A Generalization of Credal Networks

Marco E. G. V. Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

Abstract

The likelihood approach to statistics can be interpreted as a theory of fuzzy probability. This paper presents a generalization of credal networks obtained by generalizing imprecise probabilities to fuzzy probabilities; that is, by additionally considering the relative plausibility of different values in the probability intervals.

Keywords. Bayesian networks, credal networks, graphical models, d-separation, imprecise probabilities, updating, likelihood function, hierarchical model, fuzzy probabilities.

1 Introduction

A common interpretation of membership functions of fuzzy sets is as statistical likelihood functions. With this interpretation, the well-established likelihood approach to statistics appears as a theory of fuzzy probabilities. These generalize imprecise probabilities by additionally considering the relative plausibility of different values in the probability or expectation intervals. Besides the increased expressive power, the fundamental advantage of the likelihood-based fuzzy probabilities with respect to imprecise probabilities is the ability of using all the information provided by the data. In fact, the resulting hierarchical model exploits the outstanding statistical properties of the likelihood function, which makes it an ideal basis for inference and decision making (see Cattaneo, 2005, 2007).

In the present paper, the hierarchical model is used in the framework of belief networks, to describe the uncertain knowledge about the values of the involved variables. This leads to a generalization of Bayesian networks and credal networks, combining the possibility of imprecision in the probability values with the ability of using all the information provided by the data.

In Section 2 the hierarchical model is briefly introduced (see Cattaneo, 2008*a*, for a more detailed description), while in Section 3 some aspects of the model of great practical importance are presented. Finally, in Section 4 the hierarchical networks are defined and compared with credal networks.

2 Hierarchical Model

In most theories of imprecise probability, the model corresponds to a set $\mathcal{P}$ of probability measures on a measurable space $(\Omega, \mathcal{A})$. The set $\mathcal{P}$ is often assumed to be convex, and when an event $A \in \mathcal{A}$ is observed, $\mathcal{P}$ is usually updated to

$$\mathcal{P}' = \{P(\cdot \mid A) : P \in \mathcal{P},\, P(A) > 0\} \tag{1}$$

(that is, each $P \in \mathcal{P}$ is conditioned on A). The conditional probability measure $P(\cdot \mid A)$ is obtained by normalizing the "restricted" measure $P(\cdot \cap A)$, but the normalization step deletes the information about the value $P(A)$. The values $P_1(A), P_2(A)$ describe the relative ability of the probability measures $P_1, P_2 \in \mathcal{P}$ to forecast the observed event A (before observing it): the larger the probability value, the better the forecast. These values are combined in the *likelihood* function lik' on $\mathcal{P}'$ defined (up to a positive multiplicative constant) by

$$lik'(P') \propto \sup_{P \in \mathcal{P} \,:\, P(\cdot \mid A) = P'} lik(P)\, P(A) \tag{2}$$

for all $P' \in \mathcal{P}'$, where lik was the likelihood function on $\mathcal{P}$ before observing A. The likelihood function is a central concept in statistical inference: it is usually interpreted as a measure of the *relative* plausibility of the probability measures as models of the reality under consideration (proportional likelihood functions are considered equivalent). When A is the first observed event, the *prior* likelihood function $lik : \mathcal{P} \to (0, \infty)$ can be interpreted as a (subjective) measure of the relative plausibility of the elements of $\mathcal{P}$ according to the prior information (see also Dahl,

2005). In particular, prior ignorance is described by a constant prior likelihood function lik; in this case, (2) corresponds to the usual definition of statistical likelihood function induced by the observation of the event A (apart from the fact that lik' is defined on $\mathcal{P}'$ instead of $\mathcal{P}$). In general, the prior likelihood function is interpreted and used as if it were the statistical likelihood function induced by (hypothetical) past data.

In the likelihood approach to statistics (see for example Pawitan, 2001), the likelihood of a set of probability measures is usually defined as the supremum of the likelihood of its elements (this idea is used also in (2), if there are several $P \in \mathcal{P}$ such that $P(\cdot \mid A) = P'$). When lik is a likelihood function on $\mathcal{P}$, the set function LR on $2^{\mathcal{P}}$ defined by

$$LR(\mathcal{H}) = \frac{\sup_{P \in \mathcal{H}} lik(P)}{\sup_{P \in \mathcal{P}} lik(P)}$$

for all $\mathcal{H} \subseteq \mathcal{P}$ (in this paper, $\sup \varnothing = 0$) is a normalized *possibility* measure with possibility distribution proportional to lik. A possibility distribution can also be considered as the membership function of a *fuzzy* set (see Zadeh, 1978). In the present paper, possibility distributions and membership functions are interpreted as proportional to likelihood functions: this is a common interpretation (see in particular Hisdal, 1988, and Dubois, 2006). Hence, it suffices to consider normalized fuzzy sets and normalized possibility measures, but grades of membership and degrees of possibility have only a relative meaning.

The set $\mathcal{P}$ of probability measures and the likelihood function lik on $\mathcal{P}$ can be considered as the two levels of a *hierarchical* model: these two levels describe different kinds of uncertainty (probabilistic and possibilistic, respectively). When an event $A \in \mathcal{A}$ is observed, the two levels $\mathcal{P}$ and lik of the hierarchical model are updated to $\mathcal{P}'$ and lik' according to (1) and (2), respectively. The uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the induced possibility measure $LR \circ g^{-1}$ on $\mathcal{G}$ (in this paper, g^{-1} denotes the set function associating to each subset of $\mathcal{G}$ its inverse image under g); that is, by the normalized fuzzy subset of $\mathcal{G}$ with membership function proportional to the *profile* likelihood function lik_g on $\mathcal{G}$ defined (up to a positive multiplicative constant) by

$$lik_g(\gamma) \propto \sup_{P \in \mathcal{P} \,:\, g(P) = \gamma} lik(P)$$

for all $\gamma \in \mathcal{G}$. In particular, if g associates to each probability measure $P \in \mathcal{P}$ the expectation $g(P) = E_P(X)$ of a random variable X, or the probability $g(P) = P(B)$ of an event $B \in \mathcal{A}$, then the normalized fuzzy subset of $\mathbb{R}$ with membership function proportional to lik_g can be interpreted as the fuzzy expectation of X, or the fuzzy probability of B, respectively. In this sense, the likelihood approach to statistics can be interpreted as a theory of fuzzy probability. The discussion on how to evaluate by one or more real numbers the normalized fuzzy subset of $\mathbb{R}$ with membership function proportional to lik_g goes beyond the scope of the present paper (but see Cattaneo, 2007, for some interesting results): only the α-cut

$$\left\{x \in \mathbb{R} : lik_g(x) \geq \alpha \sup_{y \in \mathbb{R}} lik_g(y)\right\}$$

with $\alpha \in (0, 1)$ will be considered here. This is a likelihood-based confidence region for $g(P)$, whose coverage probability can often be approximated thanks to the result of Wilks (1938): in particular, 95% coverage probability corresponds to $\alpha = 0.1465$.

