Expanding Bayesian networks

Johan Kwisthout[1 [0000−0003−4383−7786]]

Donders Institute for Brain, Cognition, and Behaviour, Radboud University
johan.kwisthout@donders.ru.nl
https://www.socsci.ru.nl/johank

Abstract. Bayesian networks offer an efficient, factorized representation of a joint probability distribution over a set of random variables, and are as such a prominent tool for AI applications that need to represent and reason under uncertainty, such as clinical decision support systems. An obstacle towards practical applicability is the lack of a fundamental and systematic way of extending the network (and thereby the joint probability distribution) with new variables. In this short paper we introduce the concept of inverse marginalisation and propose some preliminary ideas on how to define a novel (extended) joint probability distribution as a linear program.

Keywords: Bayesian networks · model revision · probability distribution

1 Introduction

Bayesian networks are a sound and efficient way to represent a joint probability distribution over a set of stochastic variables. Given a Bayesian network, all posterior (joint) probabilities of interest can be computed in a straightforward manner; the network structure and conditional probabilities allow for human-readable and transparent representation of uncertain knowledge. These are all reasons why Bayesian networks are often the underlying computational structure in clinical decision support systems (CDSSs).

A shortcoming of Bayesian networks, particularly when they are used as a formalisation of a clinical model used by a CDSS, is that they are difficult to maintain. When a Bayesian network has been constructed, trained on data, and/or elicited from domain experts, it is far from trivial to add new knowledge while maintaining the integrity of the network. This is an important shortcoming in contexts that require flexibility when new medical guidelines arrive or new clinical studies become available. While several algorithmic approaches to (e.g.) integrating data sets with dissimilar variables have been proposed in the literature (see Section 1.1) their mathematical basis is still weak.

When it comes to specifically *expanding* a Bayesian network, we can distinguish conceptually different situations [5, 6]:

– introduce an additional value of a stochastic variable in the form of a new value in the domain of the variable (e.g., a new treatment method)

- introduce an additional value of a stochastic variable by splitting an existing value into two (e.g., start to discriminate between two related symptoms that were previously grouped)
- introduce an additional stochastic variable when a new concept needs to be introduced due to changed domain knowledge or purpose of a CDSS (e.g., include Quality of Life in addition to Survival Rate in a model)
- introduce an additional new stochastic variable to refine an existing concept with more details (e.g., include Rhesus Factor in addition to Blood Type)

In Section 4 we will further elaborate on the differences between these situations and the consequence for the possibility to formulate them in a sound and coherent way.

1.1 Related work

In [5], the authors offer, from a cognitive stance, a computational-level description of what can be revised in a generative model (formalized as a Bayesian network). Here generative models (mathematically described as Bayesian networks) are assumed to describe learned stochastic relationships between causes and effects; the authors describe in what sense these networks can be systematically adjusted (to model, for example, the effect of drug rehabilation on one's internal beliefs and assumptions). In contrast, [6] offers algorithmic approaches specific to learning network structure and parameters from distinct data sets, in the context of clinical decision support systems. The latter paper summarizes and extends earlier work where some aspects of model expansion are covered. For example, [7] merges old sufficient statistics with new sufficient statistics to facilitate incremental batch learning when new variables arise. Older work like [2] focuses on merging different sources for learning when later information has only a subset of the variables.

In contrast to these approaches, in the current short note we aim to offer a sound *mathematical* account of the effect of adding a value to the domain of a stochastic variable, or adding a stochastic variable to a joint probability distribution, with the (future) aim to interpret and compare existing and novel algorithms for model expansion within a common mathematical framework.

1.2 Our contribution

The crucial aspect in our approach is the notion of *inverse (partial) marginalisation* to capture expansions of variables and joint distributions. For this, we introduce in Section 3 the notion of a partial marginalisation as extension to the familiar marginal distribution. Expanding a joint distribution can always be formulated as taking the inverse of a marginal distribution. In contrast, adding a value to the domain of a variable can be either *invertible* or *non-invertible*. In the first case this operation can also be characterized systematically as an inverse partial marginalisation; in the latter case this is in principle not possible. Out approach is further explicated in Section 4.

In the remainder of this short paper, we introduce formal preliminaries and notation in Section 2. We introduce partial marginalisation in Section 3; formalising network extension as inverse marginalisation is introduced in Section 4. In Section 5 we discuss how to formalise constraints on the novel joint distribution as a linear program. We conclude in Section 6.

