Quasi-Bayesian Networks

Fabio G. Cozman

Escola Politécnica, Universidade de São Paulo
fgcozman@usp.br, http://www.cs.cmu.edu/~fgcozman

Abstract

This paper presents a complete theory of *Quasi-Bayesian networks*, structures that associate convex sets of probability measures to directed acyclic graphs. Quasi-Bayesian networks are graphical models for precise/imprecise beliefs. The central contribution of this work is a theory of Quasi-Bayesian networks that displays as much flexibility and representational power as the theory of standard Bayesian networks. Results in this paper show how to express judgements of irrelevance and independence in Quasi-Bayesian networks, and how to compute inferences from Quasi-Bayesian networks. The properties of the so-called type-1 extensions are clarified through a new generalization of d-separation, and novel results are presented for inferences with the so-called natural extensions. The paper describes exact and approximate methods to generate inferences for type-1 extensions and explores linear fractional programming to generate inferences for natural extensions. The paper also investigates Quasi-Bayesian networks that are defined globally through perturbations of a single network.

Key words: Graphical models of inference; Convex sets of probability measures; Bayesian networks; Lower and upper expectations; Robust Bayesian analysis; Independence relations; Graphical d-separation relations.

1 Introduction

Probabilistic reasoning has gained widespread acceptance in the Artificial Intelligence community as a methodology for the representation of beliefs [62]. But not all beliefs can be cast as sharp numeric values [6,48,65,74]. The theory

[1] This research was conducted while the author was with the School of Computer Science, Carnegie-Mellon University. The work was partially supported by NASA under Grant NAGW-1175; the author was supported under a scholarship from CNPq, Brazil.

of convex sets of probability measures, called the theory of imprecise probabilities [73] or Quasi-Bayesian theory [28], offers an alternative to standard probabilistic (Bayesian) procedures. While Bayesian theory admits a single probability measure, Quasi-Bayesian theory allows a decision-maker to represent imprecise and incomplete beliefs through a set of measures.

This paper presents a complete theory of *Quasi-Bayesian networks*, structures that associate convex sets of probability measures to directed acyclic graphs. The central contribution of this work is a theory of Quasi-Bayesian networks that displays as much flexibility and representational power as the theory of standard Bayesian networks.

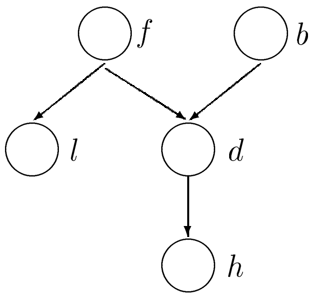
Figure 1 shows a Quasi-Bayesian network. In a standard Bayesian network, every variable is associated with a probability measure conditional on the parents of the variable. In a Quasi-Bayesian network, variables can also be associated with probability inequalities.

The key *theoretical* problem in Quasi-Bayesian networks is how to express, detect, and exploit irrelevance and independence relations; the key *practical* problem in Quasi-Bayesian networks is how to generate bounds on posterior probabilities and posterior expectations. These questions are addressed in this paper through Walley's definitions of irrelevance and independence. The properties of the so-called type-1 extensions are clarified through a new generalization of d-separation, and novel results are presented for inferences with the so-called natural extensions. The paper describes exact and approximate methods to generate inferences for type-1 extensions, and explores linear fractional programming to produce inferences for natural extensions. The paper also investigates Quasi-Bayesian networks that are defined globally through neighborhoods of a single network. Computation of bounds on variances and decision-making strategies are briefly discussed.

Quasi-Bayesian networks can be used to handle imprecise and incomplete beliefs, and to conduct robustness analysis of standard Bayesian models. Both activities can now be conducted with two freely available packages, the *JavaBayes* system and the *qb* procedure, that implement the algorithms described in this paper. Together, these algorithms and software packages open the field of graphical models to Quasi-Bayesian theory.

2 Paper roadmap

Sections 4 and 5 describe concepts and results that are used throughout the paper. Section 6 introduces locally defined Quasi-Bayesian networks and defines the concepts of type-1 extension and natural extension. Sections 7 and 8



$$
\begin{array}{cc}
0.4 \le p(f) \le 0.5 & 0.4 \le p(b) \le 0.5 \\
p(l|f) = 0.6 & p(d|f, b) = 0.8 \\
p(l|f^c) = 0.05 & p(d|f, b^c) = 0.1 \\
p(h|d) = 0.6 & p(d|f^c, b) = 0.1 \\
p(h|d^c) = 0.3 & p(d|f^c, b^c) = 0.7
\end{array}
$$

Fig. 1. Example network with five binary variables (superscript c indicates negation).

contain the main results for type-1 extensions and natural extensions respectively.

Section 9 briefly describes an alternative way to specify Quasi-Bayesian, employing neighborhoods of standard Bayesian networks. Sections 10, 11, 12, and 13 deal respectively with non-atomic inferences, expectations, variances, and decision-making. Two freely available software systems for Quasi-Bayesian network inferences are described in Section 14. Section 15 contains concluding remarks.

All proofs of lemmas and theorems are grouped in Appendix A, and a number of models employed in the paper are summarized in Appendix B.

3 Notation roadmap

Several notational conventions, grouped in this section, are used throughout the paper.

Events are denoted by the first letters of the alphabet (capitalized): A, B, C. Variables are denoted by the last letters of the alphabet (capitalized): X, Y, W, Z. A variable can be subscripted either to indicate its place in a collection of variables (X_i), or a particular value (X_j), or both (X_{ij}). Collections of variables are indicated in bold; for instance, $\mathbf{X}$. The collection of all variables that belong to $\mathbf{X}$ but do not belong to $\mathbf{Y}$ is indicated by $\mathbf{X}\backslash\mathbf{Y}$. Functions are indicated by the letters f, g and h. The symbol $\sum_{\mathbf{X}} f(\mathbf{X}, \mathbf{Y})$ indicates that all variables in $\mathbf{X}$ are summed out from the function $f(\mathbf{X}, \mathbf{Y})$.

The probability measure of an event A is indicated by $P(A)$. Because this paper focuses on discrete variables, the function defined by $P(\{X = X_j\})$ completely characterizes a probability measure [56]. The term *probability density* is used to refer to the function $P(\{X = X_j\})$ and the symbol $p(X)$ denotes the probability density of X. The expectation of a function $f(X)$ is indicated by $E[f(X)]$. A conditional probability measure is indicated by $P(A|B)$. The notation $P(A|X_j)$ indicates the probability measure conditional on the event $\{X = X_j\}$. The conditional expectation of a function $f(X)$ is denoted $E[f(X)|B]$.

Convex sets of probability measures are indicated by K. A convex set of measures is such that, if measures P_1 and P_2 belong to the set, then the measure $\alpha P_1 + (1-\alpha)P_2$ also belongs to the set. The set $K(X)$ contains all probability measures induced by densities $p(X)$; similarly, the set $K(X|B)$ contains conditional probability measures. The expression "$K(X)$ is defined by a set of densities $k(X)$" means that $K(X)$ contains all and only measures induced by densities in the set $k(X)$. The expression "$p(X)$ belongs to $K(X)$" means that the measure induced by the density $p(X)$ belongs to $K(X)$.

4 Graphical models of inference

Representations based on graphs are commonplace in Artificial Intelligence, for example in the analysis of games, search problems, and planning [62]. A popular graphical representation for probabilistic models is the *Bayesian network* formalism [39].

Any Bayesian network represents a joint probability density over a set of variables $\mathbf{X}$. The joint density is specified through a directed acyclic graph. Each node of this graph represents a variable X_i; the parents of X_i are denoted by $\mathrm{pa}(X_i)$. Each variable X_i is associated with a conditional density $p(X_i|\mathrm{pa}(X_i))$. Such a graphical structure defines a unique joint probability density through the following expression [59]:

$$p(\mathbf{X}) = \prod_i p(X_i|\mathrm{pa}(X_i)) \,. \tag{1}$$

Any directed acyclic graph displays *d-separation* relations [27]. The definition of d-separation is somewhat involved. Given three collections of variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, suppose that along every path between a variable in $\mathbf{X}$ and a variable in $\mathbf{Y}$, there is a variable W such that: either W has converging arrows and is not in $\mathbf{Z}$ and none of its descendants are in $\mathbf{Z}$, or W has no converging arrows and is in $\mathbf{Z}$. Then $\mathbf{Y}$ and $\mathbf{X}$ are d-separated by $\mathbf{Z}$. A key property of Bayesian networks is that graphical d-separation and probabilistic independence are equivalent [59, page 117]. In particular, every variable is independent of its nondescendants given its parents.

Working with d-separation relations in Bayesian network learning, Spirtes, Glymour and Scheines have introduced the concept of *faithfulness* [69]. A probability measure is *faithful* to a directed acyclic graph when all and only independence relations entailed by the measure correspond to graphical d-separation relations.

Given a Bayesian network, the event E denotes the *evidence* in the network. For example, $E = \{X_1 = X_{12}, X_3 = X_{31}\}$ fixes the values of variables X_1 and X_3. The symbol $\mathbf{X}_E$ denotes the collection of variables that have values fixed by the evidence E; for example, $\mathbf{X}_E = \{X_1, X_3\}$.

Inferences with Bayesian networks usually involve the calculation of the posterior marginal for a *queried* variable X_Q [21,58]. The symbol X_Q is used to denote both the queried variable and the event that the queried variable has a specified value. The symbol $\mathbf{X}_{QE}$ denotes the variables in $\mathbf{X}_E$ plus the queried variable X_Q.

The construction of a Bayesian network demands a number of precise probability values. Several non-Bayesian attempts have been made to relax this requirement through possibility theory [26], Dempster-Shafer theory [66] or even purely heuristic approaches [67]. These approaches depart radically from usual axioms of decision making and are not pursued further in this paper. A less radical non-Bayesian strategy is to associate interval-based probability values with variables in a network [8,16,25,34,32]. Interval representations have two problems. First, it is not always possible to apply Bayes rule to an interval-valued measure and to obtain an interval-valued posterior measure [16,32]. Second, there is no unique way to define independence for interval-valued measures [15].

An alternative to interval-based probability is the theory of closed convex sets of probability measures. In this theory, conditioning is easily defined: a set of conditional measures is obtained by applying Bayes rule to each measure in a set of probability measures. But the concept of independence is still controversial: in the most in-depth study of this matter, Campos and Moral have reviewed five different possible types of independence [20]. The lack of agreement on a definition of independence has led researchers to associate graphical models with convex sets of measures in a heuristic manner [10,12,72,51,71,81]. In previous work, concepts of independence have been proposed either to simplify computations, or to mimic concepts in standard probability theory. This has led to paradoxes and controversies regarding the semantics of graphical models, and has left the study of independence relations to the side [15]. This paper, contrary to previous work, starts from a definition of independence based on axioms of preferences and proceeds from this solid foundation. The next section presents the basic notions of the theory of convex sets of measures, including concepts of irrelevance and independence.

5 Quasi-Bayesian theory

In the real world we rarely meet all the assumptions of a Bayesian model. First, we have to face imperfections in a decision-maker's beliefs, either because the decision-maker has no time, resources, patience, or confidence to provide exact probability values. Second, we may deal with a group of disagreeing experts, each specifying a particular probability [48]. Third, we may be interested in abstracting away parts of a model and assessing the effect of this abstraction [15,33]. In such circumstances, it is advisable to consider extensions of probability theory that can handle imprecise or incomplete beliefs. One practical application of this strategy can be found in the field of Robust Bayesian Statistics, whose goal is to employ convex sets of probability measures to represent perturbations in probabilistic models [7,40,77]. Robustness analysis provides an exhaustive description of the relation between perturbations and posterior values, in contrast to the small-scale perturbations that are employed by sensitivity analysis [45,54].

Quasi-Bayesian theory employs convex sets of probability measures to represent beliefs and to evaluate decisions [28]. Several other theories use similar representations: inner/outer measures [30,34,60,70], lower probability theory [8,16,25,68], imprecise probabilities [73], convex Bayesianism [32], probability/utility sets [64]. Some theories, like Dempster-Shafer theory [65], can be cast mathematically, but not conceptually, in terms of convex sets of probability measures [34].

