Credal 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 *credal networks*, structures that associate convex sets of probability measures with directed acyclic graphs. Credal networks are graphical models for precise/imprecise beliefs. The main contribution of this work is a theory of credal 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, and how to compute inferences in credal networks. A credal network admits several extensions — several sets of probability measures comply with the constraints represented by a network. Two types of extensions are investigated. The properties of strong extensions are clarified through a new generalization of d-separation, and exact and approximate inference methods are described for strong extensions. Novel results are presented for natural extensions, and linear fractional programming methods are described for natural extensions. The paper also investigates credal 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 [65]. But not all beliefs can be cast as sharp numeric values [6,52,68,77]. 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, variously called the theory of credal sets [52], the theory of imprecise probabilities [76] or Quasi-Bayesian theory [34], offers an alternative to standard probabilistic (Bayesian) procedures. While Bayesian theory employs a single probability measure, the theory of sets of probabilities allows a decision-maker to represent imprecise and incomplete beliefs through a set of measures.

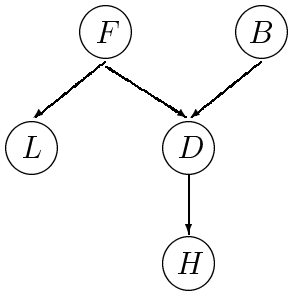
This article presents a complete theory of *credal networks*, structures that associate convex sets of probability measures with directed acyclic graphs (the author has used the term *Quasi-Bayesian networks* before, as discussed in Section 3). The main contribution of this article is a theory of credal networks that displays as much flexibility and representational power as the theory of standard Bayesian networks.

Figures 1 and 2 show two credal networks (both examples are discussed in Section 11). In a standard Bayesian network, every variable is associated with a probability measure conditional on the parents of the variable. In a credal network, variables may also be associated with probabilistic inequalities.

The key *theoretical* problem in credal networks is how to express, detect, and exploit irrelevance and independence relations; the key *practical* problem in credal networks is how to generate bounds on posterior probabilities and posterior expectations. These questions are addressed in this article through Walley's definitions of irrelevance and independence. Results in this article show how to express judgements of irrelevance and independence, and how to compute inferences in credal networks. Given a credal network, there are several sets of probability measures that comply with the constraints represented by the network; each one of these sets is an *extension* of the network. Two types of extensions are investigated. The *strong extension* of a credal network is the largest set where extreme points are Bayesian networks formed from the network. The properties of strong extensions are clarified through a new generalization of d-separation, and exact and approximate inference methods are described for strong extensions. The *natural extension* of a network is the largest set of distributions that comply with the constraints represented by the network. Novel results are presented for natural extensions, and linear fractional programming methods are described for natural extensions.

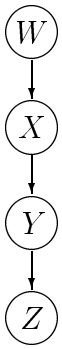
The article also investigates credal networks that are defined globally through neighborhoods of a single Bayesian network. Computation of bounds on variances and decision-making strategies are briefly discussed.

Credal 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* and the *qb* packages) that implement the algorithms described in this article. Together,



$$0.4 \leq p(f) \leq 0.5 \quad 0.4 \leq p(b) \leq 0.5$$
$$p(l|f) = 0.6 \qquad p(l|f^c) = 0.05$$
$$p(d|f,b) = 0.8 \qquad p(d|f,b^c) = 0.1$$
$$p(d|f^c,b) = 0.1 \qquad p(d|f^c,b^c) = 0.7$$
$$p(h|d) = 0.6 \qquad p(h|d^c) = 0.3$$

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



$$0.2 \leq p(w) \leq 0.3$$
$$0.1 \leq p(x|w) \leq 0.2 \quad 0.8 \leq p(x|w^c) \leq 0.9$$
$$0.4 \leq p(y|x) \leq 0.5 \quad 0.5 \leq p(y|x^c) \leq 0.6$$
$$0.7 \leq p(z|y) \leq 0.8 \quad 0.1 \leq p(z|y^c) \leq 0.2$$

Fig. 2. Example network: four binary variables (superscript c indicates negation).

these algorithms and software packages open the field of graphical models to the theory of sets of probabilities.

2 Article roadmap

Sections 4 and 5 describe concepts and results that are used throughout the article. Section 6 introduces locally defined credal networks and defines the concepts of strong extension and natural extension. Sections 7 and 8 contain the main results for strong extensions and natural extensions respectively. Section 9 briefly describes alternative ways to specify credal networks, employing neighborhoods of standard Bayesian networks. Section 10 deals with inferences for non-atomic events, computation of expectations and variances, and decision-making. Two freely available software systems for credal network inferences are described in Section 11. Section 12 contains concluding remarks. All proofs of lemmas and theorems are grouped in Appendix A.

3 Terminology and notation roadmap

There exist many theories that deal with uncertainty in various forms; unfortunately, terminology has not been properly standardized. It seems appropriate to clariy some of the terminology related to sets of probabilities even before entering into the technical contents of this article — readers that are not interested in this discussion may choose to skip this paragraph. The first and most important problem is to find a name for the theory that is of interest here. Levi has used the term *credal set* to refer to a convex set of probability measures, and the term *theory of credal sets* suggests itself [52]. Giron and Rios have used the term *Quasi-Bayesian theory* to stress the fact that the theory is not Bayesian but shares several philosophical points with Bayesian theory [34]. Walley has used the term *imprecise probabilities* in connection with a theory that deals with sets of probabilities [76]. The term Quasi-Bayesian theory has been adopted in the past, and the term *Quasi-Bayesian networks* has been used to refer to the *credal networks* of this article [20,21]. However, the development of the scope and power of the theory has suggested that the term theory of credal sets is more appropriate. The work of Zaffalon has suggested the term credal networks [84], and this term is adopted here as a better option than Quasi-Bayesian networks. There are other difficult decisions regarding terminology. The term *type-1 extensions* [20,21] is replaced by *strong extensions* in this article, as strong extensions resemble the concept of strong independence [19]. But the term strong extension may not be perfect either, as other options could have been used: decomposable extension, envelope extension, sensitivity extension. The term strong extension should be viewed as a tentative, perhaps temporary, decision. Other points are the choice between *lower envelopes* and *lower probabilities*, and the choice between *lower previsions* and *lower expectations*. It is hoped that the choices made in this article contribute to standardize terminology in a positive way.

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

Events are denoted by the first letters of the alphabet: A, B, C. The indicator function of event A defined through variable X is denoted by $I_A(X)$ (one if $x \in A$ and zero if $x \notin A$). Variables are denoted by the last letters of the alphabet (except the variables in Figure 1): W, X, Y, and Z. All variables are assumed to have finitely many values. Collections of variables are indicated in bold: $\mathbf{S}$, $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$. 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 ar-

ticle focuses on discrete variables, the function defined by $P(\{X = x_j\})$ completely characterizes a probability measure [59]. 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 $f(X)$ with respect to probability measure P is indicated by $E_P[f(X)]$. The probability of event A conditional on event B is denoted by $P(A|B)$. The notation $P(A|x_j)$ indicates the probability measure conditional on the event $\{X = x_j\}$. The expectation of a function $f(X)$ with respect to the conditional measure $P(\cdot|B)$ is denoted by $E_P[f(X)|B]$. Note that $E_P[f(X)|Y]$ represents a function of Y.

A convex set of probability measures is 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 for $\alpha \in [0, 1]$. A credal set $K(X)$ contains probability measures induced by densities $p(X)$. A credal set $K(X|B)$ contains probability measures induced by densities $p(X|B)$. The notation $K(X|Y)$ denotes a collection of credal sets indexed by Y: Each value y of Y is associated with $K(X|Y = y)$.

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 [65]. A popular graphical representation for probabilistic models is the *Bayesian network* formalism [45].

A 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 in $\mathbf{X}$; the parents of X_i are denoted by $\mathrm{pa}(X_i)$. Each variable X_i is associated with a conditional density $p(X_i|\mathrm{pa}(X_i))$, and every variable is assumed independent of its nondescendants non-parents given its parents. Such a structure induces a unique joint probability density through the following expression [62]:

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

A directed acyclic graph may display *d-separation* relations [33]. 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 two converging arrows and is not in $\mathbf{Z}$ and none of its descendants are in $\mathbf{Z}$, or W is in $\mathbf{Z}$. Then $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$. A key property of Bayesian networks is that graphical d-separation implies probabilistic independence [62, page 117].

Working with d-separation relations in Bayesian network learning, Spirtes, Glymour and Scheines have introduced the concept of *faithfulness* [72]. A probability measure is *faithful* to a directed acyclic graph when every independence relation entailed by the measure corresponds to a graphical d-separation relation. Note that Pearl used the concept of D-mappness for undirected graphs with similar meaning [62].

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 [25,61]. 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 assessments. Several non-Bayesian attempts have been made to relax this requirement through possibility theory [32], Dempster-Shafer theory [69] or even purely heuristic approaches [70]. These approaches depart radically from usual axioms of decision-making and are not pursued further in this article. A less radical non-Bayesian strategy is to associate interval-based probability values with variables in a network [8,16,31,38,41]. 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 [16,38]. 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 (both de Campos and Moral [23] and Couso et al [19] discuss several concepts of independence for credal sets). 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 [10,12,29,55,74,75]. 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]. Instead, this article starts from a definition of independence based on axioms of preferences and proceeds from this solid foundation.

5 Sets of probability measures

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 [52]. Third, we may be interested in abstracting away parts of a model and assessing the effect of this abstraction [15,39]. In such circumstances, it is advisable to consider extensions of probability theory that can handle imprecise or incomplete beliefs [24]. 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 [7,46,80]. Robustness analysis seeks an exhaustive description of the relation between perturbations and posterior values, in contrast to the small-scale perturbations that are the object of sensitivity analysis [49,57].

The theory of sets of probability measures aims at representing beliefs and evaluating decisions under incompleteness and imprecision [34,52]. Several theories use similar representations: inner/outer measures [36,41,63,73], lower probability theory [8,16,31,71], imprecise probabilities [76], convex Bayesianism [38], probability/utility sets [67]. Some theories, like Dempster-Shafer theory [68], can be cast mathematically, but not conceptually, in terms of convex sets of probability measures [41].

A closed convex set of probability measures is called a *credal set* [52]; existence of credal sets is derived from axioms about preferences [34]. [An informal introduction to aspects of the theory of credal sets can be found at http://www.cs.cmu.edu/~qBayes/Tutorial/.] 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 is indicated by $K(\mathbf{X})$.