2 Preliminaries and notation

A joint probability distribution Pr over a set of discrete random variables $\mathbf{V}$ is a function $\Pr : \xi \to [0, 1]$ on a Boolean algebra ξ of propositions spanned by $\mathbf{V}$ for which the following conditions hold [1]:

- $0 \leq \Pr(a) \leq 1$, for all $a \in \xi$;
- $\Pr(\top) = 1$;
- $\Pr(\bot) = 0$;
- for all $a, b \in \xi$, if $a \wedge b \equiv \bot$ then $\Pr(a \vee b) = \Pr(a) + \Pr(b)$.

Let $\mathbf{V}$ and $\mathbf{W}$ denote two joint probability distributions such that the variables constituting $\mathbf{W}$ form a subset of the variables constituting $\mathbf{V}$ and $\Pr(\mathbf{W} = \mathbf{w}) = \sum_{\mathbf{x} \in \Omega(\mathbf{V} \setminus \mathbf{W})} \Pr(\mathbf{V} = \mathbf{w} \wedge \mathbf{x})$. Then we will call $\Pr(\mathbf{W})$ a sub-distribution of $\Pr(\mathbf{V})$ and $\Pr(\mathbf{V})$ a super-distribution of $\Pr(\mathbf{W})$.

A (discrete) Bayesian network $\mathcal{B}$ is a graphical structure that efficiently factorizes $\Pr(\mathbf{V})$ and graphically depicts the conditional independences within $\Pr(\mathbf{V})$ [4]. $\mathcal{B}$ includes a directed acyclic graph $\mathbf{G}_{\mathcal{B}} = (\mathbf{V}, \mathbf{A})$, modeling the variables and conditional independences in the network, and a set of parameter probabilities Pr in the form of conditional probability tables (CPTs), capturing the strengths of the relationships between the variables. The network thus factorizes a joint probability distribution $\Pr(\mathbf{V}) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$ over its variables, where $\pi(V_i)$ denotes the parents of V_i in $\mathbf{G}_{\mathcal{B}}$.

Our notational convention is to use upper case letters to denote individual nodes in the network, upper case bold letters to denote sets of nodes, lower case letters to denote value assignments to nodes, and lower case bold letters to denote joint value assignments to sets of nodes. The set of values v_i that constitute the domain of a variable V is denoted as $\Omega(V)$; this notation is extended to the set of joint value assignments $\Omega(\mathbf{V})$.

3 Partial marginalisation

A marginalisation operation maps a super-distribution $\mathbf{V}$ to a sub-distribution $\mathbf{W}$ by marginalizing out the variables in $\mathbf{V} \setminus \mathbf{W}$ as per the definition above. We define a partial marginalisation as a generalisation of this operation as follows.

Definition 1 (Partial marginalisation). *Let V be a discrete stochastic variable with $\Omega(V) = \{v_1, \ldots, v_m\}$ its domain of values. Let $B_{\Omega(V)}$ be the set of partitions of $\Omega(V)$. Let $\mathcal{P} \in B_{\Omega(V)}$ be a k-partition $\{\mathbf{v_1}, \ldots, \mathbf{v_k}\}$ of $\Omega(V)$, with $\Pr(\mathbf{v_i}) = \sum_{v_j \in \mathbf{v_i}} \Pr(v_j)$. Finally, let $V^{\mathcal{P}}$ be stochastic variable formed as follows from V relative to partition $\mathcal{P}$:*

- $\Omega(V^{\mathcal{P}}) = \{v_1^{\mathcal{P}}, \ldots, v^{\mathcal{P}}\} = \{\mathbf{v_1}, \ldots, \mathbf{v_k}\}$
- $\Pr(V^{\mathcal{P}} = v_i^{\mathcal{P}}) = \Pr(\mathbf{v_i}) = \sum_{v_j \in \mathbf{v_i}} \Pr(v_j)$

We call this operation on V, resulting in $V^{\mathcal{P}}$, a partial marginalisation *of V relative to $\mathcal{P}$.*

Observe that if $\mathcal{P} = \{\{x\} : x \in \Omega(V)\}$ this transformation effectively marginalises out V.

4 Bayesian network expansion as inverse marginalisation

In marginalisation, we sum out a variable from a joint distribution to keep a distribution over variables of interest, without assigning a value to this other variable. For example, if we have a joint distribution over the variables blood type $BT = (A, B, AB, O)$ and Rhesus factor $Rh = (+, -)$ we might be interested in the relative probability of each blood type, independent of the Rhesus factor.