A closed convex set of probability measures is called a *credal set*; existence of credal sets is derived from axioms about preferences [28].[2] A credal set defined by a set of densities $p(X)$ is indicated by $K(X)$. A credal set defined by a set of joint densities $p(\mathbf{X})$ is called a *joint credal set* and indicated by $K(\mathbf{X})$.

This paper deals with credal sets that are defined as the convex hull of a finite number of probability measures; such *finitely generated credal sets* are polytopes in the space of probability measures [31]. In the context of variables with finitely many values, most of the standard models used in Robust Statistics are finitely generated credal sets (Appendix B).

Given a credal set $K(X)$, lower and upper bounds on probability can be generated for any event A:

$$\underline{P}(A) = \min_{P \in K} P(A), \quad \overline{P}(A) = \max_{P \in K} P(A).$$

[2] An introduction to technical aspects of Quasi-Bayesian theory can be found at http://www.cs.cmu.edu/~qBayes/Tutorial/.

The set-function $\underline{P}(A)$ is called a *lower envelope*; the set-function $\overline{P}(A)$ is called an *upper envelope*. *Lower densities* $\underline{p}(X)$ and *upper densities* $\overline{p}(X)$ are defined similarly, taking minima and maxima of densities in a credal set. Lower envelopes can be obtained from upper envelopes through the expression $\underline{P}(A) = 1 - \overline{P}(A^c)$. Note that a credal set always creates unique lower and upper envelopes, but lower and upper envelopes do not define a unique credal set [73, Section 2.7].

Lower and upper expectations for a function $f(X)$ are respectively defined as:

$$\underline{E}[f(X)] = \min_{P \in K} E[f(X)], \quad \overline{E}[f(X)] = \max_{P \in K} E[f(X)],$$

where $E[f(X)]$ is the standard expectation of the function $f(X)$. There is a one-to-one correspondence between lower (or upper) expectations and credal sets: the collection of lower expectations for all arbitrary functions $f(X)$ defines a credal set $K(X)$. Maxima and minima of expectations can only occur at the vertices of a credal set [73]. Lower expectations can be obtained from upper expectations through the expression $\underline{E}[f(X)] = -\overline{E}[-f(X)]$. The lower envelope for event A can be obtained by taking the lower expectation of the indicator function $I_A(X)$, which is one if $X \in A$ and zero otherwise.

A credal set can be used to generate lower and upper values for any functional. Consider the variance of a variable X, defined as $V[X] = E[X^2] - (E[X])^2$. The *lower and upper variances* are respectively:

$$\underline{V}[X] = \min_{P \in K} V[X], \quad \overline{V}[X] = \max_{P \in K} V[X].$$

Convex sets of conditional probability measures are employed to represent conditional beliefs. The *conditional credal set* $K(X|A)$ is defined by conditional densities $p(X|A)$. Quasi-Bayesian inference is performed by applying Bayes rule to each measure in a joint credal set. A *posterior credal set* is the union of all posterior measures obtained in this process. If the event A has lower envelope equal to zero, the posterior credal set $K(X|A)$ is taken (by convention) to contain every probability measure $P(X|A)$. If the event A has lower envelope larger than zero, the vertices of a posterior credal set $K(X|A)$ can be obtained by applying Bayes rule to all vertices of the corresponding joint credal set [28]. An important result is that maxima and minima values of posterior probability only occur at the vertices of a posterior credal set [73].

For an arbitrary function $f(X)$, the lower expectation of $f(X)$ conditional on an event A, $\underline{E}[f(X)|A]$, is the unique solution of the following equation in λ, called *generalized Bayes rule* [73, Section 6.4]:

$$\underline{E}[(f(X) - \lambda) I_A(X)] = 0,$$

where $I_A(X)$ is the indicator function of the event A. A bracketing algorithm that solves this equation is called *Lavine's algorithm* [46].

The concept of independence was not defined in the original formulation of Quasi-Bayesian theory by Giron and Rios [28]. Currently there is no standard way to define independence relations with credal sets [5,15,20,43,75]. The results presented in this paper adopt Walley's definition of independence because this formulation can be reduced to preference relations, following the Quasi-Bayesian philosophy [73]. Walley's original definition is stated in terms of lower expectations; to develop a theory of convex sets of measures, the next paragraphs formulate the same concept in terms of credal sets.

Consider collections of variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, and the credal sets $K(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, $K(\mathbf{X}|\mathbf{Z})$, $K(\mathbf{Y}|\mathbf{Z})$, $K(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$ and $K(\mathbf{Y}|\mathbf{X}, \mathbf{Z})$.

Two credal sets $K(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$ and $K(\mathbf{X}|\mathbf{Y})$ are said *equal* if any measure in $K(\mathbf{X}|\mathbf{Y}', \mathbf{Z}')$ (for any fixed $\mathbf{Y} = \mathbf{Y}'$, $\mathbf{Z} = \mathbf{Z}'$) also belongs to $K(\mathbf{X}|\mathbf{Z}')$, and vice-versa.

Definition 1 *Variables* $\mathbf{Y}$ *are* irrelevant *to* $\mathbf{X}$ *given* $\mathbf{Z}$ *if* $K(\mathbf{X}|\mathbf{Z})$ *is equal to* $K(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$*, regardless of the value of* $\mathbf{Z}$*. Variables* $\mathbf{X}$ *and* $\mathbf{Y}$ *are* independent *given* $\mathbf{Z}$ *if* $\mathbf{X}$ *is irrelevant to* $\mathbf{Y}$ *given* $\mathbf{Z}$ *and* $\mathbf{Y}$ *is irrelevant to* $\mathbf{X}$ *given* $\mathbf{Z}$*. Note that* $\mathbf{Z}$ *can be empty; in this case the irrelevance and independence concepts are not "conditional" on any variable.*

6 Locally defined Quasi-Bayesian networks

This section defines Quasi-Bayesian networks that are generated by associating local credal sets with a directed acyclic graph.

Definition 2 *A* locally defined Quasi-Bayesian network *is a directed acyclic graph associated with: 1) a* local credal set $K(X_i|pa(X_i))$ *for each variable* X_i*, and 2) a method for the combination of the local credal sets.*

The rationale for the previous definition is as follows. In a standard Bayesian network, Expression (1) uniquely specifies a joint density as a combination of "local" conditional densities. There is no analogue to Expression (1) in Quasi-Bayesian networks. Instead, it is more appropriate to ask a decision maker to explicitly determine how to combine the local credal sets in a given network.

The local credal sets are the *quantitative constraints* of the network. A joint credal set that satisfies all constraints (quantitative and qualitative) in a Quasi-Bayesian network is called an *extension* of the network.

The key technical difference between standard Bayesian networks and locally defined Quasi-Bayesian networks is that more than one extension may exist for a Quasi-Bayesian network. The following example suffices to illustrate this point.

Example 3 *Consider a graphical structure with three binary variables X_1, X_2 and X_3, where X_1 and X_2 have no parents and X_3 has X_1 and X_2 as parents. This network is associated with the local credal sets $K(X_1)$, $K(X_2)$ and $K(X_3|X_1, X_2)$. Suppose the set $K(X_1)$ is defined by the convex hull of two densities $p_l(X_1)$, for $l \in \{1, 2\}$. Likewise, suppose $K(X_2)$ is defined by the convex hull of two densities $p_l(X_2)$, for $l \in \{1, 2\}$. Finally, the four sets $K(X_3|X_1, X_2)$ are also defined by convex hulls of two densities each, $p_l(X_3|X_1, X_2)$, for $l \in \{1, 2\}$. There are at least two extensions that satisfy all these requirements. The first extension is defined by the convex hull of the sixty-four joint densities $p(X_1, X_2, X_3) = p_i(X_1)p_j(X_2)p_k(X_3|X_1, X_2)$, for $i, j, k \in \{1, 2\}$. The second extension is defined by the convex hull of the two joint densities $p(X_1, X_2, X_3) = p_i(X_1)p_i(X_2)p_i(X_3|X_1, X_2)$, for $i \in \{1, 2\}$.*

At first, it is tempting to adopt a methodology that uniquely selects an extension for any given Quasi-Bayesian network. The literature offers a collection of approaches that are based on this strategy [10,12,72,51,71,81]. This strategy usually fails to provide a satisfactory semantics for the selected extension, and fails to give meaning to statements of irrelevance and independence in Quasi-Bayesian networks. A more promising approach, advocated in this paper, is to start from a principled definition of irrelevance and independence and then accept that several extensions can be generated from a single directed acyclic graph.

Two types of extensions are analyzed in this paper: *type-1 extensions* and *natural extensions* (terminology adapted from Walley's [73, pp. 453, 455]). Other extensions may be considered for specific applications, but type-1 and natural extensions provide a solid foundation for the study of locally defined Quasi-Bayesian networks.

Definition 4 *The type-1 extension of a locally defined Quasi-Bayesian network is a joint credal set containing all joint measures that satisfy Expression (1) when each density $p(X_i|pa(X_i))$ is obtained from the local credal set $K(X_i|pa(X_i))$.*

Section 7 investigates the irrelevance relations that are implied by this particular method for the combination of local credal sets.

Definition 5 *The natural extension of a locally defined Quasi-Bayesian network is the largest joint credal set containing all joint measures that comply with the quantitative constraint of the network and with a collection of irrelevance relations.*

Section 8 investigates several strategies to specify well-defined natural extensions through irrelevance relations.

The choice of the "correct" extension depends on problem-specific considerations. For example, natural extensions can represent more dramatic variations in a probability measure than a type-1 extension, but natural extensions tend to be computationally more demanding than type-1 extensions.

The next sections investigate the theory and practice of type-1 and natural extensions. To simplify the presentation, algorithms are developed only for the calculation of the posterior upper envelopes of the queried variable given the evidence:

$$\overline{P}(X_Q|E) = \max\left(\frac{p(\mathbf{X}_{QE})}{p(\mathbf{X}_E)}\right) = \max\left(\frac{\sum_{\mathbf{X}\setminus\mathbf{X}_{QE}} p(\mathbf{X})}{\sum_{\mathbf{X}\setminus\mathbf{X}_E} p(\mathbf{X})}\right). \tag{2}$$

Lower envelopes can be obtained through the same operations by replacing maximization with minimization.

Fundamentally, calculation of a posterior upper envelope is an optimization problem over a closed convex set. This insight is applied in all algorithms in this paper; in independent work, Cano *et all.* [10,11] and Zaffalon [81] have explored the same observation for type-1 extensions.

7 Type-1 extensions: Theory and algorithms

A type-1 extension is the largest joint credal set where all extreme points satisfy Expression (1). Type-1 extensions are the most common type of Quasi-Bayesian network studied in the literature, due to their intuitive similarity to standard Bayesian networks [10,12,20,53,71,81]. Section 7.1 presents a novel result that ties type-1 extensions to graphical d-separation. Algorithms for type-1 extensions have been analyzed previously in a variety of fields (Section 4); the algorithms in this section unify previous work, rather than present fundamentally new results.

7.1 Irrelevance, independence, and type-1 extensions

The appeal of type-1 extensions comes from their intuitive similarity with standard Bayesian networks. This intuition can be formalized using Walley's

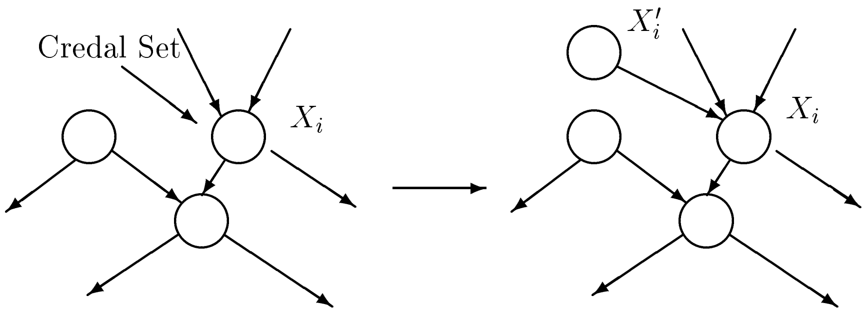


Fig. 2. The CCM transformation for variable X.

definition of independence:

Theorem 6 *Every graphical d-separation relation in a Quasi-Bayesian network corresponds to a valid conditional independence relation in the type-1 extension of the network.*

This new result demonstrates that type-1 extensions mimic the properties of standard Bayesian networks as independence-maps [59, page 119]. The theorem also complements results by Cano et al. [12]. They give conditions on independence concepts that satisfy d-separation in type-1 extensions, but they do not provide any definition of independence to illustrate their result. The theorem demonstrates that Walley's independence relations exhibit the desired correspondence with d-separation, and justify the intuitive similarity between type-1 extensions and standard Bayesian networks.

