Probabilistic-possibilistic belief networks

Marco E. G. V. Cattaneo

Department of Statistics, LMU Munich

cattaneo@stat.uni-muenchen.de

Abstract: The interpretation of membership functions of fuzzy sets as statistical likelihood functions leads to a probabilistic-possibilistic hierarchical description of uncertain knowledge. The fundamental advantage of the resulting fuzzy probabilities with respect to imprecise probabilities is the ability of using all the information provided by the data. This paper studies the possibility of using fuzzy probabilities to describe the uncertain knowledge about the values of the nodes of belief networks.

1 Introduction

In the present paper, membership functions of fuzzy sets are interpreted as statistical likelihood functions. This allows a combination of probabilistic and possibilistic uncertainty on the basis of the well-established theories of probability and likelihood. The resulting probabilistic-possibilistic hierarchical description of uncertain knowledge generalizes the description by means of imprecise probabilities, but only from the static point of view. In fact, the usual updating rule for imprecise probabilities does not use all the information provided by the data, and this waste of information can lead to statistical inconsistency and unsatisfactory results. By contrast, the probabilistic-possibilistic hierarchical model exploits the outstanding asymptotic properties of the likelihood function, which makes it an ideal basis for inference and decision making: this aspect is analyzed in Cattaneo (2005, 2007).

In the present paper, the probabilistic-possibilistic hierarchical model is combined with 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. Since simple fuzzy probability measures can be described as convex hulls of finite sets of non-normalized probability measures, the resulting probabilistic-possibilistic hierarchical networks have the same complexity as credal networks. Moreover, the graphical criterion of d-separation can be exploited, since it implies the conditional irrelevance of the involved variables.

2 Probabilistic-Possibilistic Hierarchical Model

Let P be a set of probability measures on a finite set $\Omega = \mathsf{X}_1 \times \cdots \times \mathsf{X}_n$ (for simplicity, in the present paper only the finite case is considered, but infinite sets Ω would pose no problem). Each $P \in \mathsf{P}$ is interpreted as a probabilistic model for the values of the random variables $X_i : (x_1, \ldots, x_n) \mapsto x_i$ (for all $i \in \{1, \ldots, n\}$). The interpretation of probability is not important: for instance the elements of P can be statistical models, or describe the forecasts of a group of experts.

The *likelihood function lik* on $\mathcal{P}$ induced by the observations $X_i \in A_i \subseteq \mathcal{X}_i$ (for each $i \in I \subseteq \{1,\ldots,n\}$) is defined by

$$lik(P) = P\{(x_1,\ldots,x_n) \in \Omega : x_i \in A_i \text{ for all } i \in I\};$$

lik describes the relative ability of the probabilistic models in $\mathcal{P}$ to forecast the observed data. Spurious modifications of the situation considered or of its mathematical representation can lead to likelihood functions proportional to lik (for example, if the realization of an additional random variable X_{n+1} describing the result of tossing a fair coin is also observed, then the induced likelihood function is halved). Therefore, proportional likelihood functions are considered equivalent; in fact, Fisher (1921, 1922) defined the likelihood of a statistical model as a quantity *proportional* to the probability of the observed data. Hence, only ratios $lik(P)/lik(P^{'})$ of the values of lik for different $P, P^{'} \in \mathcal{P}$ have meaning: Kullback and Leibler (1951) interpreted $\log[lik(P)/lik(P^{'})]$ as the information in the data for discrimination in favor of P against $P^{'}$, and Good (1950) considered it as the weight of evidence in favor of P against $P^{'}$ provided by the data. So the likelihood function can be interpreted as a measure of the relative plausibility of the probabilistic models in the light of the observed data alone.

The likelihood function lik measures the relative plausibility of the elements of $\mathcal{P}$, but a measure of the relative plausibility of the subsets of $\mathcal{P}$ is often needed. A simple and effective way to obtain it consists in defining the plausibility of a set of probabilistic models as the plausibility of its best element: the result is the set function

$$\mathcal{H} \mapsto \sup_{P \in \mathcal{H}} lik(P)$$

on the power set $2^{\mathcal{P}}$ of $\mathcal{P}$ (in this paper, $\sup \varnothing = 0$). Proportional set functions of this form are equivalent, since they correspond to equivalent likelihood functions: to underline this relative meaning, the expression “relative plausibility measure” is used in Cattaneo (2007) to denote an equivalence class of proportional set functions of this form. Their normalized version LR associates to each $\mathcal{H} \subseteq \mathcal{P}$ the corresponding likelihood ratio statistic

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

The likelihood ratio test discards the hypothesis that the data were generated by some $P \in \mathcal{H}$ if $LR(\mathcal{H})$ is sufficiently small. In regular problems with large samples, the critical value for $LR(\mathcal{H})$ can be obtained from the result of Wilks (1938) that $-2\log LR(H)$ is approximately χ^2 distributed under each $P \in \mathcal{H}$.

