Reliable Discretisation of Deterministic Equations in Bayesian Networks

Alessandro Antonucci
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
(IDSIA)
Manno-Lugano (Switzerland)
alessandro@idsia.ch

Abstract

We focus on the problem of modeling deterministic equations over continuous variables in discrete Bayesian networks. This is typically achieved by a discretisation of both input and output variables and a degenerate quantification of the corresponding conditional probability tables. This approach, based on classical probabilities, cannot properly model the information loss induced by the discretisation. We show that a reliable modeling of such *epistemic* uncertainty can be instead achieved by *credal sets*, i.e., convex sets of probability mass functions. This transforms the original Bayesian network in a *credal network*, possibly returning interval-valued inferences, that are robust with respect to the information loss induced by the discretisation. Algorithmic strategies for an optimal choice of the discretisation bins are also discussed.

Introduction

Bayesian networks (Koller and Friedman 2009) are popular probabilistic graphical models commonly used in AI for machine learning and to implement knowledge-based decision-support systems. Originally designed for discrete variables only (Pearl 1988), Bayesian networks have been extended to continuous variables (Lauritzen and Jensen 2001). Yet, this can be done only under some limiting assumptions (e.g., normal distributions on the nodes and linear relations). In practice, when coping with knowledge-based models, continuous variables are often discretised in order to ease the elicitation of the probabilities from the experts (Wang and Druzdzel 2000). Even in machine learning, algorithms learning Bayesian networks from data are known to provide more accurate inferences after the discretisation (Friedman and Goldszmidt 1996).

To discretise continuous variables in a Bayesian network, i.e., decide the number and the (not necessarily constant) size of the intervals, standard approaches are adopted. When coping with data, these are based on information-theoretic concepts in both supervised (Fayyad and Irani 1992) and unsupervised (e.g., quantiles) settings, while in knowledge-based systems the discretisation intervals are instead manually defined by the experts with the goal of simplifying the elicitation process as much as possible. In all these cases the need of reducing the information loss induced by the discretisation process is traded off against the number of discretisation intervals, which should be also kept as small as possible.

In this paper we focus on the particular problem of discretising Bayesian networks when the model embeds deterministic knowledge in the form of an equation constraining the values of some of its (originally continuous) variables.

A typical approach consists in discretising the variables in the equation by some of the above considered techniques and then providing a degenerate quantification (i.e., one for a state, zero for all the other ones) of the corresponding conditional probabilities. This is intended to reflect the fact that a specification of the input variables in the equation produces a single value of the output variable with no uncertainty.

Yet, the above approach does not take into account the *epistemic* uncertainty induced by the discretisation of the input variables (Tonon 2004). To do that, many authors advocate the need of non-classical models of uncertainty such as fuzzy systems (Dubois and Prade 2012), evidence theory (Tonon 2004), or imprecise probabilities (Beer, Ferson, and Kreinovich 2013).

The goal of this paper is to discuss these ideas in the framework of Bayesian networks and adopt an imprecise-probabilistic approach corresponding to a set of distributions, or *credal set* (Augustin et al. 2014), to achieve that. This basically converts the original Bayesian network into a *credal network* (Cozman 2000), for which dedicated inference algorithms (Antonucci et al. 2015) might be used to compute interval-valued posterior probabilities robust with respect to such epistemic uncertainty.

I	Category	BMI $[\mathrm{Kg/m^2}]$
i_1	Very severely underweight	$\leq$15
i_2	Severely underweight	15-16
i_3	Underweight	16-18.5
i_4	Normal	18.5-25
i_5	Overweight	25-30
i_6	Moderately obese	30-35
i_7	Severely obese	35-40
i_8	Very severely obese	$\geq$40

Table 1: BMI categories

The above example is intended to clarify these ideas.