D-separation has important algorithmic consequences. For example, consider a query involving a variable X_Q and evidence E. All variables that do not affect computation of $p(X_Q|E)$ can be obtained by d-separation in polynomial time [27].

7.2 Enumeration algorithms

One way to generate inferences with type-1 extensions is to enumerate all possible vertices of the posterior credal set $K(X_Q|E)$. The first such algorithm was proposed by Cano et al. [12,10]. The Cano/Cano/Moral (CCM) transformation takes a locally defined Quasi-Bayesian network and produces a standard Bayesian network. For each variable X_i associated with a local credal set with m vertices $p_k(X_i|\mathrm{pa}(X_i))$ (where $m > 1$), add a new variable X_i' that has X_i as its only child (Figure 2). The variable X_i' has m integer values. Now associate X_i with a family of standard conditional densities:

$$p(X_i|\mathrm{pa}(X_i), X_i' = k) = p_k(X_i|\mathrm{pa}(X_i)).$$

The variables X_i' are called *transparent* variables [10]. The posterior upper envelope $\overline{P}(X_Q|E)$ can be calculated by visiting all the values of transparent variables, because maxima of probabilities occur at the vertices of the posterior credal set. To reduce the time spent cycling through transparent variables, it is possible to perform a single standard Bayesian inference to obtain the function $p(X_Q|E, \mathbf{X}')$ (where $\mathbf{X}'$ denotes the collection of all transparent variables). Maximization with respect to the transparent variables produces the posterior upper envelope.

Inferences with type-1 extensions can be simplified by the use of convex hull algorithms. Consider an operation (summation or multiplication) that involves functions indexed by a transparent variable X_i. The operation can be performed either on the original functions or on the convex hull of the original functions. For example, take a singly connected network with binary variables; there, a convex hull operation can be taken at every step in an inference propagation algorithm [59], because the convex hull of a collection of intervals is still an interval. This property is actually the basis for Zaffalon's 2U algorithm [81].

Two factors contribute to the high computational cost of enumeration algorithms. First, inferences can be reduced by the CCM transform to standard Bayesian network inferences, which are NP-hard [18]. Second, the number of values for transparent variables generated by the CCM transform grows exponentially with the number of local credal sets. Despite these difficulties, exact inferences with type-1 extensions can be performed realistically in the *JavaBayes* engine (Section 14), because graphical d-separation leads to enormous computational savings.

7.3 Iterative and global optimization

The calculation of a posterior envelope (Expression (2)) is a convex maximization problem. Inference algorithms based on optimization have only surfaced in recent years. Cano et al. focus on stochastic global optimization [10,11]; Andersen and Hooker [1], Zaffalon [81] and Cozman [19] have explored gradient-based and deterministic global optimization. The purpose of this section is to summarize these ideas and present a number of new results: the relationship between inference type-1 in extensions and learning in Bayesian networks; the relationship between gradient calculations and graph faithfulness; and the relationship between Tessem's approximation method and global optimization. Stochastic techniques such as simulated annealing [10] or genetic algorithms [11] are not explored; such approaches deserve full treatment in future work.

7.3.1 Inferences and learning algorithms

Call Θ the vector containing all values of $p(X_i|\text{pa}(X_i))$. The search for a Θ that maximizes posterior probability is similar to the search for a maximum likelihood estimator under linear inequalities, one of the most common problems in Bayesian network learning [9]. This observation suggests that most of the effort in developing efficient algorithms for Bayesian network learning can impact Quasi-Bayesian networks and their applications.

7.3.2 Gradient-based optimization

A variety of iterative optimization algorithms employ gradient calculations in the search for a maximizing Θ [49]. Previous work has derived efficient, closed-form expressions for this gradient calculation in graphical models [45,61]. The gradient is presented here in a slightly different form to highlight the relationship between gradient calculations and the concept of *faithfulness* in a network (Section 4).

Under the assumption that $\underline{P}(E) > 0$, each component of the gradient (the derivative $\partial p(X_Q|E) / \partial p(X_{ij}|\text{pa}(X_i)))$ is equal to:

$$\frac{p(\text{pa}(X_i))}{P(E)} \left(P(X_Q, E|X_{ij}, \text{pa}(X_i)) - P(X_Q|E)\, P(E|X_{ij}, \text{pa}(X_i)) \right). \quad (3)$$

Every term in Expression (3) can be calculated by any standard Bayesian network algorithm. Note that Expression (3) is identically zero when $p(X_Q|E) = p(X_Q|E, X_{ij}, \text{pa}(X_i))$. This equality is true when the gradient is calculated for a joint density that is *not faithful* to the graph, assuming that d-separated credal sets are deleted before calculation of the gradient (Section 7.1).

7.3.3 Deterministic global optimization

Inferences with type-1 extensions can be reduced to signomial programs and solved exactly through branch-and-bound procedures [1,19,81]. Signomial programs involve optimization of polynomials [2,23], and require well-tuned programming engines [3,29,57]. Simpler global optimization approaches are possible in special cases, as long as a method that produces quick upper and lower probability bounds is available [24,41]. As an example, bounds can be calculated efficiently for singly-connected networks using Tessem's algorithm [71].

8 Natural extensions: Theory and algorithms

A natural extension is the largest credal set containing joint measures that satisfy: 1) all quantitative constraints in a Quasi-Bayesian network and 2) a given collection of irrelevance relations. A natural extension is *well-defined* when its collection of irrelevance relations can be satisfied by at least one probability measure.

The next sections describe three methods for the specification of natural extensions. These methods have the property that any irrelevance relation imposed on a network corresponds to a graphical d-separation relation in the network. Consequently, every natural extension discussed here is well-defined. The first method is to impose no irrelevance relation in the network, a strategy similar to probabilistic logic [4,50,55]. The second method is to require that the nondescendants of any variable are irrelevant to the variable given the parents of the variable. The third method is to require that the nondescendants of any variable are independent of the variable given the parents of the variable. The last two methods mimic properties of standard Bayesian networks (Section 4); together, these three methods summarize most of the technical aspects of natural extensions.

8.1 Quantitative constraints and separately specified local credal sets

The algorithms for natural extension in this paper assume that the local credal set for any variable is specified separately for each value of the variable's parents. This means that constraints on densities $p(X_i|\mathrm{pa}(X_i) = \pi_1)$ do not interfere with constraints on densities $p(X_i|\mathrm{pa}(X_i) = \pi_2)$ when $\pi_1 \neq \pi_2$. This restriction makes sense both for elicitation and representation of beliefs [73, Section 6.2].

Each finitely generated local credal set can be specified through a finite number of upper expectations (Section 5). To specify the credal set $K(X_i|\mathrm{pa}(X_i))$, take a collection of functions $f_k(X_i)$ and numbers γ_k and impose:

$$\overline{E}[f_k(X_i)] = \gamma_k \quad \Rightarrow \quad \sum_j f_k(X_{ij}) p(X_{ij}|\mathrm{pa}(X_i)) \leq \gamma_k. \tag{4}$$

Note that the functions f_k and the numbers γ_k should be indexed by a particular variable X_i and a particular value for $\mathrm{pa}(X_i)$; these indexes are suppressed to simplify notation.

The inequalities in Expression (4) can be reduced to inequalities on the values of the joint density $p(\mathbf{X})$:

$$\sum_{\mathbf{X}\setminus\mathrm{pa}(X_i)} (f_k(X_i) - \gamma_k)\, p(\mathbf{X}) \leq 0. \tag{5}$$

Apart from the collection of constraints in Expression (5), the following unitary constraint must be satisfied by any joint density $p(\mathbf{X})$:

$$\sum_{\mathbf{X}} p(\mathbf{X}) = 1. \tag{6}$$

8.2 No irrelevance relations

The simplest type of natural extension is produced when no irrelevance relation is imposed on a Quasi-Bayesian network. Denote by $K_1(\mathbf{X})$ this natural extension.

The maximization in Expression (2) subject to constraints in Expressions (5) and (6) is a linear fractional program on the values of $p(\mathbf{X})$ [37,63]. Linear fractional programs can be reduced to linear programs by a variety of methods [37,63]; consequently, inferences with $K_1(\mathbf{X})$ can be solved by linear programming techniques. The effort involved in solving such linear programs is potentially enormous for a large network, but techniques from probabilistic logic and linear programming, like column generation [38] and row suppression [17], can be used to tame this complexity.

To guarantee that only valid inferences are produced by fractional linear programming, it is necessary to check whether $\underline{P}(E)$ is non-zero. The calculation of $\underline{P}(E)$ involves a linear program subject to constraints in Expressions (5) and (6). If $\underline{P}(E) = 0$, then the posterior lower envelope $\overline{P}(X_Q|E)$ is equal to one by convention (Section 5).

8.3 Irrelevance relations for nondescendants

A more sophisticated, and perhaps the most appealing, type of natural extension is one where the nondescendants of any variable are irrelevant to the variable given the parents of the variable. Denote by $K_2(\mathbf{X})$ this natural extension.

Irrelevance of nondescendants is a key property of standard Bayesian networks [59]. In the context of Quasi-Bayesian networks, the constraint on $K_2(\mathbf{X})$ is that the credal sets $K(X_i|\mathrm{pa}(X_i))$ and $K(X_i|\mathrm{nd}(X_i))$ be equal, where $\mathrm{nd}(X_i)$

denotes the nondescendants of X_i. This constraint implies the following collection of inequalities:

$$\sum_{\mathbf{X} \backslash \mathrm{nd}(X_i)} (f_k(X_i) - \gamma_k)\, p(\mathbf{X}) \leq 0, \tag{7}$$

where all functions $f_k(X_i)$ and numbers γ_k are obtained from constraints (5). Note that Expression (7) represents a larger number of constraints than Expression (5).

Example 7 *Suppose that $nd(X_1) = \{pa(X_1), X_2\}$, where X_2 is a binary variable. The inequalities represented by Expression (7) are divided in two groups:*

$$\sum_{\substack{\mathbf{X} \backslash pa(X_1), X_2 \\ X_2 = TRUE}} (f_k(X_1) - \gamma_k)\, p(\mathbf{X}) \leq 0, \qquad \sum_{\substack{\mathbf{X} \backslash pa(X_1), X_2 \\ X_2 = FALSE}} (f_k(X_1) - \gamma_k)\, p(\mathbf{X}) \leq 0.$$

The method just described starts with a credal set and generates inequalities for another credal set with strictly more conditioning variables. Any such procedure is called a *replication procedure.*

The following result demonstrates that constraints in Expression (5) are in fact subsumed by constraints in Expression (7).

Lemma 8 *If the credal sets $K(X_i|pa(X_i))$ and $K(X_i|nd(X_i))$ are equal, then the credal sets $K(X_i|pa(X_i))$ and $K(X_i|pa(X_i), \mathbf{S})$ are equal as well, for any $\mathbf{S} \subset nd(X_i)$.*

Inferences with $K_2(\mathbf{X})$ are linear fractional programs subject to constraints in Expressions (6) and (7). Even though irrelevance relations may introduce a large number of constraints into this program, they can also introduce substantial simplifications, as demonstrated in the remainder of this section.

Consider a Quasi-Bayesian with a sub-network containing variables $\mathbf{S}$. Suppose that all ascendants of a variable in $\mathbf{S}$ are also in $\mathbf{S}$. This sub-network is at the "top" of the original network. Denote by $K_2'(\mathbf{S})$ the natural extension of the sub-network under the constraint that the nondescendants (in $\mathbf{S}$) of any variable X_i in $\mathbf{S}$ are irrelevant to X_i given the parents of X_i. The constraints on $K_2'(\mathbf{S})$ are:

$$\sum_{\mathbf{S} \backslash \mathrm{nd}(X_i)} (f_k(X_i) - \gamma_k)\, p(\mathbf{S}) \leq 0, \;\; \sum_{\mathbf{S}} p(\mathbf{S}) = 1. \tag{8}$$

Suppose that $K_2'(\mathbf{S})$ is available; how is $K_2'(\mathbf{S})$ related to the natural extension $K_2(\mathbf{X})$ of the original network? To answer this question, denote by $K_2(\mathbf{S})$ the credal set obtained by marginalization of $K_2(\mathbf{X})$.