Example 1 *Consider an urn containing* 3 *balls: one ball is white, another is black, while the third one could be white or black. We have no idea about the color (white or black) of the third ball, but we can perform a sequence of random draws with replacement from the urn, and observe the colors of the balls drawn. Conditional on the composition of the urn, these observations can be described as a sequence of independent Bernoulli trials with constant probability* $\frac{1}{3}$ *or* $\frac{2}{3}$ *of observing a black ball (depending on the color of the third ball: white or black, respectively). We shall never be able to determine with absolute certainty the composition of the urn, but if in a long sequence of draws the proportion of black balls is approximately* $\frac{2}{3}$*, then it is much more plausible that the color of the third ball is black than it is white.*

Let $\mathcal{P}$ *be the (convex) set of probability measures resulting from the only imprecise prior probability measure about the composition of the urn (that is, about the color of the third ball) such that the probability of observing a black ball in the first draw is described by the interval* $[\frac{1}{3}, \frac{2}{3}]$*. This is the vacuous imprecise prior, and therefore, if* $\mathcal{P}$ *is updated according to (1), then the (posterior) imprecise probability of observing a black ball in the next draw remains* $[\frac{1}{3}, \frac{2}{3}]$*, independently of the number and colors of the balls drawn. By contrast, the (posterior) fuzzy probability of observing a black ball in the next draw (resulting from the hierarchical model with constant prior likelihood function on* $\mathcal{P}$*) evolves as expected: it tends to concentrate on the value* $\frac{2}{3}$*, when in a sequence of draws of increasing length the proportion of black balls remains approximately* $\frac{2}{3}$*. Figure 1 shows the graphs of the membership functions of the fuzzy probability* p *of observing a black ball in the next draw: prior to any draw (dotted line), after drawing* 2 *white balls and* 5 *black balls (dashed line), and after drawing* 8 *white balls and* 15 *black balls (solid line); in particular, the*

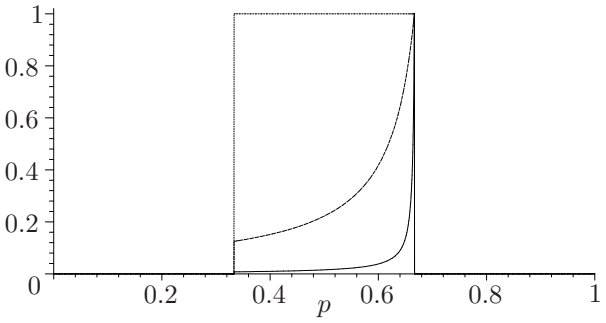

Figure 1: Membership functions of the fuzzy probability p of observing a black ball in the next draw (in the situation of Example 1): prior to any draw (dotted line), after drawing 2 white balls and 5 black balls (dashed line), and after drawing 8 white balls and 15 black balls (solid line).

corresponding α-cuts with $\alpha = 0.1465$ are the intervals $[0.333, 0.667]$, $[0.389, 0.667]$, *and* $[0.651, 0.667]$, *respectively.*

These membership functions can be easily obtained by applying the results of Section 3; the detailed calculations will be presented in Example 4.

The hierarchical model with levels $\mathcal{P}$ and lik generalizes the imprecise probability model $\mathcal{P}$, since the probabilistic level is updated in the same way (1) as the imprecise probability model, while the possibilistic level carries additional information. In particular, the fuzzy expectation of a random variable X is a fuzzy subset of the imprecise expectation

$$[\underline{E}(X), \overline{E}(X)] = [\inf_{P\in\mathcal{P}} E_P(X), \sup_{P\in\mathcal{P}} E_P(X)]\,,$$

and the fuzzy probability of an event $B \in \mathcal{A}$ is a fuzzy subset of the imprecise probability

$$[\underline{P}(B), \overline{P}(B)] = [\inf_{P\in\mathcal{P}} P(B), \sup_{P\in\mathcal{P}} P(B)]\,,$$

since their membership functions are constant equal to 0 outside these intervals (for example, the fuzzy probabilities of Figure 1 are fuzzy subsets of the imprecise probability $[\frac{1}{3}, \frac{2}{3}]$). That is, fuzzy probabilities generalize imprecise probabilities by additionally considering the relative plausibility of different values in the probability intervals (imprecise probabilities correspond to the crisp case of fuzzy probabilities). This additional information allows us in particular to get out of the state of complete ignorance; that is, to reach nontrivial conclusions also when starting with the vacuous prior, as in Example 1. Alternative updating rules for the imprecise probability model $\mathcal{P}$, making use of some information contained in the possibilistic level lik, have been proposed in particular by Moral (1992), Wilson (2001), and Held et al. (2008): these updating rules discard some of the less plausible probability measures in $\mathcal{P}$, but this can lead to important problems, since any discarded probability measure can become the most plausible one in the light of new data. To avoid these problems, it is necessary to store more information than it is possible in an imprecise probability model: the hierarchical model provides a simple solution.

When the probabilistic level of the hierarchical model is a singleton $\mathcal{P} = \{P\}$, the possibilistic level contains no information, since the likelihood function is defined only up to a positive multiplicative constant. In this case, the membership function of the fuzzy expectation of a random variable X, or of the fuzzy probability of an event $B \in \mathcal{A}$, is the indicator function of $\{E_P(X)\}$, or of $\{P(B)\}$, respectively; and when an event $A \in \mathcal{A}$ is observed, the probabilistic level is updated according to (1) by conditioning P on A. Hence, the purely probabilistic description of uncertain knowledge about $\omega \in \Omega$ (that is, the Bayesian model) is a special case of the hierarchical model. The same is true also for the purely possibilistic description of uncertain knowledge about $\omega \in \Omega$: a normalized possibility measure Π on Ω with possibility distribution π can be described by the hierarchical model with as probabilistic level the set $\mathcal{P} = \{\delta_\omega : \omega \in \Omega,\, \pi(\omega) > 0\}$ (where δ_ω is the Dirac measure on Ω concentrated on ω), and as possibilistic level the likelihood function lik on $\mathcal{P}$ defined (up to a positive multiplicative constant) by $lik(\delta_\omega) \propto \pi(\omega)$ for all $\delta_\omega \in \mathcal{P}$. In this case, $\Pi = LR \circ t^{-1}$ is the possibility measure on Ω induced by the identification of each Dirac measure $\delta_\omega \in \mathcal{P}$ with the corresponding $\omega \in \Omega$, described by the function $t : \mathcal{P} \to \Omega$ with $t(\delta_\omega) = \omega$ for all $\delta_\omega \in \mathcal{P}$. The fuzzy expectation of a random variable X corresponds then to the possibility measure $\Pi \circ X^{-1}$ on $\mathbb{R}$ induced by $X : \Omega \to \mathbb{R}$; and when an event $A \in \mathcal{A}$ is observed, the hierarchical model is updated according to (1) and (2) to the hierarchical model with levels $\mathcal{P}' = t^{-1}(A)$ and $lik' = lik|_{\mathcal{P}'}$ (the restriction of lik to $\mathcal{P}'$). That is, when A is observed, Π is updated to the normalized possibility measure Π' on Ω with possibility distribution proportional to the pointwise product of π and the indicator function of A.