A *possibility distribution* on a set $\mathcal{G}$ is a function $\pi : \mathcal{G} \to [0,1]$. The possibility measure on $\mathcal{G}$ with possibility distribution π is the set function

$$G \mapsto \sup_{\gamma \in G} \pi(\gamma)$$

on $2^{\mathcal{G}}$. A possibility distribution π on $\mathcal{G}$ can also be considered as the *membership function* of a fuzzy subset of $\mathcal{G}$ (see Zadeh, 1978); when π is *crisp* (that is, π can take only the values 0 and 1), the subset is not fuzzy and π is its indicator function on $\mathcal{G}$. The likelihood ratio statistic LR is a possibility measure on $\mathcal{P}$ with possibility distribution proportional to the likelihood function lik on $\mathcal{P}$. In fact, the membership function of a fuzzy set has often been interpreted as a likelihood function (see for example Hisdal, 1988; Dubois and Prade, 1993; Dubois, 2006), even though proportional membership functions were not always considered equivalent (see for instance Dubois et al., 1997). In the present paper, membership functions and possibility distributions are interpreted as *proportional* to likelihood functions. Hence, it suffices to consider normalized fuzzy sets and normalized possibility measures (that is, $\sup_{\gamma \in \mathcal{G}} \pi(\gamma) = 1$ is assumed), but grades of membership and degrees of possibility have only a relative meaning.

A set $\mathcal{P}$ of probability measures on Ω and a likelihood function lik on $\mathcal{P}$ can be interpreted as the two levels of a *probabilistic-possibilistic hierarchical model* for the values of the variables X_i. The two levels describe different kinds of uncertain knowledge: in the first level the uncertainty is stochastic, while in the second one it is about which of the probabilistic models in $\mathcal{P}$ is the best representation of the reality. It is important to underline that no probabilistic model in $\mathcal{P}$ is assumed to be in some sense "true": the elements of $\mathcal{P}$ are simply interpreted as more or less plausible representations of the reality. By contrast, the use of a probability measure on $\mathcal{P}$, suggested by the Bayesian approach, carries the implicit assumption that exactly one of the probabilistic models in $\mathcal{P}$ is "true" (see Cattaneo, 2007, Section 3.1). The likelihood function lik on $\mathcal{P}$ can also express subjective beliefs about the relative plausibility of the probabilistic models in $\mathcal{P}$: in this case, lik is interpreted as if it were induced by hypothetical data (see also Dahl, 2005). The choice of a subjective likelihood function on $\mathcal{P}$ seems to be better supported by intuition than the choice of a subjective probability measure on $\mathcal{P}$: in particular, a constant likelihood function describes complete ignorance (in the sense of absence of information for discrimination between the probabilistic models).

2.1 Fuzzy Probabilities and Imprecise Probabilities

Let X be a real-valued function of $X_1, \dots, X_n$, and let $g : P \mapsto E_P(X)$ be the function on $\mathcal{P}$ assigning to each probabilistic model the corresponding expectation of X. A likelihood function lik on $\mathcal{P}$ induces the (normalized) *profile likelihood function*

$$lik_g : x \mapsto LR(g^{-1}\{x\}) \propto \sup_{P \in \mathcal{P} : g(P) = x} lik(P)$$

on the set $\mathbb{R}$ of real numbers (in this paper, the exponent $^{-1}$ denotes the set function associating to a set its inverse image). The profile likelihood function lik_g measures the relative plausibility of the values of g, on the basis of the above definition of the plausibility of a set of probabilistic models as the plausibility of its best element. In fact, lik_g is the possibility distribution corresponding to the possibility measure $LR \circ g^{-1}$ induced by g on $\mathbb{R}$. Hence, the uncertain knowledge about the expectation of X is described by the fuzzy number (that is, a fuzzy subset of $\mathbb{R}$) with membership function lik_g: this fuzzy number can be interpreted as the *fuzzy expectation* of X. In particular, when X is the indicator function I_A of a set $A \subseteq \Omega$, the fuzzy expectation of I_A describes the uncertain knowledge about the probability of A, and can thus be interpreted as the *fuzzy probability* of A.

Sometimes a fuzzy number can be a satisfactory conclusion about the expectation of X, but it is often necessary to evaluate the fuzzy number by one or more real numbers. The discussion on how to do this goes beyond the scope of the present paper, but it is important to note the correspondence between some natural "defuzzification methods" and the usual likelihood-based inference methods. In fact, the *α-cut* $\{x \in \mathsf{R} : lik_g(x) \geq \alpha\}$ with $\alpha \in (0,1]$ corresponds to a likelihood-based confidence region for the expectation of X (the coverage probability of this confidence region can often be approximated thanks to the result of Wilks, 1938), and when a global maximum of lik_g exists and is unique, it corresponds to the *maximum likelihood estimate* of the expectation of X.

Since the probabilistic models outside the support $\mathsf{P}' = \{P \in \mathsf{P} : lik(P) > 0\}$ of lik have no influence on lik_g, the likelihood function can always be extended to the set of all probability measures on Ω, by defining it constant equal to 0 outside P. Hence, the hierarchical model can also be interpreted as a fuzzy probability measure on Ω, in the sense that it is a fuzzy subset of the set of all probability measures on Ω, with membership function proportional to the (extended) likelihood function. When this is crisp, it is the indicator function of the support P' of lik: there is no information for discrimination between the elements of P', and in fact the uncertain knowledge about the expectation of X is described by the set $\mathsf{G} = \{E_P(X) : P \in \mathsf{P}'\}$ (in the sense that $lik_g = I_{\mathsf{G}}$). In particular, when P' is convex and closed, the set G is the interval

$$\left[\inf_{P \in \mathsf{P}'} E_P(X), \sup_{P \in \mathsf{P}'} E_P(X)\right];$$

that is, in this case the hierarchical description of uncertain knowledge about the values of the variables X_i reduces to the description by means of *imprecise probabilities* (see Walley, 1991).

