X}^1, \mathbf{Y}^1), \ldots, R_r(\mathbf{X}^r, \mathbf{Y}^r), \mathbf{I_d})$*, where* $\mathbf{X}^i$ *and* $\mathbf{Y}^j$ *are sets of continuous variables and constants, and* $\mathbf{I_d} = \{I_{r+1}, \ldots, I_s\}$ *a set of discrete variables, can be written as*

$$\begin{aligned}&P(b(R_1(\mathbf{X}^1, \mathbf{Y}^1), \ldots, R_r(\mathbf{X}^r, \mathbf{Y}^r), \mathbf{I_d}) = e \mid \mathbf{C})\\ &\qquad = \textcircled{\scriptsize$b$} \left(f^{C_1}_{R_1(\mathbf{X}^1,\mathbf{Y}^1)}, \ldots, f^{C_r}_{R_r(\mathbf{X}^r,\mathbf{Y}^r)}, f^{C_{r+1}}_{I_{r+1}}, \ldots, f^{C_s}_{I_s}\right)(e)\end{aligned}$$

if all pairs of $\mathbf{X}^i$ *and* $\mathbf{Y}^j$ *are mutually disjoint.*

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.

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

6. SPECIAL PROBABILITY DISTRIBUTIONS

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

6.1. 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 modeled by $\le$, then the effect variable E also follows a Bernoulli distribution with parameter $p_1 - p_1 p_2 + 1$, since

$$
\begin{aligned}
P_{\le}(e \mid \mathbf{C}) &= (f^{C_1}_{I_1} \textcircled{$\le$}\ f^{C_1}_{I_2})(e)\\
&= \sum_{i_1} f_{I_1}(i_1 \mid C_1) P((i_1 \le I_2) = e \mid C_2)\\
&= (1 - p_1)(1 - p_2) + (1 - p_1)p_2 + p_1 p_2\\
&= p_1 - p_1 p_2 + 1
\end{aligned}
$$

Of course, the parameters may depend on whether C_k is true or false.

6.2. Exponential Distribution

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 $I_1 \le I_2$, meaning that intermediate effect I_1 does not occur later than I_2. Suppose we model the delay between C_1 and C_2 by δ. Then, for example, if all C_i are true, then the density of effect E is as follows:

$$
f_{I_1+\delta=I_2}(e \mid \mathbf{C}) = (f^{c_1}_{I_1+\delta} \textcircled{$=$}\ f^{c_2}_{I_2})(e)
$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} f_{I_1}(i_1 \mid c_1) f_{I_2}((i_1 + \delta = I_2) = e \mid c_2)\,\mathrm{d}i_1 \\
&= \int_{-\infty}^{\infty} \lambda_1 e^{-\lambda_1 i_1} \lambda_2 e^{-\lambda_2(i_1+\delta)}\,\mathrm{d}i_1 \\
&= \int_{-\infty}^{\infty} \lambda_1 \lambda_2 e^{-(\lambda_1+\lambda_2) i_1}\, e^{-\lambda_2 \delta}\,\mathrm{d}i_1 \\
&= \left[-\frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2}\, e^{-(\lambda_1+\lambda_2) i_1} \right]_0^{\infty} e^{-\lambda_2 \delta} \\
&= \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2}\, e^{-\lambda_2 \delta}
\end{aligned}
$$

where $\delta \geq 0$. For $\delta = 0$ only the rate parameters λ_1 and λ_2 affect the probability density $f_{I_1+\delta=I_2}$. If also $\lambda_1 = \lambda_2$, then $f_{I_1+\delta=I_2}(e \mid \mathbf{C}) = \lambda/2$. The probability mass of $I_1 \leq I_2$ is obtained by integrating out possible delays δ between I_1 and I_2:

$$
\begin{aligned}
P_{I_1 \leq I_2}(e \mid \mathbf{C}) &= \int_0^{\infty} f_{I_1+\delta=I_2}(e \mid \mathbf{C})\,\mathrm{d}\delta \\
&= \left[-\frac{\lambda_1}{\lambda_1 + \lambda_2}\, e^{-\lambda_2 \delta} \right]_0^{\infty} = \frac{\lambda_1}{\lambda_1 + \lambda_2}
\end{aligned}
$$

It follows that if $\lambda_1 = \lambda_2$, then $f_{I_1 \leq I_2}(e \mid \mathbf{C}) = 1/2$.