The hierarchical model offers a unified approach to the combination of probabilistic and possibilistic uncertainty (for instance, fuzzy data can be used without problem). Since membership functions and possibility distributions are interpreted as proportional to likelihood functions, the rules for manipulating fuzzy probabilities are implied by the well-established theories of probability and likelihood. It is important to underline that other interpretations of mem-

bership functions and possibility distributions would lead to other rules for manipulating fuzzy probabilities; in particular, the updating rule would be different. For example, Walley (1997) and De Cooman (2005) interpret possibility measures as upper probability measures: the resulting fuzzy probability model is a special case of the imprecise probability model (at least from the mathematical standpoint); in particular, constant possibility distributions remain constant independently of the data observed (that is, we cannot get out of the state of complete ignorance).

3 Convex Hierarchical Models

Let $\mathcal{M}_0$ be the set of all finite measures μ on the measurable space $(\Omega, \mathcal{A})$, and let $\mathcal{P}_0 \subset \mathcal{M}_0$ be the set of all probability measures P on $(\Omega, \mathcal{A})$. Hence, $\mathcal{M}_0$ and $\mathcal{P}_0$ are subsets of the real vector space of all finite signed measures on $(\Omega, \mathcal{A})$. Let $\mu_0 \in \mathcal{M}_0 \setminus \mathcal{P}_0$ be the measure on $(\Omega, \mathcal{A})$ with $\mu_0(\Omega) = 0$ (that is, μ_0 has constant value 0). The *normalization* function $n : \mathcal{M}_0 \setminus \{\mu_0\} \to \mathcal{P}_0$ is defined by $n(\mu) = [\mu(\Omega)]^{-1}\,\mu$ for all $\mu \in \mathcal{M}_0 \setminus \{\mu_0\}$, where the multiplication of μ with the normalization constant $[\mu(\Omega)]^{-1}$ is to be interpreted pointwise. The restriction $n|_{\mathcal{P}_0}$ of n to $\mathcal{P}_0$ is the identity function on $\mathcal{P}_0$, since $P(\Omega) = 1$ for all $P \in \mathcal{P}_0$. A set $\mathcal{M} \subset \mathcal{M}_0$ is said to be *bounded* if $\sup_{\mu\in\mathcal{M}} \mu(\Omega)$ is finite.

Each bounded set $\mathcal{M} \subset \mathcal{M}_0$ such that $\mathcal{M} \setminus \{\mu_0\}$ is not empty describes a hierarchical model: the probabilistic level

$$\mathcal{P} = \{n(\mu) : \mu \in \mathcal{M} \setminus \{\mu_0\}\}$$

is the image of $\mathcal{M}\setminus\{\mu_0\}$ under n, and the possibilistic level is the likelihood function lik on $\mathcal{P}$ defined (up to a positive multiplicative constant) by

$$lik(P) \propto \sup_{\substack{\mu\in\mathcal{M}\setminus\{\mu_0\}:\\ n(\mu)=P}} \mu(\Omega)$$

for all $P \in \mathcal{P}$. Each hierarchical model can be described in this way by a subset of $\mathcal{M}_0$: for example the hierarchical model with levels $\mathcal{P}$ and lik is described by

$$\mathcal{M} = \{lik(P)\, P : P \in \mathcal{P}\}$$

(where the multiplication of P with the constant $lik(P)$ is to be interpreted pointwise), but such a description is not unique: for instance the sets $\mathcal{M}\cup\{\mu_0\}$ and $\mathcal{M} \setminus \{\mu_0\}$ describe the same hierarchical model. The advantage of the description by a subset of $\mathcal{M}_0$ is that the updating is particularly simple: when an event $A \in \mathcal{A}$ is observed, the set $\mathcal{M}$ is updated to

$$\mathcal{M}' = \{\mu(\cdot \cap A) : \mu \in \mathcal{M}\}. \tag{3}$$

That is, the updated description $\mathcal{M}'$ is the image of $\mathcal{M}$ under r_A, where r_A is the function on $\mathcal{M}_0$ defined by $r_A(\mu) = \mu(\cdot \cap A)$. It can be easily proved that the update of $\mathcal{M}$ according to (3) corresponds to the update of the hierarchical model according to (1) and (2), because if $n(\mu) = P$ and $P(A) > 0$, then $(n \circ r_A)(\mu) = P(\cdot \,|\, A)$. In particular, when applied to the probability measures $P \in \mathcal{P}_0$ with $P(A) > 0$, the function $n \circ r_A$ describes the conditioning on A; hence, the updating (3) of the set $\mathcal{M}$ of measures corresponds to the updating (1) of the imprecise probability model, but without the normalization step (which deletes the information about the relative ability of the probability measures to forecast the observed event A). For the hierarchical model described by $\mathcal{M}$, the uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the normalized fuzzy subset of $\mathcal{G}$ with membership function proportional to the profile likelihood function lik_g on $\mathcal{G}$, which satisfies

$$lik_g(\gamma) \propto \sup_{\substack{\mu\in\mathcal{M}\setminus\{\mu_0\}:\\ (g\circ n)(\mu)=\gamma}} \mu(\Omega)$$

for all $\gamma \in \mathcal{G}$.

The imprecise probability model $\mathcal{P}$ corresponds to the hierarchical model with as probabilistic level the set $\mathcal{P}$, and as possibilistic level a constant likelihood function lik on $\mathcal{P}$; this hierarchical model is described by the set $\mathcal{M} = \mathcal{P} \subset \mathcal{M}_0$. The imprecise probability model $\mathcal{P}$ is often assumed to be convex; it can be easily proved that a set $\mathcal{M}'$ can be obtained by updating a convex set $\mathcal{M} = \mathcal{P}$ according to (3) if and only if $\mathcal{M}'$ is convex. A hierarchical model is said to be *convex* if it can be described by a convex subset of $\mathcal{M}_0$. Hence, the convex hierarchical models are the hierarchical models that can be interpreted as the result of updating (with real or hypothetical data) a convex imprecise probability model; that is, the convex hierarchical models are the direct generalizations of the convex imprecise probability models.

Let $\mathcal{L}_1, \mathcal{L}_2$ be real vector spaces, and let $\mathcal{C} \subseteq \mathcal{L}_1$ be convex. A function $f : \mathcal{C} \to \mathcal{L}_2$ is said to *maintain segments* if for all $x, y \in \mathcal{C}$, the image of the set

$$\{\lambda\, x + (1-\lambda)\, y : \lambda \in [0,1]\}$$

under f is the set

$$\{\lambda\, f(x) + (1-\lambda)\, f(y) : \lambda \in [0,1]\}\,.$$

The *convex hull* of a set $\mathcal{S} \subseteq \mathcal{L}_1$ is denoted by $\mathrm{ch}(\mathcal{S})$. The following result can be easily proved.