Theorem 9 *The extension $K_2'(\mathbf{S})$ is equal to the marginal credal set $K_2(\mathbf{S})$.*

The next theorem builds on the previous results to demonstrate that the maximization in Expression (2) can be restricted to the values $p(\mathbf{S})$.

Theorem 10 *Suppose that all variables X_i outside of $\mathbf{S}$ are associated to single conditional densities $p(X_i|pa(X_i))$. The solution for problem (2), subject to constraints in Expressions (7) and (6) is identical to:*

$$\overline{P}(X_Q|E) = \max \left(\frac{\sum_{\mathbf{S}\backslash \mathbf{X}_{QE}} h(X_Q, \mathbf{S}) p(\mathbf{S})}{\sum_{\mathbf{S}\backslash \mathbf{X}_{E}} h(X_Q, \mathbf{S}) p(\mathbf{S})} \right), \tag{9}$$

subject to constraints in Expression (8), where the function $h(X_Q, \mathbf{S})$ is:

$$h(X_Q, \mathbf{S}) = \sum_{\mathbf{X}\backslash \{\mathbf{S} \cup \mathbf{X}_{QE}\}} \left(\prod_{X_i \in \{\mathbf{X}\backslash \mathbf{S}\}} p(X_i|pa(X_i)) \right).$$

Note that all variables in $\mathbf{X}\backslash\mathbf{S}$ are associated with a unique conditional measure, so the function $h(X_Q, \mathbf{S})$ is uniquely defined. Any standard Bayesian network algorithm can generate $h(X_Q, \mathbf{S})$ by eliminating variables outside of $\mathbf{S}$. The consequence of the theorem is that networks where most local credal sets are on the "top" of the graph can profit from irrelevance constraints. This is particularly promising in practical applications, because in general the most imprecise measures are the priors, which are associated with nodes without parents.

8.4 Independence relations for nondescendants

Consider the constraint that, for every variable X_i, nondescendants of X_i are independent of X_i given the parents of X_i. Denote by $K_3(\mathbf{X})$ the natural extension that satisfies this constraint. According to Walley's definition of independence (Section 4),

- the credal sets $K(X_i|\mathrm{pa}(X_i))$ and $K_3(X_i|\mathrm{nd}(X_i))$ must be equal, and
- the credal sets $K_3(\mathrm{nD}(X_i)|\mathrm{pa}(X_i), X_i)$ and $K_3(\mathrm{nD}(X_i)|\mathrm{pa}(X_i))$ must be equal, where $\mathrm{nD}(X_i)$ denotes the nondescendants non-parents of X_i.

Example 11 *Take a Quasi-Bayesian network where a variable X_5 has nondescendants X_1, X_2, X_3, X_4 and parents X_1, X_2 (Figure 3). Independence*

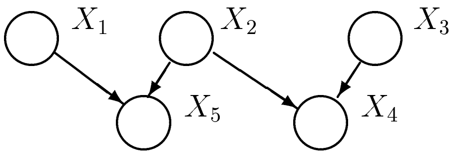


Fig. 3. Creating the natural extension for a sub-network.

of X_5 requires that $K_3(X_5|X_1, X_2, X_3, X_4)$ be equal to $K(X_5|X_1, X_2)$, and that $K_3(X_3, X_4|X_1, X_2, X_5)$ be equal to $K_3(X_2, X_4|X_1, X_2)$. The first condition leads to linear inequalities through a replication procedure (Section 8.3), because $K(X_5|X_1, X_2)$ is directly specified on the network. The second condition involves two credal sets that are not directly specified on the Quasi-Bayesian network.

Several properties of natural extensions $K_2(\mathbf{X})$ presented in Section 8.3 can be extended to natural extensions $K_3(\mathbf{X})$. The next lemma proves that irrelevance of a variable X_i to its nondescendants, given its parents, implies irrelevance of X_i to subsets of its nondescendants.

Lemma 12 *If the credal sets $K(nD(X_i)|pa(X_i), X_i)$ and $K(nD(X_i)|pa(X_i))$ are equal, then the credal sets $K(\mathbf{S}|pa(X_i), X_i)$ and $K(\mathbf{S}|pa(X_i))$ are equal as well, for any $\mathbf{S} \subset nD(X_i)$.*

The next theorem establishes the relation between the natural extension of a sub-network and the complete natural extension. Consider a Quasi-Bayesian network with a sub-network containing a collection of variables $\mathbf{S}$, with the property that all ascendants of a variable in $\mathbf{S}$ are also in $\mathbf{S}$. Denote by $K_3'(\mathbf{S})$ the natural extension of the sub-network under the constraint that the nondescendants (in $\mathbf{S}$) of any variable X_i in $\mathbf{S}$ are independent of X_i given the parents of X_i. Denote by $K_3(\mathbf{S})$ the credal set obtained by marginalization of $K_3(\mathbf{X})$.

Theorem 13 *The extension $K_3'(\mathbf{S})$ is equal to the marginal credal set $K_3(\mathbf{S})$.*

A result similar to Theorem 10 can be demonstrated using the previous lemma and theorem. Suppose that all variables outside of $\mathbf{S}$ are associated with single conditional densities $p(X_i|\mathrm{pa}(X_i))$; then $K_3(\mathbf{X})$ is obtained by combining $K_3'(\mathbf{S})$ with conditional densities for variables outside of $\mathbf{S}$. The proof of this result is similar to the proof of Theorem 10 and is omitted.

The remainder of this section develops an algorithm for the calculation of posterior envelopes from a natural extension $K_3(\mathbf{X})$. The central idea of the algorithm is to construct a sequence of sub-networks, beginning with a single variable, and ending with the complete Quasi-Bayesian network. At each step, a variable is added to the current sub-network, and linear constraints for the natural extension of the new sub-network are generated. In the end, linear fractional programming can be used to maximize posterior probability subject to a collection of linear inequalities. The complexity of this algorithm is high because a natural extension may be defined by an exponential number of linear constraints; no efficient algorithm for manipulation of independence constraints is currently known.

Algorithm BuildExtension
The construction of linear inequalities for $K_3(\mathbf{X})$ is as follows. Take a variable X_1 that has no parents and form a sub-network that contains only X_1. The marginal extension $K_3'(X_1)$ is equal to the local credal set $K(X_1)$ (Theorem 13). Now add a variable X_2 to this sub-network, such that X_2 is either a direct descendant of X_1 or X_2 has no parent. As demonstrated in the next paragraphs, all independence constraints can be generated for this new sub-network. Now form a new sub-network by adding a variable X_3, such that every ascendant of a variable in the new sub-network belongs to the new sub-network. Again, independence constraints can be generated for this new sub-network. This process continues until all variables are incorporated into a complete network.
End of algorithm BuildExtension

The algorithm BuildExtension depends on three facts. First, it is always possible to construct a sequence of sub-networks such that every ascendant of a variable in a sub-network belongs to the sub-network (property of directed acyclic graphs). Second, the natural extension of any sub-network, taking into account independence relations in the sub-network, is always equal to the marginal credal set obtained by marginalizing the complete natural extension $K_3(\mathbf{X})$ (Theorem 13). Third, it is possible to represent the credal set of a sub-network after the incorporation of a variable, as described by the next algorithm.

Algorithm IncorporateVariable
Take a sub-network with variables $\mathbf{S}$, such that all ascendants of a variable in $\mathbf{S}$ are in $\mathbf{S}$. To generate $K_3(\mathbf{S}, X_i)$ from $K_3(\mathbf{S})$:
1) The constraint that $K_3(X_i|\mathbf{S})$ and $K(X_i|\mathrm{pa}(X_i))$ be equal is satisfied through a replication procedure (Section 8.3).
2) The constraint that $K_3(\mathbf{S}\backslash\mathrm{pa}(X_i)|\mathrm{pa}(X_i), X_i)$ and $K_3(\mathbf{S}\backslash\mathrm{pa}(X_i)|\mathrm{pa}(X_i))$ be equal is satisfied in two steps. First, obtain $K_3(\mathbf{S}\backslash\mathrm{pa}(X_i)|\mathrm{pa}(X_i))$ from $K_3(\mathbf{S})$ (this step is detailed later). Then the inequalities for $K_3(\mathbf{S}\backslash\mathrm{pa}(X_i)|\mathrm{pa}(X_i), X_i)$ are obtained through a replication procedure.
3) If a variable X_j in $\mathbf{S}$ is not a direct ascendant of X_i, then the independence between X_i and X_j must be asserted. The constraint that nondescendants of X_j (in $\mathbf{S}$) plus X_i be irrelevant to X_j given the parents of X_j produces lin-

ear inequalities, again by a replication procedure, because $K(X_j|\mathrm{pa}(X_j))$ is specified by the Quasi-Bayesian network. The constraint that X_j be irrelevant to its nondescendants plus X_i given the parents of X_j is more complex. To obtain linear inequalities for this constraint, note that a sub-network $\mathbf{S}'$ can be formed by discarding X_j and its descendants from $\mathbf{S}$. Obtain the constraints for $K_3(\mathbf{S}')$ from $K_3(\mathbf{S})$ (this step is detailed later). The set $K_3(\mathbf{S}', X_i)$ is then obtained recursively by the algorithm IncorporateVariable. This recursion always terminates, as it eventually runs out of sub-networks.
End of algorithm IncorporateVariable

The algorithm IncorporateVariable assumes that linear inequalities for a marginal credal set $K(\mathbf{X})$ and a conditional credal set $K(\mathbf{X}|\mathbf{Y})$ can be generated from a given joint credal set $K(\mathbf{X}, \mathbf{Y})$. To do so, consider a credal set $K(\mathbf{X}, \mathbf{Y})$ specified through a collection of linear constraints in matricial form as $\mathbf{AP} \leq \mathbf{B}$, where $\mathbf{P}$ is a vector containing all values of the density $p(\mathbf{X}, \mathbf{Y})$. The mandatory unitary constraint in Expression (6) is written as $\sum_i \mathbf{P}_i = 1$.

- To obtain $K(\mathbf{X})$, define a vector $\mathbf{P}'$ containing all values of the density $p(\mathbf{X})$. The vectors $\mathbf{P}$ and $\mathbf{P}'$ are related by linear equalities. Now apply the Fourier-Motzkin elimination algorithm [31] to eliminate all variables $\mathbf{P}$ and to produce the linear inequalities that define the marginal credal set $K(\mathbf{X})$.
- To obtain $K(\mathbf{X}|\mathbf{Y}')$, define a vector $\mathbf{P}''$ containing all values of $p(\mathbf{X}, \mathbf{Y}')$ (with $\mathbf{Y} = \mathbf{Y}'$). Using the Fourier-Motzkin elimination algorithm, produce the linear inequalities for $\mathbf{P}''$, summarized by a matricial inequality $\mathbf{A}'\mathbf{P}'' \leq \mathbf{B}'$. Now introduce a new variable $W \geq 0$, form the matricial inequality $\mathbf{A}'\mathbf{P}'' \leq W\mathbf{B}'$, and replace the original unitary constraint by $\sum_i \mathbf{P}''_i = 1$. The variable W works as a "scale" for $\mathbf{P}''$ and creates a "cone", with apex at the origin, extending through the polytope for $\mathbf{P}''$. Now apply the Fourier-Motzkin algorithm again to eliminate W and produce the linear inequalities that define the conditional credal set $K(\mathbf{X}|\mathbf{Y}')$.

Example 14 *Take the network in Figure 3. Assume that previous computations generated the extension $K_3(X_1, X_2, X_3, X_4)$ as a collection of linear constraints. When variable X_5 is added to the network, how to compute the extension $K_3(X_1, X_2, X_3, X_4, X_5)$?*

- *Satisfy the requirement that $K_3(X_5|X_1, X_2, X_3, X_4)$ and $K(X_5|X_1, X_2)$ be equal through a replication procedure.*
- *Satisfy the requirement that $K_3(X_3, X_4|X_1, X_2, X_5)$ and $K_3(X_3, X_4|X_1, X_2)$ be equal by generating $K_3(X_3, X_4|X_1, X_2)$ directly from $K_3(X_1, X_2, X_3, X_4)$ and using a replication procedure.*
- *Satisfy the requirement that X_3 be independent of X_5. First, use $K(X_3)$ directly in a replication procedure including X_5. Second, obtain $K_3(X_1, X_2, X_5)$ recursively. Third, employ a replication procedure to generate the constraints*

for $K_3(X_1, X_2, X_5|X_3)$ based on $K_3(X_1, X_2, X_5)$.