6.3. Geometric Distribution

A geometric distribution is the discrete analogue of the exponential distribution and models the number of Bernoulli trials needed to get one success where the success probability of the trial is given by a parameter p. The probability mass function of this distribution is $P(k) = (1 - p)^k p$, with $k \geq 0$. Let I_1 be a geometric random variable with parameter p_1 and I_2 is a geometric random variable with parameter p_2. First note that $\sum_k (1 - p)^k p = 1$ (since it is a distribution) implies that $\sum_k (1 - p)^k = \frac{1}{p}$. Similarly, for the parameter $1 - (1 - p_1)(1 - p_2)$ it holds $\sum_k ((1 - p_1)(1 - p_2))^k = \frac{1}{1-(1-p_1)(1-p_2)}$. Then we have the following derivation:

$$
\begin{aligned}
f_{I_1 \leq I_2}(e \mid \mathbf{C}) &= \sum_{i_1} (1 - p_1)^{i_1} p_1 P((i_1 \leq I_2) = e \mid c_2) \\
&= \sum_{i_1} (1 - p_1)^{i_1} p_1 \sum_{k=i_1}^{\infty} P(I_2 = k) = \sum_{i_1} (1 - p_1)^{i_1} p_1 \sum_{k=i_1}^{\infty} (1 - p_2)^k p_2 \\
&= \sum_{i_1} (1 - p_1)^{i_1} p_1 (1 - p_2)^{i_1} \left(\sum_k (1 - p_2)^k \right) p_2
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i_1} (1-p_1)^{i_1} p_1 (1-p_2)^{i_1} \frac{1}{p_2} p_2 = \sum_{i_1} (1-p_1)^{i_1} p_1 (1-p_2)^{i_1} \\
&= p_1 \sum_{i_1} ((1-p_1)(1-p_2))^{i_1} = \frac{p_1}{1-(1-p_1)(1-p_2)}
\end{aligned}
$$

The probability that less trials are needed to get a success for I_1 compared to I_2 is thus the probability that the Bernoulli trials of I_1 succeed relative to the probability that at least one of the two types of trials succeed.

6.4. Conditional Gaussian Distribution

The most common hybrid distribution for Bayesian networks is the conditional Gaussian distribution.[24] We illustrate the theory for the case when a continuous interaction variable J has a continuous cause variable C_J. The distribution of J is given in this model by $f(j \mid C_J) = N(\alpha + \beta C_J, \sigma^2)$. 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 $f_{I_1 - I_2}(e \mid \mathbf{C})$ 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.

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

$$
\begin{aligned}
P_{\leq}(e \mid \mathbf{C}) &= P_{I_1 \leq I_2}((I_1 \leq I_2) = e \mid \mathbf{C}) \\
&= f_{I_1}^{C_1} \mathbin{\textcircled{$\le$}} f_{I_2}^{C_2} \\
&= \left(f_{I_1}^{C_1} \ominus f_{I_2}^{C_2} \right) \mathbin{\textcircled{$\le$}} 0 = F_J(0) \\
&= \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-(\alpha_1 + \beta_1 C_1 - \alpha_2 - \beta_2 C_2)}{\sqrt{2(\sigma_1^2 + \sigma_2^2)}} \right) \right] \\
&= P(\Theta \leq w_0 + w_1 C_1 + w_2 C_2)
\end{aligned}
$$

where $w_0 = \frac{\alpha_2 - \alpha_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$, $w_1 = \frac{-\beta_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$, $w_2 = \frac{\beta_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$, and $\Theta \sim N(0, 1)$, which is a probit 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, $(f_H^B * f_W^Y)(a)$ 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

$$
P(a) = (f_I^C \mathbin{\textcircled{$\le$}} (f_H^B * f_W^Y))(a)
$$

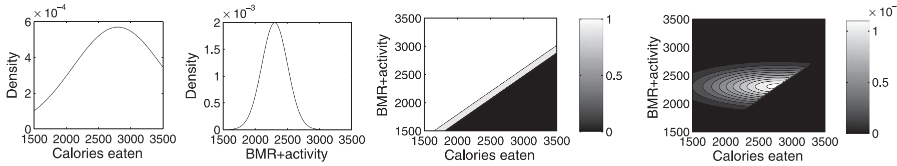

Figure 4. Example distributions, where, from left to right, the first figure shows the density of $C \sim N(2800, 700)$; the second figure shows the density of $B + Y \sim N(2300, 200)$; the third figure shows the probability distributions $P(A \mid C, B + Y)$ with $A \equiv I \leq (H + W)$ where $I \sim N(0.9 \cdot C, 200)$ and $H + W \sim N(1.1 \cdot (B + Y), 300)$; finally, the figure on the right shows the joint density of $\{A, C, B + Y\}$.

which is a probit 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 4 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.