Theorem 2 *Let $\mathcal{L}_1, \mathcal{L}_2$ be real vector spaces, and let $\mathcal{C} \subseteq \mathcal{L}_1$ be convex. If the function $f : \mathcal{C} \to \mathcal{L}_2$ main-*

tains segments, and $\mathcal{S} \subseteq \mathcal{C}$, then the image of the convex hull of $\mathcal{S}$ under f is the convex hull of the image of $\mathcal{S}$ under f; that is,

$$\{f(x) : x \in \mathrm{ch}(\mathcal{S})\} = \mathrm{ch}\left(\{f(y) : y \in \mathcal{S}\}\right).$$

The *convexification* of a hierarchical model described by the set $\mathcal{M} \subset \mathcal{M}_0$ is the convex hierarchical model described by the set $\mathrm{ch}(\mathcal{M}) \subset \mathcal{M}_0$. The function r_A on $\mathcal{M}_0$ maintains segments, since it is the restriction to $\mathcal{M}_0$ of a linear map; hence, Theorem 2 implies that if $\mathcal{M}$ is updated to $\mathcal{M}'$ according to (3), then $\mathrm{ch}(\mathcal{M})$ is updated to $\mathrm{ch}(\mathcal{M}')$ according to (3). This result is particularly useful for updating the convexification of a hierarchical model described by a finite set $\mathcal{M} \subset \mathcal{M}_0$ (such models are very important in the framework of belief networks, studied in Section 4). Since the normalization function n on $\mathcal{M}_0 \setminus \{\mu_0\}$ maintains segments, Theorem 2 can be used to prove also the well-known result that if a set $\mathcal{P}$ of probability measures is updated to $\mathcal{P}'$ according to (1), then $\mathrm{ch}(\mathcal{P})$ is updated to $\mathrm{ch}(\mathcal{P}')$ according to (1).

Let $\rho : [0,\infty] \to [0,\infty]$ be the function defined by $\rho(0) = \infty$, $\rho(\infty) = 0$, and $\rho(x) = \frac{1}{x}$ for all $x \in (0,\infty)$. The function ρ is an involution; that is, $\rho \circ \rho$ is the identity function on $[0,\infty]$. The *convex hull* of a function $\phi : \mathcal{C} \to [0,\infty]$ is denoted by $\mathrm{ch}(\phi)$; that is, $\mathrm{ch}(\phi)$ is the (pointwise) largest convex function $\gamma : \mathcal{C} \to [0,\infty]$ such that $\gamma(x) \leq \phi(x)$ for all $x \in \mathcal{C}$. The following theorem is useful because for example the functions g associating to each probability measure $P \in \mathcal{P}_0$ the expectation $g(P) = E_P(X)$ of a bounded random variable X, or the probability $g(P) = P(B)$ of an event $B \in \mathcal{A}$, are the restrictions to $\mathcal{P}_0$ of linear maps. It is a consequence of Theorem 2, since if $g : \mathcal{P}_0 \to \mathcal{G}$ is the restriction to $\mathcal{P}_0$ of a linear map, then the function $f : \mathcal{M}_0 \setminus \{\mu_0\} \to \mathcal{G} \times \mathbb{R}$ defined by

$$f(\mu) = \left((g \circ n)(\mu), [\mu(\Omega)]^{-1}\right)$$

for all $\mu \in \mathcal{M}_0 \setminus \{\mu_0\}$ maintains segments.

Theorem 3 *Let $\mathcal{G}$ be a real vector space, and let $g : \mathcal{P}_0 \to \mathcal{G}$ be the restriction to $\mathcal{P}_0$ of a linear map. If π and π_{ch} are the membership functions of the normalized fuzzy subsets of $\mathcal{G}$ describing the uncertain knowledge about the value $g(P)$ of g for a hierarchical model and its convexification, respectively, then*

$$\pi_{\mathrm{ch}} = \rho \circ \mathrm{ch}(\rho \circ \pi).$$

Theorem 3 implies in particular that for a convex hierarchical model, the membership function π of the fuzzy expectation of X, or of the fuzzy probability of B, is "reciprocally convex", in the sense that $\rho \circ \pi$ is convex (since $\pi = \pi_{\mathrm{ch}}$). Moreover, Theorem 3 implies that for the convexification of a hierarchical model described by a finite set $\mathcal{M} \subset \mathcal{M}_0$ (such models are very important in the framework of belief networks, studied in Section 4), the membership function π_{ch} of the fuzzy expectation of X, or of the fuzzy probability of B, is *piecewise hyperbolic*, in the sense that $\rho \circ \pi_{\mathrm{ch}}$ is piecewise linear; in this case, the construction of π_{ch} is particularly simple, as shown in the following example.

Example 4 *Consider the situation of Example 1. Conditional on the composition of the urn (that is, conditional on the color of the third ball: white or black), the observations about the colors of the balls drawn are modeled as a sequence of independent Bernoulli trials with constant probability $\frac{1}{3}$ or $\frac{2}{3}$ of observing a black ball, described by the probability measures $P_{\frac{1}{3}}$ and $P_{\frac{2}{3}}$, respectively. The imprecise probability model $\mathcal{P}$ resulting from the vacuous imprecise prior about the composition of the urn (that is, about the color of the third ball) is the convex hull of the finite set $\mathcal{P}_B = \{P_{\frac{1}{3}}, P_{\frac{2}{3}}\}$ of probability measures. The hierarchical model with constant prior likelihood function on $\mathcal{P}$ is described by the set $\mathcal{P} = \mathrm{ch}(\mathcal{P}_B) \subset \mathcal{M}_0$; hence, it is the convexification of the hierarchical model described by the finite set $\mathcal{M} = \mathcal{P}_B \subset \mathcal{M}_0$.*

When the colors of the balls drawn are observed, the updating to $\mathcal{M}'$ according to (3) of the hierarchical model described by the finite set $\mathcal{M} = \mathcal{P}_B$ is very simple. In fact, the updating (1) of the probabilistic level $\mathcal{P}_B = \{P_{\frac{1}{3}}, P_{\frac{2}{3}}\}$ is unimportant for the probability of observing a black ball in the next draw, because the Bernoulli trials are independent under both probability measures $P_{\frac{1}{3}}$ and $P_{\frac{2}{3}}$. The constant prior likelihood function lik on $\mathcal{P}_B$ is updated to lik' according to (2): since $\mathcal{P}_B = \{P_{\frac{1}{3}}, P_{\frac{2}{3}}\}$ has only two elements, lik' is determined (up to a positive multiplicative constant) by the likelihood ratio

$$\frac{lik'(P_{\frac{2}{3}})}{lik'(P_{\frac{1}{3}})} = \frac{(\frac{1}{3})^w\,(\frac{2}{3})^b}{(\frac{2}{3})^w\,(\frac{1}{3})^b} = 2^{b-w}$$

of $P_{\frac{2}{3}}$ and $P_{\frac{1}{3}}$, where w and b are the numbers of white and black balls observed, respectively. Assume that $b \geq w$: the hierarchical model described by the finite set $\mathcal{M}'$ simply tells us that the probability of observing a black ball in the next draw is either $\frac{1}{3}$ or $\frac{2}{3}$, with a likelihood ratio of 2^{b-w} in favor of the second value. This uncertain knowledge is described by the fuzzy probability p of observing a black ball in the next draw, whose membership function π on $[0,1]$

satisfies

$$\pi(p) = \begin{cases} (\frac{1}{2})^{b-w} & \text{if } p = \frac{1}{3}, \\ 0 & \text{if } p \in [0,1] \backslash \{\frac{1}{3}, \frac{2}{3}\}, \\ 1 & \text{if } p = \frac{2}{3}. \end{cases}$$