However, let's assume our model contains BT, but does not specify the Rhesus factor. If we were to introduce Rh as an additional variable, we would need to split, e.g., $\Pr(BT = A)$ into $\Pr(BT = A, Rh = +)$ and $\Pr(BT = A, Rh = -)$ such that $\Pr(A, +) + \Pr(A, -) = \Pr(A)$. This can be seen as the inverse of a marginalisation, notation $\Pr(BT, Rh) \stackrel{\text{def}}{=} \sum_{Rh}^{-1} \Pr(BT)$, with the above constraint on the joint probability distribution $\Pr(BT, Rh)$.

Note that we may assume that the original probability distribution $\Pr(BT)$ represents both $Rh = +$ and $Rh = -$ values, collated in a single distribution. In contrast, a conceptually different situation would be the hypothetical (and counter-factual) situation where the genetic adaptation of an absent Rh protein (i.e., negative Rhesus factor) only recently occurred and, as a consequence, the difference between $Rh = +$ and $Rh = -$ is only now relevant to model. Then, we may assume that in the original probability distribution $\Pr(BT)$ only $Rh = +$ values are represented. The bottom line here is that, while the inverse marginalisation is a one-to-many mapping, it is often possible to constrain the resulting probability distribution based on background information. In section 5 we will further explore potential requirements on $\Pr(BT, Rh)$.

What happens if we do not add a variable, but add a value to the domain of an existing variable? Here the crucial question is whether this operation can be formulated as inverse partial marginalisation or not; that is, whether the addition leads to the 'splitting' of the probability mass of a single value in two (or more) parts (the *invertible* case) or whether the addition really leads to a complete reallocation of the probability mass. See the examples below.

Example 1. Assume we have a Bayesian network with a ternary random variable Q, with values 1, 0, and X, indicating the state of an S/R latch (logical 1, logical 0, invalid $Q'Q$ combination). Assume that we want to add the latching state L which is the specific value assignment $Q' = Q = 0$ and reserve X to the illegal state $Q' = Q = 1$. Here, the original probability mass $\Pr(Q = X)$ is split into $\Pr(Q = X')$ and $\Pr(Q = L)$.

Here, we can see the addition of L (de facto splitting of X) as an inverse partial marginalisation of Q, with partition $\mathcal{P} = \{\{0\}, \{1\}, \{X, L\}\}$ of $\Omega(Q)$; notation $\Pr(Q') \stackrel{\text{def}}{=} \sum_{\mathcal{P}}^{-1} \Pr(Q)$. Now constrast this with the following example.

Example 2. Assume we have a Bayesian network with a binary random variable A, with values H and L, indicating the voltage state of an input pin A (high or low), with $\Pr(A = H) = 0.55$ and $\Pr(A = L) = 0.45$. We want to adjust this variable to include the floating state Z and adjust the distribution to $\Pr(A = H) = 0.5$, $\Pr(A = L) = 0.4$, $\Pr(A = Z) = 0.1$.

Here, the addition of Z can *not* be seen as an inverse partial marginalisation, and as a consequence this adjustment of $\Pr(A)$ cannot be characterized in a similar vein.

4.1 Factored representations

Until now we explored inverse (partial) marginalisation of a full joint probability distribution. Obviously, when adding a variable to a Bayesian network $\mathcal{B}$, we want to make use of the independences in $\mathcal{B}$ rather than revise the entire joint probability distribution Pr and then recompute the CPTs. Hence, we need a factored representation where, in addition to the constraints on the resulting probability distribution as mentioned before, also additional conditional independences (with respect to the added variable) are taken into account as Figure 1 shows.

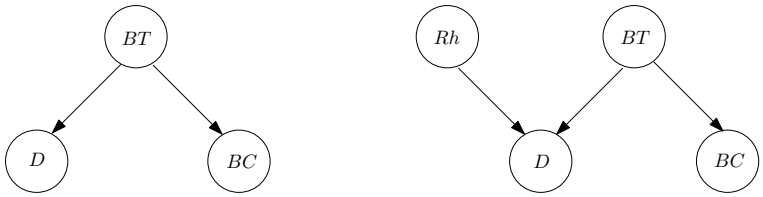

Fig. 1: (left) Example of a Bayesian network modeling the dependency of blood type (BT) on being a blood donor (D) and on risk of developing blood clots (BC). When extended with a variable Rh (right), the CPT for D should be locally adjusted (according to the inverse marginalisation concept) yet the CPT for BC should remain the same.