7. RELATED WORK

In this section, we discuss related work with respect to the types of models that can be handled by the relational language introduced in this paper. Furthermore, it is discussed how the work presented in this paper relates to previous research on causal independence models. Although our paper puts emphasis on theory, the last part of this section also indicates how the algebraic framework presented in this paper can be seen as a generalization of an existing method for improving the efficiency of probabilistic inference in causal independence models.

7.1. Mixtures of Truncated Functions

Mixtures of truncated exponentials, MTEs for short, were originally proposed to approximate continuous distributions.[25] They are closed on marginalization and conditioning, and thus efficient algorithms to reason with MTEs exist.[26] In addition, methods for parameter learning of MTEs have been developed.[27,28] In principle, the method is powerful enough for hybrid Bayesian networks, containing both discrete and continuous variables at the same time.[29] Recently, the idea to approximate probability distributions by mixtures of truncated exponential has been extended toward mixtures of polynomials,[30] and finally the method has been generalized to mixtures of truncated basis function, MTBF for short, that unifies all the different approaches.[31]

Where the algebraic framework presented in this paper is primarily meant for modeling conditional probability distributions in terms of relationship, the theory of mixtures of truncated functions is primarily meant to approximate probability distributions. Thus, both foundation and purpose of the two approaches are different. Understandably, a relational algebra to model relationships between random variables is not an integral part of the MTBF method.

7.2. Stress-Strength Models

In statistics, the probability distribution $P(X < Y)$ has been studied since the 1970s under the name of *stress-strength model* and it has attracted considerable attention;[32] it is an important special case of the framework presented in this paper. The name stress-strength model comes from engineering, where the assessment of the reliability of a component is described in terms of some "stress" X experienced by the component and Y representing the "strength" of the component to overcome that stress.[32] For example, if X represents the maximum chamber pressure by ignition in a rocket engine and Y the strength of the rocket chamber, then $P(X < Y)$ represents the probability of successful firing of the engine. Also, for medical applications this model has received considerable interest. In particular, it is the basis for the nonparametric Wilcoxon–Mann–Whitney test, which is very often used in medicine.[33] For example, when comparing two treatments, X and Y could represent remission times for each treatment. For the medical decision maker, it is then of interest to know the probability $P(X < Y)$, although in this case the name "stress-strength" is, of course, not appropriate.

Much of the work in statistics has focused on the construction of efficient and reliable estimators of parameters of $R = P(X < Y)$ based on different assumptions on the distributions of X and Y, whereas in the present paper the focus is on a more general modeling method for Bayesian networks. Typically, X and Y are chosen from the same family of distributions and independence of X and Y is assumed, similar to what is done in this paper. For a general overview of estimation methods, we refer to a review paper in *The Handbook of Statistics.*[34]

7.3. Related Causal-Independence Models

Modeling causal independence has a relatively long tradition. In early work on causality, Good[35] showed that under certain assumptions, causal influences to a variable combine in a way that is now called the noisy-OR. More than two decades later, the noisy-OR and noisy-AND were defined by Pearl in the context of Bayesian network as a canonical method to express interactions between causal influences.[36] At exactly the same time, Peng and Reggia discovered the same principles independently in the context of probabilistic abductive reasoning, extensively described in their book.[37] Pearl also showed how to exploit this structure to speed up probabilistic inference in singly connected networks. Later, this was extended for bipartite graphs with a noisy-OR interaction.[38,39] The noisy-OR gate was then generalized in several ways[10,15,40] and was also used to speed up inference in general Bayesian networks.[15,41,42]