- *Satisfy the requirement that X_4 be independent of X_5. First, use $K(X_4|X_2, X_3)$ in a replication procedure including X_5. Second, obtain $K_3(X_1, X_2, X_3, X_5)$ recursively. Third, obtain $K_3(X_1, X_5|X_2, X_3)$ from $K_3(X_1, X_2, X_3, X_5)$. Fourth, employ a replication procedure to generate $K_3(X_1, X_5|X_2, X_3, X_4)$ based on $K_3(X_1, X_5|X_2, X_3)$.*

Algorithm BuildExtension proves constructively the following non-trivial fact:

Theorem 15 *The natural extension $K_3(\mathbf{X})$ is a finitely generated credal set.*

9 Global neighborhoods of Bayesian networks

In previous sections, Quasi-Bayesian models were constructed from collections of local conditional credal sets. An alternative path is to specify Quasi-Bayesian models as neighborhoods of standard Bayesian networks. A similar strategy is employed in Robust Statistics in connection to univariate models; the purpose of this section is to briefly adapt the language and results of that field to the present context.

The central idea is to define a joint credal set through a single Bayesian network called the *base* network. The joint density specified by the base network is denoted $r(\mathbf{X})$. Four neighborhoods of $r(\mathbf{X})$ are discussed in the remainder of this section (the univariate versions of these models are presented in Appendix B). Quasi-Bayesian inference with these four neighborhoods is similar to standard Bayesian network inference, due to a *marginalization invariance property*. To define this invariance property, denote by $\Gamma\left(r(\mathbf{X})\right)$ a convex neighborhood of $r(\mathbf{X})$. This neighborhood is marginalization invariant when

$$p(\mathbf{X}) \in \Gamma\left(r(\mathbf{X})\right) \Rightarrow \sum_{\mathbf{X}\setminus\mathbf{Y}} p(\mathbf{X}) \in \Gamma\left(\sum_{\mathbf{X}\setminus\mathbf{Y}} r(\mathbf{X})\right).$$

In words: marginalization and construction of neighborhoods commute [78]. Marginal credal sets can be easily obtained for neighborhoods with marginalization invariance, because marginalization is performed only in the base network.

- An ϵ-contaminated joint credal set is defined by all joint densities $p(\mathbf{X})$ such that $p(\mathbf{X}) \geq (1-\epsilon)r(\mathbf{X})$, where $\epsilon \in (0, 1)$. The lower expectation of a function $f(X_Q)$ is obtained in closed-form:

$$\underline{E}[f(X_Q)] = \frac{(1-\epsilon)\left(\sum_{X_Q} f(X_Q) r(\mathbf{X}_{Q,E})\right) + \epsilon \min_{X_Q} f(X_Q)}{(1-\epsilon) r(E) + \epsilon}.$$

- A constant density bounded joint credal set is defined by all joint densities $p(\mathbf{X})$ such that $(1/\mu) r(\mathbf{X}) \leq p(\mathbf{X}) \leq \mu r(\mathbf{X})$, where $\mu > 1$. Expression (B.2) can then be used to generate posterior envelopes, due to the 2-monotone character of the constant density bounded model (Appendix B.3). For lower and upper expectations, Lavine's algorithm generates a convenient, iterative solution (Section 4).
- A total variation joint credal set is defined by all joint densities $p(\mathbf{X})$ such that for all events A, $|P(A) - R(A)| \leq \epsilon$, where $\epsilon \in (0, 1)$ and R is the joint probability measure induced by $r(\mathbf{X})$. Almost all the discussion for the constant density bounded model applies to the total variation class, because this model is also 2-monotone and has closed-form solutions for lower and upper envelopes (Appendix B.4). Again, lower and upper expectations can be obtained through Lavine's algorithm.
- A constant density ratio joint credal set defined by all joint densities $p(\mathbf{X})$ such that $(1/\mu) r(\mathbf{X}) \leq \alpha p(\mathbf{X}) \leq \mu r(\mathbf{X})$ for some $\alpha > 0$, where $\mu > 1$. Closed-form expressions for posterior lower and upper envelopes can be obtained because this model is invariant to marginalization and conditioning (Expression (B.3)) [76]. For lower and upper expectations, results from DeRobertis and Hartigan [22] can be used to produce bracketing algorithms (Expressions (B.4) and (B.5)).

10 Non-atomic inferences

Previous sections focused on the calculation of upper bounds for the posterior density $p(X_Q|E)$. Calculation of posterior envelopes for non-atomic events A is immediate. Algorithms for type-1 extension are identical both for enumeration or iterative algorithms. Linear programs solving inferences with natural extension must enlarge their objective functions to incorporate unions of atomic events. Finally, algorithms for global neighborhoods are identical for non-atomic events.

11 Lower and upper expectations

For the sake of simplicity, algorithms for locally defined Quasi-Bayesian networks dealt only with posterior probabilities (Section 9 describes algorithms for calculation of posterior expectations in global neighborhood models). Most algorithms for locally defined Quasi-Bayesian networks can be easily extended to deal with posterior expectations, for linear fractional programming, enumeration, or gradient-based search. Some special algorithms, for example, deterministic global optimization based on Tessem's bounds (Section 7.3.3) are specific to probability calculations and do not extend to expectations.

12 Lower and upper variances

Calculation of bounds for variances in a Quasi-Bayesian network is a great challenge because the expression for $V[X_Q]$ is quadratic on probability values. Walley's variance envelope theorem [73, theorem G2] can be used to produce a convergent algorithm for lower and upper variances. Walley demonstrates that $\underline{V}[X_Q] = \min_\mu \underline{E}[(X_Q - \mu)^2]$ and $\overline{V}[X_Q] = \min_\mu \overline{E}[(X_Q - \mu)^2]$. The calculation of lower and upper variances becomes a unidimensional optimization problem, which can be solved by discretizing μ (note that μ can only be larger than zero and smaller than the square of the largest value of X_Q). Lower and upper variances are then obtained by repeated calculation of lower and upper expectations for $(X_Q - \mu)^2$.

13 Decision making

In Quasi-Bayesian theory, it is not always possible to select a single course of action that maximizes expected utility for all measures in a credal set [48,64]. Several strategies are possible to select a "best" decision.

As an example, consider the most probable explanation problem (MPE), where the objective is to find a configuration of variables that maximizes the probability of the evidence. Generating all MPE solutions for all possible measures in the credal set may be too expensive. An alternative solution is to find the configuration that maximizes the upper posterior envelope of the evidence, $\arg\max_{X \in \{\mathbf{X} \setminus \mathbf{X}_E\}} \overline{P}(E)$. If a type-1 extension is considered, a CCM transformation can be applied to the Quasi-Bayesian network, and the problem is equivalent to a standard Bayesian network MPE problem on the transformed network. Another alternative is to find the configuration that maximizes the lower posterior envelope of the evidence, $\arg\max_{X \in \{\mathbf{X} \setminus \mathbf{X}_E\}} \underline{P}(E)$. Despite its intuitive appeal, this approach is challenging because it is a maxmin problem; at the present time no efficient algorithm for maxmin optimization in graphical models has been proposed.

14 Quasi-Bayesian networks in practice

Two freely available software packages have been developed to allow construction and manipulation of Quasi-Bayesian networks.

Type-1 extensions and global neighborhoods are available in the *JavaBayes* system, a portable and freely distributed inference engine for graphical models. *JavaBayes* is written in Java, and contains algorithms that compute posterior marginals, posterior expectations, and maximum a posteriori explanations in Bayesian networks [21]. The user interacts with the system through a graphical interface, where variables and densities can be added, deleted or edited. The user can specify local credal sets and global neighborhoods (ϵ-contaminated, density bounded, total variation and density ratio credal sets are supported), both to model imprecise beliefs or to perform robustness analysis. *JavaBayes* implements an enumeration algorithm for type-1 extensions, employing d-separation and the CCM transform to produce posterior envelopes. Documentation, code and examples for *JavaBayes* can be downloaded from http://www.cs.cmu.edu/˜javabayes/.

Natural extensions can be processed by the *qb* procedure.[3] The *qb* procedure performs linear fractional programming on a matricial description of a Quasi-Bayesian network; a system like *JavaBayes* must be used to create the network, and the user must insert the constraints that define a natural extension. Code, comments and examples for the *qb* procedure can be downloaded from http://www.cs.cmu.edu/˜qbayes/RobustInferences/Matlab/.

To illustrate the results and algorithms described previously, a simple example is discussed in the remainder of this section. This example is based on the example described by Charniak [13] and on the calculations presented by Walley [73, Section 9.3.4]; calculations were performed using *JavaBayes* and *qb*.

Consider the graph in Figure 1. There are five binary variables in the graph (the superscript c indicates negation). Note that the probabilities for f and b are not specified exactly; instead, the credal set for f and b is the interval $[0.4, 0.5]$. The question is how to evaluate the impact of this imprecision in probability values. To illustrate the various algorithms discussed in the paper, consider the calculation of $\underline{p}(d|l)$ and $\overline{p}(d|l)$.

Type-1 extension The simplest method to obtain the bounds is to identify the vertices of the local credal sets and generate a type-1 extension. The type-1 extension has four vertices, because both the credal sets $K(f)$ and $K(b)$ have

[3] The procedure *qb* is written in the Matlab™ language; information about Matlab™, a popular mathematical package, is available at http://www.matlab.com.

vertices $[0.4, 0.6]$ and $[0.5, 0.5]$. By calculating $p(d|l)$ for these four vertices (using *JavaBayes*), the bounds on $p(d|l)$ are obtained: $\underline{p}(d|l) = 0.3861$ and $\overline{p}(d|l) = 0.4461$.

Natural extension without irrelevance relations If no irrelevance relation is stated concerning the network, then the expressions in Figure 1 and the unitary constraint are the only restrictions on the natural extension. To generate lower and upper bounds on $p(d|l)$, it is necessary to write these thirteen linear constraints (nine are equality constraints and four are inequality constraints) and solve a linear fractional program with the objective function $p(d, l) / p(l)$. The solution of this program indicates that $\underline{p}(d|l) \leq 10^{-8}$ and $\overline{p}(d|l) \geq 1 - 10^{-8}$, demonstrating that the absence of irrelevance relations can lead to inferences that are essentially devoid of information.

Natural extension with irrelevance relations Consider the effect of adding irrelevance relations, in particular the statement that the nondescendants of a variable are irrelevant to the variable given the parents of the variable. A replication procedure generates the inequalities: $0.4 \leq p(f|b) \leq 0.5$, $0.4 \leq p(b|f) \leq 0.5$, $0.4 \leq p(f|b^c) \leq 0.5$ and $0.4 \leq p(b|f^c) \leq 0.5$. To simplify the calculation of lower and upper bounds, Theorem 10 can be used. The value of $\overline{p}(d|l)$ is obtained by solving the program:

$$\max \frac{0.48w_1 + 0.06w_2 + 0.005w_3 + 0.035w_4}{0.6w_1 + 0.6w_2 + 0.05w_3 + 0.05w_4}$$

$$\text{s.t.} \quad \begin{bmatrix} -3 & 0 & 2 & 0 \\ 2 & 0 & -2 & 0 \\ 0 & -3 & 0 & 2 \\ 0 & 2 & 0 & -2 \\ -3 & 2 & 0 & 0 \\ 2 & -2 & 0 & 0 \\ 0 & 0 & -3 & 2 \\ 0 & 0 & 2 & -2 \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \sum w_i = 1,$$

where $w_1 = p(f, b)$, $w_2 = p(f, b^c)$, $w3 = p(f^c, b)$, $w_4 = p(f^c, b^c)$. This program produces $\overline{p}(d|l) = 0.4509$. By minimization, $\underline{p}(d|l) = 0.3818$ is obtained. Note that these bounds are different from the bounds obtained by type-1 extension.