Here we expand $\Pr(BT, D, BC) = \Pr(BT) \times \Pr(D \mid BT) \times \Pr(BC \mid BT)$ to $\Pr(BT, Rh, D, BC) = \Pr(Rh) \times \Pr(BT) \times \Pr(D \mid BT, Rh) \times \Pr(BC \mid BT)$; using an inverse marginalisation $\Pr(Rh) \times \Pr(D \mid BT, Rh) \stackrel{\text{def}}{=} \sum_{Rh}^{-1} \Pr(D \mid BT)$.

This applies similarly for invertible value addition in a factored representation. We hypothesize, but were not able to prove in this paper, that inverse

marginalisation can be locally applied soundly in case of a factored representation, assuming knowledge on the resulting independences in the network, as in the above case.

5 Setting the novel CPTs

As indicated above the new distributions (and correspondingly, CPTs) need to be set, which is a one-to-many problem: There are infinitely many probabilities $\Pr(b_i)$ such that $\sum_{b_i \in \Omega(B)} \Pr(b_i) = 1$ holds. However, background information about the nature of the addition, and on the conditional probabilities that hold after addition of a variable to a network, can help constrain the set of possible probability mass assignments.

One way of formalising that might be in the form of a linear program, where the constraints are defined by the laws of probability theory (e.g., $\forall_i \sum_j \Pr(\mathbf{a_i}, b_j) = \Pr(\mathbf{a_i}), i = 1, \ldots, |\Omega(\mathbf{A})|, j = 1, \ldots, |\Omega(B)|$). The objective function would then be to minimize the Kullback-Leibler divergence between the distribution $\Pr(\mathbf{V}) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$ and an expected target distribution $\Pr(\mathbf{Q})$, which is based on a Dirichlet prior $\mathrm{Dir}(\alpha)$ that represents the available information, based on the nature of the adjustment, taking the maximum entropy principle into account. For example, in the absence of any information regarding $\Pr(b_j)$ we should set the hyper-parameters α of $\Pr(B)$ to 1. Additional information can then be encoded in these hyper-parameters. A completely orthogonal angle might be to encode desired properties of the new CPTs by means of algebraic constraints on structural equation models [3].

6 Conclusion

In this short note we proposed a mathematical characterization of the problem of extending a Bayesian network with a new variable or value of a variable. This, of course, is just a first initial step of a more elaborate treatment where existing (and future) actual algorithms and approaches that effectively require such expansion (as when a new data set emerges that contains more variables than the existing model with which it is to be integrated) are examined with respect to the formalism.

The approach to model CPT adjustment as a linear program, minimizing the difference between the target distribution and a Dirichlet prior that represents the available information on the desired distribution allows for representing a variety of constraints. In addition to the obvious constraints (the CPT entries must represent a valid distribution) additional desired properties with respect to, e.g., sensitivity towards an output variable of interest might be added. This, however, is left out for future research.

References

1. van der Gaag, L.C.: Probability-Based Models for Plausible Reasoning. Ph.D. thesis, Faculty of Mathematics and Computer Science, Amsterdam University, The Netherlands (1990)
2. Lam, W., Bacchus, F.: Using new data to refine a bayesian network. In: Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence. p. 383–390. UAI'94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1994)
3. van Ommen, T., Drton, M.: Graphical representations for algebraic constraints of linear structural equations models. In: Salmerón, A., Rumí, R. (eds.) Proceedings of The 11th International Conference on Probabilistic Graphical Models. Proceedings of Machine Learning Research, vol. 186, pp. 409–420. PMLR (2022)
4. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, Palo Alto (1988)
5. Rutar, D., de Wolff, E., van Rooij, I., Kwisthout, J.: Structure learning in predictive processing needs revision. Computational Brain & Behavior **5**(2), 234–243 (2022)
6. de Valk, T.: A framework for Bayesian network revisions applied to oncology. Master's thesis, Radboud University, school for AI (2022)
7. Zeng, Y., Xiang, Y., Pacekajus, S.: Refinement of Bayesian network structures upon new data. In: The 2008 IEEE International Conference on Granular Computing, GrC 2008, Hangzhou, China, 26-28 August 2008. pp. 772–777. IEEE (2008). https://doi.org/10.1109/GRC.2008.4664644, `https://doi.org/10.1109/GRC.2008.4664644`