General causal independence was first described by Heckerman and colleagues.[43,44] This early work discusses the noisy-OR, noisy-MAX, and a noisy-ADD for discrete variables as special cases. The last paper also considers a continuous version of causal independence, namely the linear Gaussian model. General properties of causal independence, in particular decomposable causal independence, are also introduced; decomposability has acted as a foundation for further generalization. Algebraic properties of general causal independence were later further studied by Lucas[13] and Van Gerven et al.[45] Since then, other specific causal-independence models have been studied, for example, causal independence combined with temporal information.[16] Furthermore, there are a number of approaches specifically designed to model the *undermining* of different causes to an effect. In many of the standard models, such as the noisy-AND and noisy-OR, this undermining cannot be modeled as in these models the causes are collectively as effective in causing the effect as some by acting by themselves. In general causal independence, however, this can be modeled by a Boolean expression that incorporates negation. The recursive noisy-OR model[46] is an approach to represent positive and negative influences, but these cannot be combined within the same model. A more general approach related to this work is the nonimpeding noisy-AND tree (NIN-AND),[47] which can be seen as a noisy-AND with negations. A similar approach is by Maaskant and Druzdzel,[48] where gates are modeled by a conjunctive normal form.

For general Bayesian networks, there have been two approaches to exploit causal independence for speeding up inference by changing the network structure, namely the *parent-divorcing method*[49] and the *temporal transformation method.*[43] Other approaches use the insight that efficient probabilistic inference is made possible by working with a factorization of the joint probability distribution, rather than working with the joint probability distribution itself. As causal independence models allow decomposition of the probability distribution beyond the factorization implied by the conditional independences derived from the associated graph, this insight can be exploited in algorithms that work with these factorizations directly such as symbolic probabilistic inference[39] and variable elimination.[50]

Right from the beginning, causal-independence models were not only used to improve probabilistic inference (by approximating the actual model) but also to facilitate the manual construction of Bayesian networks, as the number of parameters in, for example, noisy-OR models that has to be estimated is proportional rather than exponential in the number of parents.[10,49,51] This is the main reason why causal-independence models are considered as important canonical models for knowledge engineering. For a comprehensive analysis and overview of applying such models in practice, we refer to the review paper by Díez and Druzdzel.[14]

7.4. Exploiting Algebraic Properties for Efficient Probabilistic Inference

Zhang and Poole[50] introduced a combination function $\otimes$ that can be used to further factorize a discrete joint probability distribution when there is causal independence with Boolean operators which are associative and commutative (Lucas[13] analyzes such algebraic properties in his work). Zhang and Poole show that such algebraic properties can be successfully employed to speed up variable elimination in exact inference procedures. This combination function is in some sense a special case of the $\textcircled{\scriptsize b}$ operator introduced in the present paper, which we will show next.

Consider two functions g_1 and g_2, which both represent conditional probability distributions. Let $E_1, \ldots, E_n$ be effect variables, called *convergent variables* by Zhang and Poole,[50] that appear in both g_1 and g_2, and let $\mathbf{A}$ be the set of nonconvergent variables appearing in both g_1 and g_2, $\mathbf{B}$ the set of (all) variables that appears only in g_1, and $\mathbf{C}$ be the set of (all) variables that appears only in g_2. Then, Zhang and Poole define

$$\begin{aligned}
&(g_1 \otimes g_2)(e_1, \ldots, e_n, \mathbf{A}, \mathbf{B}, \mathbf{C}) \\
&\quad = \sum_{b_1(i_{11}, i_{12}) = e_1} \cdots \sum_{b_n(i_{n1}, i_{n2}) = e_n} g_1(E_1 = i_{11}, \ldots, E_n = i_{n1}, \mathbf{A}, \mathbf{B}) \\
&\qquad \times g_2(E_1 = i_{12}, \ldots, E_n = i_{n2}, \mathbf{A}, \mathbf{C})
\end{aligned}$$

The intuition behind this definition is that the intermediate variables $\mathbf{I}$ have the same domain as the effect variables $\mathbf{E}$ and the function $g_i(e, C_i) = P(I_i = e \mid C_i)$, with $i \in \{1, 2\}$. In general, it then holds for causal independence models that

$$P(e \mid C_1, \ldots, C_n) = \bigotimes_{i=1}^{n} g_i(e, C_i)$$

When $n = 2$,

$$P(e \mid C_1, C_2) = g_1 \otimes g_2(e, C_1, C_2) = \sum_{b(i_1, i_2) = e} g_1(i_1, C_1) g_2(i_2, C_2)$$