Natural extension with independence constraints The strongest statement considered here is the independence of a variable from its nondescendants given its parents. The natural extension is defined from the full joint credal set $K(f, b)$. The credal set $K(f, b)$ has has six vertices; each vertex is defined by a density $[w_1, w_2, w_3, w_4]$. The densities are: $(1/4)[1, 1, 1, 1]$, $(1/25)[4, 6, 6, 9]$, $(1/10)[2, 2, 3, 3]$, $(1/10)[2, 3, 2, 3]$, $(1/9)[2, 2, 2, 3]$, $(1/11)[2, 3, 3, 3]$. Computation of $p(d|l)$ with each of these six densities leads to $\underline{p}(d|l) = 0.3818$ and $\overline{p}(d|l) = 0.4509$.

ϵ-contaminated neighborhood A different strategy is to fix $p(f)$ and $p(b)$, say at 0.45, and use a global neighborhood to study perturbations in probability values. If an ϵ-contaminated neighborhood is taken with $\epsilon = 0.05$, calculations in *JavaBayes* yield $\underline{p}(d|l) = 0.3955$ and $\overline{p}(d|l) = 0.4455$.

15 Conclusion

The central contribution of this paper is a solid theory of graphical models of inference, a theory that combines Quasi-Bayesian theory and Walley's definitions of irrelevance and independence. Quasi-Bayesian networks offer great flexibility in dealing with imprecise probability elicitation, incomplete models, and robustness assessment.

The main technical contributions are novel results and algorithms for inferences with type-1 and natural extensions. Type-1 extensions were tied to graphical d-separation and emerged as straightforward generalizations of standard Bayesian networks. Several algorithms for type-1 extensions were unified in the context provided by Quasi-Bayesian theory. Natural extensions were analyzed through linear fractional programming, and novel properties of natural extensions were derived and used to simplify inference algorithms.

The paper also addresses a number of important issues, such as global neighborhoods, lower and upper variances, and decision-making, but it leaves a large number of questions open. How to elicit information about credal sets from experts? How to handle sequential decision-making in the framework of Quasi-Bayesian theory? These questions are to be addressed in future investigations.

The theory of locally defined Quasi-Bayesian networks presented in this paper poses an intellectually challenging question: Should we consider irrelevance or independence as a basic notion in the treatment of uncertainty? Both notions agree in standard probability theory, but they disagree in Quasi-Bayesian theory. Irrelevance can be used to define independence, and irrelevance judgements are less demanding than independence ones but still quite powerful. Should irrelevance be a more fundamental notion?

A Proofs

PROOF (Theorem 6). Consider three arbitrary disjoint sets of variables in the network, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, such that $\mathbf{X}$ is d-separated from $\mathbf{Z}$ given $\mathbf{Y}$. Take the type-1 extension $K(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ and obtain, by conditioning, $K(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$ and $K(\mathbf{X}|\mathbf{Y})$. Call $\text{ext}K$ the set of extreme points of a credal set K.

Given any bounded function $f(\mathbf{X})$, its lower expectation is attained at an extreme point of the type-1 extension:

$$\underline{E}[f(\mathbf{X})|\mathbf{Y}, \mathbf{Z}] = \min_{P \in \text{ext}\, K(\mathbf{X}|\mathbf{Y},\mathbf{Z})} \left(\sum_{\mathbf{X}} f(\mathbf{X}) p(\mathbf{X}|\mathbf{Y}, \mathbf{Z}) \right).$$

Because every such extreme point satisfies Expression (1), every extreme point satisfies $p(\mathbf{X}|\mathbf{Y}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{Y})$ (by d-separation). Then:

$$\underline{E}[f(\mathbf{X})|\mathbf{Y}, \mathbf{Z}] = \min_{P \in \text{ext}\, K(\mathbf{X}|\mathbf{Y})} \left(\sum_{\mathbf{X}} f(\mathbf{X}) p(\mathbf{X}|\mathbf{Y}) \right) = \underline{E}[f(\mathbf{X})|\mathbf{Y}].$$

Because a lower expectation uniquely defines a convex set of measures (Section 4), the lower expectation $\underline{E}[f(\mathbf{X})|\mathbf{Y}]$ uniquely defines $K(\mathbf{X}|\mathbf{Y})$ and the lower expectation $\underline{E}[f(\mathbf{X})|\mathbf{Y}, \mathbf{Z}]$ uniquely defines $K(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$. Because both lower expectations are equal for arbitrary $f(\mathbf{X})$, the underlying credal sets are equal. This argument guarantees that $\mathbf{Z}$ is irrelevant to $\mathbf{X}$ given $\mathbf{Y}$; an analogue argument proves that $\mathbf{X}$ is irrelevant to $\mathbf{Z}$ given $\mathbf{Y}$. So $\mathbf{X}$ is independent of $\mathbf{Z}$ given $\mathbf{Y}$.

PROOF (Lemma 8). Given any bounded function $f(X_i)$:

$$\begin{aligned}
\underline{E}[f(X_i)|\text{pa}(X_i)] &= \min_{P \in K(\mathbf{X})} \left(E[E[f(X_i)|\text{pa}(X_i), \mathbf{S}]] \right) \\
&\geq \underline{E}[f(X_i)|\text{pa}(X_i), \mathbf{S}] = \min_{P \in K(\mathbf{X})} \left(E[E[f(X_i)|\text{nd}(X_i)]] \right) \\
&\geq \underline{E}[f(X_i)|\text{nd}(X_i)] \, .
\end{aligned}$$

Because $\underline{E}[f(X_i)|\mathrm{pa}(X_i)] = \underline{E}[f(X_i)|\mathrm{nd}(X_i)]$ by hypothesis, $\underline{E}[f(X_i)|\mathrm{pa}(X_i)] = \underline{E}[f(X_i)|\mathrm{pa}(X_i), \mathbf{S}]$. Because these lower expectations are equal for arbitrary $f(X_i)$, the credal sets $K(X_i|\mathrm{pa}(X_i))$ and $K(X_i|\mathrm{pa}(X_i), \mathbf{S})$ are equal.

PROOF (Theorem 9). Every measure in $K_2(\mathbf{S})$ satisfies all irrelevance relations in $\mathbf{S}$ (Lemma 8), so every measure in $K_2(\mathbf{S})$ belongs to $K_2'(\mathbf{S})$ (i.e., $K_2(\mathbf{S}) \subseteq K_2'(\mathbf{S})$). To prove that every measure in $K_2'(\mathbf{S})$ belongs to $K_2(\mathbf{S})$, construct a credal set $K''(\mathbf{X})$ defined by all densities $p''(\mathbf{X})$ such that:

$$p''(\mathbf{X}) = p'(\mathbf{S}) \left(\prod_{X_i \in \{\mathbf{X} \backslash \mathbf{S}\}} p(X_i|\mathrm{pa}(X_i)) \right), \tag{A.1}$$

where $p'(\mathbf{S})$ belongs to $K_2'(\mathbf{S})$ and each $p(X_i|\mathrm{pa}(X_i))$ belongs to a local credal set $K(X_i|\mathrm{pa}(X_i))$. Expression (A.1) defines the type-1 extension of $K_2'(\mathbf{S})$ and all the $K(X_i|\mathrm{pa}(X_i))$ for $X_i \in \{\mathbf{X} \backslash \mathbf{S}\}$. Suppose that a variable X_i is in $\{\mathbf{X} \backslash \mathbf{S}\}$, and an arbitrary bounded function $f(X_i)$ is given. By Theorem 6, X_i is d-separated from $\mathbf{S}$ by the nondescendants of X_i, so the lower expectation $\underline{E}[f(X_i)|\mathrm{nd}(X_i)]$ is equal to $\underline{E}[f(X_i)|\mathrm{pa}(X_i)]$ (with respect to $K''(\mathbf{X})$). As $f(X_i)$ is arbitrary, the marginal credal set $K''(X_i|\mathrm{nd}(X_i))$ is equal to $K(X_i|\mathrm{pa}(X_i))$. Suppose now that X_i is in $\mathbf{S}$, and an arbitrary bounded function $f(X_i)$ is given. By Theorem 6, X_i is d-separated from all variables outside of $\mathbf{S}$, so $\underline{E}[f(X_i)|\mathrm{nd}(X_i)] = \underline{E}[f(X_i)|\mathrm{nd}(X_i) \cap \mathbf{S}]$ (with respect to $K''(\mathbf{X})$). By hypothesis, the last lower expectation is equal to $\underline{E}[f(X_i)|\mathrm{pa}(X_i)]$, so the credal set $K''(X_i|\mathrm{nd}(X_i))$ is equal to $K(X_i|\mathrm{pa}(X_i))$. This proves that $K''(\mathbf{X}) \subseteq K_2(\mathbf{X})$. Because $K_2'(\mathbf{S})$ is the exact marginal of $K''(\mathbf{X})$, $K_2'(\mathbf{S}) \subseteq K_2(\mathbf{S})$. This proves that $K_2'(\mathbf{S})$ and $K_2(\mathbf{S})$ are equal.

PROOF (Theorem 10). Denote by $\mathbf{W}$ the variables that are outside of $\mathbf{S}$ (i.e., $\mathbf{W} = \mathbf{X} \backslash \mathbf{S}$), and define the density

$$p(\mathbf{W}|\mathbf{S}) = \prod_{X_i \in \mathbf{W}} p(X_i|\mathrm{pa}(X_i)) \,.$$

Note that $p(\mathbf{W}|\mathbf{S})$ is the unique joint measure for $\mathbf{W}$ given $\mathbf{S}$. Uniqueness is guaranteed by the fact that the variables in $\mathbf{W}$ form a Bayesian network: 1) irrelevance is equal to independence in standard Bayesian networks; 2) Lemma 8 guarantees that all irrelevance conditions are valid when restricted to the network of $\mathbf{W}$; 3) independence of a variable from its nondescendants given its parents characterizes a unique Bayesian network [59].

By Theorem 9, the natural extension for the sub-network containing variables $\mathbf{S}$ defines the set of all $p(\mathbf{S})$; any joint density $p(\mathbf{X})$ that belongs to $K_2(\mathbf{X})$

satisfies $p(\mathbf{X}) = p(\mathbf{W}|\mathbf{S})\, p(\mathbf{S})$. Using this fact in Expression (2):

$$\overline{P}(X_Q|E) = \max \left(\frac{\sum_{\mathbf{X}\backslash\mathbf{X}_{QE}} p(\mathbf{W}|\mathbf{S})\, p(\mathbf{S})}{\sum_{\mathbf{X}\backslash\mathbf{X}_{E}} p(\mathbf{W}|\mathbf{S})\, p(\mathbf{S})} \right) = \max \left(\frac{\sum_{\mathbf{S}\backslash\mathbf{X}_{QE}} h(X_Q, \mathbf{S}) p(\mathbf{S})}{\sum_{\mathbf{S}\backslash\mathbf{X}_{E}} h(X_Q, \mathbf{S}) p(\mathbf{S})} \right).$$

PROOF (Lemma 12). Consider an arbitrary bounded function $f(\mathbf{S})$. Because every marginal measure $P(\mathbf{S}|\mathrm{pa}(X_i), X_i)$ obtained from the credal set $K(\mathrm{nD}(X_i)|\mathrm{pa}(X_i), X_i)$ belongs to the credal set $K(\mathbf{S}|\mathrm{pa}(X_i), X_i)$, the lower expectation $\underline{E}[f(\mathbf{S})|\mathrm{pa}(X_i), X_i]$ has the same value with respect to the credal sets $K(\mathbf{S}|\mathrm{pa}(X_i), X_i)$ and $K(\mathrm{nD}(X_i)|\mathrm{pa}(X_i), X_i)$. The same argument is valid for $\underline{E}[f(\mathbf{S})|\mathrm{pa}(X_i)]$. Because $\underline{E}[\mathrm{nD}(X_i)|\mathrm{pa}(X_i)]$ and $\underline{E}[\mathrm{nD}(X_i)|\mathrm{pa}(X_i), X_i]$ are equal by hypothesis and the function $f(\mathbf{S})$ is arbitrary, $K(\mathbf{S}|\mathrm{pa}(X_i), X_i)$ and $K(\mathbf{S}|\mathrm{pa}(X_i))$ are equal.

