ELSEVIER

Artificial Intelligence 83 (1996) 1-58

Artificial Intelligence

Measures of uncertainty in expert systems

Peter Walley *

Department of Mathematics, The University of Western Australia, Nedlands, WA 6907, Australia

Received April 1991; revised January 1995

Abstract

This paper compares four measures that have been advocated as models for uncertainty in expert systems. The measures are additive probabilities (used in the Bayesian theory), coherent lower (or upper) previsions, belief functions (used in the Dempster-Shafer theory) and possibility measures (fuzzy logic). Special emphasis is given to the theory of coherent lower previsions, in which upper and lower probabilities, expectations and conditional probabilities are constructed from initial assessments through a technique of natural extension. Mathematically, all the measures can be regarded as types of coherent lower or upper previsions, and this perspective gives some insight into the properties of belief functions and possibility measures. The measures are evaluated according to six criteria: clarity of interpretation; ability to model partial information and imprecise assessments, especially judgements expressed in natural language; rules for combining and updating uncertainty, and their justification; consistency of models and inferences; feasibility of assessment; and feasibility of computations. Each of the four measures seems to be useful in special kinds of problems, but only lower and upper previsions appear to be sufficiently general to model the most common types of uncertainty.

Keywords: Inference; Decision; Prevision; Bayesian theory; Dempster-Shafer theory; Belief functions; Possibility theory; Lower probability; Upper probability; Imprecise probabilities; Conditional probability; Independence

1. Introduction

My aim in this paper is to compare and evaluate mathematical measures of uncertainty that can be used in expert systems. The measures that I will consider are Bayesian probabilities, coherent lower previsions, belief functions and possibility measures. Special emphasis is given to the theory of coherent lower previsions as this approach has

* E-mail: peter@maths.uwa.edu.au.

SSDI 0004-3702(95)00009-7

received less attention than the others in the AI literature, although it is (in my view) the one that is most widely applicable. On a mathematical level, the other measures can be regarded as special types of lower or upper previsions. They can also be given the behavioural interpretation of lower or upper previsions. This point of view leads to some illuminating comparisons between the theories.

For example, the four theories differ greatly in the calculus they use for defining, updating and combining measures of uncertainty, especially the rules they use to define conditional probabilities and expectations and how they model judgements of independence. The rules used in the theory of lower previsions, which are based on a general procedure called *natural extension*, can be applied to belief functions and possibility measures, and thus they can be compared with the rules used in Dempster–Shafer theory and possibility theory.

The paper is not intended to be a survey of the four theories from a neutral position and it is certainly not intended to be a wide-ranging survey of mathematical models for uncertainty. There is a substantial and quickly growing literature on numerical measures of uncertainty in expert systems. References [46,47,88] are good introductions to the relationships between probability measures, belief functions and possibility measures. Readers can learn more about the subject by consulting the surveys in [8,31,79,86,89] and the proceedings of the annual workshops on Uncertainty in Artificial Intelligence. I shall not discuss non-numerical methods of reasoning with uncertainty such as the theory of endorsements [13], default reasoning and nonmonotonic logics [59,63,67,68], comparative probability [27,104] and modal logics [59,104]. Surveys of the non-numerical approaches can be found in [47,79,86].

I take it for granted that most practical reasoning involves uncertainty, partial ignorance and incomplete or conflicting information, and that it is often useful to formally model the uncertainty. This raises the questions (a) what is the best way to model uncertainty? (b) how should we assess, combine and update measures of uncertainty? and (c) how should we use the measures to make inferences and decisions? These questions are relevant to all kinds of reasoning with uncertainty, not only in expert systems. However expert systems are an especially good testing-ground for theories of uncertainty because they aim to formalise and automate as much as possible of the reasoning process. As an expert system is concerned only with a narrow domain of application, it is possible to formulate special assessment strategies, models and patterns of reasoning which are appropriate to that application. Many of the relevant uncertainties can be assessed by domain experts and these assessments can be encoded in the system. A user may be required to supply some further assessments but the expert system should be able to guide him through this process.

The measures of uncertainty that are combined in an expert system may come from various sources. Some may be "objective" measures, based on relative frequencies or on well established statistical models. (In medical diagnosis, for example, there may be data concerning the frequency of a disease in a population, or the statistical association between a symptom and a disease.) Other assessments of uncertainty may be supplied by the domain expert (e.g., concerning the irrelevance of specific observations in diagnosis, or the association between a symptom and a disease when there is little statistical data). Indeed several experts may have contributed in building the system. Further assessments may be made by the user of the system (e.g. about the uncertainties in the symptoms exhibited by a patient). All these measures of uncertainty need to be combined by the expert system to make inferences and decisions (e.g. to make a diagnosis for this patient).

What, then, is the meaning of the combined measures of uncertainty? *Whose* uncertainties do they measure? It is the user of the system who will act on the conclusions of the system, provided he is satisfied that its uncertainty assessments are acceptable to him. One can regard the expert system as a consultant that supplies various models and assessments, elicits others from the user, combines all the judgements, and finally informs the user "if you accept all these judgements then you should draw these conclusions". So the expert system constructs a single model for uncertainty which the user then considers adopting as a model for his own uncertainty and as a basis for action. The expert system should be able to justify its assessments of uncertainty and its reasoning procedures, when requested by the user, to make its model and conclusions more convincing. It should also be able to modify some of its assumptions and assessments if requested by the user. In the end, the uncertainty measures on which conclusions are based must be acceptable to the user.

2. Criteria for evaluating measures of uncertainty

In the rest of this paper I will compare measures of uncertainty according to the following broad criteria.

(a) *Interpretation.* The measure should have a clear interpretation that is sufficiently definite to be used to guide assessment, to understand the conclusions of the system and use them as a basis for action, and to support the rules for combining and updating measures.

(b) *Imprecision.* The measure should be able to model partial or complete ignorance, limited or conflicting information, and imprecise assessments of uncertainty.

(c) *Calculus.* There should be rules for combining measures of uncertainty, updating them after receiving new information, and using them to calculate other uncertainties, to draw conclusions and to make decisions. Some justification must be given for the rules. Special attention will be given to the rules for computing conditional probabilities and expectations from unconditional probabilities.

(d) *Consistency.* There should be methods for checking the consistency of all uncertainty assessments and default assumptions used by the system, and the rules of the calculus should ensure that the conclusions are consistent with these assessments. In the Bayesian theory and the theory of lower previsions, the intuitive notion of consistency is formalised in mathematical principles of *coherence.*

(e) *Assessment.* It should be practicable for a user of the system to make (and feel comfortable with) all the uncertainty assessments that are needed as input. The system should give some guidance on how to make the assessments. It should be able to handle judgements of various types, including expressions of uncertainty in natural language such as "if A then *probably* B", and to combine qualitative judgements with quantitative assessments of uncertainty.

(f) *Computation.* It should be computationally feasible for the system to derive inferences and conclusions from the assessments.

Readers may wish to add other desiderata to this list. It does seem to me that it is essential for a theory of uncertainty to attempt to satisfy all six criteria. We might distinguish the first four criteria, which are theoretical, from the last two, which are practical. The first four criteria are "theoretical" in the sense that one would expect an adequate theory of uncertainty to show that they can be satisfied, irrespective of the specific application. The last two criteria are "practical" in the sense that they will be satisfied in some applications but not in others, depending on the type of model involved, the number of assessments needed, practical constraints of time and computing power, and the abilities of the user.

The two practical criteria are obviously necessary if an expert system is to be implemented in practice. The first criterion, interpretation, seems essential in order to give meaning to, and to justify, the conclusions of the system. The second, imprecision, is necessary because partial ignorance and conflicting information are common in practice. A calculus is needed in order to derive conclusions from the uncertainty assessments, so the third criterion is needed. The fourth criterion, consistency, is needed to avoid erroneous and irrational conclusions. See [99] for further justification of these criteria, especially (a), (b), (d) and (e).

There is a striking divergence between workers in probability and philosophy, on the one hand, and those in expert systems, artificial intelligence, computer science and engineering, concerning their attitude to these criteria. Naturally enough, the second group has emphasised practical criteria, especially computation (f), and has mostly recognised the need for imprecise measures of uncertainty (b) and natural-language judgements (e). They have given much less attention to issues of interpretation (a), consistency (d) and justification of the calculus (c). The literature in probability and philosophy pays more attention to these theoretical issues, but less to the practical issues of assessment and computation. There is surprisingly little attention, in all the literature, to the problem of assessment.

Much of the work in expert systems is typified by MYCIN, perhaps the best-known and most influential expert system, which was designed to help physicians diagnose bacterial infections of the blood [6,47,80,81]. In MYCIN uncertainty is measured in terms of *certainty factors*. MYCIN does well on the practical criteria (e) and (f), largely because of the modularity of its inference system (dependencies amongst variables are ignored) and the use of simple rules to combine certainty factors, but it does badly on the theoretical criteria (a), (c) and (d), largely because of the lack of a clear interpretation of certainty factors and a lack of justification for the rules of combination. (An earlier version of this paper [100] contained a detailed comparison of certainty factors with other measures of uncertainty but this has been omitted on the advice of the referees.)

Although all six criteria seem essential, I regard the first criterion—the need for a clear interpretation—as the most fundamental, because an interpretation is needed to support the rules of the calculus, to formulate principles of coherence (consistency), to guide assessment, and to understand conclusions. Criterion (a) is therefore a prerequisite for criteria (c), (d) and (e). (This point is well made in [33].) The importance of interpretation tends to be underrated by workers in expert systems, perhaps because it is unclear to them how an interpretation of uncertainty measures could be used to derive and justify rules for combining and updating the measures. Readers may find it illuminating to compare the theories of de Finetti [28,29] or Walley [99], in which a behavioural interpretation of linear previsions or lower previsions is used to support principles of coherence and hence to derive the entire calculus, with the literature on certainty factors and fuzzy logic, in which the rules of the calculus seem quite arbitrary.

In the following sections I will outline de Finetti's Bayesian theory and the theory of coherent lower previsions, which emphasise interpretation and consistency, followed by the Dempster–Shafer theory of belief functions, which emphasises a simple rule of combination, and finally consider the possibility measures used in fuzzy logic, which emphasises judgements in natural language. Although the four theories differ greatly in their interpretation of uncertainty measures, their methods of assessment and their calculus, all the measures can be regarded, from a mathematical point of view, as special types of coherent lower or upper previsions, and the four theories will be compared from this perspective.

3. Bayesian probabilities

The most highly developed and best understood theory of uncertainty is the Bayesian theory. The book [56] is a good introduction to the theory and [28,29] are worthy of careful study. In the field of expert systems, [47] is a good introduction and [50,60,63,90] are important references. Bayesian probability models have been used in quite different ways in the expert systems PROSPECTOR [20,21], GLADYS [91] and MUNIN [43,47,50]. More recent applications include HUGIN [2], PATHFINDER [37], BAIES [14], and [38]. The discussion here will be brief, as the theory and its strengths and weaknesses are quite well known. The main aim is to summarise the Bayesian approach in a form that will illuminate the comparison with the other approaches.

In the theory uncertainty is measured by unconditional probabilities $P(A)$ or by conditional probabilities $P(A \mid B)$, numbers between zero and one which are interpreted as *fair betting rates*. That is, the person whose uncertainties are being modelled is taken to be willing to bet on or against event A at rates arbitrarily close to $P(A)$, or at rates arbitrarily close to $P(A \mid B)$ on condition that the bet is called off unless event B occurs. The key assumption here is that a person should always have the same marginal rate for betting on or against an event. This assumption is called the *Bayesian dogma of precision*. Bayesians have given several arguments to support the assumption, but none of the arguments is at all compelling; see [99, Chapter 5] for a thorough discussion.

Apart from countable additivity, all the familiar properties of probabilities can be derived from the behavioural interpretation and the dogma of precision, together with the coherence principle that there should be no finite combination of acceptable bets that is certain to produce a net loss. These assumptions imply that the probability function P is a normalised, nonnegative, finitely-additive set function. Further coherence principles imply that updated probabilities after observing an event B should agree with conditional probabilities $P(A \mid B)$, and that these should be related to unconditional probabilities through *Bayes' rule* $P(A \cap B) = P(A \mid B)P(B)$. If the unconditional probabilities $P(A \cap B)$ and $P(B)$ have been assessed and $P(B)$ is nonzero then the conditional probability $P(A \mid B)$ is uniquely determined through Bayes' rule. Thus Bayes' rule can be used as a rule for computing conditional probabilities from unconditional probabilities. It can also be used as a rule for computing $P(A \cap B)$ from $P(A \mid B)$ and $P(B)$, and indeed it is frequently used for that purpose.

If an unconditional probability measure P is specified, the *prevision* (or *expectation*) of a random variable X, denoted by $P(X)$, can be computed from $P(X) = \int X \, dP$. When the possibility space Ω is finite, $P(X) = \sum_{\omega \in \Omega} P(\omega)X(\omega)$. Again these should be regarded as coherence relationships. They can be used to compute the value of $P(X)$ from assessments of probabilities, but they could also be used to provide information about probabilities from assessments of previsions, as in [28]. In general, assessments of previsions determine upper and lower probabilities rather than precise values.

If coherent probabilities are specified for all events then, for every random variable X, there is a unique value of $P(X)$ that is coherent with the specified probabilities. This means that, for Bayesians, no information is lost by modelling uncertainty in terms of probabilities; previsions are uniquely determined by probabilities. This result does not carry over to imprecise probabilities: upper and lower previsions are not determined, in general, by upper and lower probabilities.

The Bayesian theory of probability is closely related to a theory of decision making. Suppose that the utility resulting from each feasible action a can be measured by a precise number $U(a, \omega)$ which depends on the unknown state ω. Define the random variable X_a by $X_a(\omega) = U(a, \omega)$. A Bayesian would compute the prevision $P(X_a)$, the expected utility of action a, for each feasible action, and attempt to choose an action to maximise expected utility.

The Bayesian theory is applied in practice by selecting some events or variables whose (conditional) probabilities or previsions can be precisely assessed, adding any judgements of independence or exchangeability, and then applying the rules of the theory to calculate other (conditional) probabilities or previsions. For example, if $\{A_1, A_2, \ldots, A_k\}$ is an exhaustive set of mutually exclusive hypotheses and B is an observable event, one might make precise assessments of the prior probabilities $P(A_i)$ and likelihoods $P(B \mid A_i)$ for $1 \leqslant i \leqslant k$, and use these to calculate the predictive probability $P(B) = \sum_{i=1}^{k} P(B \mid A_i)P(A_i)$. After event B is observed, provided $P(B)$ is not zero, one might update uncertainty about the hypotheses by calculating their posterior probabilities using Bayes' rule, $P(A_i \mid B) = P(B \mid A_i)P(A_i)/P(B)$. (There is nothing in the Bayesian theory that forces one to calculate $P(A_i \mid B)$ in this way—it might be easier to assess $P(A_i \mid B)$ directly than to assess the quantities $P(A_i)$ and $P(B \mid A_i)$.) Uncertain conclusions will typically be expressed in terms of posterior probabilities that are conditional on all the available information. When $k = 2$, Bayes' rule can be written more conveniently in the form $\rho(B) = \rho\lambda(B)$, where $\rho = P(A_1)/P(A_2)$ is the prior odds on A_1, $\rho(B) = P(A_1 \mid B)/P(A_2 \mid B)$ is the posterior odds on A_1, and $\lambda(B) = P(B \mid A_1)/P(B \mid A_2)$ is the likelihood ratio generated by B. That is, posterior odds = prior odds × likelihood ratio.

In principle, the Bayesian approach can be applied in any problem involving uncertainty. In practice, it can be difficult to make the many precise assessments of probabilities that are needed to determine a complete probability model, and to check that the assessments are coherent and that they determine a unique probability model. Unless sufficiently many assessments are made, the probabilities of interest will not be precisely determined—we obtain upper and lower probabilities [60,62]. But if a larger number of assessments are made, so the probabilities of interest are overdetermined, typically the assessments will be incoherent. There are also computational difficulties in verifying whether a given set of probability assessments and independence judgements is coherent, which is equivalent to checking whether a system of linear and quadratic equations has a solution.

To alleviate the difficulties of assessment and computation, the early expert system PROSPECTOR incorporated the simplifying assumption that separate pieces of evidence are probabilistically independent conditional on the hypotheses of interest, and used simple rules (similar to the max/min rules of fuzzy logic) to combine pieces of evidence. However these rules are inconsistent with the Bayesian calculus and they can produce incoherent probabilities.

More recently, attention has been directed to special types of models, notably the "belief networks", "causal networks" or "directed acyclic graphs" studied in [50,63,90]. These models involve judgements of *conditional independence,* based on an expert's understanding of the causal relations between variables, which can be represented graphically by directed trees and which are reasonable in many practical problems. For these models, the effort of assessment and computation is greatly reduced: assessments of conditional probabilities are needed only for the links in the tree, and the effect of new information on probabilities can be propagated locally. Belief networks can also be elaborated into "influence diagrams" by adding information about possible actions and utilities, and thereby used to make decisions.

Assuming that precise probabilities can be assessed for all events, the rules of the probability calculus are uncontroversial. It is the assumption of precision that is unacceptable. When there is little information concerning a possible event A it is inappropriate to assess any precise probability $P(A)$. (Suppose, for example, that I produce an urn containing coloured balls. Without any further information about the balls, how would you assess a precise probability that the first ball drawn from the urn will be red? See [101] for discussion of this example.) Similarly the Bayesian approach cannot deal with imprecise, qualitative or natural-language judgements such as "if A then probably B" [104].

Conclusion

The Bayesian theory does very well on criteria (a), (c) and (d), but poorly on (b) and (e). Bayesian probabilities have a simple behavioural interpretation. The rules of the probability calculus can be justified through this interpretation, and the rules guarantee consistency (coherence). Computations are feasible for some important types of models, notably for singly-connected belief networks. The theory is highly developed, especially for dealing with judgements of conditional independence [63], and useful in many practical problems.

The fundamental difficulties with the Bayesian theory concern the dogma of precision. Because they demand precise probability models, Bayesians cannot adequately model ignorance, partial information, assessments of uncertainty in natural language, or conflict between expert opinions. There is a compelling argument for allowing imprecise assessments of uncertainty (see [99] for detailed discussion), and this has been accepted in much of the expert systems literature. In the rest of this paper I consider measures of uncertainty which do admit imprecision.

4. Coherent lower previsions

This section summarises the theory of coherent lower previsions. The theory is developed in detail in [99], using mathematical concepts suggested in [28,87,110]. For related work in expert systems see [25,30,35,60,62,66,94,113], but note that these papers (except the last one) differ in some important respects from the approach outlined here; they emphasise upper and lower probabilities rather than previsions, and they adopt a different interpretation which leads to a different concept of independence. Fuller discussion of many of the ideas in this section can be found in [99].

Assessment

First consider how an expert system might, ideally, elicit assessments of uncertainty from the domain expert and the user. The system should give some guidance on how to make the assessments. It might suggest what probabilities could be assessed to constrain the probabilities of interest, or what kinds of independence judgements and probability models may be reasonable. But the user should not be forced to accept any of these judgements or to make assessments of any particular type. For example, the system may suggest *default assumptions* such as conditional independence, but if these are unacceptable to the user then the system should be able to operate without them, typically obtaining weaker conclusions.