Both the purely probabilistic and the purely possibilistic descriptions of uncertain knowledge about the values of the variables X_i appear as special cases of the probabilistic-possibilistic hierarchical description. In fact, when the support P' of lik is a singleton $\{P\}$, the description of uncertain knowledge is purely probabilistic: lik_g is the indicator function of $\{E_P(X)\}$. By contrast, when P' is a set of Dirac measures (that is, $\mathsf{P}' \subseteq \{\delta_\omega : \omega \in \Omega\}$, with $\delta_\omega\{\omega\} = 1$), the description of uncertain knowledge is purely possibilistic: it corresponds to the possibility measure $LR' = LR \circ t^{-1}$ on Ω, where t is the function $\delta_\omega \mapsto \omega$ on P'. In fact, lik_g is the possibility distribution corresponding to the possibility measure $LR' \circ X^{-1}$ induced by X on R; in particular, the support of lik_g is finite, since it is a subset of the image of X.

2.2 Updating

The definition of likelihood function implies that when $X_i \in A_i \subseteq \mathcal{X}_i$ is observed (for an $i \in \{1,\ldots,n\}$), the probabilistic level $\mathcal{P}$ of the hierarchical model is updated to the set

$$\mathcal{P}' = \{P(\cdot|X_i \in A_i) : P \in \mathcal{P}, P\{X_i \in A_i\} > 0\}$$

of conditional probability measures $P(\cdot|X_i \in A_i)$ on Ω, while the possibilistic level lik is updated to the likelihood function lik' on $\mathcal{P}'$ defined by

$$lik'(P') = \sup_{P \in \mathcal{P}: P(\cdot|X_i \in A_i) = P'} lik(P) P\{X_i \in A_i\}.$$

In fact, when interpreted as a function of P, the argument of the supremum is the new likelihood function on $\mathcal{P}$, and lik' is the corresponding (profile) likelihood function on $\mathcal{P}'$. The definition of lik' on $\mathcal{P}'$ instead of $\mathcal{P}$ can be slightly confusing (since likelihood functions are usually defined on the set $\mathcal{P}$ of unconditional probability measures, as done at the beginning of the present section), but it is necessary if the hierarchical model has to describe the available uncertain knowledge about the values of the variables X_i.

In particular, when the description of uncertain knowledge is purely probabilistic, the support of lik is a singleton $\{P\}$, and the updating corresponds to conditioning P, in accordance with the Bayesian approach. When the description of uncertain knowledge is purely possibilistic, the possibility measure LR' on Ω is updated by multiplying the corresponding possibility distribution with the indicator function of $\{X_i \in A_i\}$ and then renormalizing it. Hence, in particular, the purely probabilistic and the purely possibilistic descriptions of uncertain knowledge are maintained when updating the hierarchical model. By contrast, in general the description by means of imprecise probabilities (corresponding to the case in which the support of lik is convex and closed, and lik is constant on it) is not maintained when updating the hierarchical model, because in general lik' is not constant on its support. In fact, the usual updating rule in the theory of imprecise probabilities is the *regular extension* (see Walley, 1991, Appendix J), which corresponds to the above updating rule without the term $P\{X_i \in A_i\}$ in the argument of the supremum (so that lik' too is constant on its support).

In general, the hierarchical model with probabilistic level $\mathcal{P}$ and possibilistic level lik can be described by the set $\mathcal{M} = \{lik(P)P : P \in \mathcal{P}\}$ of non-normalized probability measures on Ω. When $X_i \in A_i \subseteq \mathcal{X}_i$ is observed (for an $i \in \{1,\ldots,n\}$), the set $\mathcal{M}$ is updated to the set $\mathcal{M}' = \{\mu(\cdot \cap \{X_i \in A_i\}) : \mu \in \mathcal{M}\}$ of non-normalized probability measures on Ω: the updating of each $\mu \in \mathcal{M}$ corresponds to the Bayesian updating without renormalization. In fact, if μ' is defined as the probability measure on Ω obtained by normalizing the non-normalized probability measure μ with $\mu(\Omega) > 0$ (that is, $\mu = \mu(\Omega)\mu'$), then the probabilistic level of the updated hierarchical model is $\mathcal{P}' = \{\mu' : \mu \in \mathcal{M}', \mu(\Omega) > 0\}$, while the possibilistic level is the likelihood function lik' on $\mathcal{P}'$ defined by

$$lik'(P') = \sup_{\mu \in \mathsf{M}' : \mu' = P'} \mu(\Omega).$$

Since the updating $\mu \mapsto \mu(\cdot \cap \{X_i \in A_i\})$ of the non-normalized probability measures on Ω is the restriction of a linear function, if M is the convex hull of a set M_0, then M' is the convex hull of the set $\mathsf{M}_0' = \{\mu(\cdot \cap \{X_i \in A_i\}) : \mu \in \mathsf{M}_0\}$. That is, the set M_0 is updated in the same way as M, and can be considered as a simpler description of the hierarchical model. Of course, this simpler description is particularly useful when M_0 is finite.