Theorem 3 allows us to easily obtain the membership function π_{ch} *of the fuzzy probability of observing a black ball in the next draw for the convexification of the hierarchical model described by* $\mathcal{M} = \mathcal{P}_B$*; that is, for the hierarchical model with constant prior likelihood function on* $\mathcal{P}$*, which was considered in Example 1. Since the function* $\rho \circ \pi$ *on* $[0,1]$ *satisfies*

$$(\rho \circ \pi)(p) = \begin{cases} 2^{b-w} & \text{if } p = \frac{1}{3}, \\ \infty & \text{if } p \in [0,1] \backslash \{\frac{1}{3}, \frac{2}{3}\}, \\ 1 & \text{if } p = \frac{2}{3}, \end{cases}$$

its convex hull $\mathrm{ch}(\rho \circ \pi)$ *is the piecewise linear function on* $[0,1]$*, whose values in* $(\frac{1}{3}, \frac{2}{3})$ *are obtained by linear interpolation of the values of* $\rho \circ \pi$ *in* $\frac{1}{3}$ *and* $\frac{2}{3}$*; that is,*

$$(\mathrm{ch}(\rho \circ \pi))(p) = \begin{cases} 2^{b-w} - 3\,(2^{b-w}-1)\,(p - \frac{1}{3}) & \text{if } p \in [\frac{1}{3}, \frac{2}{3}], \\ \infty & \text{if } p \in [0,1] \backslash [\frac{1}{3}, \frac{2}{3}]. \end{cases}$$

Hence, for the hierarchical model with constant prior likelihood function on $\mathcal{P}$ *(which was considered in Example 1), the fuzzy probability* p *of observing a black ball in the next draw, after having observed* w *white balls and* b *black balls (with* $b \geq w$*), has membership function* π_{ch} *on* $[0,1]$ *satisfying*

$$\pi_{\mathrm{ch}}(p) = \begin{cases} [2^{b-w} - 3\,(2^{b-w}-1)\,(p - \frac{1}{3})]^{-1} & \text{if } p \in [\frac{1}{3}, \frac{2}{3}], \\ 0 & \text{if } p \in [0,1] \backslash [\frac{1}{3}, \frac{2}{3}]. \end{cases}$$

Figure 1 shows the graphs of the piecewise hyperbolic function π_{ch} *when* $b - w = 0$ *(dotted line),* $b - w = 3$ *(dashed line), and* $b - w = 7$ *(solid line).*

4 Hierarchical Networks

Let $X_1, \ldots, X_k$ be some variables taking value in the finite sets $\mathcal{X}_1, \ldots, \mathcal{X}_k$, respectively. An elegant and useful way of constructing a probability measure P on $\Omega = \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$ (that is, a purely probabilistic description of uncertain knowledge about the values of the variables $X_1, \ldots, X_k$) is through a *Bayesian network* (see for example Pearl, 1988, or Jensen, 2001). This consists of a directed acyclic graph with nodes $X_1, \ldots, X_k$, such that to each node X_i is associated a stochastic kernel P_i from $\mathcal{PA}_i$ to $\mathcal{X}_i$, where $\mathcal{PA}_i$ is the image of Ω under PA_i, and PA_i is the function on Ω assigning to each $\omega = (x_1, \ldots, x_k) \in \Omega$ the vector $(x_{j_1}, \ldots, x_{j_l})$ of the values of the *parents* $X_{j_1}, \ldots, X_{j_l}$ of X_i (that is, the nodes from which start the edges pointing to X_i). The stochastic kernel P_i associates to each vector $pa_i \in \mathcal{PA}_i$ a probability measure $P_i(\cdot \,|\, pa_i)$ on $\mathcal{X}_i$; in particular, if X_i is a *root* (that is, it has no parents), then PA_i assigns the "empty vector" $()$ to all $\omega \in \Omega$, and therefore $\mathcal{PA}_i = \{()\}$ is a singleton and the stochastic kernel P_i reduces to a probability measure $P_i(\cdot \,|\, ())$ on $\mathcal{X}_i$. The probability measure $P_{P_1,\ldots,P_k}$ on Ω associated to the Bayesian network is defined by

$$P_{P_1,\ldots,P_k}\{\omega\} = \prod_{i=1}^{k} P_i(\{x_i\} \,|\, PA_i(\omega))$$

for all $\omega = (x_1, \ldots, x_k) \in \Omega$. A key property of Bayesian networks is that the graph encodes conditional independences between the variables $X_1, \ldots, X_k$: these conditional independences can be determined by the graphical criterion of *d-separation*.

Bayesian networks can be generalized to *credal networks* by associating to each node X_i a set $\mathcal{P}_i$ of stochastic kernels P_i from $\mathcal{PA}_i$ to $\mathcal{X}_i$, instead of a single stochastic kernel (see for example Cozman, 2005, or Antonucci and Zaffalon, 2008). The set $\mathcal{P}_i$ associated to a node X_i is said to be *separately specified* if for each $pa_i \in \mathcal{PA}_i$ we can specify a set $\mathcal{P}_{i,pa_i}$ of probability measures on $\mathcal{X}_i$, and obtain $\mathcal{P}_i$ as the set of all stochastic kernels P_i from $\mathcal{PA}_i$ to $\mathcal{X}_i$ such that $P_i(\cdot \,|\, pa_i) \in \mathcal{P}_{i,pa_i}$ for each $pa_i \in \mathcal{PA}_i$ (that is, $\mathcal{P}_i$ can be identified with the Cartesian product of the sets $\mathcal{P}_{i,pa_i}$). The imprecise probability model usually associated to the credal network (called *strong extension* of the credal network) is the convex hull of the set

$$\mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k} = \{P_{P_1,\ldots,P_k} : P_1 \in \mathcal{P}_1, \ldots, P_k \in \mathcal{P}_k\}.$$

In practical applications of credal networks the sets $\mathcal{P}_i$ of stochastic kernels are often finite, and thus the set $\mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k}$ of probability measures is finite too.