Generally the system should be able to work with whatever combination of judgements and expressions of uncertainty the user is able to make, including precise or imprecise assessments of unconditional or conditional probability, judgements in natural language such as "A is more probable than B", "if A then probably B" or "A is very likely", judgements of (conditional) independence, and various other kinds of judgements. The system would check whether these judgements are mutually consistent, combine them to construct an overall probability model and to draw conclusions about the questions of interest, and report the model and conclusions to the user. If the conclusions were indeterminate, the user might try to make further assessments in order to make the probability model more precise, but he may not always be able to do so.

The assessment process involves a sequence of judgements. After each judgement, the expert system could compute and display summaries of the current probability model, and analyse the model to suggest what kinds of judgements should be considered next in order to reduce the indeterminacy in inferences or decisions. In the light of the current model, the user may choose to reconsider and modify earlier judgements, make further assessments, update the model to take account of new information, refine or reformulate the possibility space, or terminate the process. Each of these steps modifies the current model in a simple way: see [99, Section 4.3] for the mathematical details. For example, the user may decide to retract some of the assumptions or judgements (a kind of nonmonotonic reasoning) because the current model is incoherent or has unacceptable implications, he may recognise that his previous possibility space is not exhaustive and decide to consider other possibilities, or he may use imprecise probabilistic information about one possibility space to provide information about a second possibility space that is related to the first space through a multivalued mapping.

The key step in this process is the construction of an overall probability model from an arbitrary combination of uncertainty judgements. This can be carried out by the expert system, without further input from a user, through a mathematical procedure called *natural extension.*

Interpretation

Before we can give a formal definition of natural extension we must characterise the kind of probability model that we aim to construct. In fact there are several types of models for imprecise probabilities that are more or less equivalent, subject to appropriate consistency requirements [99, Section 3.8]. (Lower and upper probabilities are *not* an adequate model in general, for reasons to be explained later.) The simplest model is a *lower prevision* $\underline{P}$, which is a real-valued function defined on the set $\mathcal{L}$ of all gambles. A *gamble* is a bounded mapping from the possibility space of interest, Ω (a set whose elements represent possible states of affairs or "possible worlds"), to the real numbers, and is interpreted as an uncertain reward in units of utility.

The interpretation of the quantity $\underline{P}(X)$ is that you are disposed to pay any price less than $\underline{P}(X)$ for the gamble (uncertain reward) X. Loosely, we may call $\underline{P}(X)$ a *supremum buying price* for X: it is the supremum of prices which the model asserts that you are willing to pay for X. A conjugate upper prevision $\overline{P}$ is defined by $\overline{P}(X) = -\underline{P}(-X)$. The interpretation is that you are disposed to sell the gamble X for any price greater than $\overline{P}(X)$. (The theory could be presented equivalently in terms of $\overline{P}$, but here we concentrate on $\underline{P}$.) The model does not say anything about whether you will buy or sell X if the price lies between $\underline{P}(X)$ and $\overline{P}(X)$; either course of action may be reasonable. Similarly, a conditional lower prevision $\underline{P}(X \mid B)$ is interpreted as a supremum of buying prices for X that you would be willing to pay if you learned only that event B has occurred.

Buying and selling gambles are somewhat artificial activities and they are introduced here merely to give a simple interpretation for $\underline{P}(X)$ and $\overline{P}(X)$. These quantities also have implications in more practical decision problems (outlined near the end of Section 4), and they could be interpreted in terms of their implications for other types of decisions.

This interpretation of upper and lower previsions is epistemic and behavioural, but not necessarily subjective. In some problems, $\underline{P}(X)$ and $\overline{P}(X)$ can be given a *logical* interpretation, as marginal buying and selling prices that are uniquely determined by the available evidence; an expert system that is concerned with a sufficiently narrow domain and that does not require any subjective input from users might encode a system of *inductive logic* [101]. The interpretation is *epistemic* in the sense that upper and lower previsions (like all the other measures considered in this paper) reflect a particular state of information and will usually change when more information is obtained or when information is reassessed.

The quantities $\underline{P}(X)$ and $\overline{P}(X)$ need not be maximally precise. They may be merely lower and upper bounds for quantities that could be specified more precisely, in the same way that we could give lower and upper bounds for a person's lifetime, e.g. 400 and 300 B.C. in the case of Aristotle. The values $\underline{P}(X)$ and $\overline{P}(X)$ could be generated by merely qualitative judgements, as in the football example below. This should make it clear that, contrary to common objections, the specification of $\underline{P}(X)$ and $\overline{P}(X)$ does not demand "twice as much precision" as a Bayesian model. Note also that, while it may be possible to sharpen the assessments of $\underline{P}(X)$ and $\overline{P}(X)$, there is no reason to suppose that a precise assessment $\underline{P}(X) = \overline{P}(X)$ could be made, any more than Aristotle's lifetime could be represented by a precise point in time. Imprecise (upper and lower) previsions may be needed to model incompleteness or conflict in the available information.

In particular there is no justification for a *Bayesian sensitivity analysis* interpretation of $\underline{P}(X)$ and $\overline{P}(X)$, which regards them as lower and upper bounds for some underlying linear prevision $P(X)$ that is not known precisely. Upper and lower probabilities are similarly interpreted as upper and lower bounds for an unknown, precise probability value. This interpretation seems to have been taken for granted in most previous publications on upper and lower probability, including those in the AI literature [25,30,35,36,41,60,62,66,94]. Upper and lower probabilities with this interpretation are sometimes called "probability bounds", "probability intervals" or "generalised probabilities". (The term "upper and lower probability" also may be misleading if it suggests upper and lower bounds for a precise probability; this seems to be why Shafer [71] preferred the new term "belief function".)

The Bayesian sensitivity analysis interpretation is both misleading and unnecessary. It is misleading because, in most problems, no useful meaning can be given to the "underlying linear prevision". It is unnecessary because upper and lower previsions can be given a direct behavioural interpretation, in terms of buying and selling prices for gambles or in terms of their implications in other decision problems, and the behavioural interpretation is sufficient to justify the axioms and calculus of the theory. The distinction is important because the behavioural interpretation and sensitivity analysis lead to different methods for modelling independence and other structural judgements [99].

One of the important contributions of the Dempster-Shafer theory is that it has emphasised that belief functions should not be given a sensitivity analysis interpretation [75]. Every belief function can be represented as a lower envelope of a set of probability measures. This is merely a mathematical representation, however; it is misleading and unnecessary to regard a belief function as a lower bound for some unknown probability measure. In the same way, every coherent lower prevision can be represented as a lower envelope of a set of linear previsions, but this is no reason to regard the lower prevision as a model for partial information about an unknown linear prevision.

In the theory of coherent lower previsions [99], all the axioms and rules are justified purely in terms of the behavioural interpretation, without appealing to a sensitivity analysis interpretation. (This includes the axioms for conditional previsions and the rules for conditioning or updating.) So the theory of coherent lower previsions does not rely in any way on a sensitivity analysis interpretation, and in this respect it does not differ from the Dempster–Shafer theory or possibility theory.

Of course there are some examples where $\underline{P}$ does represent partial information about an unknown Bayesian probability measure and a sensitivity analysis interpretation is justified. The behavioural interpretation still applies in these cases and they can be modelled by coherent lower previsions. But the behavioural interpretation is much more general. It can be applied, for example, to belief functions and possibility measures, for which the sensitivity analysis interpretation is usually inappropriate.

This enables us, in Sections 5 and 6, to view multivalued mappings and inexact judgements in natural language as sources of coherent lower and upper previsions. A behavioural interpretation of these kinds of models need not exclude the interpretations they are normally given as "measures of evidential support" or "degrees of possibility", although it does seem somewhat more definite and useful.

Coherence

Coherence of the lower prevision $\underline{P}$ can be characterised by three axioms:

(P1) $\underline{P}(X) \geqslant \inf\{X(\omega)\colon\ \omega \in \Omega\}$ for all $X \in \mathcal{L}$,

(P2) $\underline{P}(\lambda X) = \lambda \underline{P}(X)$ for all $X \in \mathcal{L}$, $\lambda > 0$,

(P3) $\underline{P}(X+Y) \geqslant \underline{P}(X) + \underline{P}(Y)$ for all $X, Y \in \mathcal{L}$.

There are various equivalent characterisations of coherence in [99]. Wilson and Moral [113] have expressed the coherence axioms in the form of a logic, with an associated proof theory and semantics, and this may appeal to readers who are familiar with classical logic.

Coherence implies the inequalities (for all gambles X and Y)

$$\underline{P}(X) + \underline{P}(Y) \leqslant \underline{P}(X+Y) \leqslant \underline{P}(X) + \overline{P}(Y) \leqslant \overline{P}(X+Y) \leqslant \overline{P}(X) + \overline{P}(Y).$$

The three axioms are consistency requirements which can be justified through the behavioural interpretation of $\underline{P}$. Axiom (P1) asserts a willingness to pay any price less than the infimum possible value of X for the gamble X. (Recall that X is interpreted as an uncertain reward in units of utility.) This is reasonable because such a transaction is certain to increase utility. Axiom (P2) says, in effect, that a gamble Z is acceptable if and only if λZ is acceptable for any positive constant λ. Thus the acceptability of a gamble does not depend on the unit of utility in which its rewards are expressed. To justify the superlinearity axiom (P3), consider two acceptable gambles which involve paying up to $\underline{P}(X)$ for X and up to $\underline{P}(Y)$ for Y. The combined gamble is equivalent to paying up to $\underline{P}(X) + \underline{P}(Y)$ for $X+Y$. Hence $\underline{P}(X+Y)$, the supremum buying price for $X+Y$, should be at least $\underline{P}(X) + \underline{P}(Y)$. Axioms (P2) and (P3) are compelling only if gambles are expressed in terms of a linear utility scale. They would not be reasonable if gambles were expressed in monetary units, as the acceptability of a gamble might then depend on whether its units were dollars or thousands of dollars and the combination of two acceptable monetary transactions might not be acceptable.

In general, conditional previsions will be specified as well as unconditional ones. Coherence axioms that apply to general specifications of conditional previsions are given in [99]. In the simplest case, where unconditional lower previsions $\underline{P}$ and conditional lower previsions $\underline{P}(\cdot \mid B_i)$ are specified for each event B_i in a finite partition $\{B_1, B_2, \ldots, B_k\}$, coherence can be characterised in terms of axioms such as

(C1) $\underline{P}(X) \geqslant \min\{\underline{P}(X \mid B_1), \underline{P}(X \mid B_2), \ldots, \underline{P}(X \mid B_k)\}$,
(C2) $\underline{P}(X) \leqslant \max\{\underline{P}(X \mid B_1), \overline{P}(X \mid B_2), \ldots, \overline{P}(X \mid B_k)\}$,
(C3) $\underline{P}(B_i(X - \underline{P}(X \mid B_i))) = 0$ for $i = 1, 2, \ldots, k$.

The Bayesian theory of de Finetti [28,29] is presented in terms of *linear previsions*, which are simply coherent lower previsions that satisfy the precision condition $\underline{P}(X) = \overline{P}(X)$ for all gambles X. Linear previsions are maximally precise. At the other extreme are the *vacuous previsions* $\underline{P}(X) = \inf\{X(\omega)\colon \omega \in \Omega\}$ and $\overline{P}(X) = \sup\{X(\omega)\colon \omega \in \Omega\}$, which maximise the degree of imprecision $\overline{P}(X) - \underline{P}(X)$ and model complete ignorance about Ω.

In general, the degree of imprecision in previsions can reflect both the amount of information on which they are based and the degree of conflict between different types of information (e.g. between the assessments of several experts, or between prior information and statistical data). In turn, greater imprecision in previsions leads to greater indeterminacy in conclusions (we may be unable to say which of two hypotheses is more probable) and greater indecision (we may be unable to say which of two actions is better).

For example, if several experts assess precise previsions $P_1(X), \ldots, P_n(X)$ then it is natural to define $\underline{P}(X) = \min\{P_i(X)\colon 1 \leqslant i \leqslant n\}$ and $\overline{P}(X) = \max\{P_i(X)\colon 1 \leqslant i \leqslant n\}$. Then $\underline{P}$ is the most precise model that is acceptable to every expert (it represents the behavioural dispositions that are common to all the experts, a kind of "consensus"), and the degree of imprecision $\overline{P}(X) - \underline{P}(X)$ measures the extent of conflict or disagreement amongst the experts concerning X [103]. Other ways of aggregating information from several sources are studied in [97; 99, Sections 4.3, 5.3 and 5.4; 102]. The combination of assessments from different sources is an important problem in expert systems.

Any coherent lower prevision $\underline{P}$ can be written as the lower envelope of a closed convex set $\mathcal{M}$ of linear previsions, i.e. $\underline{P}(X) = \min\{P(X)\colon P \in \mathcal{M}\}$, or as the lower envelope of the set of extreme points $\operatorname{ext}\mathcal{M}$, i.e. $\underline{P}(X) = \min\{P(X)\colon P \in \operatorname{ext}\mathcal{M}\}$. Often the simplest way to characterise the lower prevision $\underline{P}$ is by specifying the probability mass function (or probability density function) of each linear prevision in $\operatorname{ext}\mathcal{M}$.

Now we return to the problem of natural extension. Suppose that a user makes finitely many judgements of uncertainty. The problem for the expert system is to combine these with any other judgements in the system (including those supplied by domain experts) to compute a coherent lower prevision $\underline{P}$. This can be done in three steps: (1) translate the judgements into behavioural terms, hence into constraints on $\underline{P}$; (2) check that all the constraints are mutually consistent; (3) compute the minimal coherent lower prevision $\underline{P}$ that satisfies the constraints. This procedure is illustrated by the following simple example.

Example 1 (*Football game*). Consider a football game whose possible outcomes are win (W), draw (D) or loss (L) for the home team. To express his uncertainty about the outcome, the user makes the judgements:

(a) probably *not* W,

(b) W is more probable than D,

(c) D is more probable than L.

To construct the natural extension of these judgements we first translate them into behavioural terms. We take (a) to mean that the user is willing to give up one unit of utility if W occurs, provided he gains one unit if not W. We will use de Finetti's notation, in which the gamble that pays one unit if event W occurs and zero otherwise (the indicator function of W) is denoted also by W. Then (a) yields the constraint $\overline{P}(W) \leqslant \frac{1}{2}$, or equivalently $\underline{P}(D+L) \geqslant \frac{1}{2}$. According to (b), the subject is willing to pay one unit if D occurs, provided he receives one unit if W occurs. This yields the constraint $\underline{P}(W-D) \geqslant 0$. Similarly (c) yields $\underline{P}(D-L) \geqslant 0$.

The natural extension of the three judgements can be computed by finding the set $\mathcal{M}$ of all probability mass functions (w,d,l) that are consistent with the judgements, by solving the system of linear inequalities: $d+l \geqslant \frac{1}{2}$, $w \geqslant d$, $d \geqslant l$, $w \geqslant 0$, $d \geqslant 0$, $l \geqslant 0$, $w+d+l=1$. Because this system has solutions, the three judgements are consistent. The extreme points of $\mathcal{M}$ are the three probability mass functions $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$, $(\frac{1}{2},\frac{1}{2},0)$, $(\frac{1}{2},\frac{1}{4},\frac{1}{4})$. The lower prevision of any gamble X can then be calculated as the minimum expected value of X under these three mass functions. For example,

$$\underline{P}(W) = \min\left\{\tfrac{1}{3},\tfrac{1}{2},\tfrac{1}{2}\right\} = \tfrac{1}{3}, \qquad \overline{P}(W) = \max\left\{\tfrac{1}{3},\tfrac{1}{2},\tfrac{1}{2}\right\} = \tfrac{1}{2}.$$

Many other kinds of qualitative or quantitative judgements could be added to the three we have considered, for example,

(d) if *not* D then W is very likely,

(e) W is between 1 and 2 times as probable as D,

(f) I am willing to bet on L at odds of 4 to 1,

(g) W has precise probability 0.4.

Some other natural-language judgements are considered in Section 6 and further examples are given in [99]. In more complicated problems, one might also make judgements of conditional probabilities, conditional independence, positive or negative dependence, permutability or exchangeability, intervals of measures, upper and lower density functions or distribution functions or quantiles. All these judgements can be regarded as constraints on $\underline{P}$. They can be combined (in principle) by natural extension, although the linear programming problem may become intractable when many assessments are made, and then it may be necessary to restrict attention to special types of assessment or model. Note especially that ordinary-language judgements such as (a), (b) and (c) can be combined with numerical assessments; the two types of judgement occur together in many applications.

How would Bayesians deal with the football example? They would require a user to make more precise assessments in order to determine a single probability measure that is consistent with judgements (a), (b) and (c). But he may be unable to go beyond these qualitative judgements, except by choosing arbitrary numbers which would not reflect his state of uncertainty about the game, because he lacks either information about the game or expertise in assessing probabilities.

One popular method for selecting a unique probability measure from $\mathcal{M}$ is by *maximising entropy*. This gives the mass function $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in the football example. But there are alternative methods, such as assigning a second-order probability distribution on $\mathcal{M}$, which yield different answers, and any choice of a single probability measure seems arbitrary. Note that maximum entropy does not distinguish the partial information provided by the three judgements from complete ignorance: the same probability model would be selected if we had no information at all about the game. No precise probability measure can reflect the imprecision of the three judgements. For discussion of maximum entropy, see [42,52,99].

Upper and lower probability

By applying the behavioural interpretation of lower and upper previsions, the lower (or upper) probability of event A can be interpreted as specifying acceptable rates for betting *on* (or *against*) A. Consider a choice of whether to bet on or against A at betting rate x, meaning odds of x to $1-x$ on A. You will bet *on* A if x is less than $\underline{P}(A)$, you will bet *against* A if x is greater than $\overline{P}(A)$, and your choice is not determined (it may be reasonable to choose either way) if x is between $\underline{P}(A)$ and $\overline{P}(A)$.

Upper and lower probabilities have the basic properties $\underline{P}(\emptyset) = \overline{P}(\emptyset) = 0$, where $\emptyset$ denotes the empty set, $\underline{P}(\Omega) = \overline{P}(\Omega) = 1$, $\overline{P}(A) = 1 - \underline{P}(A^c)$, where A^c denotes the complement of A (hence upper probabilities are determined by lower probabilities, and vice versa), $0 \leqslant \underline{P}(A) \leqslant \overline{P}(A) \leqslant 1$, and

$$\underline{P}(A) + \underline{P}(B) \leqslant \underline{P}(A \cup B) \leqslant \underline{P}(A) + \overline{P}(B) \leqslant \overline{P}(A \cup B) \leqslant \overline{P}(A) + \overline{P}(B)$$

when A and B are disjoint.

These properties are necessary but not sufficient for coherence. The further property of *2-monotonicity*, that $\underline{P}(A \cup B) + \underline{P}(A \cap B) \geqslant \underline{P}(A) + \underline{P}(B)$ for all events A and B, is sufficient for coherence but not necessary. Some of the mathematical theory of coherent lower probabilities simplifies when the lower probabilities are 2-monotone (also known as "Choquet capacities of order 2"); see especially the simple formulas for conditional probabilities and expectations at the end of this section, and the theory in [40,96,102].

One way of constructing a lower prevision $\underline{P}$ is to assess upper and lower probabilities $\overline{P}(A)$ and $\underline{P}(A)$ for all events A and to construct $\underline{P}$ by natural extension. Not all coherent lower previsions can be constructed in this way, because many coherent lower previsions can generate the same upper and lower probabilities.