In particular, if M is closed and convex, and its elements are normalized (that is, $\mathsf{M} = \mathsf{P}$), then the hierarchical description of uncertain knowledge corresponds to the description by means of imprecise probabilities. The updating by means of regular extension consists in updating M to the set $\mathsf{M}'' = \mathsf{P}'$ of the renormalized elements of M', but the renormalization of the elements of M' deletes the information about their relative ability to forecast the observation $X_i \in A_i$. For instance, if the probabilistic models in P describe the opinions of a group of Bayesian experts, then the updating by regular extension corresponds to update the opinion of each expert without reconsidering her/his credibility, independently of how bad her/his forecasts were when compared to the forecasts of the other experts. This is not very reasonable, and in fact the updating by regular extension can lead to inconsistency, in the statistical sense of not tending to the correct conclusion, even when the amount of information provided by the data tends to infinity. The following adaptation of an example by Wilson (2001) shows that this does not only happen when the set P is too wide.

Example 1 *Let* $\mathsf{X}_1 = \ldots = \mathsf{X}_{101} = \{0,1\}$, *and let*

$$\mathsf{P}_0 = \{P_p : p \in \Delta\}$$

be a set of probability measures on $\Omega = \{0,1\}^{101}$ *such that* $\Delta = [0.1, 0.6]$ *and for all* $p \in \Delta$

$$P_p\{X_1 = 0\} = \tfrac{1}{2},$$

and conditional on the realization of X_1 *the random variables* $X_2, \ldots, X_{101}$ *are independent with*

$$P_p\{X_i = 1 | X_1 = 0\} = \tfrac{1}{2}, \quad P_p\{X_i = 1 | X_1 = 1\} = p$$

for all $i \in \{2, \ldots, 101\}$.

Consider the hierarchical model described by the set $\mathsf{M}_0 = \mathsf{P}_0$ *; that is, the probabilistic level* P *is the convex hull of* P_0, *and the possibilistic level is the likelihood function lik on* P *with constant value* 1. *Since the set* $\mathsf{M} = \mathsf{P}$ *is closed and convex, the hierarchical description of uncertain knowledge corresponds to the description by means of imprecise probabilities.*

When the realizations $X_2 = x_2, \ldots, X_{101} = x_{101}$ *are observed, the updated hierarchical model is described by the set*

$$\mathsf{M}_0^{'} = \left\{ \left(\tfrac{1}{2}\right)^{101} \delta_{\omega_0} + \tfrac{1}{2} p^x (1-p)^{100-x} \delta_{\omega_1} : p \in \Delta \right\}$$

of linear combinations of the Dirac measures δ_{ω_0} *and* δ_{ω_1}, *where* $\omega_{x_1} = (x_1, \ldots, x_{101})$ *for* $x_1 \in \{0,1\}$, *and*

$$x = \sum_{i=2}^{101} x_i .$$

Since $\mathsf{M}_0^{'}$ *is convex,* $\mathsf{M}^{'} = \mathsf{M}_0^{'}$, *and* $\mathsf{P}^{'}$ *consists of the convex combinations of* δ_{ω_0} *and* δ_{ω_1} *proportional to the linear combinations contained in* $\mathsf{M}_0^{'}$.

Figure 1 shows the graph on $[0, 5 \cdot 10^{-7}]$ *of the membership function of the fuzzy probability of* $X_1 = 0$ *according to the updated hierarchical model, when* $x = 20$. *Since* $X_1 = 1$ *is compatible with the observed data, while* $X_1 = 0$ *is not, the fuzzy probability of* $X_1 = 0$ *is extremely concentrated near* 0. *In fact, any reasonable evaluation of the fuzzy probability of* $X_1 = 0$ *by a real number (such as the maximum likelihood estimate* $0.04 \cdot 10^{-7}$, *or the midpoint* $2.13 \cdot 10^{-7}$ *of the* α*-cut with* $\alpha = 0.01$*) would be approximately* 0.