This article deals with credal sets that are specified as the convex hull of a finite number of probability measures; such *finitely generated credal sets* are polytopes in the space of probability measures [37]. In the context of variables with finitely many values, most of the standard models used in robust statistics are finitely generated credal sets.

Example 1 *An* ϵ-contaminated class *is a credal set characterized by a prob-*

ability density $r(X)$ and a real number $\epsilon \in (0,1)$; the class contains all probability densities $p(X)$ such that $p(X) = (1-\epsilon)r(X) + \epsilon q(X)$, where $q(X)$ can be any density [6]. Alternatively, an ϵ-contaminated class is a credal set containing all densities $p(X) \geq (1-\epsilon)r(X)$ [8,42]. When X has finitely many values, an ϵ-contaminated class has a finite number of 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).

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

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

The set-function $\underline{P}(A)$ is called a *lower probability*; the set-function $\overline{P}(A)$ is called an *upper probability. Lower densities* $\underline{p}(X)$ and *upper densities* $\overline{p}(X)$ are defined similarly, taking pointwise minima and maxima of densities induced by a credal set. Lower probabilities can be obtained from upper probabilities through the expression $\underline{P}(A) = 1 - \overline{P}(A^c)$. Note that a credal set always creates unique lower and upper probabilities, but lower and upper probabilities do not define a unique credal set [76, Section 2.7].

Example 2 *A* density bounded class *is a credal set containing all probability measures $P(A)$ such that $L(A) \leq P(A) \leq U(A)$; $L(A)$ and $U(A)$ are arbitrary non-negative measures so that $L(A) \leq U(A)$ for all A, $L(\cup A) \leq 1$ and $U(\cup A) \geq 1$ [50,51]. The lower and upper probabilities for an event A are respectively* $\max\left(L(A), 1 - U(A^c)\right)$ *and* $\min\left(U(A), 1 - L(A^c)\right)$.

Example 3 *A* total variation class *is a credal set containing all probability measures $P(A)$ such that $|P(A) - R(A)| \leq \epsilon$ for any event A, where $R(A)$ is a given probability measure and ϵ is a real number in the interval $(0,1)$ [80]. The lower and upper probabilities for an event A are respectively* $\max\left(R(A) - \epsilon, 0\right)$ *and* $\min\left(R(A) + \epsilon, 1\right)$.

Given a function $f(X)$, *lower and upper expectations* are respectively defined as:

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

Maxima and minima of expectations can only occur at the vertices of a credal set [76]. Lower expectations can be obtained from upper expectations through the expression $\underline{E}[f(X)] = -\overline{E}[-f(X)]$. The lower probability for an event A can be obtained by taking the lower expectation of the indicator function $I_A(X)$.

A credal set generates lower and upper values for any functional. Consider the variance of a variable X, defined as $V_P[X] = E_P[X^2] - (E_P[X])^2$. The *lower and upper variances* are respectively:

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

The *conditional credal set* $K(X|A)$ is defined by conditional densities $p(X|A)$. Inference is generally 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 probability equal to zero, the posterior credal set $K(X|A)$ is not specified by the joint credal set alone [76, Chapter 6]. The difficulties of handling zero lower probabilities are discussed in Section 6. Note that if the event A has lower probability larger than zero, the vertices of a posterior credal set $K(X|A)$ can be obtained by applying Bayes rule only to the vertices of the corresponding joint credal set [34].

Example 4 *A* density ratio class *is a credal set containing all probability measures* $P(A)$ *such that* $L(A)/U(B) \leq P(A)\,/P(B) \leq U(A)/L(B)$ *for any events* A, B, *where* $L(A)$ *and* $U(A)$ *are arbitrary positive measures such that* $L(A) \leq U(A)$ *for all* A *[26]. Bounds on the posterior probability for an event* A *given event* B *are* $\underline{P}(A|B) = L(A \cap B)/(L(A \cap B) + U(A^c \cap B))$ *and* $\overline{P}(A|B) = U(A \cap B)/(U(A \cap B) + L(A^c \cap B))$. *For an arbitrary function* $f(X)$, *the upper and lower expectations are respectively the solution of the non-linear equations in* λ *[26]:*

$$\sum_j \left(U(x_j) \max(f(x_j) - \lambda, 0)\right) + \sum_j \left(L(x_j) \min(f(x_j) - \lambda, 0)\right) = 0,$$

$$\sum_j \left(L(x_j) \max(f(x_j) - \lambda, 0)\right) + \sum_j \left(U(x_j) \min(f(x_j) - \lambda, 0)\right) = 0.$$

Given a credal set and a 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* [76, Section 6.4]:

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

A bracketing algorithm that solves this equation is *Lavine's algorithm* [50].

Currently there is no standard way to define independence relations with credal sets [5,15,23,48,78]. The results presented in this article adopt Walley's definition of independence [76, Chapter 9], because this formulation can be reduced to preference relations. 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.

Two credal sets $K(\mathbf{X}|\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z})$ and $K(\mathbf{X}|\mathbf{Y}=\mathbf{y})$ are said *equivalent* if any measure in the convex hull of $K(\mathbf{X}|\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z})$ also belongs to the convex hull of $K(\mathbf{X}|\mathbf{Z}=\mathbf{z})$, and vice-versa.

Definition 5 *Variables* $\mathbf{Y}$ *are* irrelevant *to* $\mathbf{X}$ *given* $\mathbf{Z}$ *if* $K(\mathbf{X}|\mathbf{Z}=\mathbf{z})$ *is equivalent to* $K(\mathbf{X}|\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z})$ *for all possible values of* $\mathbf{Y}$ *and* $\mathbf{Z}$.

Definition 6 *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.

It is interesting to ask what are the *graphoid* properties of the previous definitions, as these properties are important in connection with Bayesian networks [62]. Irrelevance satisfies appropriate versions of the decomposition, weak union and contraction properties; independence satisfies symmetry, decomposition and weak union [22]. Note that irrelevance does not satisfy symmetry, and independence does not satisfy the contraction property. The importance of the graphoid properties is further discussed in Section 8.4.

6 Locally defined credal networks

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

Definition 7 *A* locally defined credal network *is a directed acyclic graph where every node is associated with a variable* X_i *and a collection of* local credal sets $K(X_i|\mathrm{pa}(X_i))$*, and where a method for the combination of the local credal sets is specified.*

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 credal 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 credal network is called an *extension* of the network.

The key technical difference between standard Bayesian networks and locally defined credal networks is that more than one extension may exist for a credal

network. The following example suffices to illustrate this point.

Example 8 *Consider the network in Figure 2. It is possible to generate a joint credal set $K(W, X, Y, Z)$ with two extreme points, P_1 and P_2, that satisfies all quantitative constraints in the network. The joint distribution P_1 corresponds to Expression (1) with the lower bounds for $p(w)$, $p(x|w)$, $p(x|w^c)$, $p(y|x)$, $p(y|x^c)$, $p(z|y)$ and $p(z|y^c)$; the jont distribution P_2 corresponds to Expression (1) with the upper bounds for these probabilities. It is also possible to generate a second joint credal set $K(W, X, Y, Z)$ with 128 extreme points, using Expression (1) with every combination of lower and upper bounds for $p(w)$, $p(x|w)$, $p(x|w^c)$, $p(y|x)$, $p(y|x^c)$, $p(z|y)$ and $p(z|y^c)$.*

At first, it is tempting to adopt a methodology that uniquely selects an extension for any given credal network. The literature offers a collection of approaches that are based on this strategy [10,12,55,74,75,83]. But 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 credal networks. A more promising approach, advocated in this article, is to start from a principled definition of irrelevance and independence and then accept that several extensions may be generated from a single network.

Two types of extensions are analyzed in this article: *strong extensions* and *natural extensions*. Other extensions may be considered for specific applications, but strong and natural extensions provide a solid foundation for the study of locally defined credal networks.

Definition 9 *The* strong extension *of a locally defined credal network is the convex hull of all joint measures that satisfy Expression (1) when each density $p(X_i|\mathrm{pa}(X_i))$ (for each value of $\mathrm{pa}(X_i)$) is selected from the local credal set $K(X_i|\mathrm{pa}(X_i))$.*

A strong extension is the largest joint credal set where all extreme points satisfy Expression (1). Strong extensions are the most common type of credal network studied in the literature, due to their intuitive similarity to standard Bayesian networks [10,12,23,29,56,74]. Section 7 investigates the irrelevance relations that are implied by this particular method of combination for local credal sets. The term strong extension is inspired on the concept of strong independence proposed by Couso et al [19].

Definition 10 *The* natural extension *of a locally defined credal network is the largest joint credal set containing joint measures that comply with the quantitative constraints of the network and with a given collection of irrelevance relations.*

Section 8 investigates several strategies to specify well-defined natural extensions through irrelevance relations. The term natural extension is adapted

from Walley's terminology [76].

The choice of the "correct" extension depends on problem-specific considerations, but general guidelines can be useful to reach a decision. In short, the theory of natural extensions is more flexible and powerful, while the theory of strong extensions is quite similar to the standard theory of Bayesian networks and leads to relatively simple algorithms.

The flexibility of natural extensions comes from the fact that, when constructing the natural extension of a network, a decision-maker is free to choose the irrelevance relations that must hold in the extension. An extreme position is not to impose any irrelevance relation on the extension, adopting the view that the various credal sets in the network interact in unknown ways. Alternatively, the decision-maker may be meticulous enough to go through every possible irrelevance relation in a network, deciding which relations are valid and which relations are invalid. Finally, the decision-maker may adopt an entirely different strategy by generating irrelevance relations through some automatic scheme; for example, a decision-maker may require that all nondescendants non-parents of a variable be independent of the variable given the variable's parents (Section 8).

Natural extensions embody a least commitment strategy: a natural extension satisfies every constraint in a network, but no other constraints. As such, natural extensions can accommodate incompleteness and imprecision of beliefs more flexibly than strong extensions. Several types of uncertainty that can be represented by a natural extension cannot be represented either by standard Bayesian networks or by strong extensions.

The advantages of natural extensions come at a price, as natural extensions require significant effort during model specification and during inference. The specification effort can be minimized if the decision-maker adopts some automatic scheme to generate irrelevance relations. But the computational effort involved in manipulating natural extensions seems to be the main difficulty in practice. No algorithm is known that can bypass the potential complexity of natural extensions — the algorithms presented in this article are the first ones to deal with natural extensions of credal networks, and they can be seen as basic tools to solve small problems, helping pave the way to the development of more efficient algorithms in future research.