In the football example, suppose that the initial assessments are upper probabilities $\frac{1}{2}, \frac{1}{2}, \frac{1}{3}$ and lower probabilities $\frac{1}{3}, \frac{1}{4}, 0$ for the events W, D, L respectively. (These are the upper and lower probabilities generated by the three judgements considered earlier.) Their natural extension to a lower prevision $\underline{P}_2$, calculated as in the previous example, is the lower envelope of the five probability mass functions $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, $(\frac{1}{3}, \frac{1}{2}, \frac{1}{6})$ and $(\frac{5}{12}, \frac{1}{4}, \frac{1}{3})$. This is different from the lower prevision $\underline{P}_1$ constructed earlier, e.g. if $(X(W), X(D), X(L)) = (1, -1, 0)$ then $\underline{P}_1(X) = 0$ but $\underline{P}_2(X) = -\frac{1}{6}$, although $\underline{P}_1$ and $\underline{P}_2$ generate the same upper and lower probabilities for all events.

Thus lower previsions (defined on all gambles) contain more information than lower probabilities (defined on all events), and information may be lost when uncertainty is modelled solely in terms of upper and lower probabilities. That is why we take lower prevision to be the fundamental model, rather than lower probability. Lower (or upper) previsions are more expressive than lower (or upper) probabilities.

The extra information in lower previsions is often crucial in defining *conditional* upper and lower probabilities. Without it the conditional probabilities may be excessively imprecise. That can be seen in the football example. Taking $B = \{W, D\}$, the first (lower prevision) model gives $\underline{P}(W \mid B) = \frac{1}{2}$ and $\overline{P}(W \mid B) = \frac{2}{3}$, whereas the second (lower probability) model gives less precise values $\underline{P}(W \mid B) = \frac{2}{5}$ and $\overline{P}(W \mid B) = \frac{2}{3}$. (In both cases the conditional probabilities were calculated by conditioning the extreme points and finding upper and lower envelopes. This agrees with natural extension.)

In general, conditional upper and lower probabilities are not uniquely determined by unconditional upper and lower probabilities through the coherence axioms. But conditional upper and lower probabilities and previsions are uniquely determined by unconditional lower previsions, through axiom (C3), provided the conditioning event has nonzero lower probability.

Another reason for regarding lower previsions as fundamental is that they can often be assessed more easily than lower probabilities. If we are primarily concerned with a real-valued quantity whose value is uncertain, such as a person's age, it may be easier to assess upper and lower previsions for the quantity rather than upper and lower probabilities for the events that the quantity belongs to particular sets. Lower previsions are also needed when we start by assessing upper and lower probabilities, then observe an event and condition by natural extension; again information may be lost if only the updated upper and lower probabilities are reported [41].

Thus upper and lower probabilities are often inadequate models for uncertainty. This is a defect of all theories of upper and lower probability, including the theory of belief functions and, on the interpretation adopted here, possibility theory. A more general theory, dealing with upper and lower previsions or an equally general alternative, is needed. In particular, upper and lower probabilities are inadequate for modelling natural-language judgements of uncertainty.

Natural extension

Now consider the general procedure of natural extension. Suppose that a user makes finitely many judgements which can be translated into constraints on conditional lower previsions, say $\underline{P}(X_i \mid B_i) \geqslant \mu_i$ for $1 \leqslant i \leqslant k$, where μ_i are specified real numbers. (The case of unconditional previsions is included by taking $B_i = \Omega$, and the case of precise judgements $P(X \mid B) = \mu$ by taking $\underline{P}(X \mid B) \geqslant \mu$ and $\underline{P}(-X \mid B) \geqslant -\mu$.) Then the natural extension to any conditional lower prevision $\underline{P}(X \mid B)$ can be computed by solving a linear program, using the formula

$$\underline{P}(X \mid B) = \sup\left\{\mu\colon\ B(X-\mu) \geqslant \sum_{i=1}^{k} \lambda_i B_i(X_i-\mu_i) \text{ for some } \lambda_i \geqslant 0\right\}, \tag{1}$$

where $Y \geqslant Z$ denotes $Y(\omega) \geqslant Z(\omega)$ for all $\omega \in \Omega$, B and B_i stand for indicator functions, and μ and μ_i denote constant gambles. This definition of natural extension can be justified through the behavioural interpretation of lower previsions: if you are willing to pay up to μ_i for X_i conditional on B_i then, by combining positive multiples of such gambles, I can induce you to pay up to $\underline{P}(X \mid B)$ for X conditional on B. The resulting value of $\underline{P}(X \mid B)$ is finite provided the initial judgements are consistent in the sense that they "avoid sure loss", a much weaker requirement than coherence [99].

Natural extension summarises all the inferences that can be derived from the initial judgements through the rules of coherence. In fact, the natural extensions $\underline{P}$ are the minimal lower previsions that satisfy the initial constraints and are coherent. There may be other coherent lower previsions $\underline{P}'$ that satisfy the initial constraints, but they must dominate the natural extensions in the sense that $\underline{P}'(X \mid B) \geqslant \underline{P}(X \mid B)$ for all gambles X and all events B, and therefore they incorporate additional information that is not implied by the initial set of judgements. For example, Dempster's rule of conditioning often disagrees with natural extension because it implicitly involves judgements of conditional independence that are not implied by the initial belief function and which may not be consistent with the initial belief function.

Natural extension is a very general method of inference. Indeed the following important constructions can be regarded as special types of natural extension: construction of (upper and lower) expectations from probabilities; construction of conditional (upper and lower) probabilities from unconditional ones; the "Fundamental Theorem of Probability" of de Finetti [28]; construction of inner and outer measures; construction of joint probabilities from marginal and conditional ones; and construction of probability models from qualitative judgements. The initial constraints are quite general and this allows a user to express his uncertainty in whatever forms are most convenient, e.g. through qualitative judgements such as the natural-language expressions listed in Section 6, or through a combination of qualitative and quantitative judgements. The results of natural extension are also very general; it can be used to construct a lower prevision $\underline{P}(X \mid B)$ for any conceivable gamble X and event B, hence to construct upper and lower probabilities and preferences between gambles. Inferences can be made by computing lower previsions of important variables conditional on all available information, and decisions by computing lower previsions of differences between utility functions, both of which are linear programming problems.

Alternatively the natural extension can be computed, as in the football example, by solving a set of linear inequalities to obtain the extreme probability mass functions that are consistent with the initial constraints, and then forming their lower envelope. This is the dual linear programming problem. In the special case where all the initial probability judgements are precise and no independence judgements are made, it is equivalent to what is known as "probabilistic logic" [60,62].

This alternative procedure breaks down when independence constraints are included because these are nonlinear. On the behavioural interpretation of lower previsions, a judgement that two events are independent means that the upper and lower probabilities of one event would not change if you learned whether or not the other event occurred. This is quite different from judging that the events are independent under some Bayesian probability measure that satisfies the other constraints, which would be the appropriate definition of independence under a Bayesian sensitivity analysis interpretation [105].

Consider the simplest case where upper and lower probabilities are assessed for two events A and B that are judged to be independent, say $\underline{P}(A) = \underline{\alpha}$, $\overline{P}(A) = \bar{\alpha}$, $\underline{P}(B) = \underline{\beta}$ and $\overline{P}(B) = \bar{\beta}$. Then the behavioural interpretation of independence imposes the 12 constraints $\underline{P}(A) \geqslant \underline{\alpha}$, $\underline{P}(A \mid B) \geqslant \underline{\alpha}$, $\underline{P}(A \mid B^c) \geqslant \underline{\alpha}, \ldots, \underline{P}(B^c \mid A) \geqslant 1 - \bar{\beta}$, $\underline{P}(B^c \mid A^c) \geqslant 1 - \bar{\beta}$, and the natural extension of the judgements can be computed by natural extension of these constraints, using Eq. (1).

Bayesian sensitivity analysis would model the judgements in a different way, by finding the set of joint probability measures P which satisfy the constraints $\underline{\alpha} \leqslant P(A) \leqslant \bar{\alpha}$, $\underline{\beta} \leqslant P(B) \leqslant \bar{\beta}$ and $P(A \cap B) = P(A)P(B)$. (The independence constraint is nonlinear and this makes the computations difficult in more complicated examples.) Upper and lower previsions are then defined as upper and lower envelopes of the set of solutions. The two approaches do produce different numerical answers; see the example of two unreliable witnesses in Section 5.

All previous work in AI, e.g. [25,30,35], seems to have taken the sensitivity analysis definition of independence for granted, without considering the behavioural definition. A comparative study of the two approaches is needed, with regard to both interpretation and computational methods.

Using the behavioural definition of independence, the computation of natural extension is more complicated when many judgements of conditional independence, or other structural constraints on lower previsions, are made. Computations can be carried out, in general, by solving a finite sequence of linear programs with progressively stronger constraints. At each stage the independence and structural constraints are applied to the conditional lower previsions that were produced by natural extension in the previous stage, giving a stronger set of constraints for the next application of natural extension. It is not yet clear whether this is feasible for complex problems with moderately large numbers of conditional independence judgements, or for the types of belief networks that have been studied by Bayesians.

There has been a substantial amount of work in recent years concerning the propagation of upper and lower probabilities [25,30,94] or convex sets of Bayesian probability measures [10,94] using the sensitivity analysis definition of independence.

Natural extension can be applied to a completely arbitrary collection of assessments, but the linear programming computations will become impracticable when the number of assessments is sufficiently large, especially when many independence judgements are made. (Of course Bayesian computations also become intractable in unstructured problems.) When this happens, several approaches might be considered.

First, we could use simpler formulas, involving only local computations, to give lower bounds for $\underline{P}(X \mid B)$ and upper bounds for $\overline{P}(X \mid B)$. Some examples of such formulas are given in the following subsections (see also [66,94]). This approach produces conclusions that are always valid but less precise than those produced by natural extension. Alternatively we might try to approximate $\underline{P}(X \mid B)$ without necessarily finding a lower bound. Some approximation methods are suggested in [60] for the case in which all the probability assessments are precise. This approach is less appealing than the first as it may produce invalid conclusions.

Finally, the approach that seems most likely to be useful in practical problems is to develop special types of imprecise probability models, such as the 2-monotone lower probabilities or models defined in terms of upper and lower density functions or mass functions, for which natural extensions can be computed explicitly, without linear programming. Examples are given in the following subsections.

The expert system INFERNO [66], designed to diagnose faults on oil rigs, uses upper and lower probabilities to measure uncertainty but differs in two important ways from the approach suggested here. First, INFERNO works by propagating simple constraints on upper and lower probabilities. This has the advantage of simplifying and localising computations. However, because the propagation method and constraints are much weaker than those given by natural extension, the conclusions of the system, while valid, may often be too weak to be useful. Second, the system provides no way of updating probabilities by conditioning on new evidence. Any new information can only be regarded as specifying further constraints on a fixed, unconditional probability measure.

Calculus

All the rules of the theory follow from the principles of coherence and natural extension, and they can therefore be justified through the behavioural interpretation of lower previsions. An important example is the *generalised Bayes rule* (GBR): when $\underline{P}$ is a coherent lower prevision defined on a sufficiently large set of gambles and $\underline{P}(B) > 0$, the conditional lower prevision $\underline{P}(X \mid B)$ is the unique solution x of the equation $\underline{P}(B(X - x)) = 0$. This equation can sometimes be solved explicitly, and otherwise by simple iterative algorithms. (For example, take x_0 to be any estimate of x, and define a sequence of estimates by $x_{n+1} = x_n + 2\underline{P}(B(X - x_n))/(\overline{P}(B) + \underline{P}(B))$. Then the sequence converges to the solution x, and the error is bounded by $c\alpha^n$ where $\alpha = (\overline{P}(B) - \underline{P}(B))/(\overline{P}(B) + \underline{P}(B))$.) Thus conditional previsions are uniquely determined by unconditional ones, provided the lower previsions involved in the GBR have been specified and $\underline{P}(B) > 0$. The GBR can be used to update the initial prevision $\underline{P}$ after learning that event B has occurred. When $\underline{P}$ is a linear prevision, the GBR reduces to $P(X \mid B) = P(BX)/P(B)$, a version of the usual Bayes' rule.

As another example of natural extension, suppose $\{A_1, A_2, \dots, A_k\}$ is an exhaustive set of mutually exclusive hypotheses, B is an observable event, and upper and lower probabilities $\overline{P}(A_i), \underline{P}(A_i), \overline{P}(B \mid A_i), \underline{P}(B \mid A_i)$ are assessed for $1 \leqslant i \leqslant k$. (This generalises the Bayesian model considered in Section 3.) Assume that the assessments satisfy the coherence conditions $0 \leqslant \underline{P}(B \mid A_i) \leqslant \overline{P}(B \mid A_i) \leqslant 1$, $0 \leqslant \underline{P}(A_i) \leqslant \overline{P}(A_i) \leqslant 1$, and $\overline{P}(A_i) + \sum_{j\neq i} \underline{P}(A_j) \leqslant 1 \leqslant \underline{P}(A_i) + \sum_{j\neq i} \overline{P}(A_j)$ for every $1 \leqslant i \leqslant k$. We might wish to calculate the natural extensions of these assessments to predictive probabilities $\underline{P}(B)$ and $\overline{P}(B)$, and to posterior probabilities $\underline{P}(A_i \mid B)$ and $\overline{P}(A_i \mid B)$.

To do so, it is convenient to define the extremal probability measures P which satisfy $\underline{P}(A_j) \leqslant P(A_j) \leqslant \overline{P}(A_j)$ for $1 \leqslant j \leqslant k$. Any ordering of the hypotheses, denoted by $A_{n_1}, A_{n_2}, \dots, A_{n_k}$, determines an extremal P as follows. Set $P(A_{n_j})$ equal to $\underline{P}(A_{n_j})$ if $j < r$, equal to $\overline{P}(A_{n_j})$ if $j > r$, and intermediate between $\underline{P}(A_{n_j})$ and $\overline{P}(A_{n_j})$ if $j = r$, where the values of r and $P(A_{n_r})$ are determined by $\sum_{j=1}^{k} P(A_{n_j}) = 1$.

Using [99, Theorem 6.7.2], the natural extension can be written as $\underline{P}(B) = \sum_{j=1}^{k} \underline{P}(B \mid A_j)P_1(A_j)$, a weighted mean of the values $\underline{P}(B \mid A_j)$, where P_1 is the extremal probability measure determined by ordering the hypotheses to have decreasing values of $\underline{P}(B \mid A_j)$. Similarly $\overline{P}(B) = \sum_{j=1}^{k} \overline{P}(B \mid A_j)P_2(A_j)$, where P_2 is defined by ordering the hypotheses to have increasing $\overline{P}(B \mid A_j)$. (These and the following formulas simplify when $\underline{P}(A_j) = \overline{P}(A_j)$ for all j, as then $P_1(A_j) = P_2(A_j) = \underline{P}(A_j)$.) In general $\underline{P}(B)$ is bounded by $\sum_{j=1}^{k} \underline{P}(B \mid A_j)\underline{P}(A_j) \leqslant \underline{P}(B) \leqslant \sum_{j=1}^{k} \underline{P}(B \mid A_j)\overline{P}(A_j)$, but these bounds may be far from sharp.

Using a result from [99, 8.5.4], the posterior lower and upper probabilities after observing B are

$$\underline{P}(A_i \mid B) = \frac{\underline{P}(B \mid A_i)P_3(A_i)}{\underline{P}(B \mid A_i)P_3(A_i) + \sum_{j \neq i} \overline{P}(B \mid A_j)P_3(A_j)}, \tag{2}$$

where P_3 is defined by setting $n_1 = i$ and ordering the other hypotheses to have increasing $\overline{P}(B \mid A_j)$, and

$$\overline{P}(A_i \mid B) = \frac{\overline{P}(B \mid A_i)P_4(A_i)}{\overline{P}(B \mid A_i)P_4(A_i) + \sum_{j \neq i} \underline{P}(B \mid A_j)P_4(A_j)}, \tag{3}$$

where P_4 is defined by setting $n_k = i$ and ordering the other hypotheses to have decreasing $\underline{P}(B \mid A_j)$. In this problem the computations are quite simple. If needed, simpler bounds could be obtained by substituting either $\underline{P}(A_j)$ or $\overline{P}(A_j)$ for $P_3(A_j)$ and $P_4(A_j)$ in Eqs. (2) and (3). A special case of these results was used in [25]. More general models involving belief networks are studied in [94]. Other results, allowing general prior upper and lower previsions concerning the hypotheses A_j, are in [96,99].

These formulas simplify in the case where there are only two hypotheses, A_1 and A_2. Let $\overline{\rho} = \overline{P}(A_1)/\underline{P}(A_2)$ and $\underline{\rho} = \underline{P}(A_1)/\overline{P}(A_2)$ denote the *prior upper and lower odds* on A_1, and let $\overline{\lambda}(B) = \overline{P}(B \mid A_1)/\underline{P}(B \mid A_2)$ and $\underline{\lambda}(B) = \underline{P}(B \mid A_1)/\overline{P}(B \mid A_2)$ be the *upper and lower likelihood ratios* generated by B. The posterior upper and lower probabilities of the hypotheses are determined by the *posterior upper and lower odds* on A_1, which are given by the multiplicative formulas

$$\overline{\rho}(B) = \frac{\overline{P}(A_1 \mid B)}{\underline{P}(A_2 \mid B)} = \overline{\rho}\,\overline{\lambda}(B), \qquad \underline{\rho}(B) = \frac{\underline{P}(A_1 \mid B)}{\overline{P}(A_2 \mid B)} = \underline{\rho}\,\underline{\lambda}(B).$$

These generalise the Bayesian formula: posterior odds = prior odds × likelihood ratio.

Conditional probabilities

Suppose that coherent upper and lower probabilities, $\overline{P}$ and $\underline{P}$, are specified for all subsets of Ω, and that we wish to construct *conditional* upper and lower probabilities $\overline{P}(\cdot \mid B)$ and $\underline{P}(\cdot \mid B)$, e.g. to update beliefs after observing the event B. This problem is of interest because the other theories examined in this paper (Bayesian, belief functions and possibility theory) attempt to define conditional probabilities and expectations in terms of unconditional probabilities. (This seems to be a hangover from the classical theory of probability.) It is important to recognise, however, that this is a special case of a much more general problem. In general we will make whatever probability assessments we can and use natural extension to construct other probabilities and previsions; it may often be easier and more informative to assess some conditional probabilities and previsions directly than to assess all unconditional probabilities (e.g. consider the problem in the previous subsection). So the problem considered here, in which unconditional upper and lower probabilities are assessed for all events but no other assessments are made, should be regarded as atypical.