In the type of Bayesian network considered in this paper, the density $f_{I_1}^{C_1}$ corresponds to the function $g_1(I_1, C_1)$ and the density $f_{I_2}^{C_2}$ corresponds to the function $g_2(I_2, C_2)$. We thus obtain the following equality:

$$\begin{aligned}
g_1 \otimes g_2(e, C_1, C_2) &= \sum_{b(i_1, i_2) = e} g_1(i_1, C_1) g_2(i_2, C_2) \\
&= \sum_{i_1} f_{I_1}^{C_1}(i_1) P(b(i_1, I_2) = e \mid C_2) = (f_{I_1}^{C_1} \circledcirc_{b} f_{I_2}^{C_2})(e)
\end{aligned}$$

thereby showing the equivalence of the two operators for this specific (discrete) case. This result easily generalizes to n causal variables. Properties, such as the commutativity and associativity of the $\otimes$ operator, as proved by Zhang and Poole,[50] follow directly from the more general Proposition 1.

8. CONCLUSIONS

We presented a new algebraic framework for causal-independence modeling of Bayesian networks that goes beyond what has been available so far. In contrast to other approaches, the framework supports the modeling of discrete as well as of continuous variables, either separately or mixed, and there are no restrictive requirements with respect to algebraic properties such as associativity or commutativity. The algebraic properties that hold simply follow from the operators that the Bayesian network developer wishes to employ.

The design of the framework was inspired by the convolution theorem of probability theory, and it was shown that this theorem easily generalizes to convolution with Boolean-valued functions. We also studied a number of important modeling operators. Contrary to the commonly used 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 that might be useful when eventually building Bayesian network models for actual problems.

For inference in networks containing these causal-independence structures, one can resort to the use of approximate inference, such as discretization (e.g., using a dynamic discretization approach)[52] or sampling (e.g., using BUGS[53]). For normal distributions with the class of continuous operators as introduced in this paper, a variational approximation seems feasible, as the distribution containing relational operators is of a sigmoid shape, i.e., the probit model. Murphy[22] observed that the product of a sigmoid and a Gaussian can be approximated well by a Gaussian distribution, so the joint factorization of the probability distribution could be approximated in this way. Ideas explored by Lerner et al.[54] could also be applied to obtain an efficient variant of the join tree algorithm.

There are several other directions to extend this work. Besides approximating the distribution for the general case as discussed above, in some cases standard methods for solving the inference problem can be used, such as the probit model for the conditional Gaussian distribution. Furthermore, the algebraic manipulations could be used as a preprocessing step to simplify the inference or to further factorize the joint probability.[50] Another question is how to estimate the probabilities of the models that have been discussed in this paper. For certain models, this has been explored earlier in the context of causal independence[14] or in context of stress-strength models,[34] but a general approach is lacking. In particular, when considering the model based on a relational operator, the stress-strength literature has focused on estimating the marginal of the effect variable E based on partial information about the interacting variables, whereas in a Bayesian network approach we are typically interested in the underlying probabilities of the interacting variables. Finally, it should be noted that many bounds can be given on the probability distribution of the effect variable. These include confidence bounds when estimating the effect variable, but also bounds when approximating the marginal probability in different ways. In future research, we will address some of these issues further.