Example 1. *The* body mass index *(BMI) I of a person whose weight (in* Kg*) is W and height (in* m*) is H is defined as $I := W/H^2$ (Flegal et al. 2012). Consider a person with (W, H) equal to $(89, 1.71)$ and another one with values $(86, 1.74)$. Following the canonical BMI categories in Table 1, the first is considered* moderately obese *(BMI $\simeq$ 30.4) while the second is just* overweight *(BMI $\simeq$ 28.4). Assuming measurements in ranges of 5 Kg and 5 cm, we might not distinguish the two persons, thus being unable to assign them a single BMI category. Table 2 shows other combinations of ranges of W and H leading to a similar indecision.*

W—H	h_1 (1.55-1.60)	h_2 (1.60-1.65)	h_3 (1.65-1.70)	h_4 (1.70-1.75)	h_5 (1.75-1.80)	h_6 (1.80-1.85)	h_7 (1.85-1.90)	h_8 (1.90-1.95)	h_9 (1.95-2.00)
w_1 (55-60)	4	4	4	3-4	3-4	3-4	2-3	1-2	1-2
w_2 (60-65)	4-5	4-5	4	4	4	3-4	3-4	2-3	2-3
w_3 (65-70)	5	4-5	4-5	4	4	4	3-4	3-4	3
w_4 (70-75)	5-6	5	4-5	4-5	4	4	4	3-4	3-4
w_5 (75-80)	5-6	5-6	5	4-5	4-5	4	4	4	4
w_6 (80-85)	6-7	5-6	5-6	5	4-5	4-5	4	4	4
w_7 (85-90)	6-7	6-7	5-6	5-6	5	4-5	4-5	4	4
w_8 (90-95)	7	6-7	6	5-6	5-6	5	4-5	4-5	4
w_9 (95-100)	7	6-7	6-7	6	5-6	5-6	5	4-5	4-5
w_{10} (100-105)	7	7	6-7	6-7	6	5-6	5-6	5	5
w_{11} (105-110)	7	7	7	6-7	6-7	6	5-6	5-6	5
w_{12} (110-115)	7	7	7	7	6-7	6-7	6	5-6	5-6

Table 2: BMI for different combinations of W and H

The term *epistemic* uncertainty is used in the literature to contrast that of *aleatory* uncertainty. The first refers to the uncertainty affecting the subjective way a process is modeled, while the latter refers to the intrinsic randomness of the process (Dubois and Guyonnet 2011).

The uncertainty about W and H in Example 1 is epistemic as we decide to measure weight and height with a tolerance, which in principle can be reduced. The example shows that such epistemic uncertainty induced by the discretisation of the input variables W and H propagates through the deterministic relation defining BMI and this might lead to a non-unique identification of the output.

Basics

Discretisation. Let X denote a (real) variable defined on the interval $[a, b] \subseteq \mathbb{R} \cup \{+\infty, -\infty\}$. A *discretisation* Δ of X is a collection of values $\{x_i\}_{i=0}^n$ such that: $x_i \in [a, b]$ for each i, $x_{i'} < x_{i''}$ for each $i' < i''$, $x_0 = a$ and $x_n = b$. The discretisation Δ induces the specification of a categorical variable, denoted as $\tilde{X}$, with n possible values $\{\tilde{x}_i\}_{i=1}^n$. Given a possible value of X, say $x \in [a, b]$, the corresponding value of $\tilde{X}$ depends on which interval of the partition of $[a, b]$ induced by the discretisation Δ contains x, i.e.,

$$X = x \in [x_{i-1}, x_i) \Rightarrow \tilde{X} = \tilde{x}_i \,. \tag{1}$$

for each $i = 1, \ldots, n$:[1]

[1]For $\tilde{x}_n$ both the extreme values of the interval are included.

The inner points of Δ, i.e., $\{x_i\}_{i=1}^{n-1}$ are called *bins*, while x_0 and x_n are called *extremes*. If Δ' is a discretisation with same extremes of Δ, same bins, and one or more additional bin, we call Δ' a *refinement* of Δ. The refinement of Δ obtained by adding a new bin $x \in (a, b)$ will be denoted as Δ^{+x}. We call *coarsening* the inverse operation consisting in the removal of one or more bins and denote as Δ^{-x} the coarsening of Δ obtained by removing the bin x.