PROOF (Theorem 13). Every measure in $K_3(\mathbf{S})$ satisfies all independence relations in $\mathbf{S}$ (Lemmas 8 and 12), so every measure in $K_3'(\mathbf{S})$ belongs to $K_3(\mathbf{S})$ (i.e., $K_3'(\mathbf{S}) \subseteq K_3(\mathbf{S})$). To prove that every measure in $K_3'(\mathbf{S})$ belongs to $K_3(\mathbf{S})$, construct a credal set $K''(\mathbf{X})$ following Expression (A.1). Suppose that a variable X_i is in $\{\mathbf{X}\backslash\mathbf{S}\}$, and arbitrary bounded functions $f(X_i)$ and $g(\mathrm{nD}(X_i))$ are given. By Theorem 6, X_i is d-separated from $\mathbf{S}$ by the nondescendants of X_i, so the lower expectation $\underline{E}[f(X_i)|\mathrm{nd}(X_i)]$ is equal to $\underline{E}[f(X_i)|\mathrm{pa}(X_i)]$ and the lower expectation $\underline{E}[g(\mathrm{nD}(X_i))|\mathrm{pa}(X_i), X_i]$ is equal to $\underline{E}[g(\mathrm{nD}(X_i))|\mathrm{pa}(X_i)]$. As $f(X_i)$ is arbitrary, the credal set $K''(X_i|\mathrm{nd}(X_i))$ is equal to $K(X_i|\mathrm{pa}(X_i))$; as $g(\mathrm{nD}(X_i))$ is arbitrary, the credal set $K''(\mathrm{nD}(X_i)|\mathrm{pa}(X_i), X_i)$ is equal to $K''(\mathrm{nD}(X_i)|\mathrm{pa}(X_i))$. Suppose now that X_i is in $\mathbf{S}$, and arbitrary bounded functions $f(X_i)$ and $g(\mathrm{nD}(X_i))$ are given. By Theorem 9, the nondescendants of X_i are irrelevant to X_i given the parents of X_i. It remains to be shown that for X_i in $\mathbf{S}$, X_i is irrelevant to its nondescendants given its parents. To prove that, divide $\mathrm{nD}(X_i)$ in two collections of variables: $\mathbf{S}'(X_i) = \mathrm{nD}(X_i) \cap \mathbf{S}$ (the elements of $\mathrm{nD}(X_i)$ that belong to $\mathbf{S}$) and $\mathbf{W}'(X_i) = \mathrm{nD}(X_i) \cap \{\mathbf{X}\backslash\mathbf{S}\}$ (the elements of $\mathrm{nD}(X_i)$ that do not belong to $\mathbf{S}$). By Theorem 6, X_i is d-separated from all variables outside of $\mathbf{S}$, so:

$$\begin{aligned}
&\underline{E}[g(\mathrm{nD}(X_i))|\mathrm{pa}(X_i), X_i] \\
&\quad = \min_{p''(\mathbf{X}) \in K''(\mathbf{X})} E[E[g(\mathbf{S}'(X_i), \mathbf{W}'(X_i))|\mathbf{S}'(X_i), \mathrm{pa}(X_i), X_i]\,|\mathrm{pa}(X_i), X_i] \\
&\quad = \min_{p''(\mathbf{X}) \in K''(\mathbf{X})} E[E[g(\mathbf{S}'(X_i), \mathbf{W}'(X_i))|\mathbf{S}'(X_i), \mathrm{pa}(X_i)]\,|\mathrm{pa}(X_i), X_i]\,.
\end{aligned}$$

Note that $E[g(\mathbf{S}'(X_i), \mathbf{W}'(X_i))|\mathbf{S}'(X_i), \mathrm{pa}(X_i)]$ is itself a function of $\mathbf{S}'(X_i)$. By hypothesis:

$$
\begin{aligned}
&\underline{E}[g(\mathrm{nD}(X_i))|\mathrm{pa}(X_i), X_i] \\
&\quad = \min_{p''(\mathbf{X}) \in K''(\mathbf{X})} E[E[g(\mathbf{S}'(X_i), \mathbf{W}'(X_i))|\mathbf{S}'(X_i), \mathrm{pa}(X_i)]\,|\mathrm{pa}(X_i)] \\
&\quad = \underline{E}[g(\mathrm{nD}(X_i)|\mathrm{pa}(X_i)]\,.
\end{aligned}
$$

Consequently, $K''(\mathrm{nd}(X_i)|\mathrm{pa}(X_i), X_i)$ is equal to $K''(\mathrm{nd}(X_i)|\mathrm{pa}(X_i))$. This proves that $K''(\mathbf{X}) \subseteq K_3(\mathbf{X})$. Because $K_3'(\mathbf{S})$ is the exact marginal of $K''(\mathbf{X})$, $K_3'(\mathbf{S}) \subseteq K_3(\mathbf{S})$. This proves that $K_3'(\mathbf{S})$ and $K_3(\mathbf{S})$ are equal.

B Classes of finitely generated credal sets

This section collects results related to several classes of finitely generated credal sets. All results are known but are scattered throughout the literature.

A *Choquet capacity* is a set-function with values from 0 to 1, such that the empty set has value 0 and the union of all sets has value 1 [36]. A Choquet capacity is *2-monotone* if $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B) - \underline{P}(A \cap B)$. Any 2-monotone Choquet capacity defines a credal set containing all probability measures such that $P(A) \geq \underline{P}(A)$ for all events A. All credal sets in this section, except the density ratio class, are 2-monotone Choquet capacities.

When a credal set generates a 2-monotone Choquet capacity and has the property that any measure P such that $P(A) \geq \underline{P}(A)$ belongs to the set, lower and upper expectations for any bounded function $f(X)$ are produced by the *Choquet integrals* [73, Section 3.2.4]. Suppose $f(X)$ is a bounded function in a discrete domain and $f(X)$ has m_1 non-negative values and m_2 non-positive values. Define f_i' as the ith smallest non-negative value of $f(X)$, f_i'' as the ith largest non-positive value of $f(X)$, and $f_0' = 0$, $f_0'' = 0$. The Choquet integrals of $f(X)$ are:

$$
\begin{aligned}
\underline{E}[f(X)] &= \sum_{i=0}^{m_1-1} (f_{i+1}' - f_i')\underline{P}(f(X) > f_i') + \sum_{i=0}^{m_2-1} (f_{i+1}'' - f_i'')\overline{P}(f(X) < f_i'')\,, \\
\overline{E}[f(X)] &= \sum_{i=0}^{m-1} (f_{i+1}' - f_i')\overline{P}(f(X) > f_i') + \sum_{i=0}^{m_2-1} (f_{i+1}'' - f_i'')\underline{P}(f(X) < f_i'')\,.
\end{aligned}
\tag{B.1}
$$

Posterior envelopes can be calculated in closed-form for 2-monotone lower envelopes [14,34,80]:

$$
\underline{P}(A|B) = \frac{\underline{P}(A,B)}{\underline{P}(A,B) + \overline{P}(A^c,B)}, \quad \overline{P}(A|B) = \frac{\overline{P}(A,B)}{\overline{P}(A,B) + \underline{P}(A^c,B)}. \tag{B.2}
$$

B.1 The ϵ-contaminated class

An ϵ-contaminated class is characterized by a probability density $r(X)$ and a real number $\epsilon \in (0, 1)$; the class is defined by all probability densities $p(X)$ such that $p(X) = (1 - \epsilon)r(X) + \epsilon q(X)$, where $q(X)$ can be any density. An ϵ-contaminated class defines a credal set containing all densities $p(X) \geq (1 - \epsilon)r(x)$ [8,35]. An ϵ-contaminated class has vertices defined by the densities $(1 - \epsilon)r(X) + \epsilon I_{X_j}(X)$, for all the values X_j that the variable X may assume ($I_{X_j}(X)$ is the indicator function of X_j).

B.2 The belief function and sub-sigma classes

Consider a discrete variable X with a finite set of values $\hat{X}$; call $\mathcal{X}$ the set of subsets of $\hat{X}$. A *basic mass assignment* is a probability measure over $\mathcal{X}$. A *belief function* is defined from a basic mass assignment as $\mathrm{Bel}(A) = \sum_{B \subset A} m(B)$ [65]. Conversely, a basic mass assignment can be extracted from a belief function using the expression $m(A) = \sum_{B \subset A} (-1)^{|A-B|} \mathrm{Bel}(B)$ (called the Möbius transform [42]). A belief function represents a credal set containing all probability measures such that $P(A) \geq \mathrm{Bel}(A)$ for all A [79]. To generate the finitely many vertices of this credal set, concentrate the non-zero basic mass assignment of a set A into each one of the subsets of A. Lower and upper expectations for a function $f(X)$ can be easily obtained for a belief function class [79]:

$$\underline{E}[f(X)] = \sum_{A \in \mathcal{X}} \left(\min_{X \in A} f(X) \right) m(A), \quad \overline{E}[f(X)] = \sum_{A \in \mathcal{X}} \left(\max_{X \in A} f(X) \right) m(A).$$

A *sub-sigma class* is a credal set generated by a particular type of basic mass assignment [7,44,52]. Consider a variable X with values $\hat{X}$; a sub-sigma class is generated by an assignment that only puts non-zero probability masses to non-overlapping subsets of $\hat{X}$.

B.3 The density bounded class

A density bounded class is the set of all probability measures $P(A)$ so that $L(A) \leq P(A) \leq U(A)$; $L(A)$ and $U(A)$ are arbitrary non-negative measures so that $L(\cup A) \leq 1$ and $U(\cup A) \geq 1$ [46,47]. The lower and upper envelopes for an event A are respectively $\max(L(A), 1 - U(A^c))$ and $\min(U(A), 1 - L(A^c))$.

B.4 The total variation class

The total variation class is the set of probability measures $P(A)$ so that [77]: $|P(A) - R(A)| \leq \epsilon$ for any event A, where $R(A)$ is a given probability measure and ϵ is a real number in the interval $(0, 1)$. The lower and upper envelopes for an event A are respectively $\max(P(A) - \epsilon, 0)$ and $\min(P(A) + \epsilon, 1)$.

B.5 The density ratio class

A density ratio class is the set of probability measures $P(A)$ so that [22]:

$$\frac{L(A)}{U(B)} \leq \frac{P(A)}{P(B)} \leq \frac{U(A)}{L(B)}, \quad \text{for any events } A,\ B,$$

where $L(A)$ and $U(A)$ are arbitrary positive measures such that $L(A) \leq U(A)$ for any event A. An equivalent formulation for this class is the set of probability measures such that $L(A) \leq \alpha P(A) \leq U(A)$, for some $\alpha > 0$. Any density ratio class can be expressed as the union of a collection of density bounded classes [78]. Bounds on the posterior probability for an event A are [22]:

$$\underline{P}(A) = \frac{L(A)}{L(A) + U(A^c)}, \quad \overline{P}(A) = \frac{U(A)}{U(A) + L(A^c)}. \tag{B.3}$$

For an arbitrary function f, the upper and lower expectations are respectively the solution of the non-linear equations in λ [22]:

$$\sum_j \left(U(X_j) \max(f(X_j) - \lambda, 0)\right) + \sum_j \left(L(X_j) \min(f(X_j) - \lambda, 0)\right) = 0, \tag{B.4}$$

$$\sum_j \left(L(X_j) \max(f(X_j) - \lambda, 0)\right) + \sum_j \left(U(X_j) \min(f(X_j) - \lambda, 0)\right) = 0. \tag{B.5}$$

A bracketing algorithm can be constructed for these expressions because these equations have unique roots in the interval between the smallest and largest values of $f(X)$.

Acknowledgements

I thank Eric Krotkov for substantial support during the research that led to this work, Lonnie Chrisman for reading an earlier draft and suggesting substantial improvements, and Teddy Seidenfeld for teaching me how to understand the theory of convex sets of probability measures. Thanks to Bruce D'Ambrosio, Sreekanth Nagarajan, Chao-Lin Liu, Rina Dechter, and Akihiro Shinmori for important help with the *JavaBayes* system.

References

[1] K. A. Andersen and J. N. Hooker, Bayesian logic, *Decision Support Systems*, **11** (1994) 191–210.

[2] M. Avriel, *Advances in Geometric Programming* (Plenum Press, New York, 1980).

[3] M. Avriel, R. Dembo, and U. Passy, Solution of generalized geometric programs, in: M. Avriel, ed., *Advances in Geometric Programming* (Plenum Press, New York, 1980) 203–226.