Let B be any subset of Ω such that $\underline{P}(B) > 0$, and let $\mathcal{M}$ denote the set of all probability measures P such that $P(C) \geqslant \underline{P}(C)$ for all $C \subseteq \Omega$. Then lower probabilities conditional on B can be computed by applying the general formula for natural extension, giving

$$\begin{aligned}\underline{P}(A \mid B) &= \sup\left\{\mu\colon\ B(A-\mu) \geqslant \sum_{i=1}^{n} \lambda_i(A_i - \underline{P}(A_i)) \right.\\ &\qquad\qquad \left. \text{for some } n \geqslant 0,\ A_i \subseteq \Omega,\ \lambda_i \geqslant 0\right\}\\ &= \inf\{P(A\cap B)/P(B)\colon\ P \in \mathcal{M}\}. \qquad (4)\end{aligned}$$

Provided Ω is finite, $\underline{P}(A \mid B)$ can be computed by linear programming techniques. (Much simpler formulas can be used in the special case where $\underline{P}$ is 2-monotone, discussed below.) A more general formula, which applies whenever $\overline{P}(B) > 0$, is

$$\underline{P}(A \mid B) = \inf\{P(A\cap B)/P(B)\colon\ P \in \mathcal{M},\ P(B) > 0\}.$$

The conditional upper probabilities are defined by

$$\overline{P}(A \mid B) = \sup\{P(A\cap B)/P(B)\colon\ P \in \mathcal{M},\ P(B) > 0\}.$$

The conditional probabilities $\underline{P}(\cdot \mid B)$ and $\overline{P}(\cdot \mid B)$ defined by these formulas are always coherent lower and upper probabilities. Thus conditioning by natural extension preserves coherence. Moreover, $\underline{P}(\cdot \mid B)$ and $\overline{P}(\cdot \mid B)$ are always coherent with the unconditional probabilities $\underline{P}$ and $\overline{P}$. We will see that the families of 2-monotone lower probabilities, belief functions and possibility measures are also closed under conditioning by natural extension.

Provided $\underline{P}(B) > 0$, the conditional probabilities satisfy the following inequalities:

$$\frac{\underline{P}(A\cap B)}{\underline{P}(A\cap B) + \overline{P}(A^{c}\cap B)} \leqslant \underline{P}(A \mid B) \leqslant \min\left\{\frac{\underline{P}(A\cap B)}{\underline{P}(B)}, \frac{\overline{P}(A\cap B)}{\overline{P}(B)}\right\},$$

$$\frac{\overline{P}(A\cap B)}{\overline{P}(A\cap B) + \underline{P}(A^{c}\cap B)} \geqslant \overline{P}(A \mid B) \geqslant \max\left\{\frac{\underline{P}(A\cap B)}{\underline{P}(B)}, \frac{\overline{P}(A\cap B)}{\overline{P}(B)}\right\}.$$

These hold whenever $\underline{P}$ and $\overline{P}$ are coherent, but the lower bound for $\underline{P}(A \mid B)$ is actually achieved whenever $\underline{P}$ has the stronger property of *2-monotonicity*. In that case the natural extensions are

$$\underline{P}(A \mid B) = \frac{\underline{P}(A \cap B)}{\underline{P}(A \cap B) + \overline{P}(A^c \cap B)},$$

$$\overline{P}(A \mid B) = \frac{\overline{P}(A \cap B)}{\overline{P}(A \cap B) + \underline{P}(A^c \cap B)}, \tag{5}$$

provided $\underline{P}(B) > 0$ [9,15,23,96]. (The same formulas apply when $\overline{P}(B) > 0$ and $\underline{P}(B) = 0$, provided the denominators are nonzero; set $\underline{P}(A \mid B) = 1$ if the first denominator is zero, and $\overline{P}(A \mid B) = 0$ if the second denominator is zero. This case is uninteresting because $\underline{P}(A \mid B) = 0$ and $\overline{P}(A \mid B) = 1$ for every A that satisfies $\overline{P}(A \cap B) > 0$ and $\overline{P}(A^c \cap B) > 0$.) Numerical examples of this rule will be discussed in Sections 5 and 6. It has been generalised in [96,102,108] to characterise the posterior probabilities generated by a statistical likelihood function and prior lower probabilities that are 2-monotone.

When $\underline{P}$ is 2-monotone, the conditional lower probabilities $\underline{P}(\cdot \mid B)$ are 2-monotone as well as coherent. Thus conditioning by natural extension preserves 2-monotonicity. (See [96] for a proof.)

Expectations

Again suppose that coherent upper and lower probabilities are specified for all subsets of Ω. We wish to construct upper and lower previsions or "expectations", $\overline{P}(X)$ and $\underline{P}(X)$, for gambles (bounded random variables) X. For example, in order to make decisions we would need to compute upper and lower previsions of differences between utility functions. But again we must point out that the case considered here, in which unconditional probabilities are assessed for all events but no other previsions are directly assessed, is atypical; especially as some judgements, such as the natural-language judgements in the football example, cannot be modelled adequately in terms of upper and lower probabilities.

Again the upper and lower previsions can be computed from the general formula for natural extension (1), giving

$$\underline{P}(X) = \sup\left\{\mu:\ X - \mu \geqslant \sum_{i=1}^{n} \lambda_i(A_i - \underline{P}(A_i)) \text{ for some } n \geqslant 0,\ A_i \subseteq \Omega,\ \lambda_i \geqslant 0\right\}$$
$$= \inf\left\{\int X \,\mathrm{d}P:\ P \in \mathcal{M}\right\}, \tag{6}$$

$$\overline{P}(X) = \inf\left\{\mu:\ X - \mu \leqslant \sum_{i=1}^{n} \lambda_i(A_i - \overline{P}(A_i)) \text{ for some } n \geqslant 0,\ A_i \subseteq \Omega,\ \lambda_i \geqslant 0\right\}$$
$$= \sup\left\{\int X \,\mathrm{d}P:\ P \in \mathcal{M}\right\}. \tag{7}$$

Again these formulas simplify in the special case where the lower probability $\underline{P}$ is 2-monotone. Define $\overline{F}_X$ and $\underline{F}_X$, the upper and lower distribution functions of X, by

$\overline{F}_X(x) = \overline{P}(\{\omega: X(\omega) \leqslant x\})$ and $\underline{F}_X(x) = \underline{P}(\{\omega: X(\omega) \leqslant x\})$. Provided $\underline{P}$ is 2-monotone, the natural extensions can be written as Choquet integrals [96]

$$\underline{P}(X) = \int_{-\infty}^{\infty} x \, \mathrm{d}\overline{F}_X(x), \qquad \overline{P}(X) = \int_{-\infty}^{\infty} x \, \mathrm{d}\underline{F}_X(x). \tag{8}$$

Decision making

Upper and lower previsions are used to make decisions in the following way. Suppose we need to choose an action from a finite set of possible actions $\{a_1, a_2, \ldots, a_k\}$, where the utility $U(a,\omega)$ of action a depends on the unknown ω. (We assume that utilities are specified precisely; otherwise the decision problem is more complicated.) Define a gamble X_j by $X_j(\omega) = U(a_j,\omega)$ for each $j = 1,2,\ldots,k$. To compare two actions a_i and a_j we compute the upper and lower previsions $\overline{P}(X_i - X_j)$ and $\underline{P}(X_i - X_j)$ based on all available information. Then action a_i is preferred to a_j if $\underline{P}(X_i - X_j) > 0$, a_j is preferred to a_i if $\overline{P}(X_i - X_j) < 0$, and if neither condition holds there is insufficient information to determine a preference. Say that action a_i is *maximal* if there is no other action a_j that is preferred to a_i. The non-maximal actions can be eliminated, and it is reasonable to choose any of the maximal actions. For more details and applications to real decision problems see [99, Sections 3.9 and 5.6; 101; 103].

This method produces only a partial preference ordering in general, and there may be more than one maximal action. Methods for generating complete preferences are discussed in [57,85,99]. All such methods seem somewhat arbitrary. It should be acknowledged that, when probability judgements are imprecise, there may be more than one reasonable course of action.

Imprecise conclusions

The inferences produced by natural extension are often imprecise. That can be seen in simple examples, e.g. if A and B are logically independent events and an expert assesses that each has precise probability $\frac{1}{2}$ but makes no judgement about their degree of dependence, then the natural extension to upper and lower probabilities for their intersection is $\overline{P}(A \cap B) = \frac{1}{2}$, $\underline{P}(A \cap B) = 0$. When natural extension is used to compute conditional upper and lower probabilities from unconditional ones, the conditional probabilities may be highly imprecise, especially when the initial probabilities are 2-monotone. (Two examples are discussed later in the paper.) In such cases, other methods of conditioning such as Dempster's rule often produce conditional probabilities that are more precise. Advocates of the alternative methods have suggested that natural extension tends to produce *excessive* imprecision.

Wilson and Moral [113] have given an interesting example of this, also discussed in [120]. Suppose that an expert makes two assessments of conditional lower probabilities: $\underline{P}(B \mid A) = 1$ and $\underline{P}(C \mid B) \geqslant 0.999$, where A, B and C are logically independent events. What do these judgements imply about $\underline{P}(C \mid A)$?

It may seem that we can make a fairly strong inference in this case, because the second judgement is "close to" $\underline{P}(C \mid B) = 1$ which, together with $\underline{P}(B \mid A) = 1$, would yield the inference $\underline{P}(C \mid A) = 1$ by natural extension. (To see that, apply the general formula for natural extension (1), and use $A \cap C^c \subseteq (A \cap B^c) \cup (B \cap C^c)$ to show that $A(C-1) \geqslant A(B-1)+B(C-1)$.) However, natural extension of the expert's judgements produces only the trivial inferences $\underline{P}(C \mid A) = 0$ and $\overline{P}(C \mid A) = 1$.

These inferences may indeed seem "excessively imprecise", but I believe that they are reasonable because nothing more can be derived from the two explicit judgements without making further assumptions. Suppose that an alternative method of inference produced the conclusion $\underline{P}(C \mid A) \geqslant \delta$ from the two judgements, where δ is a specified positive number. The expert might then make some further assessments about the events. Suppose he makes the further judgements that $\underline{P}(A) \geqslant 0.001$ and $\overline{P}(A \cap C) = 0$. These judgements are perfectly consistent with the two initial judgements, since all four judgements are consistent with a Bayesian probability measure P that satisfies $P(B) = 1$, $P(A \cap B) = P(A) = 0.001$, $P(B \cap C) = P(C) = 0.999$ and $P(A \cap C) = 0$. However the new judgements are not consistent with the inference $\underline{P}(C \mid A) \geqslant \delta$ because they imply $\overline{P}(C \mid A) = 0$. This shows that the inference goes beyond the information contained in the two initial judgements; it relies on extra (implicit) assumptions which may be inconsistent with the expert's other beliefs.

This argument can be generalised as follows. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ denote two sets of judgements and let $\mathcal{E}_1$ denote a set of inferences produced from $\mathcal{D}_1$ by applying the rules of the calculus. Then we require the following consistency principle: if the overall set of judgements $(\mathcal{D}_1 \cup \mathcal{D}_2)$ is consistent then $(\mathcal{E}_1 \cup \mathcal{D}_2)$ should be consistent.

The force of this principle depends on the technical meaning that is given to "consistency". In the theory of lower previsions "consistency" is identified with "avoiding sure loss", the rules of natural extension do satisfy the consistency principle, and they are the *strongest* rules which do so. That is, the inferences given by natural extension are the most precise inferences possible, if the consistency principle is to be satisfied.

It seems, therefore, that natural extension produces exactly the inferences that are implied by the explicit judgements and assumptions. Inferences may be excessively imprecise, not because natural extension is the wrong method of inference, but rather because the judgements and assumptions are excessively imprecise. Indeed the computation of natural extensions will often reveal indeterminacy in conclusions or decisions that compels us to make our judgements more precise. For example, the problem that unconditional upper and lower probabilities tend to generate very imprecise conditional probabilities can sometimes be resolved by assessing upper and lower previsions, which are more informative than probabilities and therefore generate more precise inferences.

One way to sharpen inferences in many problems is to add judgements of conditional independence or dependence. Consider again the judgements $\underline{P}(B \mid A) = 1$ and $\underline{P}(C \mid B) \geqslant 0.999$. The natural extensions must satisfy the coherence conditions

$$\underline{P}(C \mid A) \geqslant \underline{P}(B \cap C \mid A) \geqslant \underline{P}(C \mid A \cap B)\underline{P}(B \mid A),$$

hence $\underline{P}(C \mid A) \geqslant \underline{P}(C \mid A \cap B)$ in this case. This is useful only if we can relate $\underline{P}(C \mid A \cap B)$ to $\underline{P}(C \mid B)$. One way to do so is to judge that A and C are *conditionally independent* given B, which gives $\underline{P}(C \mid A \cap B) = \underline{P}(C \mid B) \geqslant 0.999$ and hence

$\underline{P}(C \mid A) \geqslant 0.999$, a very strong conclusion. The same inference is produced by the weaker judgement that A and C are *nonnegatively correlated* conditional on B, so that $\underline{P}(C \mid A \cap B) \geqslant \underline{P}(C \mid B)$. Either condition suffices to rule out the possibility that A and C are essentially incompatible events, and this must be ruled out before any nontrivial inference can be obtained.

Another way to sharpen inferences in this problem is to make a further numerical assessment of $\underline{P}(A \mid B)$. Coherence requires that

$$\underline{P}(C \mid A) \geqslant \underline{P}(C \mid A \cap B) \geqslant \underline{P}(A \cap C \mid B)/(\underline{P}(A \cap C \mid B) + \overline{P}(A \cap C^c \mid B)).$$

Here

$$\underline{P}(A \cap C \mid B) \geqslant \underline{P}(A \mid B) + \underline{P}(C \mid B) - 1 \geqslant \underline{P}(A \mid B) - 0.001,$$
$$\overline{P}(A \cap C^c \mid B) \leqslant \overline{P}(C^c \mid B) = 1 - \underline{P}(C \mid B) \leqslant 0.001.$$

Hence $\underline{P}(C \mid A) \geqslant 1 - 0.001/\underline{P}(A \mid B)$, and any assessment of $\underline{P}(A \mid B)$ greater than 0.001 will produce a nontrivial lower bound for $\underline{P}(C \mid A)$. For example, the judgement that $\underline{P}(A \mid B) \geqslant 0.01$ gives the strong conclusion $\underline{P}(C \mid A) \geqslant 0.9$.

It appears that, in moderately complex expert systems, judgements or assumptions of conditional independence are needed to reduce the effort of assessment and produce useful conclusions. Many expert systems use expert knowledge about causal relationships to build "belief networks" based on assumptions of conditional independence. In other systems independence constraints are taken as *default assumptions*; if the expert or user supplies no information about the relationship between two events or variables then they are assumed to be independent.

Such default assumptions may sometimes be needed to produce useful conclusions, but it is important that they always be made as explicit as possible (they may be inconsistent with an expert's other beliefs), that users be encouraged to consider whether they are reasonable in a particular application, and that they can be easily retracted if the overall set of judgements becomes inconsistent. Ideally, inferences should be computed both with and without default assumptions so that a user can compare their effects. A logic of default assumptions that is compatible with the theory of lower previsions is outlined in [113].

Conclusion

The theory of coherent lower previsions is a general theory of reasoning in the presence of uncertainty and partial ignorance. Lower previsions are more general and more expressive than lower and upper probabilities. They have a simple behavioural interpretation as supremum buying prices for gambles and they should not be interpreted, in general, as lower bounds for an unknown Bayesian prevision.

The theory certainly satisfies criteria (a)-(d) of Section 2. Lower previsions have a clear behavioural interpretation which supports the principles of coherence and natural extension. All the rules of the theory can be derived from these principles, and they can be used to check consistency of the initial assessments and to ensure consistency of assessments with conclusions. The imprecision of lower previsions can be used to model a lack of information, conflict between several types of information or between expert opinions, or the vagueness of probability judgements in natural language. Coherent models can be produced by combining qualitative (natural-language) judgements with precise or imprecise numerical assessments. Because lower previsions have a behavioural interpretation, it is quite easy to understand the practical meaning of conclusions that are expressed in terms of them, and to use them in making decisions.

The task of assessment can be handled, in principle, by allowing the user to make whatever judgements he finds most comprehensible and natural from a wide variety of admissible judgements. In particular he can express his uncertainties in ordinary language; some ways of doing so are discussed in Section 6. Other important sources of coherent lower previsions include multivalued mappings (Section 5), partial information about precise probabilities, combination of expert opinions [103] and various models based on statistical data [99,101,102]. It seems that, in complex problems, assumptions of independence or conditional independence will be needed to reduce the effort of assessment and produce useful conclusions. Further work is needed to compare several ways of modelling independence judgements, to study how independence can be used as a default assumption in expert systems, and to determine its effect on the precision of conclusions.

The general method for making inferences and decisions in the theory is natural extension. In general, the computation of natural extension can be reduced to a linear programming problem, or (in the case of independence judgements) to a finite sequence of linear programs. Again, these problems may be intractable in moderately large expert systems when many assessments are made, and then (as in the Bayesian theory) special types of models are required. Again further work is needed to find computationally efficient methods for computing natural extensions, especially when independence judgements are involved, to develop tractable types of models (e.g. using 2-monotonicity or upper and lower densities), and to develop efficient methods for propagating lower previsions in belief networks.

5. Belief functions

The theory of belief functions was initiated by Dempster in a series of papers in the 1960s and developed by Shafer [71]. Its relevance to expert systems is discussed in [34,74]. For more recent developments see [64,75] and the ensuing discussion [19,65,76,84,92,107,112]. Smets has developed an interesting variant of the theory called "the transferable belief model" [82–85]. Applications of belief functions in expert systems include OASES [5] and its shell [4], MacEvidence [39], PSEIKI [44,47], PULCINELLA [69] and [49,58].

A *belief function* $\underline{P}$ is a real-valued function, defined on all subsets of a possibility space Ω, which can be written in the form $\underline{P}(A) = \sum_{B\subseteq A} m(B)$ for all subsets A, where m is a probability mass function on subsets of Ω, i.e. $m(\emptyset) = 0$, $m(B) \geqslant 0$ for all $B \subseteq \Omega$, and $\sum_{B\subseteq\Omega} m(B) = 1$. Here $\subseteq$ denotes set inclusion, not necessarily strict. (In Smets' theory $m(\emptyset)$ may be positive.) The conjugate upper probabilities are defined by $\overline{P}(A) = 1 - \underline{P}(A^c) = \sum_{B\cap A\neq\emptyset} m(B)$.

The mass function m is called the *probability assignment* for $\underline{P}$. It is determined by $\underline{P}$ through the Möbius inversion formula $m(B) = \sum_{A\subseteq B}(-1)^{|B-A|}\underline{P}(A)$. Any lower probability function $\underline{P}$ determines a function m through this formula, and $\underline{P}$ is a belief function if and only if m is a probability mass function. One can think of $m(B)$ as a fluid probability mass that is free to move to any element of B. Bayesian probability measures are a special type of belief function for which $m(B) = 0$ unless B is a singleton set. The vacuous lower probability is a belief function, defined by $m(\Omega) = 1$.

Interpretation

The natural extension of a belief function $\underline{P}$ to a lower prevision is defined for all gambles X by $\underline{P}(X) = \sum_{B\subseteq\Omega} m(B)\inf\{X(\omega)\colon\ \omega \in B\}$. It is easily verified that $\underline{P}$ satisfies the coherence axioms (P1)-(P3) in Section 4, and it follows that every belief function is a coherent lower probability function. So belief functions can be given the behavioural interpretation of lower probabilities: $\underline{P}(A)$ is a supremum of acceptable rates for betting on event A.