Credal networks can be generalized to *hierarchical networks* by associating to each node X_i also a (prior) likelihood function lik_i on the set $\mathcal{P}_i$ of stochastic kernels associated to X_i. When the set $\mathcal{P}_i$ associated to a node X_i is separately specified by the sets $\mathcal{P}_{i,pa_i}$ of probability measures on $\mathcal{X}_i$ (where $pa_i \in \mathcal{PA}_i$), the likelihood function lik_i on $\mathcal{P}_i$ associated to X_i is said to be *separately specified* if for each $pa_i \in \mathcal{PA}_i$ we can specify a likelihood function lik_{i,pa_i} on $\mathcal{P}_{i,pa_i}$, and obtain lik_i as the function on $\mathcal{P}_i$ defined by

$$lik_i(P_i) = \prod_{pa_i \in \mathcal{PA}_i} lik_{i,pa_i}(P_i(\cdot \,|\, pa_i))$$

for all $P_i \in \mathcal{P}_i$ (that is, lik_i can be interpreted as the independent combination of the marginals lik_{i,pa_i}). A node X_i is said to be *Bayesian* if the set $\mathcal{P}_i$ of stochastic kernels associated to X_i is a singleton;

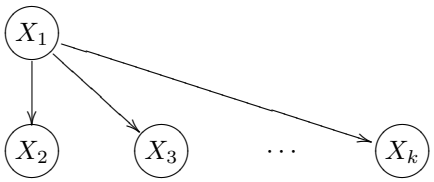

Figure 2: Directed acyclic graph of the hierarchical network of Example 5.

that is, the uncertain knowledge about the value of a Bayesian node conditional on the values of its parents is purely probabilistic. A node X_i is said to be *fuzzy* if $P_i(\cdot \,|\, pa_i)$ is a Dirac measure on $\mathcal{X}_i$ for all $pa_i \in \mathcal{PA}_i$ and all stochastic kernels P_i in the set $\mathcal{P}_i$ associated to X_i; that is, the uncertain knowledge about the value of a fuzzy node conditional on the values of its parents is purely possibilistic. The hierarchical model associated to the hierarchical network has as probabilistic level the set $\mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k}$, and as possibilistic level the likelihood function lik on $\mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k}$ defined (up to a positive multiplicative constant) by

$$lik(P) \propto \sup_{\substack{P_1 \in \mathcal{P}_1,\ldots,P_k \in \mathcal{P}_k : \\ P_{P_1,\ldots,P_k} = P}} \prod_{i=1}^{k} lik_i(P_i)$$

for all $P \in \mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k}$. Hence, the hierarchical model associated to the hierarchical network is described by the set $\mathcal{M} \subset \mathcal{M}_0$ consisting of all measures $\mu_{P_1,\ldots,P_k}$ on Ω with $P_1 \in \mathcal{P}_1, \ldots, P_k \in \mathcal{P}_k$, where

$$\mu_{P_1,\ldots,P_k}\{\omega\} = \prod_{i=1}^{k} \left[lik_i(P_i)\, P_i\left(\{x_i\} \,|\, PA_i(\omega)\right)\right]$$

for all $\omega = (x_1, \ldots, x_k) \in \Omega$. If only convexifications of hierarchical models are considered, then credal networks correspond to the hierarchical networks with constant likelihood functions lik_i, and it often suffices to use finite sets $\mathcal{P}_i$ of stochastic kernels, so that the set $\mathcal{M}$ of measures is finite and the results of Section 3 can be exploited, as in the following examples.

Example 5 *Consider a hierarchical network about the value of the binary variables $X_1, \ldots, X_k \in \{0,1\}$. The directed acyclic graph is plotted in Figure 2. The root X_1 is Bayesian with uniform probability; that is, $\mathcal{P}_1 = \{P_1\}$ with $P_1(\{0\} \,|\, ()) = P_1(\{1\} \,|\, ()) = \frac{1}{2}$. For each $i \geq 2$ the set $\mathcal{P}_i$ associated to the node X_i consists of all stochastic kernels P_i from $\mathcal{PA}_i = \{0,1\}$ to $\mathcal{X}_i = \{0,1\}$ such that $P_i(\{x\} \,|\, (x)) \geq 0.9$ for both $x \in \{0,1\}$. All (prior) likelihood functions lik_i on the sets $\mathcal{P}_i$ are constant. Hence, the hierarchical network corresponds to a credal network with separately specified sets $\mathcal{P}_i$. It can be interpreted as follows: X_1 is the unobservable variable of interest, and for each $i \geq 2$ the variable X_i describes the observation returned by a sensor with a probability of being correct of at least 90%. We want to describe the uncertain knowledge about the value of X_1 that we gain from the observations returned by the $k-1$ sensors, which are assumed to be independent conditional on X_1.*

The case with $k = 3$ (interpreted as a credal network) was studied by Antonucci et al. (2007, Example 1): they showed that if the observations x_2, x_3 returned by the two sensors are equal, then the posterior imprecise probability that X_1 has value $x_2 = x_3$ is $[0.988, 1]$, while if the observations x_2, x_3 are different, then the posterior imprecise probability about the value of X_1 is vacuous. This can be reasonable, but the problem is that the model behaves in the same way in the cases with $k > 3$: it suffices that one of the observations $x_2, \ldots, x_k$ returned by the $k-1$ sensors is different from the others, in order for the posterior imprecise probability about the value of X_1 to be vacuous, independently of the number of sensors. The reason is that for each $i \geq 2$ it is considered possible that the sensor returning the observation X_i is perfect (that is, always correct) while all others are not (that is, they can be wrong), and in this case the posterior probability that X_1 has value x_i is 1, even when all observations returned by the other sensors are different from x_i. However, even if the sensor returning the observation X_i is always correct while all others can be wrong, it is extremely improbable that all others are wrong at the same time. Hence, when the observation returned by a sensor is different from all others, it is extremely implausible that this sensor is perfect. This information about plausibility is described by the likelihood function, and in fact the problem disappears when the network is interpreted as a hierarchical network instead of a credal network.

The convexification of the hierarchical model associated to the hierarchical network can be easily updated thanks to the results of Section 3: for instance, in the case with $k = 5$, when 3 of the observations $x_2, \ldots, x_5$ returned by the 4 sensors are equal x and one is different from x, the membership function of the posterior fuzzy probability p that X_1 has value x is plotted in Figure 3; in particular, the α-cut with $\alpha = 0.1465$ is the interval $[0.932, 1]$. As expected, this fuzzy probability is very high, although no probability value in the interval $[0,1]$ is completely excluded.

To solve the above problem in the framework of credal networks, we should exclude the possibility of perfect sensors by bounding from above the probability that sensors are correct. That is, we should choose a small $\varepsilon > 0$, and for each $i \geq 2$ replace the set $\mathcal{P}_i$ by the set of all stochastic kernels P_i from $\mathcal{PA}_i = \{0,1\}$ to $\mathcal{X}_i = \{0,1\}$ such that $P_i(\{x\} \,|\, (x)) \in [0.9, 1-\varepsilon]$

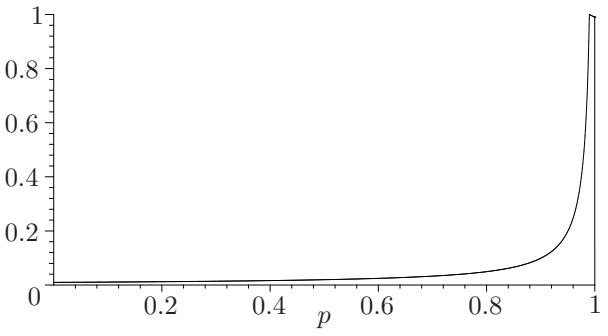

Figure 3: Membership function of the posterior fuzzy probability p that X_1 has value x, when 3 of the observations $x_2, \dots, x_5$ returned by the 4 sensors are equal x and one is different from x (for the hierarchical network of Example 5 with $k = 5$).