Bayesian networks. A probability *mass function* $P(\tilde{X})$ over a discrete variable $\tilde{X}$ is a non-negative real function over the values of $\tilde{X}$, that is also normalised to one. The mass function is *degenerate* if all the mass is assigned to a single state of $\tilde{X}$. Let $\tilde{X}'$ denote a second discrete variable. A *conditional probability table* (CPT) $P(\tilde{X}|\tilde{X}')$ is a collection of probability mass functions over $\tilde{X}$, one for each possible value of $\tilde{X}'$. We call $\tilde{X}$ the *output* variable of the CPT and $\tilde{X}'$ the *input* (that might also be a joint variable). The mass functions in a CPT, that are indexed by the states of the input variable, are called *columns*.

CPTs are the key to define Bayesian networks. Given a collection of discrete variables, say $\tilde{X}^1, \ldots, \tilde{X}^m$, and a (directed acyclic) graph whose nodes are in one-to-one correspondence with these variables, a *Bayesian network* is a collection of CPTs $\{P(\tilde{X}^j|\mathrm{Pa}^j)\}_{j=1}^m$, where Pa^j denotes the *parents* of X^j, i.e., the variables associated to the predecessors of the node associated to $\tilde{X}^j$ in the graph. A Bayesian network defines a joint probability mass function factorising as follows: $P(\tilde{x}^1, \ldots, \tilde{x}^m) = \prod_{j=1}^m P(\tilde{x}^j|\mathrm{pa}^j)$. The factorisation follows from the *Markov condition*, i.e., the conditional independence of each variable from the non-descendants non-parents given the parents. These concepts are clarified in the following example.

Example 2. *Consider a discrete Bayesian network over four variables associated to the graph in Figure 1. Following the Markov condition we observe the conditional independence between gender and BMI conditionally on weight and height. The CPTs of this Bayesian network are $P(I|W, H)$, $P(W|G)$, $P(H|G)$ and the (unconditional, as G has no parents) $P(G)$. The BMI equation, whose (continuous) input and output coincide with the (discrete) input and output of the CPT associated to the I, can be therefore used for the quantification of $P(I|W, H)$.*

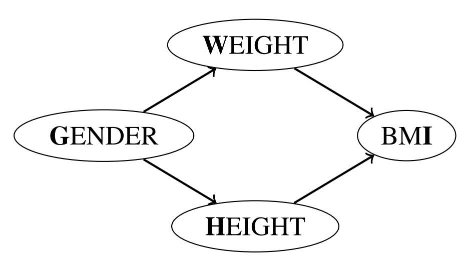

Figure 1: A Bayesian network over four variables

Credal sets. A *credal set* $K(\tilde{X})$ over $\tilde{X}$ is defined as a convex set of probability mass functions over $\tilde{X}$. This represents a generalisation of classical uncertainty models corresponding to a (single) probability mass function $P(\tilde{X})$. The *vacuous* credal set is the largest credal set we can define, and it includes all the probability mass functions over $\tilde{X}$. This is commonly regarded as a model of *ignorance* about the state of $\tilde{X}$. More informative credal sets are obtained by adding constraints to the vacuous credal set.

A special class of credal sets is associated to *events*. Let $\tilde{\mathcal{X}}$ denote the set of possible values of $\tilde{X}$. Given a subset $\tilde{\mathcal{X}}'$ of $\tilde{\mathcal{X}}$, the set of all the probability mass functions assigning probability one to $\tilde{\mathcal{X}}'$ and zero to its complement is denoted as $K_{\tilde{\mathcal{X}}'}(X)$ and it is called the credal set of the event $\tilde{\mathcal{X}}'$. In particular the vacuous credal set is the credal set associated to the universe, i.e., $K_{\tilde{\mathcal{X}}}(\tilde{X})$, while the credal sets associated to singletons are degenerate probability mass functions. Credal sets associated to events can be analogously defined for continuous variables. Figure 2 depicts a geometrical view of the vacuous credal set for a ternary variable (gray triangle), together with a more informative model (white triangle), and the credal set associated to event $\tilde{\mathcal{X}}' = \{\tilde{x}_2, \tilde{x}_3\}$ (dotted segment).