Various other interpretations of belief functions have been discussed in [36,63-65,72,75,76,85,104]; see [83] for a survey. Shafer prefers to interpret belief functions by drawing an analogy with a "canonical example" of a "randomly coded message" [72,75]. In Shafer's canonical example, the belief function is generated from an underlying precise probability measure through a multivalued mapping (defined below). The underlying probability measure does appear to have a behavioural interpretation in terms of rational betting rates, and the belief function inherits this behavioural interpretation through the multivalued mapping [99, p. 182]. So it seems to me that a behavioural interpretation of belief functions is not only compatible with Shafer's interpretation, but also a necessary consequence of his interpretation.

Indeed, I regard Shafer's semantics and the other interpretations of belief functions as possible ways of elaborating the behavioural interpretation. On [99, pp. 20 and 61], I call the behavioural interpretation "minimal" because it is compatible with a wide variety of elaborations. It requires that belief functions have certain implications for betting and other decisions, but it does not exclude other semantics which relate belief functions to the evidence on which they are based.

Shafer [75] describes the random coding example as a "metaphor" which may provide some guidance in *constructing* belief functions. The behavioural interpretation is concerned with how belief functions are *used* in making decisions. The two types of interpretation are compatible and I think that both are needed. (I prefer to call Shafer's canonical example an "assessment strategy" rather than an "interpretation" of belief functions, but that is not to deny its utility.) In practice we need to construct belief functions from evidence, but we also need to use belief functions to make decisions and this seems to require some kind of behavioural interpretation [33]. Without one, the practical meaning of inferences that are expressed in terms of belief functions is somewhat unclear. See [107] for an interesting comparison of the two interpretations.

In [72] Shafer does accept the behavioural interpretation of belief functions, although he argues that other aspects of their meaning are more important. In [75] he argues vehemently against a Bayesian sensitivity analysis interpretation of belief functions, but that argument is irrelevant to the present discussion as I also reject the sensitivity analysis interpretation. The theory of coherent lower previsions and the rules of natural extension rely only on a behavioural interpretation.

Of course I am not claiming that the whole Dempster–Shafer theory is compatible with a behavioural interpretation of belief functions. Much of the theory is based on Dempster's rule for combining belief functions, and in many problems Dempster's rule produces inferences that are unacceptable under a behavioural interpretation. Some presentations of the theory [75,76] give the impression that Dempster's rule follows naturally from the "random coding" or "multivalued mapping" semantics for belief functions, together with an inocuous assumption of unconditional independence. In fact, even under Shafer's own semantics, the justification of Dempster's rule relies on stronger assumptions of conditional independence which seem to be unreasonable in many applications. (These assumptions are discussed later in this section.) Nor do the other interpretations of belief functions provide a convincing justification for Dempster's rule. Shafer's semantics and the concept of a multivalued mapping are compatible with a behavioural interpretation, but the indiscriminate use of Dempster's rule is not.

Assessment

Belief functions are often generated by a *multivalued mapping* [15], which is a mapping A from points of an underlying space $\Psi = \{\psi_1, \dots, \psi_n\}$ to subsets of Ω. For simplicity I assume that the sets $A(\psi_1), \dots, A(\psi_n)$ are distinct. It may be possible to assess a Bayesian probability measure P on Ψ, using either frequency information or subjective judgement, and this induces a belief function on Ω through $m(A(\psi_i)) = P(\psi_i)$.

Example 2 (*An unreliable witness*). In this simplest example, an unreliable witness claims that he observed an event C. Either (ψ_1) he did observe C, so $A(\psi_1) = C$, or (ψ_2) he observed nothing, so $A(\psi_2) = \Omega$. Suppose that, after hearing his report, we judge the witness to have credibility α, so $P(\psi_1) = \alpha$ and $P(\psi_2) = 1-\alpha$. This generates the probability assignment $m(C) = \alpha$ and $m(\Omega) = 1 - \alpha$, and the corresponding belief function has $\underline{P}(C) = \alpha$ and $\overline{P}(C) = 1$. Thus there is some evidence in favour of C (we would bet on C at any odds better than $1-\alpha$ to α), but no evidence against C (we are not prepared to bet against C at any odds).

The imprecision of this belief function simply reflects the absence of information about the "base rate" frequency of C, $P(C \mid \psi_2)$. A Bayesian would need to make a precise assessment of $P(C \mid \psi_2)$ and then compute $P(C) = \alpha + (1-\alpha)P(C \mid \psi_2)$. Depending on the practical context, there may or may not be information on which to assess $P(C \mid \psi_2)$. In general, we may only be able to assess lower probabilities $\underline{P}(C \mid \psi_2)$ and $\underline{P}(\psi_1) = \underline{\alpha}$, and then we obtain $\underline{P}(C) = \underline{\alpha} + (1-\underline{\alpha})\underline{P}(C \mid \psi_2)$ and $\overline{P}(C) = 1$ by natural extension. Note that Bayesians require two precise assessments, of α and $P(C \mid \psi_2)$, the multivalued mapping requires one precise assessment (α), but we can reach useful conclusions without making *any* precise assessments.

Belief functions can often be assessed by using the model of a multivalued mapping and assessing underlying Bayesian probabilities. (Although it is somewhat ironic that precise assessments should be needed in a theory of imprecise probabilities! A more general model, which does not assume that precise probabilities can be specified on the underlying space Ψ, is given in [99, Section 4.3.5].) In other problems we may be able to assess the probability assignment m more directly. For example there may be frequency information about the outcomes $B_1, \ldots, B_n$ of previous trials, where the outcome is recorded as a subset B_i of Ω rather than a single element. We might then take the past relative frequency of B as our assessment of $m(B)$ [19,104]. Some other assessment strategies have been suggested in [71,73,75].

Example 3 (*Football example*). It is important to recognise, however, that there are many types of information which cannot be modelled by belief functions. One simple example is the football example of Section 4, where three judgements were expressed in ordinary language. Let $\underline{P}$ be the lower probability function constructed by natural extension of the three judgements, which is the lower envelope of the three probability mass functions $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ on $\Omega = \{W, D, L\}$. By the Möbius inversion formula,

$$\begin{aligned} m(\Omega) &= \underline{P}(\Omega) - \underline{P}(W \cup D) - \underline{P}(W \cup L) - \underline{P}(D \cup L) + \underline{P}(W) + \underline{P}(D) + \underline{P}(L) \\ &= 1 - \frac{2}{3} - \frac{1}{2} - \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + 0 = -\frac{1}{12}. \end{aligned}$$

As $m(\Omega)$ is negative, m cannot be a probability mass function so $\underline{P}$ is not a belief function.

It appears that probability judgements in natural language cannot be modelled, in general, by belief functions. Note also that, even when such judgements do yield a belief function, this may be less informative than the appropriate lower prevision. (As explained in Section 4, lower probabilities are less expressive than lower previsions.) Even when a belief function $\underline{P}$ is generated by a multivalued mapping, we may be able to make further assessments to sharpen $\underline{P}$, and the resulting lower probability may not be a belief function. There are other examples in [48,64,96,99] of models and assessment strategies which produce coherent lower probabilities that are not belief functions. (Consider, for example, two tosses of a fair coin with unknown correlation between the outcomes.) There seems to be no good reason to restrict attention to belief functions, rather than coherent lower probabilities or coherent lower previsions. Nor is it clear why *unconditional* belief functions are taken to be the fundamental measures of uncertainty. Direct assessments of *conditional* lower and upper probabilities are often needed to measure the uncertainty associated with the "if-then" rules that are prevalent in expert systems [64].

Dempster's rule of combination

Dempster's rule is extensively used in the theory to combine and update belief functions. Let m_1 and m_2 be probability assignments based on separate ("independent") bodies of evidence. A combined probability assignment m is defined by

$$m(C) = \rho^{-1} \sum_{A \cap B = C} m_1(A) m_2(B) \quad \text{for all non-empty sets } C, \tag{9}$$

where

$$\rho = \sum_{A \cap B \neq \emptyset} m_1(A) m_2(B)$$

is a normalizing constant, provided ρ is nonzero. It appears to be computationally feasible to use Dempster's rule to combine belief functions that are defined on certain kinds of tree structures [77,78,114]. Other computational results are discussed in [34,61,75,111,112].

Example 4 (*Two unreliable witnesses*). As a simple example of Dempster's rule, suppose that there are two unreliable witnesses of the type considered earlier. The first witness, who has credibility α_1, reports that he observed event C_1. The second witness, with credibility α_2, reports C_2. So m_1 and m_2 are defined by $m_1(C_1) = \alpha_1$, $m_1(\Omega) = 1 - \alpha_1$, $m_2(C_2) = \alpha_2$, $m_2(\Omega) = 1 - \alpha_2$. Suppose that C_1 and C_2 are logically independent events. Then $\rho = 1$ in Dempster's rule and the combined probability assignment [71] is $m(C_1 \cap C_2) = \alpha_1\alpha_2$, $m(C_1) = \alpha_1(1 - \alpha_2)$, $m(C_2) = (1 - \alpha_1)\alpha_2$, $m(\Omega) = (1 - \alpha_1)(1 - \alpha_2)$.

When should beliefs be combined by Dempster's rule? The rule appears to give reasonable answers in some problems, but it produces unsatisfactory and seemingly erroneous conclusions in other problems (an example is given below). Its applicability seems to rely, in general, on some rather delicate judgements of conditional independence. In the example, the rule is applicable provided that beliefs, based on the report of one witness, about whether her report is correct would be unchanged by further information which specified both the report of the other witness and whether his report was correct.

This is a type of *conditional independence*, conditional on the reports of both witnesses. It is quite different from the type of *unconditional independence* that is suggested by Shafer [72,75,76] as a justification for Dempster's rule. (A similar distinction is made in [95].) In the example, each witness sometimes reports correctly and sometimes incorrectly, and the two witnesses may be unconditionally independent in the sense that learning whether one witness reported correctly (without knowing what the report was) would not change our belief that the other witness reported correctly.

Unconditional independence is a simple judgement and often a reasonable one, for instance when the witnesses are known to have no interaction. But it is not sufficient to justify the use of Dempster's rule, because the belief functions involved in Dempster's rule are based on (or "conditional on") the specific reports of the two witnesses. The credibility of witness i (α_i) must be a posterior probability, conditional on the report of witness i, in order to determine beliefs based on all the available evidence. (Shafer [72, p. 5], equates this posterior probability with the prior probability that witness i will report correctly, but Levi [55] has shown that this is unjustified.)

Dempster's rule therefore relies on an assumption of conditional independence and this is harder to justify. If the two witnesses gave detailed reports which agreed in most details and were otherwise compatible, for example, then learning that one report is correct would certainly give us more confidence that the other report is correct and conditional independence would fail. It would fail also if the two reports were inconsistent, even though unconditional independence might be reasonable in these two cases.

I am not suggesting that Dempster's rule of combination is always inappropriate in such examples, merely that the implicit judgements of conditional independence should be made explicit and carefully considered. But conditional independence seems less reasonable in most other, more complex examples than in the case of the two witnesses, because learning one body of evidence and which mechanism or "code" produced it will typically provide information about the true state ω and hence change beliefs about which mechanism produced the other body of evidence. That is especially clear in cases of *conditioning*, where one body of evidence actually restricts ω to a subset of the initial possibility space Ω.

Dempster's rule is therefore unreliable unless careful attention is given to the implicit assumptions of conditional independence. For a general mathematical formulation of these assumptions and for further discussion, see [99, Section 5.13]. The conditions given in [99] are similar to, but slightly different from, those given by Voorbraak [95, p. 188]. Voorbraak reaches a similar conclusion about the limited applicability of Dempster's rule.

Example 5 (*Independent witnesses*). Even in the simple example of the two witnesses, there are other ways of modelling independence. One way is to make the assessments $\underline{P}(C_1) = \alpha_1$, $\underline{P}(C_2) = \alpha_2$, and judge that events C_1 and C_2 are independent (conditional on the testimony of both witnesses), meaning that upper and lower probabilities for one event would not change if we learned whether or not the other event occurred. Coherent lower previsions can then be computed by natural extension of the judgements $\underline{P}(C_1) = \underline{P}(C_1 \mid C_2) = \underline{P}(C_1 \mid C_2^c) = \alpha_1$ and $\underline{P}(C_2) = \underline{P}(C_2 \mid C_1) = \underline{P}(C_2 \mid C_1^c) = \alpha_2$. Natural extension produces a lower probability model that is not a belief function (it is not even 2-monotone), and which is more precise than the belief function produced by Dempster's rule. For example, let A denote the event that C_1 and C_2 have the same truth value (i.e. both occur or neither occurs). Then Dempster's rule gives $\underline{P}_D(A) = \alpha_1\alpha_2$, whereas natural extension gives $\underline{P}_E(A) = \alpha_1\alpha_2/(\alpha_1 + \alpha_2 - \alpha_1\alpha_2)$ which is strictly larger than $\underline{P}_D(A)$ provided $0 < \alpha_1 < 1$ and $0 < \alpha_2 < 1$. (Both rules give $\overline{P}(A) = 1$.) This is one type of application in which natural extension produces inferences that are more precise than Dempster's rule. (Compare with conditioning.)

A third approach, suggested by Bayesian sensitivity analysis, is to look for all Bayesian probability measures P which satisfy $P(C_1) \geqslant \alpha_1$ and $P(C_2) \geqslant \alpha_2$, and under which C_1 and C_2 are independent [105]. By forming the lower envelope of all such measures, we obtain a lower probability model that is more precise than the two previous models and again is not a belief function. For this model

$$\underline{P}_B(A) = \min\{\alpha_1, \alpha_2, \alpha_1\alpha_2 + (1-\alpha_1)(1-\alpha_2)\}$$

and $\underline{P}_B(A) > \underline{P}_E(A) > \underline{P}_D(A)$, provided $0 < \alpha_1 < 1$ and $0 < \alpha_2 < 1$. If each witness has credibility $\frac{1}{2}$, for example, we obtain $\underline{P}_B(A) = \frac{1}{2}$, $\underline{P}_E(A) = \frac{1}{3}$ and $\underline{P}_D(A) = \frac{1}{4}$.

The three models do agree on some probabilities, in particular $\underline{P}(C_1 \cap C_2) = \alpha_1\alpha_2$ and $\overline{P}(C_1 \cap C_2) = 1$ for each model. (A fourth model, based on possibility theory, will be given in Section 6. This gives $\underline{P}(C_1 \cap C_2) = \min\{\alpha_1, \alpha_2\}$, which seems less reasonable than the other solutions.) The differences between the four models may be important because judgements of independence are common in expert systems. It is not clear which model is most appropriate for practical applications, but see [99, Chapter 9;105] for some comparisons.

Dempster's rule of conditioning

Suppose that the second probability assignment in Dempster's rule is defined by $m_2(B) = 1$, representing knowledge that event B has occurred. Dempster's rule of combination then reduces to Dempster's rule of conditioning, which can be written most simply in terms of upper probabilities as $\overline{P}_D(A \mid B) = \overline{P}(A \cap B)/\overline{P}(B)$, defined whenever $\overline{P}(B) > 0$, with $\underline{P}_D(A \mid B) = 1 - \overline{P}_D(A^c \mid B)$. This rule is used in the theory of belief functions to update beliefs after receiving new information.

Compare this rule with the rule of natural extension given in Section 4. Because every belief function $\underline{P}$ is 2-monotone, the conditional probabilities defined by natural extension are given by the simple formulas

$$\underline{P}_E(A \mid B) = \frac{\underline{P}(A \cap B)}{\underline{P}(A \cap B) + \overline{P}(A^c \cap B)},$$

$$\overline{P}_E(A \mid B) = \frac{\overline{P}(A \cap B)}{\overline{P}(A \cap B) + \underline{P}(A^c \cap B)}, \tag{10}$$

provided the denominators are nonzero. Several authors [23,41,93] have shown that if $\underline{P}$ is a belief function then so is $\underline{P}_E(\cdot \mid B)$. (This is true whenever $\overline{P}(B) > 0$; see [41].) Thus conditioning a belief function by natural extension produces another belief function.

The conditional probabilities defined by Dempster's rule are always at least as precise as those defined by natural extension [15], in the sense that

$$\underline{P}_E(A \mid B) \leqslant \underline{P}_D(A \mid B) \leqslant \overline{P}_D(A \mid B) \leqslant \overline{P}_E(A \mid B). \tag{11}$$

Both rules yield the same conditional probabilities when the initial probability measure is precise (and then both agree with Bayes' rule), but, in other cases, Dempster's rule typically yields conditional probabilities that are more precise than natural extension. (See [19,23,48] for interesting comparisons of the two rules.) Natural extension produces conditional upper and lower probabilities that are coherent with the unconditional upper and lower probabilities. In some cases the conditional probabilities defined by Dempster's rule may also be coherent with the unconditional probabilities, even when Dempster's rule and natural extension produce different answers. For instance, Dempster's rule seems to produce reasonable inferences when it is used to update possibility measures (see the examples in Section 6). But there are examples, such as the following, in which Dempster's rule produces inferences that seem to be seriously wrong.

Example 6 (*ε-contamination model*). To compare the two rules of conditioning, consider the following statistical model. A random variable X takes values in the sample space $\mathcal{X} = \{1,2,3,\ldots,N\}$. To be specific we will take $N = 10^4$. (In fact Dempster's rule will produce the same type of inconsistency for any value of N greater than 1, but the degree of inconsistency increases with N.) With probability 0.99, X is generated by the uniform probability distribution on $\mathcal{X}$. With probability 0.01, X is generated by another, completely unknown, probability distribution on $\mathcal{X}$. Thus almost all observations follow a uniform distribution but one observation in a hundred is expected to be a "gross error", in the sense that it is generated by a completely unknown (but possibly drastically different) mechanism. This sampling model is called an *ε-contamination neighbourhood* of the uniform distribution (here $\varepsilon = 0.01$) [40].

Let B_x denote the event that $X = x$, let A denote the event that X is generated by the uniform distribution, and define $y = 1$ if A occurs and $y = 0$ otherwise. There is uncertainty about both the value of X and whether A occurs, so we take the possibility space to be $\Omega = \{(x,y){:}\ x = 1,2,\ldots,N;\ y = 0,1\}$.

The uncertainty about Ω can be modelled by a belief function whose probability assignment m is defined by $m(A \cap B_x) = 0.99/N$ for $x = 1,2,\ldots,N$, $m(A^c) = 0.01$, and $m(C) = 0$ for all other sets C. This model is imprecise because we do not know how to distribute the probability mass 0.01 amongst the alternatives to a uniform distribution.

Before observing X, the probability of A is precisely 0.99. Now suppose we make a single observation $X = x$. How should we update our uncertainty about A?

Dempster's rule of conditioning produces

$$\overline{P}_D(A \mid B_x) = \frac{\overline{P}(A \cap B_x)}{\overline{P}(B_x)} = \frac{m(A \cap B_x)}{m(A \cap B_x) + m(A^c)} = \frac{0.99/N}{0.99/N + 0.01}$$
$$= \frac{99}{99+N} < 0.01 \quad \text{when } N = 10^4.$$

Similarly $\overline{P}_D(A^c \mid B_x) = N/(99+N)$, hence

$$\underline{P}_D(A \mid B_x) = 1 - \overline{P}_D(A^c \mid B_x) = \overline{P}_D(A \mid B_x).$$

Thus $\overline{P}_D(A \mid B_x) = \underline{P}_D(A \mid B_x) < 0.01$ for every possible value of x. After observing x, the updated probability of A is precise and smaller than 0.01, whatever the value of x.