Strong extensions are computationally more tractable than natural extensions; several methods of standard Bayesian networks can be adapted to work with strong extensions. In addition, strong extensions can be justified from the point of view of robustness analysis. Here the decision-maker is trying to construct a set of models that contains some underlying Bayesian network; the objective is to study how the model variations affect inferences [7]. Strong extensions

always have extreme points represented by standard Bayesian networks, so they are well suited to robustness analysis.

The next sections investigate the theory and practice of strong and natural extensions. To simplify the presentation, algorithms are developed only for the calculation of the posterior upper probabilities:

$$\begin{aligned}\overline{P}(X_Q = x_Q|E) &= \max\left(\frac{p(X_Q = x_Q, E)}{p(E)}\right) \\ &= \max\left(\frac{\sum_{\mathbf{X}} I_{\{X_Q = x_Q, E\}}(\mathbf{X}_{QE})p(\mathbf{X})}{\sum_{\mathbf{X}} I_E(\mathbf{X}_E)p(\mathbf{X})}\right) \\ &= \max\left(\frac{\sum_{\mathbf{X}\setminus\mathbf{X}_{QE}, \mathbf{X}_{QE}=\mathbf{x}_{QE}} p(\mathbf{X})}{\sum_{\mathbf{X}\setminus\mathbf{X}_E, \mathbf{X}_E=\mathbf{x}_E} p(\mathbf{X})}\right). \end{aligned} \quad (2)$$

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

To simplify the discussion, it will be assumed throughout that every combination of variables has positive lower probability. A similar assumption has been advocated for standard probability models, where it has been argued that every event should receive at least some small probability [28]. Zero probabilities are even more delicate in the theory of credal sets. It should be noted that "lower probability zero" is quite a weak statement, as an event with zero lower probability may occur with a positive upper probability. What are the coherency conditions that must hold in this situation, as automatic application of elementwise Bayes rule certainly fails? There is no complete agreement on how to proceed in such situations, so the positivity conditions seems to be necessary at this point; future research must investigate the consequences of abandoning the positivity condition.

Fundamentally, calculation of a posterior upper probability is an optimization problem over a closed convex set. This insight is applied in all algorithms in this article; in independent work, Cano et al [10,11] and Zaffalon [83] have explored the same observation for strong extensions.

7 Strong extensions: Theory and algorithms

Section 7.1 presents a novel result that ties strong extensions to graphical d-separation. Algorithms for strong extensions have been analyzed previously in a variety of fields; the algorithms in Sections 7.2 and 7.3 unify previous work, rather than present fundamentally new results.

7.1 Irrelevance, independence, and strong extensions

The appeal of strong extensions comes from their intuitive similarity with standard Bayesian networks. This intuition can be formalized using Walley's definition of independence:

Theorem 11 *Given a credal network where every combination of variables has positive lower probability, any graphical d-separation relation in the credal network corresponds to a valid conditional independence relation in the strong extension of the network.*

This new result demonstrates that strong extensions mimic the properties of standard Bayesian networks as independence-maps [62, page 119]. The theorem also complements results by Cano et al in their study of independence concepts that satisfy d-separation [12]. Theorem 11 demonstrates that Walley's independence relations exhibit the desired correspondence with d-separation, and justify the intuitive similarity between strong extensions and standard Bayesian networks.

Theorem 11 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 $p(X_Q = x_Q|E)$ can be obtained by d-separation in polynomial time [33]. Every algorithm in Sections 7.2 and 7.3 must start only after d-separation is used to remove unnecessary variables.

7.2 Enumeration algorithms

One way to generate inferences with strong 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 [10,12]. The Cano/Cano/Moral (CCM) transformation takes a locally defined credal network and produces a standard Bayesian network. Suppose that the local credal sets $K(X_i|\text{pa}(X_i))$ can be combined in m ways (where $m > 1$); each combination is a function $p_l(X_i|\text{pa}(X_i))$ for $1 \leq l \leq m$. The CCM transform adds a new variable X_i' to the network, where

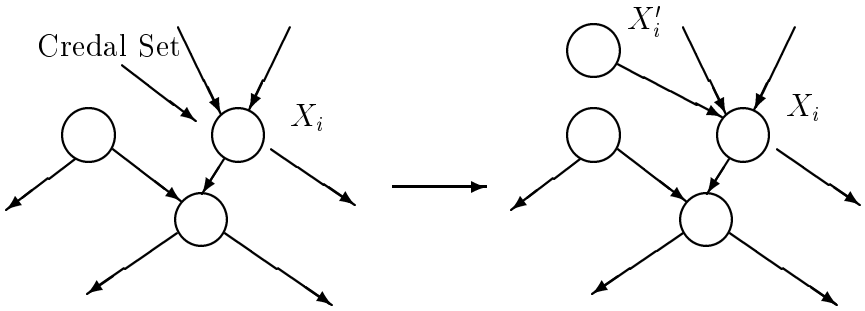


Fig. 3. The CCM transformation for variable X.

X_i' has X_i as its only child (Figure 3) and has m integer values. The variable X_i is then associated with a family of standard conditional densities:

$$p(X_i|\text{pa}(X_i), X_i' = l) = p_l(X_i|\text{pa}(X_i)).$$

The variables X_i' are called *transparent* variables [10]. The posterior upper probability $\overline{P}(X_Q = x_Q|E)$ can be calculated by visiting all the values of transparent variables, because posterior probabilities are maximized 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 probability.

Two factors contribute to the high computational cost of enumeration algorithms. First, inferences are reduced, by the CCM transform, to standard Bayesian network inferences, which are NP-hard [18]. Second, the number of transparent variables generated by the CCM transform is equal to the number of local credal sets. Despite these difficulties, exact inferences with strong extensions can be performed realistically in the *JavaBayes* engine (Section 11), because graphical d-separation leads to enormous computational savings.

Inferences with strong extensions can be simplified by the use of convex hull algorithms. To see that, note that the generation of the function $p(X_Q|E, \mathbf{X}')$ can follow any of the standard algorithms for Bayesian networks [45]. In the process of generating $p(X_Q|E, \mathbf{X}')$ by any standard inference algorithm, several intermediate functions are generated. Consider a function $f(X_1, X_2, X_1', X_2', X_3')$: For each fixed combination of values of the transparent variables X_1', X_2', X_3', f is a function of X_1, X_2. Consequently, the function f can be viewed as a *set* of functions of X_1, X_2 indexed by X_1', X_2', X_3'. In general, each intermediate function in an inference algorithm can be viewed as a set F of functions indexed by transparent variables. As any function of transparent variables is useful so long as it represents an extreme point of the strong extension, the set F can be replaced by its convex hull. The removal of extreme points of F may lead to exponential savings, as each extreme point of F may potentially

be combined with an exponential number of distributions. As an example of this procedure, take a singly connected network with binary variables; there, a convex hull operation can be taken at every step in an inference propagation algorithm [62], because the convex hull of a collection of intervals is still an interval. This property is actually the basis for the efficiency of Zaffalon's 2U algorithm [29].

7.3 Iterative and global optimization

The calculation of a posterior upper probability (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 [83] and Cozman [20] have independently explored gradient-based and deterministic global optimization. The purpose of this section is to summarize these ideas and to discuss the relationship between inference in strong extensions and learning in Bayesian networks, and the relationship between gradient calculations and graph faithfulness. Stochastic techniques such as simulated annealing [10] and genetic algorithms [11] are not explored; such approaches deserve full treatment in future work.

Call Θ a vector containing all values of $p(X_i|\mathrm{pa}(X_i))$ for all variables X_i. Denote by θ_{ijk} the probability of the jth value of X_i given the kth value of $\mathrm{pa}(X_i)$: $\theta_{ijk} = p(X_i = x_{ij}|\mathrm{pa}(X_i) = \pi_k)$. The search for a vector 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]. A variety of iterative optimization algorithms employ gradient calculations in the search for a maximizing Θ [53]. Previous work has derived efficient, closed-form expressions for the gradient of Θ in Bayesian networks [49,64]; this 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 is:

$$\begin{aligned}\frac{\partial p(X_Q = x_Q|E)}{\partial \theta_{ijk}} &= \frac{\partial p(X_Q = x_Q|E)}{\partial p(X_i = x_{ij}|\mathrm{pa}(X_i) = \pi_k)} \\ &= \frac{p(\mathrm{pa}(X_i) = \pi_k)}{P(E)} \times \\ &\quad (P(X_Q = x_Q, E|X_i = x_{ij}, \mathrm{pa}(X_i) = \pi_k) - \\ &\quad P(X_Q = x_Q|E)\, P(E|X_i = x_{ij}, \mathrm{pa}(X_i) = \pi_k))\,. \end{aligned} \tag{3}$$

Every term in Expression (3) can be calculated by any standard Bayesian net-

work algorithm. Note that Expression (3) is identically zero for all x_Q when $p(X_Q|E) = p(X_Q|E, X_i = x_{ij}, \mathrm{pa}(X_i) = \pi_k)$. This equality is true when the gradient is calculated for a joint density that is *not faithful* to the graph, assuming that d-separated variables are deleted before calculation of the gradient (Section 7.1).

Moving away from iterative optimization schemes, inferences with strong extensions can be reduced to *signomial programs* and solved exactly through branch-and-bound procedures [1,20,83]. *Signomial programming* is a branch of optimization theory that deals with optimization of polynomials [2,3,27,35,60]. Deterministic global optimization usually requires methods that produce lower and upper bounds for the objective function [30,47]. Such methods can be available in some cases; for example, probability bounds can be calculated efficiently for singly-connected networks using Tessem's algorithm [74]. So far no global optimization method has been implemented for strong extensions.

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 credal network and 2) a given collection of irrelevance relations. A collection of irrelevance relations is *well-defined* when it can be satisfied by at least one probability measure. Sections 8.2 to 8.4 describe three methods for the specification of natural extensions. These methods share the property that any irrelevance relation imposed on a network corresponds to a graphical d-separation relation in the network; consequently, the methods generate well-defined collections of constraints.

The first method is to impose no irrelevance relation in the network, a strategy similar to probabilistic logic [4,54,58]. The second method is to require that the nondescendants non-parents of any variable be *irrelevant* to the variable given the parents of the variable. The third method is to require that the nondescendants non-parents of any variable be *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 article assume that the local credal set for any variable in a network is specified separately for each value of the variable's parents. A collection of credal sets $K(X|Y)$ is *separately specified*

when all constraints that specify $K(X|y)$ (for a particular value y of Y) involve only the density $p(X|y)$. This restriction makes sense both for elicitation and representation of beliefs, and is quite a natural consequence of specifications based on lower expectations [76, Section 6.2].