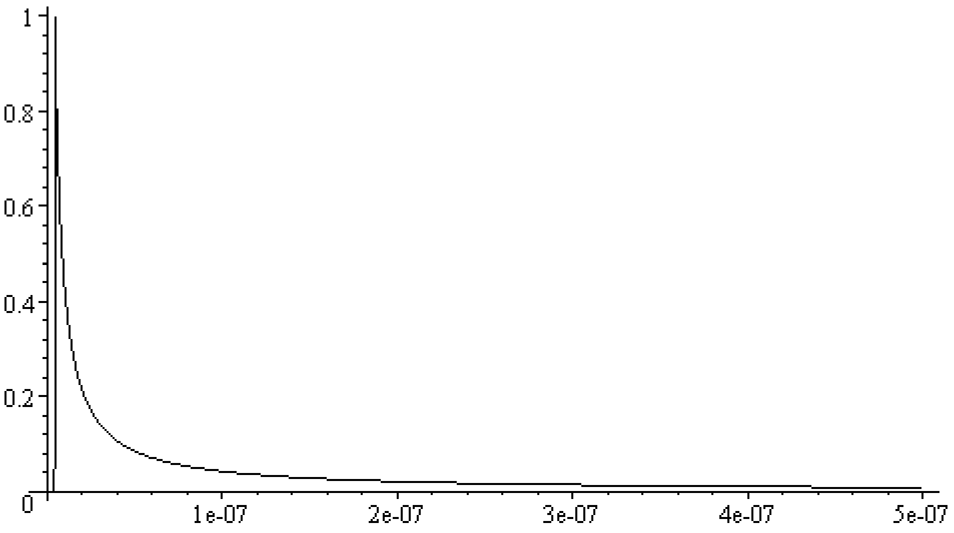

Figure 1 Membership function of the fuzzy probability of $X_1 = 0$ when $x = 20$.

However, the updating of $\mathsf{M} = \mathsf{P}$ *by means of regular extension is* $\mathsf{M}^{''} = \mathsf{P}^{'}$; *that is, each element of* $\mathsf{M}^{'} = \mathsf{M}_0^{'}$ *is renormalized, without considering how improbable the observed data were for the corresponding probabilistic models in* P. *For the probability of* $X_1 = 0$ *this simply means forcing the crispness of the membership function, by making it constant equal to* 1 *on its support: when* $x = 20$, *the resulting uncertain knowledge about the probability of* $X_1 = 0$ *is described by the interval* $[4.26 \cdot 10^{-9}, 1 - 6.77 \cdot 10^{-7}] \approx [0,1]$. *That is, despite the overwhelming information in favor of* $X_1 = 1$ *against* $X_1 = 0$, *almost complete ignorance about the probability of* $X_1 = 0$ *is obtained when updating the imprecise probabilities by means of regular extension.*

3 Belief Networks

An elegant and useful way of constructing a probabilistic model for the values of the variables X_i is through a *Bayesian network* (see Pearl, 1988; Jensen, 2001). This consists of a directed acyclic graph with nodes $X_1, \dots, X_n$, such that to each node X_i is associated a stochastic kernel assigning a probability measure $P_i\left(\cdot | pa_i(\omega) \right)$ on $\mathcal{X}_i$ to each possible vector $pa_i(\omega)$ of values for the *parents* of X_i (that is, the nodes from which start the edges pointing to X_i). For each $i \in \{1, \dots, n\}$, the function pa_i on Ω assigns to each $\omega = (x_1, \dots, x_n)$ the vector $(x_{j_1}, \dots, x_{j_m})$ of the values of the parents $X_{j_1}, \dots, X_{j_m}$ of X_i; when X_i is a *root* (that is, it has no parents), pa_i assigns the empty set to all $\omega \in \Omega$, and the stochastic kernel associated to X_i reduces to a probability measure $P_i\left(\cdot | \varnothing \right)$ on $\mathcal{X}_i$. The probability measure P on Ω associated to the Bayesian network is defined by

$$P\{\omega\} = \prod_{i=1}^{n} P_i\left(\{x_i\} | pa_i(\omega) \right)$$

for all $\omega = (x_1, \dots, x_n) \in \Omega$.

The probability measures on Ω *compatible* with a directed acyclic graph with nodes $X_1, \dots, X_n$ are those that can be constructed as above by a suitable choice of the stochastic kernels. A key property of Bayesian networks is that the graph encodes conditional independencies between the variables $X_1, \dots, X_n$: these conditional independencies can be determined by the graphical criterion of *d-separation* (see Pearl, 1988).

In the theory of imprecise probabilities, Bayesian networks have been generalized to *credal networks* by associating to each node X_i a closed convex set $\mathcal{P}_i$ of stochastic kernels, instead of a single stochastic kernel P_i (see Cozman, 2000, 2005). There are several ways of associating to a credal network a closed convex set $\mathcal{P}$ of probability measures on Ω; the simplest one is to define $\mathcal{P}$ as the convex hull of the set $\mathcal{P}_0$ of all probability measures on Ω that can be constructed as above by all possible choices of the stochastic kernels $P_i \in \mathcal{P}_i$. Hence, all elements of $\mathcal{P}_0$ are compatible with the directed acyclic graph considered, but in general not all elements of $\mathcal{P}$ are compatible with it.

When some data are observed, the imprecise probabilities described by $\mathcal{P}$ are usually updated by means of regular extension; but the results of Section 2.2 show that regular extension can lead to unsatisfactory conclusions. A simple solution consists in considering $\mathcal{P}$ as the description $\mathcal{M} = \mathcal{P}$ of a hierarchical model (the one with $\mathcal{P}$ as probabilistic level and the likelihood function *lik* on $\mathcal{P}$ with constant value 1 as possibilistic level), and updating it to the set $\mathcal{M}'$ of non-normalized probability measures on Ω, as described in Section 2.2. In general, the resulting conclusions are then fuzzy expectations and fuzzy probabilities, but if necessary these fuzzy numbers can be evaluated by intervals, for instance by considering their α-cuts (see also Moral, 1992; Cano and Moral, 1996).