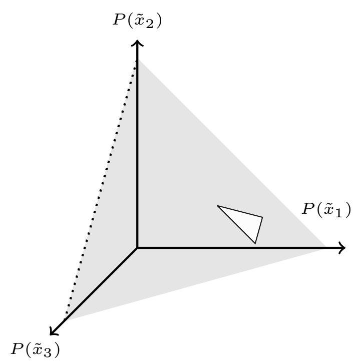

Figure 2: Credal sets for a ternary variable

Credal networks. Credal networks (Cozman 2000) are an extension of Bayesian networks based on credal sets. Each CPT column $P(\tilde{X}^i|\mathrm{pa}^i)$ is replaced by a credal set $K(\tilde{X}^i|\mathrm{pa}^i)$ A credal network defines a joint credal set $K(\tilde{X}^1, \ldots, \tilde{X}^m)$ whose elements are joint mass functions associated to Bayesian networks with the same graph with each CPT column $P(\tilde{X}^i|\mathrm{pa}^i)$ taking values from $K(\tilde{X}^i|\mathrm{pa}^i)$. The typical inference tasks for Bayesian networks consists in the computation of the conditional probability for a variable of interest given evidence about other variables (e.g., $P(i|g)$ for the Bayesian network in Example 2), the analogous problem for a credal network consists in the computation of the lower and upper bounds of this conditional probabilities, e.g., $[\underline{P}(i|g), \overline{P}(i|g)]$, with respect to any Bayesian network specification consistent with the credal sets of the credal network.

Discretising Equations

In this section we present a possible solution to the problem of quantifying a CPT in a Bayesian network on the basis of a deterministic equation whose input and output variables are the same as in CPT (as in Example 2). Again we use the BMI example to introduce the discussion.

Example 3. *Consider the CPT $P(I|W,H)$ of the discrete Bayesian network in Example 2. The quantification of this CPT might be based on the deterministic equation (for the corresponding continuous variables) $I = W/H^2$. As shown in Table 2, the intervals spanned by the BMI for the ranges of weight and height associated to some discrete values of these input variables are entirely included in one of the BMI ranges in Table 1. E.g., a person whose weight is in the range* 85-90 *and height in* 1.75-1.80 *has a BMI in the range* 26.23-29.39*, i.e., certainly belonging to the category* over-weight *(i.e., i_5). The corresponding column $P(I|w_7, h_5)$ can be therefore safely quantified as a degenerate mass function having all its mass in the state i_5. As already discussed in Example 1, this is not the case for $P(I|w_7, h_4)$, for which the interval spanned by the BMI overlaps with both i_5 and i_6 (see Figure 3).*

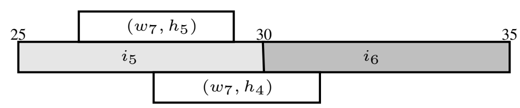

Figure 3: BMI ranges and spanned intervals

Strategies to solve the indecision between i_5 and i_6 can be obtained within the framework of Bayesian networks and classical probabilities. For instance, we can split the probability mass between i_5 and i_6 in equal parts or according to some weighting scheme (e.g., *proportional* to the overlap between the BMI range and each interval). Alternatively, we can adopt *representative* values for the input ranges (e.g., midpoints), compute the corresponding output with the equation, and concentrate the mass on the category including the output. In the example, as BMI$(87.5, 1.725) \simeq 29.41$, this would be i_4. All these approaches are based on implicit heuristic assumptions we make explicit by the following result.