Every finitely generated local credal set can be specified through a finite number of upper expectations; a credal set $K(X_i|\mathrm{pa}(X_i) = \pi_k)$ can be specified through a collection of functions $f_l(X_i)$ and numbers γ_l:

$$\overline{E}[f_l(X_i)|\mathrm{pa}(X_i) = \pi_k] = \gamma_l \;\Rightarrow\; \sum_j f_l(x_{ij}) p(X_i = x_{ij}|\mathrm{pa}(X_i) = \pi_k) \leq \gamma_l. \quad (4)$$

Note that the functions f_l and the numbers γ_l should be indexed by the variable X_i and by the value of $\mathrm{pa}(X_i)$; these indexes are suppressed to simplify notation.

The inequality in Expression (4) can be reduced to an inequality on the values of the joint density $p(\mathbf{X})$. For an arbitrary value π_k of the parents of X_i:

$$\sum_{\mathbf{X}\setminus\mathrm{pa}(X_i),\mathrm{pa}(X_i)=\pi_k} (f_l(X_i) - \gamma_l)\, p(\mathbf{X}) \leq 0. \quad (5)$$

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

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

8.2 No irrelevance relations

The simplest type of natural extension is produced when no irrelevance relation is imposed on a credal network. Denote by $K_o(\mathbf{X})$ 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 $p(\mathbf{X})$ [43,66]. Linear fractional programs can be reduced to linear programs by a variety of methods [43,66]; consequently, inferences with $K_o(\mathbf{X})$ 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 [44] and row suppression [17], can be used to tame this complexity. To guarantee that only valid inferences are produced by linear fractional programming, it is necessary to check whether $\underline{P}(E)$ is non-zero (this is satisfied when all combinations of variables have positive lower probability).

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

The constraint on $K_r(\mathbf{X})$ is that $K(X_i|\mathrm{pa}(X_i))$ and $K(X_i|\mathrm{nd}(X_i))$ are equal for any variable X_i. This constraint implies the following collection of inequalities, for an arbitrary value π_k of $\mathrm{pa}(X_i)$:

$$\sum_{\mathbf{X}\setminus\mathrm{nd}(X_i),\mathrm{pa}(X_i)=\pi_k} (f_l(X_i) - \gamma_l)\, p(\mathbf{X}) \leq 0, \tag{7}$$

where all functions $f_l(X_i)$ and numbers γ_l are obtained from constraints (5). Note that Expression (7) represents a larger number of constraints than Expression (5), as the next example illustrates.

Example 12 *Suppose that* $\mathrm{nd}(X_1) = \{\mathrm{pa}(X_1), X_2\}$*, where* X_2 *is a binary variable. The inequalities represented by Expression (7) are divided in two groups:*

$$\sum_{\substack{\mathbf{X}\setminus\{\mathrm{pa}(X_1),X_2\}\\ X_2=\mathrm{TRUE}}} (f_l(X_1) - \gamma_l)\, p(\mathbf{X}) \leq 0, \qquad \sum_{\substack{\mathbf{X}\setminus\{\mathrm{pa}(X_1),X_2\}\\ X_2=\mathrm{FALSE}}} (f_l(X_1) - \gamma_l)\, p(\mathbf{X}) \leq 0.$$

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

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

Lemma 13 *Consider a credal network where every combination of variables has positive lower probability. If* $K(X_i|\mathrm{pa}(X_i))$ *and* $K(X_i|\mathrm{nd}(X_i))$ *are equal for any value of* $\mathrm{nd}(X_i)$*, then* $K(X_i|\mathrm{pa}(X_i))$ *and* $K(X_i|\mathrm{pa}(X_i), \mathbf{S})$ *are equal for any value of* $\{\mathrm{pa}(X_i), \mathbf{S}\}$*, for* $\mathbf{S} \subset \mathrm{nd}(X_i)$*.*

Note that Expressions (6) and (7) are not only a minimal collection of constraints on $K_r(\mathbf{X})$, but these constraints in fact imply all required irrelevance relations. This is guaranteed by the existence of the strong extension of the underlying network, as the strong extension satisfies each one of the constraints with equality.

Inferences with $K_r(\mathbf{X})$ 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. The idea here is to study the properties of sub-networks of a credal network; namely, those sub-networks at the "top" of the original network.

Definition 14 *Given a credal network with variables* $\mathbf{X}$*, a* top sub-network *is a network formed by variables* $\mathbf{S}$ *in* $\mathbf{X}$ *such that if a variable* X_i *belongs to* $\mathbf{S}$*, all ascendants of* X_i *in* $\mathbf{X}$ *also belong to* $\mathbf{S}$*.*

Consider a credal network with a top sub-network $\mathbf{S}$. Denote by $K'_r(\mathbf{S})$ the natural extension of the top sub-network $\mathbf{S}$ under the constraint that the nondescendants non-parents (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 $K'_r(\mathbf{S})$ are:

$$\sum_{\mathbf{S}\setminus \mathrm{nd}(X_i), \mathrm{pa}(X_i)=\pi_k} (f_l(X_i) - \gamma_l)\, p(\mathbf{S}) \leq 0, \quad \sum_{\mathbf{S}} p(\mathbf{S}) = 1. \tag{8}$$

Suppose that $K'_r(\mathbf{S})$ is available; how is $K'_r(\mathbf{S})$ related to the natural extension $K_r(\mathbf{X})$ of the original network?

Theorem 15 *Given a credal network with variables* $\mathbf{X}$ *with separately specified local credal sets, where every combination of variables has positive lower probability, and a top sub-network* $\mathbf{S}$ *with natural extension* $K'_r(\mathbf{S})$ *(nondescendants non-parents in* $\mathbf{S}$ *of a variable in* $\mathbf{S}$ *are irrelevant to the variable given the variable's parents), denote by* $K_r(\mathbf{S})$ *the credal set obtained by marginalization of the extension* $K_r(\mathbf{X})$*. Then* $K'_r(\mathbf{S})$ *is equal to* $K_r(\mathbf{S})$*.*

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 16 *Given a credal network with variables* $\mathbf{X}$ *with separately specified local credal sets, where every combination of variables has positive lower probability, and given a top sub-network* $\mathbf{S}$ *with extension* $K'_r(\mathbf{S})$*, suppose that every variable* X_i *not in* $\mathbf{S}$ *is associated with a single conditional density* $p(X_i|\mathrm{pa}(X_i))$*. The solution for problem (2), subject to constraints in Expressions (6) and (7), is identical to:*

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

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

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

Note that all variables in $\mathbf{X}\setminus\mathbf{S}$ 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 at 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, the nondescendants non-parents of X_i are independent of X_i given the parents of X_i. Denote by $K_n(\mathbf{X})$ the natural extension that satisfies this constraint, and denote by $\mathrm{nD}(X_i)$ the nondescendants non-parents of variable X_i. According to Walley's definition of independence (Section 4),

- the credal sets $K(X_i|\mathrm{pa}(X_i))$ and $K_n(X_i|\mathrm{nd}(X_i))$ must be equal for all values of $\mathrm{nd}(X_i)$, and
- the credal sets $K_n(\mathrm{nD}(X_i)|\mathrm{pa}(X_i), X_i)$ and $K_n(\mathrm{nD}(X_i)|\mathrm{pa}(X_i))$ must be equal for all values of $\mathrm{pa}(X_i)$.

Example 17 *Take a credal network where variable* X_5 *has nondescendants* X_1, X_2, X_3, X_4 *and parents* X_1, X_2 *(Figure 4). Independence of* X_5 *and* $\mathrm{nD}(X_5)$ *given* $\mathrm{pa}(X_5)$ *requires that the credal set* $K_n(X_5|X_1, X_2, X_3, X_4)$ *be equal to* $K(X_5|X_1, X_2)$*, and that the credal set* $K_n(X_3, X_4|X_1, X_2, X_5)$ *be equal to* $K_n(X_3, X_4|X_1, X_2)$*. 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 in the network. The second condition involves two credal sets that are not directly specified in the credal network.*

Several properties presented in Section 8.3 have analogues that are useful in connection to $K_n(\mathbf{X})$. The next lemma proves that irrelevance of a variable X_i to its nondescendants non-parents given its parents implies irrelevance of X_i to subsets of its nondescendants non-parents given its parents.

Lemma 18 *If the credal sets* $K(\mathrm{nD}(X_i)|\mathrm{pa}(X_i), X_i)$ *and* $K(\mathrm{nD}(X_i)|\mathrm{pa}(X_i))$ *are equal for all values of* $\{X_i, \mathrm{pa}(X_i)\}$*, then the credal sets* $K(\mathbf{S}|\mathrm{pa}(X_i), X_i)$

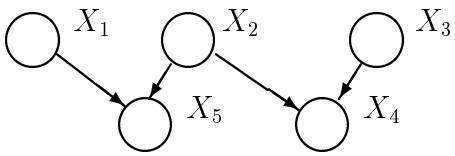


Fig. 4. Creating the natural extension for a sub-network.

and $K(\mathbf{S}|\mathrm{pa}(X_i))$ *are equal as well, for any* $\mathbf{S} \subset \mathrm{nD}(X_i)$.

The next theorem establishes the relation between the complete natural extension of a network and the natural extension of a top sub-network.

Theorem 19 *Given a credal network with variables* $\mathbf{X}$ *with separately specified local credal sets, where every combination of variables has positive lower probability, and a top sub-network* $\mathbf{S}$ *with natural extension* $K'_n(\mathbf{S})$ *(nondescendants non-parents in* $\mathbf{S}$ *of a variable in* $\mathbf{S}$ *are independent of the variable given the variable's parents), denote by* $K_n(\mathbf{S})$ *the credal set obtained by marginalization of the extension* $K_n(\mathbf{X})$. *Then* $K'_n(\mathbf{S})$ *is equal to* $K_n(\mathbf{S})$.

A result similar to Theorem 16 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 $K_n(\mathbf{X})$ is obtained by combining $K'_n(\mathbf{S})$ with the conditional densities for variables outside of $\mathbf{S}$. The proof of this result is similar to the proof of Theorem 16 and is omitted.

The remainder of this section develops an algorithm for the calculation of posterior probabilities induced by a natural extension $K_n(\mathbf{X})$. The main idea of the algorithm is to construct a sequence of top sub-networks, beginning with a single variable, and ending with the complete credal network. At each step, a variable is added to the current top 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 the manipulation of independence constraints is currently known.