References

1. Friedman N. Inferring cellular networks using probabilistic graphical models. Science 2004;303(5659):799–805.

2. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003;302(5644):449–453.
3. Leibovici L, Paul M, Nielsen AD, Tacconelli E, Andreassen S. The TREAT project: decision support and prediction using causal probabilistic networks. Int J Antimicrob Agents 2007;30.
4. Velikova M, Samulski M, Lucas PJF, Karssemeijer N. Improved mammographic CAD performance using multi-view information: a Bayesian network framework. Phys Med Biol 2009;54:1131–1147.
5. Visscher S, Kruisheer E, Schurink C, Lucas PJF, Bonten M. Predicting pathogens causing ventilator-associated pneumonia using a Bayesian network model. J Antimicrob Chemother 2008;62:184–188.
6. Holický M. Probabilistic risk optimization of road tunnels. Struct Saf 2009;31(3):260–266.
7. Hommersom A, Lucas PJF. Using Bayesian networks in an industrial setting: Making printing systems adaptive. In: Coelho H, Struder R, Wooldridge M, editors, Proceedings of ECAI-2010. Amsterdam; IOS Press; 2010. pp 401–406.
8. Koller D, Friedman N. Probabilistic graphical models: Principles and techniques. Cambridge, MA: MIT Press; 2009.
9. Pearl J. Causality: Models, reasoning and inference. Cambridge, MA: MIT Press; 2000.
10. Henrion M. Some practical issues in constructing belief networks. In: Lemmer JF, Kanal LN., editors. Uncertainty in artificial intelligence. Amsterdam: Elsevier; 1989. p. 161–173.
11. Pearl J. Probabilistic reasoning in intelligent systems: Networks of plausible inference. Palo Alto, CA: Morgan Kaufmann; 1988.
12. Heckerman D, Breese JS. Causal independence for probabilistic assessment and inference using Bayesian networks. IEEE Trans Syst Man Cybern 1996;26(6):826–831.
13. Lucas PJF. Bayesian network modelling through qualitative patterns. Artif Intell 2005;163:233–263.
14. Díez FJ, Druzdzel MJ. Canonical probabilistic models for knowledge engineering. Technical Report CISIAD-06-01, UNED, Madrid, Spain; 2006.
15. Díez FJ. Parameter adjustment in Bayes networks: the generalized noisy OR-gate. In: Proc Ninth Annual Conf on Uncertainty in Artificial Intelligence, (UAI'93), Washington, DC; July 9-11, 1993; pp. 99–105.
16. Galán SF, Díez FJ. Modeling dynamic causal interactions with Bayesian networks: temporal noisy gates. In: 2nd Int Workshop on Causal Networks (CaNew2000), Berlin, August 20; 2000. pp 1–5.
17. Bishop CM. Pattern recognition and machine learning. Berlin: Springer; 2006.
18. Hommersom A, Lucas PJF. Generalising the interaction rules in probabilistic logic. In Proceedings IJCAI-2011. AAAI, 445 Burgess Drive, Suite 100 Menlo Park CA 94025, USA; 2011. pp 912–917.
19. Grimmett G, Stirzaker D. Probability and random processes. Oxford, UK: Oxford University Press; 2001.
20. Birkhoff G, Mac Lane S. A survey of modern algebra, 4th ed. New York: Macmillan; 1977.
21. Wegener I. The complexity of Boolean functions. New York: Wiley; 1987.
22. Murphy KP. A variational approximation for Bayesian networks with discrete and continuous latent variables. In: 12th Conf on Uncertainty in Artificial Intelligence (UAI'96), Portland, OR; July 31–August 4, 1999. pp 457–466.
23. Williamson RC. Probabilistic arithmetic. PhD thesis, University of Queensland, St. Lucia, Australia; 1989.
24. Lauritzen SL, Wermuth N. Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann Stat 1989;17:31–57.
25. Moral S, Rumí R, Salmerón A. Mixtures of truncated exponentials in hybrid Bayesian networks. In: Proc 6th European Conf on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'01), Toulouse, France. Lecture Notes in Artificial Intelligence, Vol. 2143, Berlin: Springer; 2001. pp 135–143.
26. Cobb BR, Shenoy BR. Inference in hybrid Bayesian networks with mixtures of truncated exponentials. Int J Approx Reason 2006;41(3):257–286.