Proposition 1. *The CPT $P(\tilde{Y}|\tilde{X}^1, \ldots, \tilde{X}^m)$ in a Bayesian network is quantified by an equation $Y = F(X^1, \ldots, X^m)$ whose (continuous) input and output variables coincide (after discretisation) with those of the CPT. The conditional probability $P(\tilde{y}_i|\tilde{x}^1_{j_1}, \ldots, \tilde{x}^{(m)}_{j_m})$ corresponds to the integral:*

$$\int_{y_{i-1}}^{y_i} \int_{x^1_{j_1-1}}^{x^1_{j_1}} \ldots \int_{x^m_{j_m-1}}^{x^m_{j_m}} \delta[y - F(x^1, \ldots, x^m)] \prod_{k=1}^{m} \left[\pi_k(x^k) dx^k\right] dy \quad (2)$$

where $\pi_k(x^k)$ is the (continuous) conditional probability distribution for X^k given that $\tilde{X}^k = \tilde{x}^k_{j_k}$.

Proof. From Equation (2) remove the integration over y and the constraints on the integration bounds for the other integration variables. By the total probability theorem, the resulting integral is the (continuous) posterior probability distribution for Y given $(\tilde{x}^1_{j_1}, \ldots, \tilde{x}^m_{j_m})$. By inverting the implication in Equation (1):

$$\tilde{X}^k = \tilde{x}^k_{j_k} \Rightarrow x^k \in [x^k_{j_k-1}, x^k_{j_k})\,, \tag{3}$$

i.e., the support of π_k is $[x^k_{j_k-1}, x^k_{j_k})$. Thus, the previously obtained integral can be restricted to the same integration bounds as in Equation (2) and the probability for $\tilde{y}_i$ is just the integral over the corresponding range $[y_{i-1}, y_i)$. □

Note also that the conditional distributions in Equation (2) can be obtained from the marginal distribution $\hat{\pi}(x^k)$ as:

$$\pi_k(x^k) = \frac{\hat{\pi}(x_k) I_{[x_{j_k-1}, x_{j_k})}(x_k)}{\int_{x^k_{j_k-1}}^{x^k_{j_k}} \hat{\pi}(x_k)\, dx^k}\,, \tag{4}$$

where I the indicator of the corresponding interval.

Proposition 1 makes clear that, despite the determinism provided by the equation, a proper quantification on the CPT requires the assessment of the marginal distributions of the (continuous) input variables. Having data available for these variables, a statistical procedure can be used to assess the marginal and, after the transformation in Equation (4), Equation (2) might be eventually used for the CPT quantification.

If no data are available (which is the typical case for knowledge-based systems), the only information we might rely on is that the support of the conditional distributions is the interval in Equation (3). The approach taking *representative* midpoints of the ranges corresponds to:

$$\pi_k(x^k) := \delta\left(x^k - \frac{x^k_{j_k-1} + x^k_{j_k}}{2}\right)\,. \tag{5}$$

This appears as a very unrealistic assumption of a marginal distribution made of a combination of improper functions. The *proportional* approach corresponds instead to the case of a uniform marginal (leading to uniform conditionals) which, again, seems very unrealistic in many scenarios. As noticed in (Walley 1996), uniform distributions are modeling of condition of *indifference* among the different options, while in our case we cope with a condition of *ignorance*, which should be better modeled by a *vacuous* credal set, i.e., *all* the possible specification for the marginal distributions $\hat{\pi}$. The corresponding conditional distributions returned by Equation (4) are all the distributions whose support is as in Equation (3). In this setting, Equation (2) returns different values for different specifications of π_k, this transforming the original Bayesian network specification into a credal network.

The extreme values of this specification can be equivalently obtained by considering extreme (i.e., degenerate) distributions only (Benavoli and Noack 2012). Accordingly, we rewrite Equation (2) with, for each $k = 1, \ldots, m$, $\pi_k(x^k) := \delta(x^k - \hat{x}^k)$, with $\hat{x}^k \in [x^k_{j_k-1}, x^k_{j_k})$, and obtain that $P(\tilde{y}_i|\tilde{x}^1, \ldots, \tilde{x}^m)$ is:

$$\int_{y_{i-1}}^{y_i} \delta(y - F(\hat{x}^1, \ldots, \hat{x}^m)) = \begin{cases} 1 & \text{if } F(\hat{x}^1, \ldots, \hat{x}^m) \in [y_{i-1}, y_i), \\ 0 & \text{otherwise}\,. \end{cases} \tag{6}$$

This corresponds to a degenerate credal set assigning zero probability to all the discrete states of $\tilde{Y}$ non-overlapping the interval $[\underline{F}, \overline{F}]$, where:

$$\underline{F} := \min_{\substack{x_k \in [x^k_{j_k-1}, x^k_{j_k}) \\ k=1,\ldots,m}} F(\hat{x}^1, \ldots, \hat{x}^m) \tag{7}$$

and analogously for $\overline{F}$.