The algorithm BuildExtension (Figure 5) depends on three facts. First, it is always possible to construct a sequence of top sub-networks as required by the algorithm (property of directed acyclic graphs). Second, the natural extension of a top sub-network, taking into account independence relations in the top sub-network, is always equal to the marginal credal set obtained by marginalizing the complete natural extension $K_n(\mathbf{X})$ (Theorem 19). Third,

(1) Take a variable X_1 that has no parents and form a sub-network that contains only X_1. The marginal extension $K'_n(X_1)$ is equal to the local credal set $K(X_1)$ (Theorem 19).
(2) 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. All independence constraints can be generated for this new sub-network (algorithm IncorporateVariable).
(3) Form a new top sub-network by adding a variable X_3. Again, independence constraints can be generated for this new top sub-network (algorithm IncorporateVariable).
(4) This process continues until all variables are incorporated into a complete network.

Fig. 5. BuildExtension: The construction of linear inequalities for $K_n(\mathbf{X})$.

it is possible to represent the credal set of a sub-network after the incorporation of a variable into a top sub-network, as described by the algorithm IncorporateVariable (Figure 6).

Example 20 *Take the network in Figure 4. Assume that previous computations generated the extension $K_n(X_1, X_2, X_3, X_4)$ as a collection of linear constraints. When variable X_5 is added to the network, how to compute the extension $K_n(X_1, X_2, X_3, X_4, X_5)$?*

- *Satisfy the requirement that $K_n(X_5|X_1, X_2, X_3, X_4)$ and $K(X_5|X_1, X_2)$ be equal through a replication procedure.*
- *Satisfy the requirement that $K_n(X_3, X_4|X_1, X_2, X_5)$ and $K_n(X_3, X_4|X_1, X_2)$ be equal by generating $K_n(X_3, X_4|X_1, X_2)$ directly from $K_n(X_1, X_2, X_3, X_4)$ and then use a replication procedure.*
- *Satisfy the requirement that X_3 be independent of X_1, X_2, and X_5. First, use $K(X_3)$ directly in a replication procedure including X_1, X_2, and X_5. Then obtain $K_n(X_1, X_2, X_5)$ recursively and employ a replication procedure to generate the constraints for $K_n(X_1, X_2, X_5|X_3)$ based on $K_n(X_1, X_2, X_5)$.*
- *Satisfy the requirement that X_4 be independent of X_1 and X_5 given X_2 and X_3. First, use $K(X_4|X_2, X_3)$ in a replication procedure including X_1 and X_5. Second, obtain $K_n(X_1, X_2, X_3, X_5)$ recursively. Third, obtain the credal set $K_n(X_1, X_5|X_2, X_3)$ from the credal set $K_n(X_1, X_2, X_3, X_5)$. Fourth, employ a replication procedure to generate $K_n(X_1, X_5|X_2, X_3, X_4)$ based on $K_n(X_1, X_5|X_2, X_3)$.*

Algorithm BuildExtension proves constructively the following non-trivial fact:

Theorem 21 *The natural extension $K_n(\mathbf{X})$ of a credal network where every combination of variables has positive lower probability is a finitely generated*

Take a top sub-network with variables $\mathbf{S}$ and a variable X_i such that $\mathbf{S}$ and X_i form a top sub-network. To generate $K_n(\mathbf{S}, X_i)$ from $K_n(\mathbf{S})$:

(1) The constraint that $K_n(X_i|\mathbf{S})$ and $K(X_i|\text{pa}(X_i))$ be equal is satisfied through a replication procedure (Section 8.3).

(2) The constraint that $K_n(\mathbf{S}\backslash\text{pa}(X_i)|\text{pa}(X_i), X_i)$ and $K_n(\mathbf{S}\backslash\text{pa}(X_i)|\text{pa}(X_i))$ be equal is satisfied in two steps:
   (a) Obtain $K_n(\mathbf{S}\backslash\text{pa}(X_i)|\text{pa}(X_i))$ from $K_n(\mathbf{S})$ (Appendix B).
   (b) Obtain the inequalities for $K_n(\mathbf{S}\backslash\text{pa}(X_i)|\text{pa}(X_i), X_i)$ through a replication procedure.

(3) If X_i is nondescendant of a variable X_j in $\mathbf{S}$, then the independence between X_j and $\text{nD}(X_j)$ given $\text{pa}(X_j)$ must be asserted.
   (a) The constraint that nondescendants non-parents of X_j (in $\mathbf{S}$) plus X_i be irrelevant to X_j given the parents of X_j produces linear inequalities by a replication procedure, because $K(X_j|\text{pa}(X_j))$ is specified by the credal network.
   (b) The constraint that X_j be irrelevant to its nondescendants non-parents (including X_i) given the parents of X_j is more complex:
      (i) Form a sub-network $\mathbf{S}'$ discarding X_j and its descendants from $\mathbf{S}$, and obtain the constraints for $K_n(\mathbf{S}')$ from $K_n(\mathbf{S})$ (Appendix B).
      (ii) The set $K_n(\mathbf{S}', X_i)$ is then obtained recursively by the algorithm IncorporateVariable. This recursion always terminates, as it eventually runs out of sub-networks.

Fig. 6. IncorporateVariable: Adding a variable to the natural extension.

credal set.

Independence judgements are stronger than irrelevance judgements, and the former produce a larger number of constraints than the latter in natural extensions. It might seem that independence judgements would lead to significant computational simplifications, as independence judgements in standard Bayesian networks lead to simplifications due to d-separation. But independence relations in the theory of credal sets do not satisfy the contraction property that is crucial to prove d-separation in standard Bayesian networks (Section 5). No proof of d-separation properties (or any other separation property) is known at this point for natural extensions.

9 Global neighborhoods of Bayesian networks

In previous sections, joint credal sets were constructed from collections of local conditional credal sets. An alternative path is to specify joint credal sets as neighborhoods of standard Bayesian networks. Similar strategies are employed in robust statistics; the purpose of this section is to briefly adapt the language and results of that field to the present context.

The main idea is to define a joint credal set as a neighborhood of 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 (univariate versions were used as examples in Section 5). 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 $r(\mathbf{X})$. A neighborhood is marginalization invariant when

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

In words: marginalization and construction of neighborhoods commute [81]. For neighborhoods with marginalization invariance, marginalization is performed only in the base network. Upper and lower expectations are not dependent on marginalization of high dimensional credal sets, because the neighborhood of a marginal density is the marginal neighborhood of the base network. Consequently, calculation of upper and lower expectations are reduced to low dimensional operations.

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

$$\overline{E}[f(X_Q)|E] = \frac{(1-\epsilon)\left(\sum_{X_Q} f(x_Q)r(X_Q,E)\right) + \epsilon \max_{X_Q} f(x_Q)}{(1-\epsilon)r(E)+\epsilon}.$$

- A *constant* density bounded joint credal set contains all joint densities $p(\mathbf{X})$ such that $(1/\mu)r(\mathbf{X}) \leq p(\mathbf{X}) \leq \mu r(\mathbf{X})$, where $\mu > 1$. This model is a 2-monotone capacity (that is, it satisfies $\underline{P}(A\cap B)+\underline{P}(A\cup B) \geq \underline{P}(A)+\underline{P}(B)$ for events A and B); consequently, posterior lower and upper probabilities can be easily computed as $\underline{P}(A|B) = \underline{P}(A,B)\,/(\underline{P}(A,B)+\overline{P}(A^c,B))$ and $\overline{P}(A|B) = \overline{P}(A,B)\,/(\overline{P}(A,B)+\underline{P}(A^c,B))$ [14,41,82]. For lower and upper expectations, Lavine's algorithm generates a convenient, iterative solution (Section 4), as upper and lower expectations can be computed using Choquet integrals [76, Section 3.2.4].

- A total variation joint credal set contains all joint densities $p(\mathbf{X})$ such that for all events A, $|P(A) - R(A)| \leq \epsilon$, where $\epsilon \in (0, 1)$ and $R(A)$ is the 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.
- A *constant* density ratio joint credal set contains 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 probabilities can be computed because this model is invariant to marginalization and conditioning [79]. For lower and upper expectations, results from DeRobertis and Hartigan [26] can be used to produce bracketing algorithms (Example 4).

10 Other inferences and decision-making

Previous sections focused on the calculation of upper bounds for the posterior density $p(X_Q = x_Q|E)$. Calculation of posterior probabilities for non-atomic events is immediate: Algorithms for strong extension are identical both for enumeration and iterative approaches; linear programs solving inferences for natural extension must only enlarge their objective functions to incorporate unions of atomic events; finally, algorithms for global neighborhoods are identical for non-atomic events.

For the sake of simplicity, algorithms for locally defined credal networks dealt only with posterior probabilities. Most algorithms for locally defined credal networks can be easily extended to deal with posterior expectations using linear fractional programming, enumeration schemes, or gradient-based search.

Calculation of lower and upper variances in a credal network is a great challenge because the expression for $V_P[X_Q]$ is quadratic on probability values. Walley's variance probability theorem [76, 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 must be between the smallest and largest values of X_Q). Lower and upper variances are then obtained by repeated calculation of lower and upper expectations for $(X_Q - \mu)^2$.

Decision-making is a somewhat complex activity in the theory of credal sets. It is not always possible to select a single course of action that maximizes expected utility for all measures in a credal set [52,67]. Several strategies are possible when selecting a "best" decision. As an example, consider the *most probable explanation* problem (MPE), where the objective is to find a configu-

ration 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 posterior upper probability of the evidence, $\arg\max_{X \in \{\mathbf{X} \setminus \mathbf{X}_E\}} \overline{P}(X|E)$. If a strong extension is built, a CCM transformation can be applied to the credal network and the problem is similar to a standard Bayesian network MPE problem on the transformed network. Another alternative is to find the configuration that maximizes the lower posterior probability of the evidence, $\arg\max_{X \in \{\mathbf{X} \setminus \mathbf{X}_E\}} \underline{P}(X|E)$. 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.

11 Credal networks in practice

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

Strong extensions and global neighborhoods are available in the *JavaBayes* system, a portable 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 [25]. The user interacts with the system through a graphical interface, where variables and densities can be added, deleted or edited. The user can also specify local credal sets and global neighborhoods (ϵ-contaminated, constant density bounded, total variation and constant density ratio credal sets are supported), both to model imprecise beliefs or to perform robustness analysis. *JavaBayes* implements an enumeration algorithm for strong extensions, employing d-separation and the CCM transform to produce posterior probabilities. Documentation, code and examples for the *JavaBayes* system can be downloaded from http://www.cs.cmu.edu/~javabayes/.