Credal networks can be generalized by allowing the use of fuzzy probabilities also before the updating. In the resulting *probabilistic-possibilistic hierarchical networks*, to each node X_i is associated a fuzzy stochastic kernel; that is, a fuzzy subset of the set of all possible stochastic kernels P_i, with membership function π_i. Let P_0 be the set of all probability measures on Ω compatible with the directed acyclic graph considered: a fuzzy subset of P_0 can be constructed on the basis of the hierarchical network by defining the degree of membership $lik_0(P)$ of $P \in \mathsf{P}_0$ as the supremum of $\prod_{i=1}^{n} \pi_i(P_i)$ over all choices of the stochastic kernels P_i such that P is associated to the corresponding Bayesian network. Since the membership functions π_i are interpreted as proportional to likelihood functions, the use of their product is implied by the implicit assumption that these likelihood functions have been induced by independent (hypothetical) observations. Under an analogous assumption, the membership function π_i of the fuzzy stochastic kernel associated to a node X_i can be defined as the product of the membership functions of the fuzzy probability measures on X_i corresponding to each possible vector of values for the parents of X_i.

The set $\mathsf{M}_0 = \{lik_0(P)P : P \in \mathsf{P}_0\}$ of non-normalized probability measures on Ω describes the hierarchical model with P_0 as probabilistic level and the function lik_0 on P_0 as possibilistic level. The hierarchical model associated to the hierarchical network can be defined as the one described by M_0 or as the one described by the convex hull M of M_0. Let P and lik be the two levels of the probabilistic-possibilistic hierarchical model described by M. As for credal networks, in general not all elements of the support of lik are compatible with the directed acyclic graph considered; in fact, credal networks correspond to the special case in which all membership functions π_i are crisp with closed convex support.

When some data are observed, the hierarchical model described by M_0 or M can be updated to the one described by $\mathsf{M}_0^{'}$ or $\mathsf{M}^{'}$ (that is, the convex hull of $\mathsf{M}_0^{'}$), as considered in Section 2.2. This is particularly simple when M_0 is finite, or when M is the convex hull of another finite set of non-normalized probability measures on Ω. This is the case in particular when all fuzzy probability measures appearing in the hierarchical network can be described by finite sets of non-normalized probability measures, as in the following simple example.

Example 2 *Let $\mathsf{X}_1 = \mathsf{X}_2 = \mathsf{X}_3 = \{0,1\}$, and consider the Bayesian network consisting of the directed acyclic graph $X_1 \leftarrow X_2 \rightarrow X_3$ and of the stochastic kernels P_1, P_2, P_3 defined by*

$$P_2(\{0\}|\varnothing) = P_1(\{0\}|(0)) = P_1(\{1\}|(1)) = P_3(\{1\}|(0)) = P_3(\{0\}|(1)) = 0.9.$$

Let P be the probability measure on $\Omega = \{0,1\}^3$ associated to the Bayesian network:

$$P\{X_3 = 1\} = 0.82, \quad P\{X_3 = 1|X_1 = 0\} \approx 0.890, \quad \text{and} \quad P\{X_3 = 1|X_1 = 1\} = 0.5.$$

Since $P\{X_1 = 0\} = 0.82$, the observation $X_1 = 0$ is not surprising, and does not have much influence on the probability of $X_3 = 1$. By contrast, the observation $X_1 = 1$ is more *surprising, and has a larger influence on the probability of* $X_3 = 1$*; in fact, the probability* 0.5 *of* $X_3 = 1$ *conditional on* $X_1 = 1$ *can be considered as arising from conflicting evidence.*

Consider the hierarchical network obtained by modifying in the following way each one of the 5 *probability measures on* $\{0,1\}$ *that define the above Bayesian network: assume that any probability measure on* $\{0,1\}$ *is at least* 0.01 *times as plausible as the one used in the Bayesian network, and take the convex hull of the corresponding set of non-normalized probability measures on* $\{0,1\}$*. When building each one of these* 5 *convex hulls, it suffices to consider the two extreme non-normalized probability measures* $0.01\delta_0$ *and* $0.01\delta_1$ *on* $\{0,1\}$*, besides the probability measure used in the Bayesian network; for example, the fuzzy probability measure associated to the node* X_2 *is described by the convex hull of the set* $\{0.01\delta_0, 0.9\delta_0 + 0.1\delta_1, 0.01\delta_1\}$ *of non-normalized probability measures on* $\{0,1\}$.

When the set M_0 *of non-normalized probability measures on* Ω *is defined as above, its convex hull* M *is the convex hull of the set of the* 99 *non-normalized probability measures on* Ω *resulting from all possible combinations of the elements of the* 5 *sets of* 3 *non-normalized probability measures on* $\{0,1\}$ *that generate the* 5 *fuzzy probability measures defining the hierarchical network (because of the two extreme non-normalized probability measures* $0.01\delta_0$ *and* $0.01\delta_1$ *associated to the node* X_2*, there are only* $3^4 + 2 \cdot 3^2 = 99$ *possible combinations, and not* $3^5 = 243$*). As considered in Section 2.2, when data are observed, the hierarchical model described by* M *is updated to the hierarchical model described by the convex hull* M' *of the set obtained by updating each one of the* 99 *non-normalized probability measures on* Ω.