Assuming the function associated to the deterministic equation easy to minimize/maximize on hypercubes, we have a clear procedure to embed equations in CPTs. The approach is reliable in the sense that the resulting credal networks, to be queried by dedicated inference algorithm (Antonucci et al. 2015), return interval-valued inferences corresponding to the union of the inferences obtained for any possible quantification of the marginals of the input variables.

Discretisation Strategies

In the previous section we assumed the discretisation of both the input and the output variables given. In this section we discuss possible strategies to identify *ad hoc* discretisations reducing the information loss in the CPTs embedding deterministic equations. This can be intended as an optimisation task requiring the minimisation of the information loss induced by the discretisation with respect to some freedom in the choice of the position and the number of bins.

Following (Abellán and Moral 2005) we can adopt the *upper entropy* of credal sets as an information-theoretic measure. In the special case of credal sets associated to events, the upper entropy takes a particularly simple form, being simply the logarithm of the number of ranges with non-zero probability. For CPTs we can simply sum of the upper entropies of all the columns. The information loss/gain obtained when moving from a discretisation to another one is therefore the difference between the two sums. Such a simple descriptor might already trade off against the number of bins for the input variables, as the highest is this number, the highest is the number of columns in the CPT (and then, possibly, the total entropy of the CPT). To eliminate possible ties, we also take into account the total size of the ranges associated to the events. Optimal discretisations for the setup considered in this paper should therefore minimise the total entropy of the CPT.

The simplest situation consists in having a fixed discretisation of the input variables and complete freedom in the choice of the discretisation of the output variable. A simple strategy inspired by the *interval partitioning problem* (Sahni 1976) is depicted by Algorithm 1. The idea is to iteratively add bins to the discretisation (up to a bound on the maximum size $n_{\max}$ of the discretisation). The choice is restricted to the bounds of the intervals $[\underline{F}, \overline{F}]$ spanned by the function in the deterministic equation defined as in Equation (7) when the input variables vary in their ranges. By taking these intervals for the different columns of the CPT as the input for Algorithm 1, we obtain a discretisation minimising the information loss. Remember that in case of ties in the identification of $\hat{y}$, we take the y leading to the set of intervals associated to the smallest range by only considering the elements of the tie. Such an algorithm has worst-case quadratic complexity with respect to the number of intervals which, in turn, corresponds to the CPT size.

Algorithm 1 Optimal discretisation of the output variable

```
input: I ← {[l_i, u_i]}_{i=1}^q // set of intervals
Δ = {a, b} // initialise discretisation
γ ← true
while γ and |Δ| < n_max do
   // minimum overlap with other intervals
   ŷ ← arg min_{y ∈ {l_1,...,l_n,u_1,...,u_n}\Δ} |{I ∈ I : y ∈ I}|
   if E(Δ^{+ŷ}) < E(Δ) then
      Δ ← Δ ∪ {y} // add bin if entropy decrease
   else
      γ ← false // stop otherwise
   end if
end while
output: Δ // discretisation
```

For a very preliminary validation of the algorithm, we obtained an alternative BMI discretisation different from that in Table 1 but with the same number of bins. The new discretisation induces a more informative CPT compared to that in Table 2 because of an increased number of precise columns. Yet, to prevent *concentration* effects (i.e., a single large range and all the other ones very small) additional constraints such a minimum width for the discretisation intervals should be better added.