Natural extensions can be processed by the *qb* package, a set of Matlab™ procedures (information on Matlab™ is available at http://www.matlab.com). The *qb* package performs linear fractional programming on a matricial description of a credal network; a system like *JavaBayes* must be used to create the network, and the user must insert the constraints that define the natural extension. Code, comments and examples for the *qb* package can be downloaded from http://www.cs.cmu.edu/~qbayes/RobustInferences/Matlab/.

To illustrate the results and algorithms described previously, the examples in Figures 1 and 2 are discussed in the remainder of this section.

Consider the graph in Figure 1. There are five binary variables in the graph (the superscript c indicates negation). This example is based on the examples described by Charniak [13] and by Walley [76, Section 9.3.4]; calculations were performed using *JavaBayes* and *qb*. Note that probabilities for F and B are not specified exactly; instead, the probabilities for $\{F = f\}$ and for $\{B = b\}$ are in 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 article, consider the calculation of $\underline{p}(D = d|L = l)$ and $\overline{p}(D = d|L = l)$.

Strong extension The simplest method to obtain the bounds is to identify the vertices of the local credal sets and generate a strong extension. The strong extension has four vertices, because both the credal sets $K(F)$ and $K(B)$ have two vertices. By calculating $p(D = d|L = l)$ for these four vertices (using *JavaBayes*), the bounds on $p(D = d|L = l)$ are obtained: $\underline{p}(D = d|L = l) = 0.3861$ and $\overline{p}(D = d|L = l) = 0.4461$.

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 $p(D = d|L = l)$, 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 $p(D = d, L = l)/p(L = l)$. The solution of this program indicates that $\underline{p}(D = d|L = l) \leq 10^{-8}$ and $\overline{p}(D = d|L = l) \geq 1 - 10^{-8}$, 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 non-parents of a variable are irrelevant to the variable given the parents of the variable. A replication procedure generates: $0.4 \leq p(F = f|B = b) \leq 0.5$, $0.4 \leq p(B = b|F = f) \leq 0.5$, $0.4 \leq p(F = f|B = b^c) \leq 0.5$ and $0.4 \leq p(B = b|F = f^c) \leq 0.5$. To simplify the calculation of lower and upper bounds, Theorem 16 can be used. The value of $\overline{p}(D = d|L = l)$ is then obtained by solving the program:

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

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

where $w_1 = p(F = f, B = b)$, $w_2 = p(F = f, B = b^c)$, $w_3 = p(F = f^c, B = b)$, $w_4 = p(F = f^c, B = b^c)$. This program produces $\overline{p}(D = d|L = l) = 0.4509$. By minimization, $\underline{p}(D = d|L = l) = 0.3818$ is obtained. Note that these bounds are different than the bounds obtained by strong extension.

Natural extension with independence relations 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 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 $p(D = d|L = l)$ with these six densities leads to $\underline{p}(D = d|L = l) = 0.3818$ and $\overline{p}(D = d|L = l) = 0.4509$.

ϵ-contaminated neighborhood A different strategy is to fix $p(F = f)$ and $p(B = b)$, 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 $\underline{p}(D = d|L = l) = 0.3955$ and $\overline{p}(D = d|L = l) = 0.4455$.

11.2 Example: Markov chain

Figure 2 displays the structure of a Markov chain with four binary variables. The strong extension of this network has 128 extreme points and can be easily computed using *JavaBayes*. It is more interesting to focus on the computation of the natural extension $K_n(W, X, Y, Z)$ for this network. The qualitative constraints are: Y and W are independent given X; Z and (W, X) are independent given Y. The computations that follow were performed using the freely

Table 1
Extreme points for $K(W, X)$.

	$p(x, w)$	$p(x, w^c)$	$p(x^c, w)$	$p(x^c, w^c)$	$p(w)$	$p(x\|w)$	$p(x\|w^c)$
p_1	3/50	63/100	6/25	7/100	0.3	0.2	0.9
p_2	1/25	18/25	4/25	2/25	0.2	0.2	0.9
p_3	3/50	14/25	6/25	7/50	0.3	0.2	0.8
p_4	1/25	16/25	4/25	4/25	0.2	0.2	0.8
p_5	3/100	14/25	27/100	7/50	0.3	0.1	0.8
p_6	1/50	16/25	9/50	4/25	0.2	0.1	0.8
p_7	3/100	63/100	27/100	7/100	0.3	0.1	0.9
p_8	1/50	18/25	9/50	2/25	0.2	0.1	0.9

available lrs package, generously produced by David Avis and distributed at ftp://mutt.cs.mcgill.ca/pub/C/lrs.html.

Consider the application of algorithm BUILDEXTENSION to this problem. The algorithm starts with the credal set $K(W)$, given in the specification of the problem. The second step is to add the variable X; at this point there is no independence constraint to satisfy. There are 6 constraints to be satisfied by $K(W, X)$ plus the unitary constraint. This produces a credal set with 8 extreme points shown in Table 1. The table also shows the marginal probability for W and the probabilities for X conditional on W. Note that every extreme point of $K(W, X)$ corresponds to a choice of extreme points of $K(W)$ and $K(X|W)$.

The third step of the algorithm is to add the variable Y; now the independence of Y and W given X must be enforced. To do so, Fourier-Motzkin elimination must be used to obtain $K(W|X)$ (Appendix B). This produces $1/37 \leq p(W = w|X = x) \leq 3/31$ and $1/2 \leq p(W = w|X = x^c) \leq 27/34$. These constraints are used to form the 22 constraints that characterize $K(W, X, Y)$. The credal set $K(W, X, Y)$ has 292 extreme points.

Finally, $K(W, X|Y)$ must be computed from $K(W, X, Y)$, again using Fourier-Motzkin elimination, and the linear constraints on $K(W, X, Y, Z)$ must be generated. There are 110 such constraints. From them, lower and upper expectations on W, X, Y and Z can be computed. Table 2 shows some lower and upper probabilities obtained in this manner. Note the fact that W is irrelevant to Z given X, a d-separation property that cannot be proved directly without the contraction property. This suggests that some form of separation may be valid for natural extensions, even though the contraction property is violated by independence in credal sets.

Table 2
Lower and upper probabilities.

Probability of	Lower value	Upper value
$Z = z$	443/1250	2623/5000
$Z = z \mid X = x$	17/50	1/2
$Z = z \mid X = x^c$	2/5	14/25
$Z = z \mid X = x, W = w$	17/50	1/2
$Z = z \mid X = x, W = w^c$	17/50	1/2
$Z = z \mid X = x^c, W = w$	2/5	14/25
$Z = z \mid X = x^c, W = w^c$	2/5	14/25

12 Conclusion

The main contribution of this article is a solid theory of graphical models of inference based on the theory of credal sets and Walley's definitions of irrelevance and independence. Credal networks offer great flexibility in dealing with imprecise probability elicitation, incomplete models, and robustness assessment.

The main technical contributions are novel results for inferences with strong and natural extensions, and novel algorithms to deal with natural extensions. Strong extensions were tied to graphical d-separation and emerged as straightforward generalizations of standard Bayesian networks. Natural extensions were analyzed through linear fractional programming, and novel properties of natural extensions were derived and used to simplify inference algorithms. The inherent complexity of natural extensions limits the applicability of the algorithms to small networks, but the algorithms produce exact solutions to be used to test and verify more efficient, perhaps approximate, algorithms.

The article 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. What is the best concept of independence for credal sets, and the best method of extension to use in practice? What are the separation properties of natural extensions? How to elicit information about credal sets from experts? These questions are to be addressed in future investigations.

The theory of natural extensions presented in this article poses a 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 the theory of credal sets. Irrelevance can be used to define independence, and irrelevance judgements are less demanding than indepen-

dence ones but still quite powerful. Should irrelevance be a more fundamental notion?

A Proofs

PROOF (Theorem 11). 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 strong 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 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 strong extension:

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

Given the positivity assumption, and because every 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 collection of lower expectations uniquely defines a convex set of measures, 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 13). Given any bounded function $f(X_i)$, the assumptions of the theorem and the properties of credal sets imply that

$$\underline{E}[f(X_i)|\text{pa}(X_i), \mathbf{S}] \geq \min_{\text{nd}(X_i) \setminus \{\text{pa}(X_i), \mathbf{S}\}} \underline{E}[f(X_i)|\text{nd}(X_i)] = \underline{E}[f(X_i)|\text{pa}(X_i)],$$

and

$$\begin{aligned}\underline{E}[f(X_i)|\mathrm{pa}(X_i)] &\geq \underline{E}[\underline{E}[f(X_i)|\mathrm{pa}(X_i),\mathbf{S}]\,|\mathrm{pa}(X_i)] \\ &\geq \underline{E}[\underline{E}[f(X_i)|\mathrm{pa}(X_i)]\,|\mathrm{pa}(X_i)] = \underline{E}[f(X_i)|\mathrm{pa}(X_i)]\,.\end{aligned}$$

Consequently, $\underline{E}[\underline{E}[f(X_i)|\mathrm{pa}(X_i),\mathbf{S}]\,|\mathrm{pa}(X_i)] = \underline{E}[f(X_i)|\mathrm{pa}(X_i)]$. Using the positivity assumption and that $\underline{E}[f(X_i)|\mathrm{pa}(X_i),\mathbf{S}] \geq \underline{E}[f(X_i)|\mathrm{pa}(X_i)]$, it must be the case that $\underline{E}[f(X_i)|\mathrm{pa}(X_i),\mathbf{S}] = \underline{E}[f(X_i)|\mathrm{pa}(X_i)]$. 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 15). The credal set $K_r(\mathbf{S})$ must satisfy irrelevance relations in $\mathbf{S}$ (Lemma 13), so every measure in $K_r(\mathbf{S})$ belongs to $K'_r(\mathbf{S})$ (i.e., $K_r(\mathbf{S}) \subseteq K'_r(\mathbf{S})$).

To prove that every measure in $K'_r(\mathbf{S})$ belongs to $K_r(\mathbf{S})$, construct a credal set $K''(\mathbf{X})$ defined by all densities $p''(\mathbf{X})$ such that:

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

for all $p'(\mathbf{S})$ that belong to $K'_r(\mathbf{S})$ and where each $p(X_i|\mathrm{pa}(X_i))$ is selected from a local credal set $K(X_i|\mathrm{pa}(X_i))$.