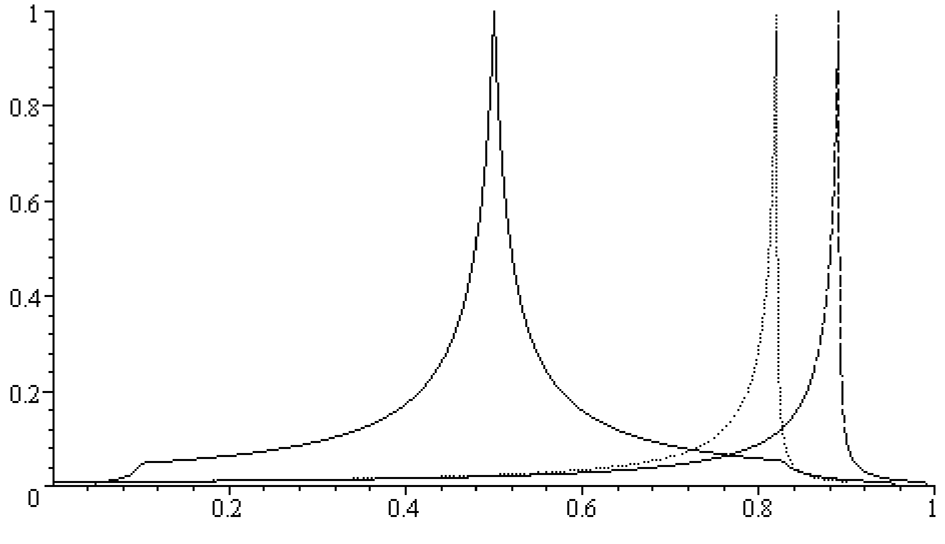

Figure 2 Membership functions of the fuzzy probability of $X_3 = 1$: unconditional (dotted line), conditional on $X_1 = 0$ (dashed line), and conditional on $X_1 = 1$ (solid line).

Note that every probability measure $P^{'}$ on Ω compatible with the graph considered has positive degree of membership in the fuzzy probability measure on Ω described by M. When data are observed, each $P^{'}$ is updated and its degree of membership is modified in accordance with the relative ability of $P^{'}$ to forecast the observed data. For example, Figure 2 shows the graphs of the membership functions of the fuzzy probability of $X_3 = 1$ according to the hierarchical model described by M (dotted line) and to the updated hierarchical model described by $\mathsf{M}^{'}$, when $X_1 = 0$ is observed (dashed line), or when $X_1 = 1$ is observed (solid line). The maximum likelihood estimates are the probabilities resulting from the Bayesian network, while the α-cuts with $\alpha = 0.1$ are $[0.746, 0.828]$, $[0.809, 0.899]$, and $[0.311, 0.671]$, respectively. That is, the possibilistic uncertainty about the value of the probability of $X_3 = 1$ remains more or less constant when the unsurprising realization $X_1 = 0$ is observed, while it clearly increases when the more surprising realization $X_1 = 1$ is observed: the possibilistic uncertainty is larger for probability values arising from conflicting evidence.

3.1 Irrelevance and D-separation

Let $X, Y, Z \subseteq \{X_1, \ldots, X_n\}$ be three disjoint sets of variables. Y is said to be *irrelevant* to X given Z (with respect to a probabilistic-possibilistic hierarchical model for the values of the variables X_i) 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. Note that in general the conditional independence of X and Y given Z under each probabilistic model in the probabilistic level of the hierarchical model does not suffice for the irrelevance of Y to X given Z, because the possibilistic level can be influenced by the observations about the variables in Y. However, when the hierarchical model is constructed through a hierarchical network, the following result holds.

Theorem 3 *If X and Y are d-separated by Z in the directed acyclic graph of a probabilistic-possibilistic hierarchical network, then Y is irrelevant to X given Z, with respect to the hierarchical model associated to the hierarchical network.*

The theorem can be proved as follows. Let $Y^{'}$ be the set of all variables $X_i \notin X \cup Z$ such that X and $\{X_i\}$ are d-separated by Z in the graph considered, and let $X^{'}$ be the complement of $Y^{'} \cup Z$ in $\{X_1, \ldots, X_n\}$. Then let Z_1 be the set of all variables in Z that have no parents in $Y^{'}$, and let Z_2 be the complement of Z_1 in Z; the sets $X^{'}, Y^{'}, Z_1, Z_2$ build a partition of $\{X_1, \ldots, X_n\}$. The definition of d-separation implies that for each probability measure on Ω associated to a Bayesian network on the graph considered, the value of the probability of the observed realization of the variables in Z factorizes in two parts: one depending only on the stochastic kernels P_j for the variables $X_j \in X^{'} \cup Z_1$, and the other depending only on the stochastic kernels P_k for the variables $X_k \in Y^{'} \cup Z_2$. Moreover, the value of the probability of an observation about the variables in X conditional on the observed realization of the variables in Z depends only on the stochastic kernels P_j for the variables $X_j \in X' \cup Z_1$, while the value of the probability of an observation about the variables in Y conditional on the observed realization of the variables in Z depends only on the stochastic kernels P_k for the variables $X_k \in Y' \cup Z_2$.