for both $x \in \{0,1\}$. However, the resulting posterior imprecise probabilities can depend strongly on the choice of ε: for instance, in the case with $k = 5$, when 3 of the observations $x_2, \dots, x_5$ returned by the 4 sensors are equal x and one is different from x, the posterior imprecise probability that X_1 has value x is $[0.422, 1.000]$ when $\varepsilon = 0.001$, $[0.786, 1.000]$ when $\varepsilon = 0.005$, and $[0.880, 1.000]$ when $\varepsilon = 0.01$. By contrast, in these cases the membership functions of the posterior fuzzy probability that X_1 has value x are (almost) equal to the pointwise product of the indicator function of the corresponding posterior imprecise probability and the membership function for the case with $\varepsilon = 0$ (plotted in Figure 3). Hence, this fuzzy probability does not change much when ε is varied, since only rather implausible probability values are excluded; in particular, the α-cuts with $\alpha = 0.1465$ for the cases with $\varepsilon = 0.001$, $\varepsilon = 0.005$, or $\varepsilon = 0.01$ are practically equal to the α-cut $[0.932, 1]$ for the case with $\varepsilon = 0$.

The possibilistic level of the hierarchical model associated to the hierarchical network of Example 5 contains no information before the updating, because the (prior) likelihood functions lik_i on the sets $\mathcal{P}_i$ of stochastic kernels associated to the nodes X_i are constant. But also hierarchical networks such that the possibilistic levels of the associated hierarchical models contain some prior information (that is, some of the likelihood functions lik_i are not constant) can be useful. In particular, when the stochastic kernels of the network are learned from training data, it is not necessary to reduce the likelihood function to the maximum likelihood estimates (and thus discard the information about the uncertainty of these estimates): the whole likelihood function induced by the training data can be maintained as the possibilistic level of the hierarchical model associated to the hierarchical network. This is a very interesting topic, but goes beyond the scope of the present paper.

Another useful application of hierarchical networks with nonconstant (prior) likelihood functions lik_i is the contamination of a Bayesian (or credal) network: for each node X_i we can give high relative plausibility to the original stochastic kernels P_i associated to X_i, and low relative plausibility to all (or a subset of) other stochastic kernels P_i from $\mathcal{PA}_i$ to $\mathcal{X}_i$. A similar contamination would be possible also in the framework of credal networks (by considering neighborhoods of the original stochastic kernels), but we could not include all possible stochastic kernels (since otherwise the resulting imprecise probability model would be useless), and the final considerations of Example 5 suggest that the resulting posterior imprecise probabilities would be much more sensitive than the posterior fuzzy probabilities to the exact choice of the contamination. In a certain sense, in the framework of hierarchical networks the contamination can be at the possibilistic level, while in the framework of credal networks it must be at the probabilistic level, and this can lead to instability.

Example 6 *Consider the Bayesian network obtained from the hierarchical network of Example 5 by selecting, for each $i \geq 2$, the stochastic kernel P_i from $\mathcal{PA}_i = \{0,1\}$ to $\mathcal{X}_i = \{0,1\}$ such that $P_i(\{x\} \,|\, (x)) = 0.95$ for both $x \in \{0,1\}$. We can contaminate this Bayesian network by choosing a small $\gamma > 0$ and associating to each node X_i the (separately specified) set $\mathcal{P}_i$ of all stochastic kernels P_i from $\mathcal{PA}_i = \{0,1\}$ to $\mathcal{X}_i = \{0,1\}$ and the (prior) likelihood function lik_i on $\mathcal{P}_i$ separately specified by the likelihood functions $lik_{i,(x)}$ on the set of all probability measures on $\{0,1\}$ such that $lik_{i,(x)}(P_i(\cdot \,|\, (x))) = 1$ if $P_i(\cdot \,|\, (x))$ is the corresponding conditional probability in the Bayesian network, and $lik_{i,(x)}(P_i(\cdot \,|\, (x))) = \gamma$ otherwise, for both $x \in \{0,1\}$. The resulting hierarchical network describes the situation in which there is some uncertainty about the conditional probabilities of the Bayesian network; it is useful because it tells us how robust against modifications of the conditional probabilities are the conclusions of the Bayesian network.*

The convexification of the hierarchical model associated to the hierarchical network can be easily updated thanks to the results of Section 3: for instance, Figure 4 shows the graphs of the membership functions of the fuzzy probability p of $X_1 = 1$ in the case with $k = 3$ and $\gamma = 0.05$: prior to any observation (dashed line), after observing $X_2 = X_3 = 0$ (solid line with maximum near 0), after observing $X_2 = 1$ and $X_3 = 0$ or vice versa (dotted line), and after observing $X_2 = X_3 = 1$ (solid line with maximum near 1); in particular, the corresponding α-

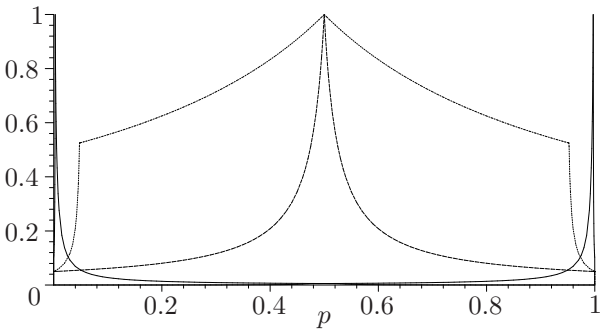

Figure 4: Membership functions of the fuzzy probability p of $X_1 = 1$ (for the hierarchical network of Example 6 with $k = 3$ and $\gamma = 0.05$): prior to any observation (dashed line), after observing $X_2 = X_3 = 0$ (solid line with maximum near 0), after observing $X_2 = 1$ and $X_3 = 0$ or vice versa (dotted line), and after observing $X_2 = X_3 = 1$ (solid line with maximum near 1).

cuts with $\alpha = 0.1465$ *are the intervals* $[0.347, 0.653]$, $[0.001, 0.019]$, $[0.035, 0.965]$, *and* $[0.981, 0.999]$, *respectively. Hence, the conclusions of the Bayesian network are pretty robust when the two sensors agree (the uncertainty about the probability of* $X_1 = 1$ *decreases), while they are not robust at all when the two sensors disagree (the uncertainty about the probability of* $X_1 = 1$ *increases).*

When $X, Y, Z \subseteq \{X_1, \ldots, X_k\}$ are three disjoint sets of variables, Y is said to be *irrelevant* to X given Z (with respect to a hierarchical model on Ω) if the fuzzy probability distribution for the variables in X conditional on any realization of the variables in Z does not change when also something about the variables in Y is observed. This definition of conditional irrelevance is stronger than the corresponding one for imprecise probability models, since the invariance of both levels of the hierarchical model is required. However, when the hierarchical model is constructed through a hierarchical network, the following fundamental result holds (for a sketch of the proof see Cattaneo, 2008*b*, Subsection 3.1).