Initially we are very confident that X will be generated by the uniform distribution. But if we use Dempster's rule to update our beliefs then we will become very confident, whatever value of X we observe, that X was *not* generated by the uniform distribution!

Intuitively there is a strong inconsistency between the initial and updated probabilities. Indeed an observer who knew that Dempster's rule would be used to update probabilities could exploit the inconsistency to make a sure gain, by initially betting against event A and, after x is observed, betting on A at a more favourable rate. The initial and updated probabilities violate the coherence axioms (C1) and (C2) of Section 4 and they "incur sure loss" in the mathematical sense of [99].

In simple terms, Dempster's rule produces the wrong answer because it treats the probability mass $m(A^c) = 0.01$, which is spread over all possible values of x, as if it were entirely focused on the x that is actually observed. Indeed Dempster's rule produces the inferences that a Bayesian would obtain from the precise probability assessments $P(A \cap B_x) = 0.99/N$ and $P(A^c \cap B_x) = 0.01$, which are incoherent when asserted for every possible x. I mentioned earlier that Dempster's rule is applicable when a particular type of conditional independence holds. In this problem the required condition is that, for each x, learning that $X = x$ would not change the relative likelihood of $A \cap B_x$ and A^c. Of course this condition is unreasonable.

There appear to be many problems in which Dempster's rule of conditioning or combination produces incoherent or unacceptable inferences; other examples can be found in [99, Section 5.13] and in [36,48,64,65,98].

Compare Dempster's rule of conditioning with the method of natural extension, which always produces coherent inferences. The conditional upper and lower probabilities generated by natural extension can be easily computed as follows:

$$\overline{P}_E(A \mid B_x) = \frac{\overline{P}(A \cap B_x)}{\overline{P}(A \cap B_x) + \underline{P}(A^c \cap B_x)} = \frac{0.99/N}{0.99/N + 0} = 1,$$

$$\underline{P}_E(A \mid B_x) = \frac{\underline{P}(A \cap B_x)}{\underline{P}(A \cap B_x) + \overline{P}(A^c \cap B_x)} = \frac{0.99/N}{0.99/N + 0.01} = \frac{99}{99 + N} < 0.01.$$

In this problem, these are the *unique* updated upper and lower probabilities that are coherent with the initial belief function.

The two rules of conditioning produce the same updated lower probability for A, but they produce very different upper probabilities. The updated probabilities given by Dempster's rule are precise, whereas those given by natural extension are highly imprecise. The reason is that, for each possible i, the initial model is consistent with the hypothesis H_i that, whenever A^c occurs, X takes the value i for certain. If we knew H_x to be true we would obtain precise probability $P(A \mid B_x) = 99/(99 + N)$, but if we knew H_i to be true (where $i \neq x$) we would obtain $P(A \mid B_x) = 1$. The range of lower to upper updated probabilities must cover both values. The initial probability of A is precisely 0.99 under all the hypotheses H_i, but the updated probabilities are very different under different hypotheses and this produces imprecision in the updated upper and lower probabilities.

Despite this explanation, some readers may feel that observation of x should have absolutely no effect on uncertainty about A and that the updated probabilities should be $P(A \mid B_x) = 0.99$ (precisely) for every possible x to agree with the initial probability $P(A) = 0.99$. However these updated probabilities are not coherent with the initial belief function. If you take the updated probabilities to be precisely 0.99 then you must modify the initial probability model to make it precise. Indeed, we can use natural extension to compute the unconditional probabilities that are generated by the assessments $P(A \mid B_x) = 0.99$ and $P(A \cap B_x) = 0.99/N$ for all x, and we obtain $P(A^c \cap B_x) = 0.01/N$ (precisely) for all x. This is tantamount to assuming that the probability mass $m(A^c) = 0.01$ can be distributed uniformly over $\mathcal{X}$, i.e. that X is generated by a uniform distribution when A^c occurs, as well as when A occurs! If there is really no information about how X is generated when A^c occurs then the initial probability model should be imprecise, and this inevitably leads to imprecision in the updated probabilities.

Expectations

It is assumed in the Dempster-Shafer theory that uncertainties are modelled in terms of upper and lower probabilities. As discussed in Section 4, upper and lower probabilities may not be sufficient. In many problems upper and lower previsions or conditional probabilities are needed and information is lost if these are defined in terms of unconditional probabilities.

Assuming that beliefs are specified in terms of a belief function $\underline{P}$ with probability assignment m, lower and upper previsions of a gamble X can be computed by natural extension, through the formulas

$$\underline{P}(X) = \int_{-\infty}^{\infty} x \, \mathrm{d}\overline{F}_X(x) = \sum_{B \subseteq \Omega} m(B) \inf\{X(\omega) : \omega \in B\},$$

$$\overline{P}(X) = \int_{-\infty}^{\infty} x \, \mathrm{d}\underline{F}_X(x) = \sum_{B \subseteq \Omega} m(B) \sup\{X(\omega) : \omega \in B\}, \tag{12}$$

where $\overline{F}_X$ and $\underline{F}_X$ are the upper and lower distribution functions of X under $\underline{P}$ (see Eq. (8)). Preferences between actions can be constructed by computing upper and lower previsions of differences between utility functions, as outlined in Section 4. Other ways of using belief functions in decision making are described in [85,92].

Conclusion

Belief functions are a special type of coherent lower probability. They can model various types of partial ignorance and limited or conflicting evidence. Because they can be represented in terms of a probability mass function m, belief functions appear to be mathematically and computationally simpler, in some ways, than the general class of coherent lower previsions. Computationally efficient methods have been developed for combining and propagating belief functions.

Belief functions can be assessed through multivalued mappings or in other ways, and the assessment strategies suggested by the theory are useful in many applications. (Shafer's emphasis on constructing belief functions from simple evaluations of evidence is especially valuable.) However, many other assessment strategies produce coherent lower probabilities that are not belief functions. Some important types of uncertainty, e.g. judgements of probability in ordinary language, cannot be adequately modelled by belief functions. In fact belief functions are much less expressive than coherent lower previsions. The theory gives no justification for restricting attention to belief functions rather than coherent lower probabilities, or to lower probabilities rather than lower previsions.

The calculus of belief functions relies heavily on Dempster's rule of combination. Dempster's rule may be useful in some problems where it is supported by explicit judgements of conditional independence but it is unreliable in general and it can produce inferences that are intuitively inconsistent and formally incoherent. More attention needs to be given to the exact conditions under which Dempster's rule is applicable.

Apart from the fundamental role it gives to Dempster's rule, the Dempster-Shafer theory appears to be broadly compatible with the theory of coherent lower previsions. Belief functions that are constructed through a multivalued mapping have a behavioural interpretation and can be regarded as coherent lower probabilities, and they could be combined and updated through the rules of natural extension rather than by Dempster's rule.

6. Possibility measures

The theory of fuzzy sets was introduced by Zadeh [116] for the purpose of modelling the imprecision and ambiguity of ordinary language. Uncertainty is often measured in this theory by *possibility measures*, which are described in [16,17,46,115,118]. Concerning the use of possibility measures and fuzzy logic in expert systems, see [17,120,121]. Specific applications to expert systems include DIABETO [7], CADIAG-2 [1], SPHINX [26], TAIGER [24], SPII-2 [51], PULCINELLA [53,69].

Zadeh [121] argues that "classical probability is insufficiently expressive to cope with the multiplicity of kinds of uncertainty which one encounters in AI and, more particularly, in expert systems". In particular, Bayesian probabilities are inadequate for dealing with vague (inexactly defined) events or predicates, such as "young" or "tall", or with natural-language expressions of probability, such as "probably". These two types of vagueness occur together in the expression "it is likely that Mary is young". Many of the production rules in expert systems such as MYCIN are fuzzy in these ways. (See [22,120] for many examples.) If expert systems are to be widely used, it does seem necessary to provide mathematical models for vague expressions in natural language.

Definitions

Consider the expression "Mary is young". According to Zadeh, this provides some information about Mary's precise age X, and the information can be modelled by a *possibility distribution function* π_X, defined on the set Ω of possible ages. The number $\pi_X(\omega)$ lies between zero and one and is read as "the degree to which it is possible that Mary has precise age ω, given that she is young". It is assumed that $\sup\{\pi_X(\omega):\ \omega \in \Omega\} = 1$, that is, the function π_X is normalised to have supremum value one. (Zadeh [118] did not require this normalisation but it has been assumed by most later authors.) Zadeh identifies the function π_X with the membership function for the fuzzy set "young" on Ω. Thus a possibility distribution function is a type of membership function, and this enables the basic theory of fuzzy sets to be carried over to possibility distributions.

A *possibility measure* π is defined for all subsets A of Ω by $\pi(A) = \sup\{\pi_X(\omega):\ \omega \in A\}$, with $\pi(\emptyset) = 0$. It follows that π has the properties (for all subsets A and B): $0 \leqslant$

$\pi(A) \leqslant 1$, $\pi(\Omega) = 1$, $\max\{\pi(A), \pi(A^c)\} = 1$, and $\pi(A \cup B) = \max\{\pi(A), \pi(B)\}$. In general $\pi(A \cap B) \leqslant \min\{\pi(A), \pi(B)\}$ but there is equality in the important special case where A and B are "non-interactive" events [117]. Non-interaction appears to correspond to independence in probability theory. A possibility distribution function is analogous to a probability mass function or density function, and a possibility measure is analogous to a probability measure.

Interpretation

Thus Zadeh translates natural-language expressions into the mathematical formalism of possibility measures. But how should we interpret these measures? Zadeh clearly wishes to distinguish degrees of possibility from degrees of probability: "A basic assumption which underlies our approach to approximate reasoning is that the imprecision which is intrinsic in natural languages is, in the main, possibilistic rather than probabilistic in nature. ... possibility relates to our perception of the degree of feasibility or ease of attainment, whereas probability is associated with the degree of belief, likelihood, frequency, or proportion." [119]

It is widely recognised that possibility is distinct from probability and this distinction is central to all theories of probability, e.g. [28]. Degrees of probability can be interpreted as betting rates or as relative frequencies, but it is less clear that it is meaningful to talk of "degrees of possibility". Zadeh [118] refers to expressions such as "slightly possible" and "quite possible" but admits that these usually indicate degrees of probability rather than degrees of possibility.

Zadeh's explication of degrees of possibility [118,119], and the example he gives to illustrate it, takes them to be measures of how *easy* or *feasible* it is to perform an action. This is somewhat different from the way that degrees of possibility are used in fuzzy logic, where they seem to measure how *plausible* it is that a proposition is true. Given the information "Mary is young", for example, $\pi_X(20)$ might measure how plausible it is that Mary is aged 20, but it is difficult to understand it as a measure of "how easy" or "how feasible" it is for Mary to be aged 20. The general usage is consistent with the following interpretation of degrees of possibility as upper probabilities, which are sometimes called "degrees of plausibility". Various other interpretations of degrees of possibility have been suggested [16,17], but I do not know of any that justifies the usual min-max rules of combination.

It is essential to have a clear interpretation of "degree of possibility" for the reasons mentioned earlier. If we do not understand the meaning of the numbers $\pi_X(\omega)$ then it will be difficult to assess them in practical problems, to understand conclusions that are expressed in terms of them, and to justify the calculus that is used to manipulate them.

Interpretation as upper probability

In fact, as suggested by Giles [32,33], possibility measures can be given a behavioural interpretation as *upper probabilities*. That is, we regard a possibility measure π as an upper probability measure, $\overline{P}(A) = \pi(A) = \sup\{\pi_X(\omega)\colon\ \omega \in A\}$, and interpret $\pi(A)$ as a marginal acceptable betting rate for betting *against* A. The corresponding lower probabilities are defined by $\underline{P}(A) = 1 - \sup\{\pi_X(\omega)\colon \omega \in A^c\} = 1 - \pi(A^c)$. The lower probabilities are sometimes called *necessity measures* [17,46]. They have the properties (for all subsets A, B of Ω): $\overline{P}(A \cup B) = \max\{\overline{P}(A), \overline{P}(B)\}$, $\underline{P}(A \cap B) = \min\{\underline{P}(A), \underline{P}(B)\}$, $\underline{P}(A) \leqslant \overline{P}(A)$, $\min\{\underline{P}(A), \underline{P}(A^c)\} = 0$, and either $\underline{P}(A) = 0$ or $\overline{P}(A) = 1$.

Suppose, for example, that A is the event that Mary's age is at least 30 years. Given the information "Mary is young", we might assess $\pi(A) = 0.4$ and $\pi(A^c) = 1$, so that $\overline{P}(A) = 0.4$ and $\underline{P}(A) = 0$. The behavioural interpretation is that we would bet *against* A at odds of 3 to 2 against, but we would not bet *on* A at any finite odds. (As noted in Section 4, a more general behavioural interpretation could be given in terms of the implications of possibility measures for decision making. Decisions would be made, as in Section 4, by computing upper and lower previsions of differences between utility functions, using Eqs. (13) and (14) below.)

All possibility measures are *coherent* upper probabilities. One way to see that is to construct their natural extensions to lower and upper previsions. If the possibility measure has possibility distribution π_X, its natural extensions are (for any gamble Z defined on Ω)

$$\underline{P}(Z) = \int_0^1 \inf\{Z(\omega)\colon \pi_X(\omega) \geqslant u\}\,\mathrm{d}u = \sup Z - \int_{\inf Z}^{\sup Z} \sup\{\pi_X(\omega)\colon Z(\omega) \leqslant y\}\,\mathrm{d}y \tag{13}$$

$$\overline{P}(Z) = \int_0^1 \sup\{Z(\omega)\colon \pi_X(\omega) \geqslant u\}\,\mathrm{d}u = \inf Z + \int_{\inf Z}^{\sup Z} \sup\{\pi_X(\omega)\colon Z(\omega) > y\}\,\mathrm{d}y \tag{14}$$

where $\inf Z$ and $\sup Z$ denote the infimum and supremum values of $Z(\omega)$ over all $\omega \in \Omega$. The second versions of each formula follow from Eqs. (8). See [102, Lemma 1] for a derivation of these formulas.

From the first expression for $\underline{P}(Z)$, it is easy to verify that the lower prevision $\underline{P}$ satisfies axioms (P1)-(P3) of Section 4 and is therefore coherent. Then verify that its restrictions to upper and lower probabilities are defined by $\overline{P}(A) = \sup\{\pi_X(\omega)\colon \omega \in A\}$ and $\underline{P}(A) = 1 - \sup\{\pi_X(\omega)\colon \omega \in A^c\}$, hence these upper and lower probabilities are coherent. (Alternatively, verify that $\underline{P}$ is 2-monotone and use the result that all 2-monotone lower probabilities are coherent.)

There are great advantages in adopting this interpretation of possibility measures, especially as the theory of coherent upper and lower previsions could then be used to guide the assessment of possibility distributions and to derive rules for combining them. But note that possibility measures are a very special type of coherent upper probability. Their characteristic property $\overline{P}(A \cup B) = \max\{\overline{P}(A), \overline{P}(B)\}$ is not necessary for coherence, e.g. if A and B are logically independent events and you assess $\overline{P}(A) = \overline{P}(B) = \frac{1}{2}$, the natural extension is $\overline{P}(A \cup B) = 1$, and any assessment of $\overline{P}(A \cup B)$ between $\frac{1}{2}$ and 1 would be coherent with the initial assessments, whereas $\overline{P}(A \cup B) = \frac{1}{2}$ is necessary for $\overline{P}$ to be a possibility measure.

For finite Ω, a possibility measure π is just a *consonant plausibility function* in the sense of Shafer [71], and the corresponding lower probability $\underline{P}(A) = 1 - \pi(A^c)$ is a special type of belief function, characterised by the property that the sets B for which $m(B) > 0$ form a nested sequence [46]. Indeed it seems more appropriate to call $\pi(A)$ a *degree of plausibility* rather than a degree of possibility. Event A is more or less plausible according to the amount of evidence pointing *against* A, e.g. if there is no evidence against A then A is fully plausible, $\pi(A) = 1$, and there is no reason to bet against A at any odds. We could also interpret $\underline{P}(A) = 1 - \pi(A^c)$ as a "degree of potential surprise" that we would experience if A failed to occur [54,70], or as a "degree of provability" of A [12]. The uncertainty measures proposed by Shackle [70] and Cohen [12] appear to be mathematically equivalent to possibility measures and thus they are special types of coherent lower probabilities.

Imprecision

Possibility measures can be used to model some types of imprecise or partial information. For example, complete ignorance about a variable ω can be modelled by the possibility distribution $\pi_X(\omega) = 1$ for all ω in Ω, which corresponds to the *vacuous* upper and lower probabilities. The degree of imprecision concerning an event A can be measured in general by $\overline{P}(A) - \underline{P}(A) = \pi(A) + \pi(A^c) - 1 = \min\{\pi(A), \pi(A^c)\}$. Some natural-language judgements, such as "Mary is young" and the variants discussed in the following subsections, can be modelled by possibility measures.

However, first-order possibility measures do not seem to be sufficiently flexible to model many common types of uncertainty. The football example and other examples involving natural-language judgements of uncertainty cannot be adequately modelled by belief functions, and certainly not by first-order possibility measures which correspond to a special type of belief function. (Second-order possibility measures, considered later, can be used to model natural-language judgements of uncertainty, but they are considerably more complicated than first-order measures.) Most examples of multivalued mappings and belief functions, such as the ε-contamination model in Section 5, involve coherent upper probabilities that are not possibility measures. Nor can (first-order) possibility measures model precise probability assessments. Bayesian probability measures are a special type of belief function or coherent lower probability, but not a special type of possibility measure. The upper and lower probabilities defined by a non-degenerate possibility distribution are always imprecise and usually very imprecise, i.e. $\overline{P}(A) - \underline{P}(A)$ is large for many events A.

Example 7 (*Modelling vague predicates*). Zadeh [121] argues that Bayesian probabilities are unable to model judgements like "Mary is young" or "it is likely that Mary is young". He shows how these judgements can be modelled by possibility distributions, although he does not give numerical assessments of the required distributions. I want to indicate how these judgements can be modelled using the theory of coherent lower previsions and the extent to which this is consistent with fuzzy reasoning. Other models for the same judgements are proposed in [11,18].

Consider the proposition "Mary is young". This provides some (but not much) information about Mary's age. We could model our uncertainty about Mary's age, given the information that she is young, by a coherent lower prevision. There are many possible ways of constructing a lower prevision.

One way, which seems suitable for relating terms such as "young", "old", "tall" and "rich" to an appropriate numerical scale, is through assessments of *upper* and *lower distribution functions* [99].

Let X denote Mary's precise age. The upper and lower distribution functions of X are defined by $\overline{F}(\omega) = \overline{P}(X \leqslant \omega)$ and $\underline{F}(\omega) = \underline{P}(X \leqslant \omega)$ for all $\omega > 0$. Given only that Mary is young, it is entirely plausible that $X \leqslant \omega$ for any specific ω, so it is natural to take $\overline{F}(\omega) = 1$ for all ω. (This is a vacuous assessment, so $\overline{F}$ could be ignored altogether.) But "Mary is young" does provide some evidence that she is under 30, and strong evidence that she is under 40. On this basis, one might assess $\underline{F}(\omega) = 0$ for $0 < \omega \leqslant 15$, $\underline{F}(\omega) = (\omega - 15)/25$ for $15 < \omega < 40$, $\underline{F}(\omega) = 1$ for $w \geqslant 40$. (Alternatively, one might make a few qualitative judgements such as "probably $X \leqslant 25$" and "very probably $X \leqslant 30$" which can be translated into constraints on $\underline{F}$. Note that the information provided by the assertion "Mary is young" is highly sensitive to the context in which it is made—is she "young to be walking" or "young to be retiring"?—and it is debatable whether a useful model can be given that is independent of context. We must assume, at least, that Mary is human!)