The strategy of the proof is to consider the irrelevance relations satisfied by $K''(\mathbf{X})$. For any given X_i, consider the conventions:

$$\begin{aligned}\mathbf{S}' &= \mathrm{nD}(X_i) \cap \mathbf{S}, \\ \mathbf{S}'' &= \mathbf{S} \setminus \{\mathbf{S}', X_i, \mathrm{pa}(X_i)\}\,, \\ \mathbf{W} &= \mathrm{nD}(X_i) \setminus \mathbf{S}', \\ \mathbf{W}' &= \text{descendants of } X_i \text{ outside of } \mathbf{S}.\end{aligned}$$

Consider first a variable $X_i \notin \mathbf{S}$ and an arbitrary function $f(X_i)$; the value of $\underline{E}[f(X_i)|\mathrm{nd}(X_i)]$ is the minimum of

$$\frac{\sum_{X_i,\mathbf{W}',\mathbf{S}''} f(X_i) p(\mathbf{W}'|X_i,\mathrm{pa}(X_i),\mathbf{W},\mathbf{S})\, p(X_i|\mathrm{pa}(X_i))\, p(\mathrm{pa}(X_i),\mathbf{W},\mathbf{S})}{\sum_{X_i,\mathbf{W}',\mathbf{S}''} p(\mathbf{W}'|X_i,\mathrm{pa}(X_i),\mathbf{W},\mathbf{S})\, p(X_i|\mathrm{pa}(X_i))\, p(\mathrm{pa}(X_i),\mathbf{W},\mathbf{S})}.$$

Because $\sum_{\mathbf{W}'} p(\mathbf{W}'|X_i,\mathrm{pa}(X_i),\mathbf{W},\mathbf{S}) = 1$ and $\sum_{X_i} p(X_i|\mathrm{pa}(X_i)) = 1$:

$$\underline{E}[f(X_i)|\mathrm{nd}(X_i)] = \min \frac{\sum_{\mathbf{S}''} \left(\sum_{X_i} f(X_i) p(X_i|\mathrm{pa}(X_i)) \right) p(\mathrm{pa}(X_i),\mathbf{W},\mathbf{S})}{\sum_{\mathbf{S}''} p(\mathrm{pa}(X_i),\mathbf{W},\mathbf{S})}$$

$$
\begin{aligned}
&= \min \sum_{X_i} f(X_i) p(X_i|\mathrm{pa}(X_i)) \\
&= \underline{E}[f(X_i)|\mathrm{pa}(X_i)] \,.
\end{aligned}
$$

As $f(X_i)$ is arbitrary, the credal set $K''(X_i|\mathrm{nd}(X_i))$ is equal to $K''(X_i|\mathrm{pa}(X_i))$.

Consider now a variable $X_i \in \mathbf{S}$:

$$
\begin{aligned}
\underline{E}[f(X_i)|\mathrm{nd}(X_i)] &= \underline{E}[f(X_i)|\mathrm{pa}(X_i), \mathbf{S}', \mathbf{W}] \\
&= \min \left(\frac{\sum_{\mathbf{W}', X_i, \mathbf{S}''} f(X_i) p(\mathbf{W}'|\mathbf{W}, \mathbf{S})\, p(\mathbf{W}|\mathbf{S})\, p(\mathbf{S})}{\sum_{\mathbf{W}', X_i, \mathbf{S}''} p(\mathbf{W}'|\mathbf{W}, \mathbf{S})\, p(\mathbf{W}|\mathbf{S})\, p(\mathbf{S})} \right) \\
&= \min \left(\frac{p(\mathbf{W}|\mathbf{S}') \sum_{X_i, \mathbf{S}''} f(X_i) p(\mathbf{S})}{p(\mathbf{W}|\mathbf{S}') \sum_{X_i, \mathbf{S}''} p(\mathbf{S})} \right) \\
&= \min \left(\frac{\sum_{X_i, \mathbf{S}''} f(X_i) p(\mathbf{S})}{\sum_{X_i, \mathbf{S}''} p(\mathbf{S})} \right) \\
&= \underline{E}[f(X_i)|\mathrm{pa}(X_i), \mathbf{S}'] \,.
\end{aligned}
$$

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 $K''(\mathbf{X}) \subseteq K_r(\mathbf{X})$. Because $K'_r(\mathbf{S})$ is the exact marginal of $K''(\mathbf{X})$, $K'_r(\mathbf{S}) \subseteq K_r(\mathbf{S})$. This proves that $K'_r(\mathbf{S})$ and $K_r(\mathbf{S})$ are equal.

PROOF (Theorem 16). 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 identical to independence in standard Bayesian networks; 2) Lemma 13 guarantees that all irrelevance conditions are valid when restricted to the network of $\mathbf{W}$; 3) independence of a variable from its nondescendants non-parents given its parents characterizes a unique joint probability density [62].

By Theorem 15, the natural extension for the top sub-network containing variables $\mathbf{S}$ defines the set of all $p(\mathbf{S})$; any joint density $p(\mathbf{X})$ that belongs to $K_r(\mathbf{X})$ satisfies $p(\mathbf{X}) = p(\mathbf{W}|\mathbf{S})\, p(\mathbf{S})$. Using this fact in Expression (2):

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

PROOF (Lemma 18). By hypothesis, $\underline{E}[f(\mathrm{nD}(X_i))|\mathrm{pa}(X_i), X_i]$ is equal to $\underline{E}[f(\mathrm{nD}(X_i))|\mathrm{pa}(X_i)]$ for any bounded $f(\mathrm{nD}(X_i))$. In particular, $f(\mathrm{nD}(X_i)) = g(\mathbf{S})$ leads to $\underline{E}[g(\mathbf{S})|\mathrm{pa}(X_i), X_i] = \underline{E}[g(\mathbf{S})|\mathrm{pa}(X_i)]$; and because $g(\mathbf{S})$ is an arbitrary function, $K(\mathbf{S}|\mathrm{pa}(X_i), X_i)$ and $K(\mathbf{S}|\mathrm{pa}(X_i))$ are equal.

PROOF (Theorem 19). The credal set $K_n(\mathbf{S})$ must satisfy all independence relations in $\mathbf{S}$ (Lemmas 13 and 18), so every measure in $K_n(\mathbf{S})$ belongs to $K'_n(\mathbf{S})$ (i.e., $K_n(\mathbf{S}) \subseteq K'_n(\mathbf{S})$). To prove that every measure in $K'_n(\mathbf{S})$ belongs to $K_n(\mathbf{S})$, construct a credal set $K''(\mathbf{X})$ following Expression (A.1). By Theorem 15, the nondescendants non-parents of X_i are irrelevant to X_i given the parents of X_i. It remains to be shown that X_i is irrelevant to its nondescendants given its parents. Consider the same conventions for $\mathbf{W}$, $\mathbf{W}'$, $\mathbf{S}'$ and $\mathbf{S}''$.

Consider first a variable $X_i \notin \mathbf{S}$ and an arbitrary function $f(X_i)$; the value of $\underline{E}[\mathrm{nD}(X_i)|\mathrm{pa}(X_i), X_i]$ is the minimum of

$$\frac{\sum_{\mathbf{W},\mathbf{W}',\mathbf{S}'} f(\mathrm{nD}(X_i))p(\mathbf{W}'|X_i, \mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})\,p(X_i|\mathrm{pa}(X_i))\,p(\mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})}{\sum_{\mathbf{W},\mathbf{W}',\mathbf{S}'} p(\mathbf{W}'|X_i, \mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})\,p(X_i|\mathrm{pa}(X_i))\,p(\mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})}.$$

Because $\sum_{\mathbf{W}'} p(\mathbf{W}'|X_i, \mathrm{pa}(X_i), \mathbf{W}, \mathbf{S}) = 1$:

$$\begin{aligned}
\underline{E}[f(\mathrm{nD}(X_i))|\mathrm{pa}(X_i), X_i] = & \\
& \min \frac{\sum_{\mathbf{W},\mathbf{S}'} f(\mathrm{nD}(X_i))p(X_i|\mathrm{pa}(X_i))\,p(\mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})}{\sum_{\mathbf{W},\mathbf{S}'} p(X_i|\mathrm{pa}(X_i))\,p(\mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})} \\
& = \min \frac{p(X_i|\mathrm{pa}(X_i)) \sum_{\mathbf{W},\mathbf{S}'} f(\mathrm{nD}(X_i))p(\mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})}{p(X_i|\mathrm{pa}(X_i)) \sum_{\mathbf{W},\mathbf{S}'} p(\mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})} \\
& = \min \frac{\sum_{\mathbf{W},\mathbf{S}'} f(\mathrm{nD}(X_i))p(\mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})}{\sum_{\mathbf{W},\mathbf{S}'} p(\mathrm{pa}(X_i), \mathbf{W}, \mathbf{S})} \\
& = \underline{E}[f(\mathrm{nD}(X_i))|\mathrm{pa}(X_i)]\,.
\end{aligned}$$

As $f(\mathrm{nD}(X_i))$ is arbitrary, the credal set $K''(\mathrm{nD}(X_i)|\mathrm{pa}(X_i), X_i)$ is equal to $K''(\mathrm{nD}(X_i)|\mathrm{pa}(X_i))$.