27. Moral S, Rumí R, Salmerón A. Estimating mixtures of truncated exponentials from data. In: Benferhat S, Besnard Ph, editors. Proc ESCQUARU 2001; LNCS 2143, Springer, Berlin, 2001. pp 156–167.
28. Moral S, Rumí R, Salmerón A. Approximating conditional MTE distributions by means of mixed trees. In: Proc 7th European Conf on Symbolic and Quantitative Approaches to Reasoning with Uncertainty(ECSQARU'03). Lecture Notes in Artificial Intelligence, Vol 2711. Berlin: Springer; 2003. pp 173–183.
29. Cobb BR, Rumi R, Salmerón A. Bayesian network models with discrete and continous variables. In Lucas PJF, Gámez JA, Salmerón A, editors. Advances in probabilistic graphical models, Vol StudFuzz 213. Berlin: Springer; 2007. pp 81–102.
30. Shenoy P, West J. Inference in hybrid Bayesian networks using mixtures of polynomials. Int J Approx Reason 2011;52:614–657.
31. Langseth H, Nielsen TD, Rumí R, Salmerón A. Mixtures of truncated basis functions. Int J Approx Reason 2012;53:212–227.
32. Kotz S, Lumelskii Y, Pensky M. The stress-strength model and its generalizations. Singapore: World Scientific; 2003.
33. Kühnast C, Neuhäuser M. A note on the use of the non-parametric Wilcoxon–Mann–Whitney test in the analysis of medical studies. GMS Ger Med Sci 2008;6:Doc02.
34. Johnson RA. Stress-strength models for reliability. In; Krishnaia PR, Rao CR, editors. Handbook of statistics. Amsterdam: Elsevier Science; 1988. Vol 7, pp 27–54.
35. Good IJ. A causal calculus (i). Br J Phil Sci 1961;XI:305–318.
36. Pearl J. Fusion, propagation and structuring in belief networks. Artif Intell 1986;29(3):241–288.
37. Peng Y, Reggia JA. Abductive inference models for diagnostic problem solving. New York: Springer-Verlag; 1990.
38. D'Ambrosio B. Symbolic probabilistic inference in large BN20 networks. In Proc UAI '94, Seattle, Washington, July 29–31, 1994. pp 128–135.
39. D'Ambrosio B. Local expression languages for probabilistic dependence. Int J Approx Reason 1995;13(1):61–81.
40. Srinivas S. A generalization of the noisy-OR model. In: Proc Ninth Annual Conf on Uncertainty in Artificial Intelligence, (UAI'93), Washington, DC, July 9–11, 1993. pp 208–215.
41. Takikawa M, DAmbriosio B. Multiplicative factorization of the noisy-MAX. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI99). Palo Alto, CA: Morgan Kaufmann, 1999. pp 622–630.
42. Díez FJ, Galán SF. Efficient computation for the noisy MAX. Int J Intell Syst 2003;18(2):165–177.
43. Heckerman D. Causal independence for knowledge acquisition and inference. In: Proc Ninth Annual Conf on Uncertainty in Artificial Intelligence, (UAI'93), Washington, DC, July 9–11, 1993. pp 122–127.
44. Heckerman D, Breese JS. A new look at causal independence. In Tenth Annual Conf on Uncertainty in Artificial Intelligence (UAI'94), Seattle, WA; July 29–31, 1994. pp 286–292.
45. van Gerven M, Lucas PJF, van der Weide ThP. A generic qualitative characterization of causal independence models. Int J Approx Reason 2008;48(1):214–236.
46. Lemmer J, Gossink D. Recursive noisy OR-a rule for estimating complex probabilistic causal interactions. IEEE Trans Syst Man Cybernet B 2004;34(6):2252–2261.
47. Xiang Y, Jia N. Modeling causal reinforcement and undermining for efficient CPT elicitation. IEEE Trans Knowl Data Eng 2007;19(12):1708–1718.
48. Maaskant PP, Druzdzel MJ. An independence of causal interactions model for opposing influences. In: Fourth European Workshop on Probabilistic Graphical Models; Hirtshals, Denmark, September 17–19; 2008. pp 185–192.
49. Olesen KG, Kjærulff U, Jensen F, Jensen FV, Falck B, Andreassen S, Andersen SK. A MUNIN network for the median nerve—a case study on loops. Appl Artif Intell 1989; 3(2-3):385–403.
50. Zhang NL, Poole D. Exploiting causal independence in Bayesian network inference. J Artif Intell Res 1996;5:301–328.

51. Olesen KG, Andreassen S. Specification of models in large expert systems based on causal probabilistic networks. Artifi Intell Med 1993;5(3):269–281.
52. Neil M, Tailor M, Marquez D. Inference in hybrid Bayesian networks using dynamic discretization. Stat Comput 2007;17:219–233, September.
53. Gilks WR, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian modelling. J R Stat Soc Series D (The Statistician) 1994;43(1):169–177.
54. Lerner U, Segal E, Koller D. Exact inference in networks with discrete children of continuous parents. In: 17th Annual Conf on Uncertainty in Artificial Intelligence (UAI), 2001.