A coherent lower prevision is then constructed by natural extension of these judgements. For any set A of possible ages (positive real numbers), we obtain upper and lower probabilities $\overline{P}(A) = \sup\{1 - \underline{F}(\omega)\colon\ \omega \in A\}$, $\underline{P}(A) = \inf\{\underline{F}(\omega)\colon\ w \in A^c\}$. This probability model is highly imprecise, as one would expect.

Here the upper probability $\overline{P}$ is actually a possibility measure, with possibility distribution function $\pi_X(\omega) = 1 - \underline{F}(\omega) = \overline{P}(X > \omega)$, so $\pi_X(\omega) = 1$ if $0 < \omega \leqslant 15$, $\pi_X(\omega) = (40 - \omega)/25$ if $15 < \omega < 40$, and $\pi_X(\omega) = 0$ if $\omega \geqslant 40$. In this case $\pi_X(\omega)$, the degree to which it is possible that Mary has precise age ω, can be identified with the upper probability that Mary's age exceeds ω, and the analysis based on natural extension of the lower distribution function $\underline{F}$ agrees with the analysis based on the possibility distribution π_X; both generate the same possibility measure $\overline{P}$. In general, when both the upper and lower distribution functions are non-vacuous, the upper probability produced by natural extension will not be a possibility measure.

Conditional possibilities

Suppose that unconditional upper probabilities are defined through a possibility distribution π by $\overline{P}(A) = \sup\{\pi(\omega)\colon\ \omega \in A\}$ and we wish to construct conditional upper probabilities $\overline{P}(\cdot \mid B)$. Assume that $\pi(\omega) > 0$ for some ω in B, so that $\overline{P}(B) > 0$. (Otherwise there is no useful information in π from which to construct $\overline{P}(\cdot \mid B)$.) The conditional probabilities can be constructed by natural extension, using Eq. (5).

Because the corresponding lower probabilities are 2-monotone,

$$\begin{aligned}\overline{P}(A \mid B) &= \frac{\overline{P}(A \cap B)}{\overline{P}(A \cap B) + \underline{P}(A^c \cap B)} \\ &= \frac{\sup\{\pi(\omega):\ \omega \in A \cap B\}}{\sup\{\pi(\omega):\ \omega \in A \cap B\} + 1 - \sup\{\pi(\omega):\ \omega \in A \cup B^c\}} \\ &= \sup\{\pi(\omega \mid B):\ \omega \in A\} \end{aligned} \tag{15}$$

where the *conditional possibility distribution* $\pi(\cdot \mid B)$ is defined by

$$\pi(\omega \mid B) = \begin{cases} \dfrac{\pi(\omega)}{\pi(\omega) + 1 - \max\{\pi(\omega), \beta\}}, & \text{if } \omega \in B \text{ and } \pi(\omega) > 0, \\ 0, & \text{if } \omega \in B^c \text{ or } \pi(\omega) = 0, \end{cases} \tag{16}$$

and $\beta = \overline{P}(B^c)$.

It can be verified that $\sup\{\pi(\omega \mid B):\ \omega \in \Omega\} = 1$ so that $\pi(\cdot \mid B)$ is a possibility distribution. This shows that if the unconditional upper probability $\overline{P}$ is a possibility measure then so is the conditional upper probability $\overline{P}(\cdot \mid B)$ defined by natural extension, provided $\overline{P}(B) > 0$. Thus conditioning a possibility measure by natural extension produces another possibility measure.

Example 8 (*Examples of conditioning*). Consider the model for the judgement "Mary is young", characterised by the possibility distribution $\pi(\omega) = 1$, if $0 < \omega \leqslant 15$, $\pi(\omega) = (40 - \omega)/25$ if $15 < \omega < 40$, $\pi(\omega) = 0$ if $\omega \geqslant 40$. Suppose we learn the additional information that Mary's age belongs to a specified set of real numbers, B. We can update upper and lower probabilities concerning Mary's age simply by updating the possibility distribution π to $\pi(\cdot \mid B)$.

First let B be the event that Mary is no older than 30 years. Using Eq. (16) we find that $\pi(\omega \mid B) = \pi(\omega)$ if $0 < \omega \leqslant 30$, $\pi(\omega \mid B) = 0$ if $\omega > 30$. Here π is updated simply by truncating it at age 30; the plausibility of ages below 30 does not change.

As a second example, if B is the event that Mary's age is not between 20 and 30 years, we find that $\pi(\omega \mid B) = \pi(\omega)$ if $0 < \omega < 20$, $\pi(\omega \mid B) = 0$ if $20 \leqslant \omega \leqslant 30$ or $\omega \geqslant 40$, and $\pi(\omega \mid B) = (40 - \omega)/(45 - \omega)$ if $30 < \omega < 40$. Here the plausibility of ages below 20 does not change, but the new information makes ages between 30 and 40 more plausible than before.

Finally, notice that if B does not contain the entire interval $(0, 15]$ then the updated probabilities will be vacuous, in the sense that $\pi(\omega \mid B) = 1$ if $\omega \in B \cap (0, 40)$ and $\pi(\omega \mid B) = 0$ otherwise. If we learn, for example, that Mary is at least 5 years old then $\pi(\omega \mid B) = 1$ if $5 \leqslant \omega < 40$ and $\pi(\omega \mid B) = 0$ otherwise; all ages between 5 and 40 become fully plausible. More generally, if B^c contains a state ω that is fully plausible (i.e. $\pi(\omega) = 1$) then $\overline{P}(B^c) = 1$ and hence the probabilities conditional on B are vacuous, in that $\pi(\omega \mid B) = 1$ if $\omega \in B$ and $\pi(\omega) > 0$, while $\pi(\omega \mid B) = 0$ otherwise.

In this last case, where $\underline{P}(B) = 0$, the conditional probabilities defined by natural extension are essentially vacuous. This is generally the case when natural extension is used to condition on an event with lower probability zero. It can be understood through the consistency principle stated near the end of Section 4. In the example let A denote "Mary's age is less than 30" and let B denote "Mary's age is at least 5". Then $\underline{P}(B) = 0$, and it would be consistent with the initial possibility measure to make a further judgement that $P(A \cap B) = 0$ precisely, which would produce $P(A \mid B) = 0$. To be consistent with this we need $\underline{P}(A \mid B) = 0$. The same argument applies when A is any proper subset of the interval $[5,40)$.

So natural extension cannot produce any nontrivial inferences when the conditioning event B has $\underline{P}(B) = 0$. (Unfortunately, for possibility measures there are many such events B. In fact, for every event B, either $\underline{P}(B) = 0$ or $\underline{P}(B^c) = 0$.) The problem is that the initial model is too imprecise to determine conditional probabilities. One way to avoid this is to specify a more precise model, in particular one in which $\underline{P}(B) > 0$. Of course the corresponding upper probability may no longer be a possibility measure.

Dempster's rule of conditioning

Another option is to use a different rule of conditioning when $\underline{P}(B) = 0$. When applied to a possibility measure $\overline{P}$ for which $\overline{P}(B) > 0$, Dempster's rule of conditioning gives the conditional upper probabilities $\overline{P}_D(A \mid B) = \sup\{\pi^D(\omega \mid B):\ \omega \in A\}$, where the conditional possibility distribution $\pi^D(\cdot \mid B)$ is defined by

$$\pi^D(\omega \mid B) = \begin{cases} k^{-1}\pi(\omega), & \text{if } \omega \in B, \\ 0, & \text{if } \omega \in B^c, \end{cases} \tag{17}$$

and $k = \sup\{\pi(\omega):\ \omega \in B\} = \overline{P}(B)$. The conditional upper probability $\overline{P}_D(\cdot \mid B)$ is a possibility measure. Writing $\pi^E(\cdot \mid B)$ and $\overline{P}_E(\cdot \mid B)$ for the conditional possibility distribution and conditional upper probability defined by natural extension, we have $\pi^D(\omega \mid B) \leqslant \pi^E(\omega \mid B)$ for all ω, and $\overline{P}_D(A \mid B) \leqslant \overline{P}_E(A \mid B)$ for all sets A. Thus the conditional probabilities determined by Dempster's rule are always at least as precise as those defined by natural extension.

Consider the three examples of conditioning the information about Mary's age. In the first example, where B is the interval $(0,30]$, $\pi^D(\cdot \mid B) = \pi^E(\cdot \mid B)$ and the two rules produce the same solution. In the second example, where $B = [20,30]^c$, $\pi^D(\omega \mid B)$ agrees with $\pi^E(\omega \mid B)$ except when $30 < \omega < 40$, in which case $\pi^D(\omega \mid B) = \pi(\omega)$ is smaller than $\pi^E(\omega \mid B)$. The biggest difference between the two rules occurs in the third example, where $B = [5,\infty)$: $\pi^D(\omega \mid B)$ agrees with the initial degree of possibility $\pi(\omega)$ if $5 \leqslant \omega < 40$, whereas $\pi^E(\omega \mid B) = 1$. In this case conditioning by Dempster's rule preserves the information in the initial possibility distribution but natural extension does not, and Dempster's rule is arguably more reasonable. (In fact $\pi^D(\omega \mid B) = \pi(\omega)$ for all ω in B in all three examples, because $k = \overline{P}(B) = 1$ in each case.)

Dempster's rule for conditioning possibility distributions (17) can also be derived within possibility theory. The information that event B has occurred is represented by a possibility distribution π' which is simply the indicator function of B. This is combined with the initial possibility distribution π using the minimum rule, producing the conditional possibility distribution $\pi(\omega \mid B) \propto \min\{\pi(\omega), \pi'(\omega)\}$, so $\pi(\omega \mid B)$

$\propto \pi(\omega)$ if $\omega \in B$ and $\pi(\omega \mid B) = 0$ if $\omega \in B^c$. The proportionality constant is determined by the fact that $\pi(\cdot \mid B)$ is a possibility distribution, and the solution agrees with Dempster's rule (17). A different rule for conditioning possibility distributions is proposed in [17].

Example 9 (*Probabilistic qualification*). Now consider the qualified judgement "it is likely that Mary is young". Zadeh [121] treats this as information about an underlying probability density function for Mary's age, whereas I regard it simply as information about Mary's age. Let B denote the event that Mary is young, according to the criteria used by the speaker. The conditional probabilities $\overline{P}(A \mid B)$ and $\underline{P}(A \mid B)$ can be identified with the (unconditional) probabilities based on the information "Mary is young" which were defined previously. The judgement that B is likely is translated as $\underline{P}(B) \geqslant \frac{1}{2}$. The natural extensions of these assessments are

$$\underline{P}(A) = \tfrac{1}{2}\underline{P}(A \mid B) = \tfrac{1}{2}\inf\{\underline{F}(\omega)\colon\ \omega \in A^c\} = \tfrac{1}{2} - \tfrac{1}{2}\sup\{\pi(\omega)\colon\ \omega \in A^c\},$$

$$\overline{P}(A) = \tfrac{1}{2} + \tfrac{1}{2}\overline{P}(A \mid B) = 1 - \tfrac{1}{2}\inf\{\underline{F}(\omega)\colon\ \omega \in A\} = \tfrac{1}{2} + \tfrac{1}{2}\sup\{\pi(\omega)\colon\ \omega \in A\},$$

for any set A of possible ages.

Again the upper probability is a possibility measure, with possibility distribution function $\pi(\omega) = 1 - \frac{1}{2}\underline{F}(\omega)$. The effect of the qualification "it is likely that" is simply to reduce the lower distribution function $\underline{F}$ by a factor of $\frac{1}{2}$. (Equivalently, replace π by $\frac{1}{2} + \frac{1}{2}\pi$.) Although it produces a possibility distribution for Mary's age X, this analysis is much simpler than that of Zadeh [121], who requires a possibility distribution to be defined on the space of all probability density functions for X.

More generally, suppose that $\overline{P}(\cdot \mid B)$ and $\underline{P}(\cdot \mid B)$ are upper and lower probabilities based on information B, and this information is qualified in some way that can be translated as $\underline{P}(B) \geqslant \beta$, e.g. by asserting that B is "very probable" or by using one of the other natural-language expressions listed later in this section. Then upper and lower probabilities based on the qualified information can be computed by $\underline{P}(A) = \beta\underline{P}(A \mid B)$ and $\overline{P}(A) = \beta\overline{P}(A \mid B) + 1 - \beta$. This is a simple method of *discounting* information.

To see that such an analysis will not always produce a possibility measure, consider the upper probabilities generated by natural extension from the two judgements "it is likely that Mary is young" and "it is likely that Mary is older than 10 years". Let A and B denote the events "Mary is younger than 10 years" and "Mary is older than 40 years". Then $\overline{P}(A) = \overline{P}(B) = \frac{1}{2}$ but $\overline{P}(A \cup B) = 1 > \max\{\overline{P}(A), \overline{P}(B)\}$, so $\overline{P}$ is not a possibility measure.

Cheeseman [11] suggests several Bayesian models for the uncertainty about Mary's age, given the information "Mary is probably young". He would first model the uncertainty given "Mary is young" by specifying a probability density function for Mary's age, and similarly for the uncertainty given "Mary is *not* young". To model "probably" he would construct a second-order probability density on the unit interval. This is analogous to Zadeh's second-order possibility distribution for "likely".

Thus Cheeseman's analysis requires second-order assessments of similar complexity to Zadeh's. However Cheeseman claims that "For the accuracy appropriate to this type of linguistic information, it is sufficient to extract a single estimate (probability)" as an approximation to the second-order density, and he chooses the precise probability 0.9 (the mean of his second-order density) to represent "probably". So Cheeseman, given only the information "probably A", would be willing to bet on A at odds of 9 to 1 on! Both the second-order density and the precise value 0.9 have strong implications concerning betting rates and other decisions, and they are therefore inadequate models for an imprecise term like "probably". Cheeseman's overall model for the uncertainty about Mary's age is a precise probability density function, which is similarly inadequate to model the vagueness of the information on which it is based.

Vague probability judgements (second-order possibility measures)

Now consider the translation of vague probability judgements such as "event A is likely" or "A is very likely" into possibility distributions. Zadeh [117,121] treats such a judgement as information about the precise probability (p) of event A, and models it through a possibility distribution function $\pi_{\rm p}$ defined on the probability interval $[0,1]$. For instance, he models "it is likely that A" by a possibility distribution function $\pi_{\rm p}(p)$ that is zero when $0 \leqslant p \leqslant \frac{1}{2}$ and increases nonlinearly on $\frac{1}{2} < p \leqslant 1$ to a maximum $\pi_{\rm p}(1) = 1$. He models "it is very likely that A" by squaring the function $\pi_{\rm p}$. Watson, Weiss and Donnell [109] model "pretty likely" by a possibility distribution function that is zero on the interval $0 \leqslant p \leqslant 0.55$ and roughly constant on $0.65 \leqslant p \leqslant 0.9$. More generally, a judgement such as "it is likely that Mary is young" would be translated by assigning a degree of possibility $\pi(f)$ to every conceivable probability density function f for Mary's age.

It is strange that the theory of possibility measures, apparently designed to model the imprecision in ordinary-language judgements of uncertainty, needs to refer to underlying probabilities that are *precise*. The number $\pi_{\rm p}(p)$ measures the degree of possibility that the true probability of A is precisely p (given the judgement "it is likely that A"), i.e. it measures second-order uncertainty about a first-order probability. I have already remarked that it is unclear what is meant by "degree of possibility", but there is a more basic difficulty here: it is not even clear what is meant by "the true probability of A", and why this should be assumed to have the properties of a Bayesian probability measure.

The most natural interpretation is that the "true probability of A" is the subjective probability assigned to A by the speaker, who provides only partial information about it when he asserts "it is likely that A". But there is no reason to suppose that the speaker has made (or could make) a precise assessment of this probability—the vagueness of his assertion suggests just the opposite. (This issue is discussed at length in [99].) Similarly, there will rarely be a "true probability" in any objective sense. If we do not understand what is meant by "the true probability of A", how can we hope to assess the degree of possibility that p is the true probability?

Compare Zadeh's translation of "it is likely that A" with the behavioural translation of "probably A" that was suggested in Section 4. (Zadeh [117] regards "likely" and "probable" as "more or less synonymous".) The behavioural translation is that the speaker is willing to accept an even-money bet on A, which is modelled by $\underline{P}(A) \geqslant \frac{1}{2}$.

This model does not presuppose that A has any "true" precise probability, whether subjective or objective. It seems also that the behavioural translation matches more closely the qualitative form of the assertion "probably A". The behavioural model says *only* that the speaker would bet on A at even stakes, whereas the fuzzy model π_p is a mathematical function defined on the unit interval, which seems overly complicated to model such a simple judgement.

If a possibility distribution π_p concerning an unknown probability $P(A)$ is interpreted as a second-order upper probability function, it generates first-order upper and lower probabilities $\overline{P}(A)$ and $\underline{P}(A)$ by a result in [102]. Let $Z(p)$ denote the probability of A conditional on the hypothesis that $P(A) = p$, so Z is simply the identity function $Z(p) = p$. Then $\underline{P}(A)$, the first-order lower probability of A, is defined to be the second-order lower prevision of Z, which is constructed by natural extension of the possibility measure using Eq. (13). Hence the first-order probabilities are

$$\underline{P}(A) = 1 - \int_0^1 \sup\{\pi_p(p):\ p \leqslant y\}\,dy,$$

$$\overline{P}(A) = \int_0^1 \sup\{\pi_p(p):\ p > y\}\,dy. \tag{18}$$

For example, if "it is likely that A" is translated into the possibility distribution $\pi_p(p) = 1$ when $\frac{1}{2} \leqslant p \leqslant 1$, $\pi_p(p) = 0$ otherwise, we obtain $\underline{P}(A) = \frac{1}{2}$ and $\overline{P}(A) = 1$. If the same judgement is modelled by $\pi_p(p) = 2p - 1$ when $\frac{1}{2} \leqslant p \leqslant 1$, $\pi_p(p) = 0$ otherwise, then $\underline{P}(A) = \frac{3}{4}$ and $\overline{P}(A) = 1$. The possibility distribution for "likely" suggested by Zadeh [117, Fig. 1] yields approximately these values. The lower probability seems somewhat too high.

More generally, if π_p is unimodal with mode m, we find that the induced lower and upper probabilities are

$$\underline{P}(A) = m - \int_0^m \pi_p(y)\,dy, \qquad \overline{P}(A) = m + \int_m^1 \pi_p(y)\,dy. \tag{19}$$

Most of the possibility distributions that have been suggested are unimodal. In this case the degree of imprecision concerning A is $\overline{P}(A) - \underline{P}(A) = \int_0^1 \pi_p(y)\,dy$, which is just the area under the graph of π_p.