Consider now a variable $X_i \in \mathbf{S}$:

$$\begin{aligned}
\underline{E}[f(\mathrm{nD}(X_i))|\mathrm{pa}(X_i), X_i] &= \underline{E}[f(\mathbf{S}', \mathbf{W})|\mathrm{pa}(X_i), X_i] \\
&= \min\left(\frac{\sum_{\mathbf{W},\mathbf{W}',\mathbf{S}\setminus\{X_i,\mathrm{pa}(X_i)\}} f(\mathbf{S}', \mathbf{W}) p(\mathbf{W}'|\mathbf{W}, \mathbf{S})\, p(\mathbf{W}|\mathbf{S})\, p(\mathbf{S})}{\sum_{\mathbf{W},\mathbf{W}',\mathbf{S}\setminus\{X_i,\mathrm{pa}(X_i)\}} p(\mathbf{W}'|\mathbf{W}, \mathbf{S})\, p(\mathbf{W}|\mathbf{S})\, p(\mathbf{S})}\right) \\
&= \min\left(\frac{\sum_{\mathbf{S}\setminus\{X_i,\mathrm{pa}(X_i)\}} \left(\sum_{\mathbf{W}} f(\mathbf{S}', \mathbf{W}) p(\mathbf{W}|\mathbf{S})\right) p(\mathbf{S})}{\sum_{\mathbf{S}\setminus\{X_i,\mathrm{pa}(X_i)\}} \left(\sum_{\mathbf{W}} p(\mathbf{W}|\mathbf{S})\right) p(\mathbf{S})}\right) \\
&= \min\left(\frac{\sum_{\mathbf{S}\setminus\{X_i,\mathrm{pa}(X_i)\}} \left(\sum_{\mathbf{W}} f(\mathbf{S}', \mathbf{W}) p(\mathbf{W}|\mathbf{S})\right) p(\mathbf{S})}{\sum_{\mathbf{S}\setminus\{X_i,\mathrm{pa}(X_i)\}} p(\mathbf{S})}\right) \\
&= \underline{E}\left[\sum_{\mathbf{W}} f(\mathbf{S}', \mathbf{W}) p(\mathbf{W}|\mathbf{S}')\,|\mathrm{pa}(X_i), X_i\right] \\
&= \underline{E}\left[\sum_{\mathbf{W}} f(\mathbf{S}', \mathbf{W}) p(\mathbf{W}|\mathbf{S}')\,|\mathrm{pa}(X_i)\right] \text{ (by hypothesis).}
\end{aligned}$$

Because $p(\mathbf{W}|\mathbf{S}')$ is a unique probability density in the credal set $K''(\mathbf{X})$, $\underline{E}[\sum_{\mathbf{W}} f(\mathbf{S}', \mathbf{W}) p(\mathbf{W}|\mathbf{S}')\,|\mathrm{pa}(X_i)] = \underline{E}[f(\mathbf{S}', \mathbf{W})|\mathrm{pa}(X_i)]$ [76], and then

$$\underline{E}[f(\mathrm{nD}(X_i))|\mathrm{pa}(X_i), X_i] = \underline{E}[f(\mathrm{nD}(X_i))|\mathrm{pa}(X_i)]\,.$$

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 $K''(\mathbf{X}) \subseteq K_n(\mathbf{X})$. Because $K'_n(\mathbf{S})$ is the exact marginal of $K''(\mathbf{X})$, $K'_n(\mathbf{S}) \subseteq K_n(\mathbf{S})$. This proves that $K'_n(\mathbf{S})$ and $K_n(\mathbf{S})$ are equal.

B Constraints for marginal and conditional credal sets in natural extensions

The algorithm IncorporateVariable assumes that linear inequalities for a marginal credal set $K(\mathbf{X})$ and a conditional credal set $K(\mathbf{X}|\mathbf{Y} = \mathbf{y})$ can be generated from a given finitely generated 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:

$$\mathbf{A}[p(\mathbf{X}, \mathbf{Y})] \leq \mathbf{B}, \tag{B.1}$$

where $[p(\mathbf{X}, \mathbf{Y})]$ denotes a vector containing all values of the density $p(\mathbf{X}, \mathbf{Y})$. The mandatory unitary constraint, $\sum p(\mathbf{X}, \mathbf{Y}) = 1$, is assumed to be part of Expression (B.1).

It is possible to obtain $K(\mathbf{X})$ and $K(\mathbf{X}|\mathbf{Y} = \mathbf{y})$ from Expression (B.1) using Fourier-Motzkin elimination [37]. This elimination procedure is a classic algo-

rithm for the solution of systems of inequalities, and variants of it have been used in probabilistic logic since the work of Boole [40]. The procedure starts with a collection of linear inequalities and eliminates one variable at a time; each elimination step is accomplished with the addition of new linear inequalities. The main difficulty of Fourier-Motzkin elimination is that a quadratic number of inequalities may be introduced at each elimination step.

To obtain the marginal credal set $K(\mathbf{X})$, denote by $[p(\mathbf{X})]$ the vector of all values of a generic density $p(\mathbf{X})$. Note that an element of $[p(\mathbf{X})]$ is related to $[p(\mathbf{X}, \mathbf{Y})]$ by the basic probability rule:

$$p(\mathbf{X}) = \sum_{\mathbf{Y}} p(\mathbf{X}, \mathbf{Y}) \,. \tag{B.2}$$

Take the joint collection of constraints in Expressions (B.1) and (B.2). Now apply the Fourier-Motzkin elimination algorithm to eliminate all variables in $[p(\mathbf{X}, \mathbf{Y})]$ and to produce the linear inequalities for $[p(\mathbf{X})]$.

To obtain the conditional credal set $K(\mathbf{X}|\mathbf{Y} = \mathbf{y})$, denote by

- $[p(\mathbf{X}, \mathbf{Y} = \mathbf{y})]$ the vector of all values of a generic density $p(\mathbf{X}, \mathbf{Y} = \mathbf{y})$, and
- $[p(\mathbf{X}|\mathbf{Y} = \mathbf{y})]$ the vector of all values of a generic density $p(\mathbf{X}|\mathbf{Y} = \mathbf{y})$.

First obtain the inequalities for the joint credal set $K(\mathbf{X}, \mathbf{Y} = \mathbf{y})$. Note that $[p(\mathbf{X}, \mathbf{Y} = \mathbf{y})]$ is a subset of $[p(\mathbf{X}, \mathbf{Y})]$, so the constraints on $[p(\mathbf{X}, \mathbf{Y} = \mathbf{y})]$ are obtained by application of Fourier-Motzkin elimination in Expression (B.1). The resulting constraints can be written as a matricial inequality:

$$\mathbf{A}'[p(\mathbf{X}, \mathbf{Y} = \mathbf{y})] \leq \mathbf{B}'. \tag{B.3}$$

Now introduce a new variable $W \geq 0$ and form the matricial inequality

$$\mathbf{A}'[p(\mathbf{X}, \mathbf{Y} = \mathbf{y})] \leq W\mathbf{B}',$$

and add the unitary constraint

$$\sum_{\mathbf{X}} p(\mathbf{X}, \mathbf{Y} = \mathbf{y}) = 1.$$

The variable W works as a "scale" for $[p(\mathbf{X}, \mathbf{Y} = \mathbf{y})]$ and creates a "cone", with apex at the origin, extending through the polytope of possible values for $[p(\mathbf{X}, \mathbf{Y} = \mathbf{y})]$. Now apply the Fourier-Motzkin algorithm again to eliminate W and produce the linear inequalities that define the conditional credal set $K(\mathbf{X}|\mathbf{Y} = \mathbf{y})$.

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. I benefited from joint work with Peter Walley on graphoid properties of irrelevance; the proof of Lemma 13 is based on his statement of my original proof. Thanks to David Avis for freely distributing the lrs package. I thank one of the anonymous reviewers, who detected a flaw in my first version of Theorem 11 and suggested several improvements.

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: G. Goos, J. Hartmanis, and J. van Leeuwen, eds., *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (Paris, France, July 1994) 4–8.

[11] A. Cano and S. Moral, A genetic algorithm to approximate convex sets of probabilities, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (Granada, Spain, 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: E. Horvitz and F. Jensen, eds., *XII Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, California, 1996) 169–177.

[16] L. Chrisman, Propagation of 2-monotone lower probabilities on an undirected graph, in: E. Horvitz and F. Jensen, eds., *XII Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, California, 1996) 178–186.

[17] K. L. Clarkson, K. Mehlhorn, and R. Seidel, Four results on randomized incremental constructions, in: *International Symposium on Theoretical Aspects of Computer Science (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] I. Couso, S. Moral, and P. Walley, Examples of independence for imprecise probabilities, in: G. de Cooman, F. G. Cozman, S. Moral, and P. Walley, eds., *Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications* (Imprecise Probabilities Project, Universiteit Gent, Ghent, Belgium, 1999) 121–130.

[20] F. G. Cozman, Robustness analysis of Bayesian networks with local convex sets of distributions, in: D. Geiger and P. Shenoy, eds., *XIII Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, California, 1997) 108-115.

[21] F. G. Cozman, Irrelevance and independence in Quasi-Bayesian networks, in: G. Cooper and S. Moral, eds., *XIV Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, California, 1998) 89–96.

[22] F. G. Cozman, Irrelevance and independence axioms in quasi-Bayesian theory, in: A. Hunter and S. Parsons, eds., *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)* (Springer, London, England, 1999) 128–136.

[23] L. de Campos and S. Moral, Independence concepts for convex sets of probabilities, in: P. Besnard and S. Hanks, eds., *XI Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, California, 1995) 108–115.

[24] G. de Cooman, F. G. Cozman, S. Moral, and P. Walley (eds.), *Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications* (Imprecise Probabilities Project, Universiteit Gent, Ghent, Belgium, 1999).

[25] R. Dechter, Bucket elimination: A unifying framework for probabilistic inference, in: E. Horvitz and F. Jensen, eds., *XII Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, California, 1996) 211–219.

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

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

[28] J. Earman, *Bayes or Bust?* (The MIT Press, Cambridge, Massachussets, 1992).

[29] E. Fagiuoli and M. Zaffalon, 2U: An exact interval propagation algorithm for polytrees with binary variables, *Artificial Intelligence* **106** (1) (1998) 77–107.

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

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

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

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

[34] 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.

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

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

[37] 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.

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

[39] V. Ha and P. Haddawy, Theoretical foundations for abstraction-based probabilistic planning, in: E. Horvitz and F. Jensen, eds., *XII Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, California, 1996) 291–298.

[40] T. Hailperin, *Sentential Probability Logic* (Lehigh University Press, Bethlehem, 1996).

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

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

[43] 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.

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

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

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

[47] 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.

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

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

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

[51] 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.

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

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

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

[55] 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.

[56] S. Moral, A formal language for convex sets of probabilities, in: M. R. B. Clarke, R. Kruse and S. Moral, *Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU)* (Springer-Verlag, 1993) 274–281.

[57] R. Musick, Minimal assumption distribution propagation in belief networks, in: D. Heckerman and A. Mamdani, eds., *IX Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, California, 1993) 251–258.

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

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

[60] 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.

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

[62] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Mateo, California, 1988).

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

[64] 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.

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

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

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

[68] G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, 1976).

[69] 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.

[70] 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).

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

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

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

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

[75] 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.

[76] P. Walley, *Statistical Reasoning with Imprecise Probabilities* (Chapman and Hall, London, 1991).

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

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

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

[80] L. A. 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.

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

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

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

[84] M. Zaffalon. A credal approach to naive classification. in: G. de Cooman, F. G. Cozman, S. Moral, and P. Walley, eds., *Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications* (Imprecise Probabilities Project, Universiteit Gent, Ghent, Belgium, 1999) 405–414.