Theorem 7 *Let* $X, Y, Z \subseteq \{X_1, \ldots, X_k\}$ *be three disjoint sets of variables. If* X *and* Y *are d-separated by* Z *in the directed acyclic graph of a hierarchical network, then* Y *is irrelevant to* X *given* Z*, with respect to the hierarchical model associated to the hierarchical network.*

Theorem 7 is of crucial importance for the meaning and usefulness of hierarchical networks: conditional irrelevances between the variables $X_1, \ldots, X_k$ are encoded in the graph and can be determined by the graphical criterion of d-separation. Together with the results of Section 3, Theorem 7 allows the calculation of exact inferences in simple hierarchical networks.

Any probability measure on Ω can be constructed through a Bayesian network with nodes $X_1, \ldots, X_k$. By contrast, not all hierarchical models on Ω can be constructed through hierarchical networks with nodes $X_1, \ldots, X_k$. However, any hierarchical model describing the uncertain knowledge about the values of the variables $X_1, \ldots, X_k$ can be constructed through a hierarchical network with nodes $X_1, \ldots, X_{k+1}$: it suffices to add a root X_{k+1}, which in general is a parent of all other nodes, and which indexes the probability measures in the probabilistic level $\mathcal{P}$ of the hierarchical model. Hence, the variable X_{k+1} takes values in the set $\mathcal{P}$, which can be infinite, but this is unimportant, since the root X_{k+1} is fuzzy (with likelihood function lik_{k+1} corresponding to the possibilistic level lik of the hierarchical model); by contrast, the nodes $X_1, \ldots, X_k$ are Bayesian.

More generally, we can easily transform any hierarchical network with nodes $X_1, \ldots, X_k$ into a larger hierarchical network which describes the same uncertain knowledge about the values of the variables $X_1, \ldots, X_k$, but such that each node is either Bayesian or fuzzy (and we can also require that only roots can be fuzzy). In fact, when a node X_i is neither Bayesian nor fuzzy (or it is fuzzy but not a root), we can simply add a root which is a parent of X_i only, and which indexes the set $\mathcal{P}_i$ of stochastic kernels associated to X_i. This additional root is fuzzy (with likelihood function corresponding to the likelihood function lik_i on $\mathcal{P}_i$ associated to X_i), while the node X_i becomes Bayesian. In particular, we can always obtain a hierarchical network such that each node X_i is either Bayesian or fuzzy and both the set $\mathcal{P}_i$ of stochastic kernels and the likelihood function lik_i on $\mathcal{P}_i$ associated to X_i are separately specified (since this is always the case for roots and Bayesian nodes).

From the above considerations it follows easily the result (showed by Antonucci and Zaffalon, 2008) that we can transform any credal network with nodes $X_1, \ldots, X_k$ into a larger credal network which describes the same uncertain knowledge about the values of the variables $X_1, \ldots, X_k$, but such that each node X_i is either Bayesian or the set $\mathcal{P}_i$ of stochastic kernels associated to X_i is separately specified by vacuous imprecise probability models. More specifically, we can always obtain a credal network such that each node X_i is either Bayesian or it is a root and the set $\mathcal{P}_i$ of probability measures on $\mathcal{X}_i$ is the vacuous imprecise probability model. The difference between the hierarchical model and the imprecise probability model is in the way in which such roots X_i are updated when data

are observed (since the Bayesian nodes are updated in the same way in both models): in the framework of credal networks we remain in the state of complete ignorance about the value of X_i (apart from when we get deterministic information about it), while in the framework of hierarchical networks the possibilistic level allows us to get out of the state of complete ignorance about the value of X_i.

This shows in particular that hierarchical networks cannot be described by possibly larger credal networks (for instance by interpreting possibility measures as upper probability measures), because these could not display the same behavior when data are observed, not even with an alternative updating rule.

5 Conclusion

In the present paper, the use of fuzzy probabilities to describe the uncertain knowledge about the values of the nodes of belief networks has been studied. The increased expressive power, the ability of using all the information provided by the data, and the increased robustness of the conclusions are important advantages over credal networks. The possibility of using the whole likelihood function induced by training data (and not only the maximum likelihood estimates) seems very promising and deserves further study. The description of convex hierarchical models by finite sets of measures and the validity of the criterion of d-separation allow the calculation of the desired inferences in simple hierarchical networks. However, approximation algorithms are necessary for the calculation of inferences in more complex networks: some algorithm for credal networks can probably be adapted to hierarchical networks, thanks to the strong similarity between the descriptions of the hierarchical model and of the imprecise probability model as convex sets of measures.

References

Antonucci, A., Brühlmann, R., Piatti, A., and Zaffalon, M. (2007). Credal networks for military identification problems. In *ISIPTA '07*. SIPTA, 1–10.

Antonucci, A., and Zaffalon, M. (2008). Decision-theoretic specification of credal networks: A unified language for uncertain modeling with sets of Bayesian networks. *Int. J. Approx. Reasoning* 49, 345 – 361.

Cattaneo, M. (2005). Likelihood-based statistical decisions. In *ISIPTA '05*. SIPTA, 107–116.

Cattaneo, M. (2007). *Statistical Decisions Based Directly on the Likelihood Function*. PhD thesis, ETH Zurich.

Cattaneo, M. (2008*a*). Fuzzy probabilities based on the likelihood function. In *Soft Methods for Handling Variability and Imprecision*. Springer, 43–50.

Cattaneo, M. (2008*b*). Probabilistic-possibilistic belief networks. Technical Report 32. Department of Statistics, LMU Munich.

Cozman, F. G. (2005). Graphical models for imprecise probabilities. *Int. J. Approx. Reasoning* 39, 167–184.

Dahl, F. A. (2005). Representing human uncertainty by subjective likelihood estimates. *Int. J. Approx. Reasoning* 39, 85–95.

De Cooman, G. (2005). A behavioural model for vague probability assessments. *Fuzzy Sets Syst.* 154, 305–358.

Dubois, D. (2006). Possibility theory and statistical reasoning. *Comput. Stat. Data Anal.* 51, 47–69.

Held, H., Augustin, T., and Kriegler, E. (2008). Bayesian learning for a class of priors with prescribed marginals. *Int. J. Approx. Reasoning* 49, 212 – 233.

Hisdal, E. (1988). Are grades of membership probabilities? *Fuzzy Sets Syst.* 25, 325–348.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*. Springer.

Moral, S. (1992). Calculating uncertainty intervals from conditional convex sets of probabilities. In *UAI '92*. Morgan Kaufmann, 199–206.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Pearl, J. (1988). *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann.

Walley, P. (1997). Statistical inferences based on a second-order possibility distribution. *Int. J. Gen. Syst.* 26, 337–383.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* 9, 60–62.

Wilson, N. (2001). Modified upper and lower probabilities based on imprecise likelihoods. In *ISIPTA '01*. Shaker, 370–378.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst.* 1, 3–28.