Consider the hierarchical model described by the set M_0 of non-normalized probability measures on Ω defined as above on the basis of the hierarchical network. The degree of membership of a probability distribution for the variables in X conditional on the observed realization of the variables in Z is proportional to the supremum (over all choices of the stochastic kernels P_i leading to this probability distribution) of the product of $\prod_{i=1}^{n} \pi_i(P_i)$ with the corresponding value of the probability of the observed realization of the variables in Z. Hence, the argument of the supremum factorizes in two parts as above, and the part depending only on the stochastic kernels P_k for the variables $X_k \in Y' \cup Z_2$ disappears in the proportionality constant. Since X and Y are conditionally independent given Z under each probabilistic model compatible with the graph considered, the same supremum is obtained when considering also the observation about the variables in Y: the additional factor in the argument depends only on the stochastic kernels P_k for the variables $X_k \in Y' \cup Z_2$, and therefore it disappears in the proportionality constant too. This result for the hierarchical model described by M_0 implies the same result for the hierarchical model described by the convex hull M of M_0.

3.2 Probabilistic-Possibilistic Belief Networks

Any probabilistic model for the values of the variables $X_1, \ldots, X_n$ can be constructed through a Bayesian network with nodes $X_1, \ldots, X_n$. By contrast, not all closed convex sets of probability measures on Ω can be constructed through credal networks with nodes $X_1, \ldots, X_n$, and not all hierarchical models on Ω can be constructed through hierarchical networks with nodes $X_1, \ldots, X_n$. For instance, the hierarchical model of Example 1 cannot be constructed through a hierarchical network with nodes $X_1, \ldots, X_{101}$.

However, any hierarchical model for the values of the variables $X_1, \ldots, X_n$ can be constructed through a hierarchical network with nodes $X_1, \ldots, X_{n+1}$: it suffices to add a root X_{n+1}, which in general is a parent of all other nodes, and which indexes the probabilistic models in the probabilistic level of the hierarchical model. Hence, in general X_{n+1} is infinite, but this is unimportant, because the uncertain knowledge about the value of X_{n+1} is purely possibilistic (with possibility distribution corresponding to the possibilistic level of the hierarchical model). By contrast, the uncertain knowledge about the value of any other node X_i, given the values of its parents, is purely probabilistic. For instance, the hierarchical model of Example 1 can be constructed through a hierarchical network with nodes $X_1, \ldots, X_{102}$, where $\mathsf{X}_{102} = \Delta$, the nodes X_1 and X_{102} are roots, and they are the only two parents of all other nodes. The uncertain knowledge about the value p of the variable X_{102} is purely possibilistic (with possibility distribution constant equal to 1 on Δ), while the uncertain knowledge about the value of the variable X_1 is purely probabilistic, and the same is true for the uncertain knowledge about the values of the variables $X_2, \ldots, X_{101}$, conditional on the values of X_1 and X_{102}.

In general, every hierarchical network with nodes $X_1, \ldots, X_n$ can be easily transformed into a larger hierarchical network which describes the same uncertain knowledge about the values of the variables $X_1, \ldots, X_n$, but such that the uncertain knowledge about the value of each node, given the values of its parents, is either purely probabilistic or purely possibilistic. In fact, when the uncertain knowledge about the value of a node X_i, given the values of its parents, is not purely probabilistic or purely possibilistic, it suffices to add a root which is a parent of X_i only, and which indexes the possible stochastic kernels P_i. The uncertain knowledge about the value of this additional root is purely possibilistic (with possibility distribution corresponding to π_i), while the uncertain knowledge about the value of the node X_i, given the values of its parents, is now purely probabilistic.

4 Conclusion

The use of fuzzy probabilities to describe the uncertain knowledge about the values of the nodes of belief networks seems very promising. The description of fuzzy probability measures as convex hulls of finite sets of non-normalized probability measures and the exploitation of the criterion of d-separation allow the use of fuzzy probabilities in those belief networks that can be afforded by imprecise probabilities. The resulting probabilistic-possibilistic hierarchical models can also be interpreted as a simple generalization of Bayesian networks, in which the uncertainty about the values of some nodes can be possibilistic instead of probabilistic.

References

Cano, A., Moral, S. (1996). A genetic algorithm to approximate convex sets of probabilities. In *IPMU '96*. Vol. 2. Universidad de Granada, 859-864.

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. Available online at e-collection.ethz.ch.

Cozman, F. G. (2000). Credal networks. *Artif. Intell.* 120, 199-233.

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.

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

Dubois, D., Moral, S., Prade, H. (1997). A semantics for possibility theory based on likelihoods. *J. Math. Anal. Appl.* 205, 359-380.

Dubois, D., Prade, H. (1993). Fuzzy sets and probability: Misunderstandings, bridges and gaps. In *Second IEEE International Conference on Fuzzy Systems*. Vol. 2. IEEE Service Center, 1059-1068.

Fisher, R. A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron* 1, 3-32.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond., Ser. A* 222, 309-368.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin.

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

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

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79-86.

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

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

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

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.