[4] F. Bacchus, *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach* (MIT Press, Cambridge, 1990).

[5] J. Berger and E. Moreno, Bayesian robustness in bidimensional model: Prior independence, *Journal of Statistical Planning and Inference* **40** (1994) 161–176.

[6] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer-Verlag, 1985).

[7] J. O. Berger, Robust Bayesian analysis: Sensitivity to the prior, *Journal of Statistical Planning and Inference* **25** (1990) 303–328.

[8] J. S. Breese and K. W. Fertig, Decision making with interval influence diagrams, in: P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, eds., *Uncertainty in Artificial Intelligence 6* (Elsevier Science, North-Holland, 1991) 467–478.

[9] W. L. Buntine, Operations for learning with graphical models, *Journal of Artificial Intelligence Research* **2** (1994) 159–225.

[10] A. Cano, J. E. Cano, and S. Moral, Convex sets of probabilities propagation by simulated annealing, in: *Proceedings IPMU* (1994) 4–8.

[11] A. Cano and S. Moral, A genetic algorithm to approximate convex sets of probabilities, in: *Proceedings IPMU* (1996) 859–864.

[12] J. Cano, M. Delgado, and S. Moral, An axiomatic framework for propagating uncertainty in directed acyclic networks, *International Journal of Approximate Reasoning* **8** (1993) 253–280.

[13] E. Charniak, Bayesian networks without tears, *AI Magazine* (Fall 1991) 50–63.

[14] L. Chrisman, Incremental conditioning of lower and upper probabilities, *International Journal of Approximate Reasoning* **13** (1) (1995) 1–25.

[15] L. Chrisman, Independence with lower and upper probabilities, in: *Proceedings XII Uncertainty in Artificial Intelligence Conference* (1996) 169–177.

[16] L. Chrisman, Propagation of 2-monotone lower probabilities on an undirected graph, in: *Proceedings XII Uncertainty in Artificial Intelligence Conference* (1996) 178–186.

[17] K. L. Clarkson, K. Mehlhorn, and R. Seidel, Four results on randomized incremental constructions, in: *Proceedings STACS* (1992) 463–474.

[18] G. F. Cooper, The computational complexity of probabilistic inference using Bayesian belief networks, *Artificial Intelligence* **42** (1990) 393–405.

[19] F. G. Cozman, Robustness analysis of Bayesian networks with local convex sets of distributions, in: *Proceedings XIII Uncertainty in Artificial Intelligence Conference* (1997) 108-115.

[20] L. de Campos and S. Moral, Independence concepts for convex sets of probabilities, in: *Proceedings XI Uncertainty in Artificial Intelligence Conference* (1995) 108–115.

[21] R. Dechter, Bucket elimination: A unifying framework for probabilistic inference, in: *Proceedings XII Uncertainty in Artificial Intelligence Conference* (1996) 211–219.

[22] L. DeRobertis and J. A. Hartigan, Bayesian inference using intervals of measures, *The Annals of Statistics* **9** (2) (1981) 235–244.

[23] R. J. Duffin and E. L. Peterson, Geometric programming with signomials, *Journal of Optimization Theory and Applications* **11** (1) (1973) 3–35.

[24] J. Falk and R. Soland, An algorithm for separable nonconvex programming problems, *Management Science* **15** (1969) 550–569.

[25] T. L. Fine, Lower probability models for uncertainty and nondeterministic processes, *Journal of Statistical Planning and Inference* **20** (1988) 389–411.

[26] J. Gebhardt and R. Kruse, Learning possibilistic networks from data, in: *Proceedings Fifth International Workshop on Artificial Intelligence and Statistics* (1995) 233–243.

[27] D. Geiger, T. Verma, and J. Pearl, d-separation: from theorems to algorithms, in: M. Henrion and R. D. Shachter and L. N. Kanal and J. F. Lemmer, eds., *Uncertainty in Artificial Intelligence 5* (Elsevier Science, 1990) 139–148.

[28] F. J. Giron and S. Rios, Quasi-Bayesian behaviour: A more realistic approach to decision making? in: J. M. Bernardo, J. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds., *Bayesian Statistics* (University Press, Valencia, Spain, 1980) 17–38.

[29] W. Gochet and Y. Smeers, A brach-and-bound method for reversed geometric programming, *Operations Research* **27** (5) (1979) 983–996.

[30] I. J. Good, *Good Thinking: The Foundations of Probability and its Applications* (University of Minnesota Press, Minneapolis, 1983).

[31] P. Gritzmann and V. Klee, Mathematical programming and convex geometry, in: P. Gruber and J. M. Wills, eds., *Handbook of Convex Geometry A* (Elsevier Science Publishers, 1993) 627–674.

[32] H. E. Kyburg Jr, Bayesian and non-Bayesian evidential updating, *Artificial Intelligence* **31** (1987) 271–293.

[33] V. Ha and P. Haddawy, Theoretical foundations for abstraction-based probabilistic planning, in: *Proceedings XII Uncertainty in Artificial Intelligence Conference* (1996) 291–298.

[34] J. Y. Halpern and R. Fagin, Two views of belief: Belief as generalized probability and belief as evidence, *Artificial Intelligence* **54** (1992) 275–317.

[35] T. Herron, T. Seidenfeld, and L. Wasserman, Divisive conditioning: Further results on dilation, Technical Report 585, Department of Statistics, Carnegie Mellon University (1993).

[36] P. J. Huber, *Robust Statistics* (Wiley, New York, 1980).

[37] T. Ibaraki, Solving mathematical programming problems with fractional objective functions, in: S. Schaible and W. T. Ziemba, eds., *Generalized Concavity in Optimization and Economics* (Academic Press, 1981) 440–472.

[38] B. Jaumard, P. Hansen, and M. P. de Aragão, Column generation methods for probabilistic logic, *ORSA Journal on Computing* **3** (2) (1991) 135–148.

[39] F. V. Jensen, *An Introduction to Bayesian Networks* (Springer Verlag, New York, 1996).

[40] J. B. Kadane, *Robustness of Bayesian Analyses*, volume 4 of *Studies in Bayesian econometrics* (Elsevier Science Pub. Co., New York, 1984).

[41] R. B. Kearfott, A review of techniques in the verified solution of constrained global optimization problems, in: R. B. Kearfott and V. Kreinovich, eds., *Applications of Interval Computations* (Kluwer, Dordrecht, Netherlands, 1996) 23–60.

[42] R. Kennes, Computational aspects of the Möbius transformation of graphs, *IEEE Transactions on Systems, Man and Cybernetics* **22** (2) (1992) 201–223.

[43] H. E. Kyburg and M. Pittarelli, Set-based Bayesianism, *IEEE Transactions on Systems, Man and Cybernetics A* **26** (3) (1996) 324–339.

[44] D. Lambert and G. T. Duncan, Single-parameter inference based on partial prior information, *The Canadian Journal of Statistics* **14** (4) (1986) 297–305.

[45] K. B. Laskey, Sensitivity analysis for probability assessments in Bayesian networks, *IEEE Transactions on Systems, Man, and Cybernetics* **25** (6) (1995) 901–909.

[46] M. Lavine, Sensitivity in Bayesian statistics, the prior and the likelihood, *Journal of the American Statistical Association* **86** (414) (1991) 396–399.

[47] J. F. Lemmer and H. E. Kyburg Jr., Conditions for the existence of belief functions corresponding to intervals of belief, in: *Proceedings 9th National Conference on Artificial Intelligence* (1991) 488–493.

[48] I. Levi, *The Enterprise of Knowledge* (The MIT Press, Cambridge, Massachusetts, 1980).

[49] D. G. Luenberger, *Linear and Nonlinear Programming* (Addison-Wesley, Reading, Mass., 1989).

[50] T. Lukasiewicz, Uncertain reasoning in concept lattices, in: *Proceedings 3rd European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty* (1995) 293–300.

[51] C. Luo, C. Yu, J. Lobo, G. Wang, and T. Pham, Computation of best bounds of probabilities from uncertain data, *Computational Intelligence* **12** (4) (1996) 541–566.

[52] C. F. Manski, Learning and decision making when subjective probabilities have subjective domains, *The Annals of Statistics* **9** (1) (1981) 59–65.

[53] S. Moral, A formal language for convex sets of probabilities, in: *Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU)* (1993) 274–181.

[54] R. Musick, Minimal assumption distribution propagation in belief networks, in: *Proceedings IX Uncertainty in Artificial Intelligence Conference* (1993).

[55] N. J. Nilsson, Probabilistic logic, *Artificial Intelligence* **28** (1986) 71–87.

[56] A. Papoulis, *Probabilities, Random Variables and Stochastic Processes* (McGraw-Hill, New York, 1991).

[57] U. Passy, Global solutions of mathematical programs with intrisically concave functions, in: M. Avriel, ed., *Advances in Geometric Programming* (Plenum Press, New York, 1980) 355–373.

[58] J. Pearl, On probability intervals, *International Journal of Approximate Reasoning* **2** (1988) 211–216.

[59] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kauffman, San Mateo, CA, 1988).

[60] E. H. Ruspini, The logical foundations of evidential reasoning, Technical Report SRIN408, SRI International (1987).

[61] S. Russell, J. Binder, D. Koller, and K. Kanazawa, Local learning in probabilistic networks with hidden variables, in: *Proceedings Fourteenth International Joint Conference on Artificial Intelligence* (1995) 1146-1152.

[62] S. J. Russell and P. Norvig, *Artificial Intelligence: a Modern Approach* (Prentice Hall, New Jersey, 1995).

[63] S. I. Schaible and W. T. Ziemba, *Generalized Concavity in Optimization and Economics* (Academic Press, 1981).

[64] T. Seidenfeld, Outline of a theory of partially ordered preferences, *Philosophical Topics* **21** (1) (1993) 173–188.

[65] G. Shafer, *A mathematical theory of evidence* (Princeton University Press, 1976).

[66] P. P. Shenoy and G. Shafer, Axioms for probability and belief-function propagation, in: R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, eds., *Uncertainty in Artificial Intelligence 4*, (Elsevier Science Publishers, North-Holland, 1990) 169–198.

[67] E. H. Shortliffe and B. G. Buchanan, *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project* (Addison-Wesley, Reading, Mass., 1985).

[68] C. A. B. Smith, Consistency in statistical inference and decision, *Journal Royal Statistical Society B* **23** (1961) 1–25.

[69] P. Spirtes, C. Glymour, and R. Scheines, *Causality, Prediction, and Search* (Springer-Verlag, New York, 1993).

[70] P. Suppes, The measurement of belief, *Journal Royal Statistical Society B* **2** (1974) 160–191.

[71] B. Tessem, Interval probability propagation, *International Journal of Approximate Reasoning* **7** (1992) 95–120.

[72] L. van der Gaag, Computing probability intervals under independency constraints, in: P. P. Bonissone, M. Henrion, and J. F. Lemmer, eds., *Uncertainty in Artificial Intelligence 6* (Elsevier Science, 1991) 457–466.

[73] P. Walley, *Statistical Reasoning with Imprecise Probabilities* (Chapman and Hall, New York, 1991).

[74] P. Walley, Measures of uncertainty in expert systems, *Artificial Intelligence* **83** (1996) 1–58.

[75] P. Walley and T. L. Fine, Towards a frequentist theory of upper and lower probability, *The Annals of Statistics* **10** (3) (1982) 741–761.

[76] L. Wasserman, Invariance properties of density ratio priors, *The Annals of Statistics* **20** (4) (1992) 2177–2182.

[77] L. Wasserman, Recent methodological advances in robust Bayesian inference, in: J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 4* (Oxford University Press, 1992) 483–502.

[78] L. Wasserman and J. B. Kadane, Computing bounds on expectations, *Journal of the American Statistical Association* **87** (418) (1992) 516–522.

[79] L. A. Wasserman, Prior envelopes based on belief functions, *The Annals of Statistics* **18** (1) (1990) 454–464.

[80] L. A. Wasserman and J. B. Kadane, Bayes' theorem for Choquet capacities, *The Annals of Statistics* **18** (3) (1990) 1328–1339.

[81] M. Zaffalon, *Inferenze e Decisioni in Condizioni di Incertezza con Modelli Grafici Orientati*, PhD thesis (in Italian), Università di Milano (1997).