Conclusions

A novel approach to the embedding of deterministic equation in discrete Bayesian networks has been proposed. This allows for a robust modeling of the epistemic uncertainty induced by the discretisation. A clear semantics for the proposed approach has been provided. As a future work we intend to study the relation between the particular credal networks considered in this paper and the *possibilistic networks* (Benferhat et al. 2000) in order to devise dedicated inference algorithms.[2]

[2]Some of the basic ideas of this paper have been already presented by the authors of this paper during the XXII World Congress of the International Measurement Confederation (IMEKO 2018).

References

[Abellán and Moral 2005] Abellán, J., and Moral, S. 2005. Upper entropy of credal sets. applications to credal classification. *International Journal of Approximate Reasoning* 39(2-3):235–255.

[Antonucci et al. 2015] Antonucci, A.; de Campos, C. P.; Huber, D.; and Zaffalon, M. 2015. Approximate credal network updating by linear programming with applications to decision making. *International Journal of Approximate Reasoning* 58:25–38.

[Augustin et al. 2014] Augustin, T.; Coolen, F. P.; de Cooman, G.; and Troffaes, M. C. 2014. *Introduction to imprecise probabilities*. John Wiley & Sons.

[Beer, Ferson, and Kreinovich 2013] Beer, M.; Ferson, S.; and Kreinovich, V. 2013. Imprecise probabilities in engineering analyses. *Mechanical systems and signal processing* 37(1-2):4–29.

[Benavoli and Noack 2012] Benavoli, A., and Noack, B. 2012. Pushing Kalman's idea to the extremes. In *Proceedings of the 15th International Conference on Information Fusion*, 1202–1209. IEEE.

[Benferhat et al. 2000] Benferhat, S.; Dubois, D.; Garcia, L.; and Prade, H. 2000. Directed possibilistic graphs and possibilistic logic. In *Information, Uncertainty and Fusion*. Springer. 365–379.

[Cozman 2000] Cozman, F. G. 2000. Credal networks. *Artificial intelligence* 120(2):199–233.

[Dubois and Guyonnet 2011] Dubois, D., and Guyonnet, D. 2011. Risk-informed decision-making in the presence of epistemic uncertainty. *International Journal of General Systems* 40(02):145–167.

[Dubois and Prade 2012] Dubois, D., and Prade, H. 2012. Gradualness, uncertainty and bipolarity: making sense of fuzzy sets. *Fuzzy Sets and Systems* 192:3–24.

[Fayyad and Irani 1992] Fayyad, U. M., and Irani, K. B. 1992. On the handling of continuous-valued attributes in decision tree generation. *Machine learning* 8(1):87–102.

[Flegal et al. 2012] Flegal, K. M.; Carroll, M. D.; Kit, B. K.; and Ogden, C. L. 2012. Prevalence of obesity and trends in the distribution of body mass index among us adults, 1999-2010. *Jama* 307(5):491–497.

[Friedman and Goldszmidt 1996] Friedman, N., and Goldszmidt, M. 1996. Discretizing continuous attributes while learning Bayesian networks. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, 157–165.

[Koller and Friedman 2009] Koller, D., and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT Press.

[Lauritzen and Jensen 2001] Lauritzen, S. L., and Jensen, F. 2001. Stable local computation with conditional Gaussian distributions. *Statistics and Computing* 11(2):191–203.

[Pearl 1988] Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Mateo, California: Morgan Kaufmann.

[Sahni 1976] Sahni, S. K. 1976. Algorithms for scheduling independent tasks. *Journal of the ACM (JACM)* 23(1):116–127.

[Tonon 2004] Tonon, F. 2004. Using random set theory to propagate epistemic uncertainty through a mechanical system. *Reliability Engineering & System Safety* 85(1-3):169–181.

[Walley 1996] Walley, P. 1996. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)* 3–57.

[Wang and Druzdzel 2000] Wang, H., and Druzdzel, M. J. 2000. User interface tools for navigation in conditional probability tables and elicitation of probabilities in Bayesian networks. In *Proceedings of the Sixteenth conference on Uncertainty in Artificial Intelligence*, 617–625. Morgan Kaufmann Publishers Inc.