Example 10 (*Football game*). Consider the football example from Section 4. A subject makes three probability judgements about the result of the game. The first judgement, "probably *not* W", might be translated into the possibility distribution $\pi_1(w) = \max\{1 - 2w, 0\}$, where (w, d, l) will be used to denote the precise probabilities of (W, D, L). Similarly, "W is more probable than D" might be translated into $\pi_2(w-d) = \max\{w - d, 0\}$, and "$D$ is more probable than L" into $\pi_3(d - l) = \max\{d - l, 0\}$.

How should the three judgements be combined to form a joint possibility distribution π for (w, d, l)? Zadeh [117, Example 1.2] suggests that such judgements can be regarded as "non-interactive" and a joint distribution formed by taking minima. (This is questionable because the first judgement effectively restricts the range of possible values for $w - d$. But this kind of "interaction" between judgements is very common, and it is not clear how possibility theorists would combine the possibility distributions if not by the minimum rule.) Thus, assuming $w + d + l = 1$ so that (w, d, l) is a probability distribution,

$$\begin{aligned}\pi(w, d, l) &\propto \min\{\pi_1(w), \pi_2(w - d), \pi_3(d - l)\} \\ &\propto \max\{0, \min\{1 - 2w, w - d, d - l\}\}.\end{aligned}$$

It is necessary to renormalise the right-hand expression to have maximum value one, and this yields the joint possibility distribution

$$\pi(w, d, l) = 9\max\{0, \min\{1 - 2w, w - d, d - l\}\}.$$

The maximum possibility is achieved by $w = \frac{4}{9}$, $d = \frac{1}{3}$, $l = \frac{2}{9}$.

Compare this with the lower prevision constructed in Section 4, which is the lower envelope of the set $\mathcal{M}$ of all probability distributions (w, d, l) that are consistent with the three judgements, i.e. satisfy $w \leqslant \frac{1}{2}$, $w \geqslant d$, $d \geqslant l$. The joint possibility $\pi(w, d, l)$ is positive when (w, d, l) lies in the interior of $\mathcal{M}$, it increases with distance from the boundary of $\mathcal{M}$, and it is zero otherwise.

Marginal possibility distributions can be computed from π by maximisation. For example, after some calculations the marginal possibility distribution for l is found to be $\pi_L(l) = \max\{\pi(\omega, d, l)\colon 0 \leqslant \omega \leqslant 1,\ 0 \leqslant d \leqslant 1,\ \omega + d = 1 - l\} = 9\max\{0, \min\{\frac{1}{2}l, \frac{1}{3} - l\}\}$, which is unimodal with mode $m = \frac{2}{9}$. Hence, using Eqs. (19), this model generates upper and lower probabilities

$$\overline{P}(L) = m + \int_m^1 \pi_L(y)\,dy = \frac{5}{18} = 0.278,$$

$$\underline{P}(L) = m - \int_0^m \pi_L(y)\,dy = \frac{1}{9} = 0.111.$$

These probabilities are more precise than the values $\overline{P}(L) = \frac{1}{3}$ and $\underline{P}(L) = 0$ generated by the model in Section 4; here $\overline{P}(L) - \underline{P}(L) = \int_0^1 \pi_L(y)\,dy = \frac{1}{6}$, compared with $\frac{1}{3}$ for the earlier model.

Calculus

As seen in the preceding example, the rules for combining and manipulating possibility distributions involve operations of forming maxima and minima. These rules are derived from the basic rules of fuzzy sets proposed by Zadeh [116]. Without a definite interpretation of possibility measures the rules appear quite arbitrary. But if possibility measures are interpreted as coherent upper probabilities, as suggested earlier, the rules need to be evaluated according to whether they *preserve coherence*, in the sense that, after applying the rules, the overall upper probability model is coherent. In some special cases, such as (b) below, the rules agree with natural extension and then they do preserve coherence. This gives some justification for these rules.

But other rules, such as (c), disagreee with natural extension in general. It needs to be investigated whether such rules can produce upper probabilities that are incoherent or incur sure loss. I am currently studying these questions and the results will be reported elsewhere.

The most important rules are as follows.

(a) Given a joint possibility distribution $\pi_{X,Y}$ for two variables X and Y, a *marginal possibility distribution* for X is defined by $\pi_X(x) = \sup\{\pi_{X,Y}(x,y):\ y \in \Omega_x\}$ where $\Omega_x = \{y:\ (x,y) \in \Omega\}$ and Ω is the joint possibility space. This rule is essentially built into the definition of a possibility measure. It does preserve coherence of the corresponding upper probabilities.

(b) Given two marginal possibility distributions π_X and π_Y, where X and Y are logically independent variables, a *joint possibility distribution* for X and Y is defined by $\pi_{X,Y}(x,y) = \min\{\pi_X(x), \pi_Y(y)\}$. This implies, in terms of the corresponding upper probabilities, that $\overline{P}_{X,Y}(A \times B) = \min\{\overline{P}_X(A), \overline{P}_Y(B)\}$ for product sets $A \times B$, and this rule agrees with natural extension of the marginal upper probabilities to a joint upper probability. However the two rules can disagree for sets that are not products.

(c) Given two possibility distributions π_1 and π_2, concerning the same unknown ω but based on "non-interactive" bodies of evidence, a *combined possibility distribution* is defined by $\pi(\omega) \propto \min\{\pi_1(\omega), \pi_2(\omega)\}$, where the normalising factor is chosen so that π has supremum value 1. This rule was used to combine judgements in the football example. Rule (b) can be regarded as a special case of (c) with $\omega = (x,y)$. Another special case is the rule of conditioning discussed earlier, where π_2 is the indicator function of the conditioning event, which agrees with Dempster's rule of conditioning (17).

The more general rule (c) appears to have a similar role in combining possibility distributions to Dempster's rule for combining belief functions. It is used in expert systems to combine information from different sources. However the rule disagrees with natural extension in general (although one special case of agreement was noted in (b)), and it need not produce coherent inferences. The next example shows that rule (c) and Dempster's rule of combination can produce quite different answers.

Example 11 (*Two unreliable witnesses*). Consider the example in Section 5 of two unreliable witnesses. The first witness has credibility α_1 and reports that event C_1 occurred. This can be modelled by a marginal possibility distribution π_1 that assigns $\pi_1(C_1) = 1$ and $\pi_1(C_1^c) = 1 - \alpha_1$. This generates the upper and lower probabilities assumed in Section 5, $\overline{P}_1(C_1) = 1$ and $\underline{P}_1(C_1) = \alpha_1$. Similarly the information provided by the second witness is modelled by a possibility distribution $\pi_2(C_2) = 1$ and $\pi_2(C_2^c) = 1 - \alpha_2$.

If we combine the two marginal possibility distributions by the minimum rule, we obtain degrees of possibility 1, $1-\alpha_2$, $1-\alpha_1$ and $1-\max\{\alpha_1,\alpha_2\}$ for the elementary events $C_1 \cap C_2$, $C_1 \cap C_2^c$, $C_1^c \cap C_2$ and $C_1^c \cap C_2^c$. This generates upper and lower probabilities $\overline{P}(C_1 \cap C_2) = 1$ and $\underline{P}(C_1 \cap C_2) = \min\{\alpha_1, \alpha_2\}$. The lower probability is quite different from the value $\alpha_1 \alpha_2$ which is given by all three models in Section 5, based on different ways of formalising the judgement that the two reports are independent. There is also a discrepancy between the values of $\overline{P}(C_1^c \cap C_2^c)$, which are $1 - \max\{\alpha_1, \alpha_2\}$ for this model and $(1 - \alpha_1)(1 - \alpha_2)$ for the three earlier models.

In this example the combined possibility distribution seems less satisfactory than the three earlier models. When $\alpha_1 = \alpha_2 = \frac{1}{2}$, for example, the earlier models give $\underline{P}(C_1 \cap C_2) = \frac{1}{4}$ whereas the rule for combining possibility measures gives $\underline{P}(C_1 \cap C_2) = \frac{1}{2}$. The latter value seems unreasonable; one would expect $\underline{P}(C_1 \cap C_2)$ to be somewhat less than the marginal lower probabilities $\underline{P}(C_1)$ and $\underline{P}(C_2)$, which are each $\frac{1}{2}$, unless there is evidence that the two reports are correlated in a particular way.

In fact, for any joint possibility measure, the corresponding lower probability must satisfy $\underline{P}(C_1 \cap C_2) = \min\{\underline{P}(C_1), \underline{P}(C_2)\}$ since this is a general property of necessity measures. But this property seems inappropriate when the two sources of information are judged to be independent. So it does not seem possible to adequately model independence judgements using possibility measures.

Computation

The computation of inferences and decisions from possibility distributions requires, in general, the solution of a *nonlinear* programming problem. (Compare with the computation of natural extension suggested in Section 4, which involves only *linear* programming.) For example, decisions are made from second-order possibility distributions by computing "fuzzy expectations", i.e. degrees of possibility for all possible expected values x of a random variable X, by maximising $\pi(P)$ over all probability distributions P which satisfy $P(X) = x$. Because $\pi(P)$ will generally be a nonlinear function of the probability masses, this entails a nonlinear programming problem—in fact, a separate problem for each possible value of x! (See [109,117,121] for more details.) Similarly the computation of a marginal possibility distribution involves, in general, many nonlinear programs. So computations will often be difficult, despite the apparent simplicity of the calculus. The computations are generally easier for first-order possibility distributions than for second-order distributions because the optimisations are generally over lower-dimensional spaces.

Assessment

How difficult is it to make the assessments required for a fuzzy analysis? The aim is to allow users of the system (or domain experts) to express their uncertainty in natural language, in whatever forms are most convenient. This is valuable, as it makes the task of assessing uncertainty relatively simple for the system user. The difficulties are passed on to the experts on fuzzy logic, who must translate the natural-language judgements into possibility distributions. For example, they must translate "it is likely that A" into a function π defined on the probability interval $[0, 1]$. More complex judgements of uncertainty concerning Ω must be translated into a function π defined on the space of all probability distributions on Ω. A great deal of input is needed to specify these functions, and there seems to be a great deal of arbitrariness in selecting them. Again it is generally easier to assess a first-order possibility distribution than a second-order distribution because the former is usually defined on a lower-dimensional space.

There seems to be little guidance in the literature on possibility theory about how to select the required functions. (But see [17, p. 19], for some suggestions.) Specific assessments of possibility distributions are sometimes given to illustrate the methodology, but these appear quite arbitrary and no justification is offered. For example, Zadeh [117] specifies two quite different possibility distributions to model the judgement "likely" in successive examples: his Example 1.1 has $\pi(0.8) = 0.9$ while Example 1.2 has $\pi(0.8) = 0.5$. Another translation, shown graphically in his Fig. 1, seems to be different again. No reasons are given to support any of these assessments, and it is hard to see how they could be supported until the meaning of the numbers $\pi(p)$ is clarified.

In practice one might use standard translations for common terms such as "likely", "very likely" or "more likely than". That is, all instances of the term "very likely" would be translated into the same possibility distribution. The choice of this standard distribution may still be somewhat arbitrary but no further input would be needed.

The danger of using standard translations is that different people seem to use expressions like "very likely" in different ways and usage varies with context [3,106]. Ideally, one would like to use a different translation for each person and context. That would be impracticable in most cases, as it would require too much input from the person. However one could require the standard translation of "very likely" to *encompass* the meanings intended by most speakers in most contexts. We can illustrate the idea as follows, using the behavioural interpretation of upper and lower probabilities.

Standard translations into lower probabilities

Consider the expression "probably A". Most people who use this expression would be willing to bet on A at even stakes. Hence they would accept the behavioural translation $\underline{P}(A) \geqslant 0.5$. We could, in principle, determine a lower probability $\underline{P}_i(A)$ for person i by finding the least favourable odds at which he is willing to bet on A. Provided $\underline{P}_i(A) \geqslant 0.5$ for all persons i, the behavioural translation $\underline{P}(A) \geqslant 0.5$ is acceptable to all, and we may say that it *encompasses* the meanings of "probably A" for all the persons. It could then be used as a standard ("cautious" or "default") translation of "probably A" that is acceptable in the great majority of problems. In any particular problem we would try to (a) check with the person that he accepts the behavioural translation of his judgement, and (b) encourage him to make this more precise, e.g. if he is willing to bet on A at odds of 3 to 2 on then he will accept the more precise judgement $\underline{P}(A) \geqslant 0.6$. If he was unwilling to do this, or if the checks could not be carried out, then the standard translation would be used.

The standard translations should be based on empirical studies of the meanings of terms such as "probably". It seems, for example, that most people do not apply the term "probable" to events with very high probability, so that one might include the constraint $\overline{P}(A) \leqslant 0.9$, as well as $\underline{P}(A) \geqslant 0.5$, in the standard translation of "A is probable". In the light of empirical studies such as [3,45,106], I suggest the following translations of common judgements in natural language. In all these expressions I take "probable" to be synonymous with "likely". I identify negative expressions such as "A is improbable" with positive expressions such as "A^c is probable". The latter is translated into $\underline{P}(A^c) \geqslant 0.5$, which is equivalent to $\overline{P}(A) \leqslant 0.5$. This translation is somewhat cautious, as there is evidence that most people would accept a stronger translation $\overline{P}(A) \leqslant 0.4$. Many of the other translations could be strengthened considerably in suitable contexts.

- A is extremely probable $\rightarrow \underline{P}(A) \geqslant 0.98$.
- A has very high probability $\rightarrow \underline{P}(A) \geqslant 0.9$.
- A is highly probable $\rightarrow \underline{P}(A) \geqslant 0.85$.
- A is very probable $\rightarrow \underline{P}(A) \geqslant 0.75$.
- A has a very good chance $\rightarrow \underline{P}(A) \geqslant 0.65$.
- A is quite probable $\rightarrow \underline{P}(A) \geqslant 0.6$.
- A is probable [likely] $\rightarrow \underline{P}(A) \geqslant 0.5$.
- A is improbable [unlikely] $\rightarrow \overline{P}(A) \leqslant 0.5$.
- A is somewhat unlikely [quite improbable] $\rightarrow \overline{P}(A) \leqslant 0.4$.
- A is very unlikely $\rightarrow \overline{P}(A) \leqslant 0.25$.
- A has little chance $\rightarrow \overline{P}(A) \leqslant 0.2$.
- A is highly improbable $\rightarrow \overline{P}(A) \leqslant 0.15$.
- A has very low probability $\rightarrow \overline{P}(A) \leqslant 0.1$.
- A is extremely unlikely $\rightarrow \overline{P}(A) \leqslant 0.02$.
- A has a good chance $\rightarrow \underline{P}(A) \geqslant 0.4$, $\overline{P}(A) \leqslant 0.85$.
- The probability of A is about $\alpha \rightarrow \underline{P}(A) \geqslant \alpha - 0.1$, $\overline{P}(A) \leqslant \alpha + 0.1$.
- A is more probable than $B \rightarrow \underline{P}(A - B) \geqslant 0$.

There is some degree of arbitrariness in choosing a single number (0.85) to translate an expression such as "highly probable", but this is much less than the arbitrariness in choosing a possibility distribution function π, i.e. a degree of possibility $\pi(p)$ for every value of p between 0 and 1. Similar translations, from imprecise probabilities to linguistic expressions, could be used to make a system's conclusions and reasoning more comprehensible to a user.

Conclusion

Possibility theory and fuzzy logic have made a substantial contribution to our understanding of uncertainty by drawing attention to the important problem of modelling natural-language judgements and suggesting possibility measures as suitable models. Provided second-order measures are allowed, possibility measures can model a wide variety of uncertainty judgements, including imprecise judgements in natural language. Bayesian probabilities are inadequate to model vague predicates and vague probability judgements, as in "Mary is probably young", but it appears that upper and lower previsions may be adequate, especially as possibility measures can be regarded as a special type of coherent upper probability.

It is important to distinguish between first- and second-order possibility measures. First-order possibility measures, which are used to model the uncertainty generated by vague predicates like "young" and "tall", can be interpreted as coherent upper probabilities. They are mathematically simpler than the other types of coherent upper probabilities as they can be defined in terms of possibility distributions, functions whose domain is Ω rather than the power set of Ω. This may simplify computations and assessment, e.g. possibility measures can sometimes be assessed through upper or lower distribution functions. First-order possibility measures are likely to be useful models in many expert systems where information is elicited in terms of vague predicates.

The main defect of first-order possibility measures is that they cannot model many of the common types of uncertainty. In particular they cannot model natural-language judgements of uncertainty or precise probability assessments. Possibility measures are a very special type of upper probability (in fact they correspond to a special type of belief function), and upper probability is itself inadequate in many problems; upper and lower previsions are needed in general.

Second-order possibility measures, defined on the set of all probability distributions, are much more expressive than first-order measures. They can model both precise and imprecise judgements of uncertainty. But they are much more complicated than first-order measures and they have some serious defects. Second-order models are difficult to interpret and assess and they seem overly complicated to model qualitative judgements. Theoretical papers on fuzzy logic tend to ignore the practical problem of assessing second-order functions, each defined on a space of probability distributions. Computations, e.g. of marginal possibility distributions or fuzzy expected values, generally require nonlinear programming.

The rules for combining possibility measures are simple (they involve operations of maximising and minimising), but they do not appear to have any compelling justification. If possibility measures are interpreted as coherent upper probabilities then the rules can be compared with the rules of natural extension and we can investigate whether they preserve coherence. They appear to do so in some cases but not in general.

7. Comparison and evaluation

Most practical reasoning involves uncertainty. In expert systems, as in many other fields, it is frequently necessary to measure the uncertainty. What measure should we use? Four measures have been considered in this paper. Here is a summary of the extent to which they satisfy the criteria proposed in Section 2.

Interpretation, calculus and consistency

These criteria are satisfied only by the theories of Bayesian probability and coherent lower previsions, which start with a simple behavioural interpretation and use this to justify principles of coherence and to derive rules for combining and updating probabilities. There is a general method, called *natural extension*, for computing new previsions from an arbitrary set of judgements. Natural extension can be used to make inferences and decisions. There are general methods for checking consistency of the initial judgements, and the rules of natural extension ensure that conclusions will be consistent with the judgements.

Possibility theory and the Dempster–Shafer theory do not do so well on these criteria. These theories propose mathematical properties to characterise their uncertainty measures and simple rules for combining measures, but they fail to give any compelling justification for their properties and rules. Belief functions can be interpreted in terms of multivalued mappings but this supports Dempster's rule of combination only when some strong assumptions of conditional independence are added. Both theories lack methods for checking the consistency of models and conclusions, and their rules can produce conclusions that are intuitively inconsistent with the initial model.

Imprecision

Bayesian probabilities cannot adequately model ignorance, imprecise or qualitative judgements of uncertainty, or vague predicates in natural language. The other measures can do so to some extent, but belief functions and first-order possibility measures are not sufficiently general to model common types of imprecise judgements.

Assessment

Insufficient guidance is given in these theories (especially possibility theory) about how to make assessments of uncertainty, although the theories of belief functions and lower previsions do take this problem seriously. All the theories seem to need judgements of independence or non-interaction, in multivariate problems, to reduce the number of assessments. Lower previsions and second-order possibility measures can model a wide variety of uncertainty judgements, including qualitative judgements in ordinary language, although assessments of second-order possibility distributions seem complicated and arbitrary. Assessment is onerous for Bayesians because they require precise assessments and a complete probability model; the other theories are more flexible.

Computation

For all the measures, computational feasibility will depend on the type and complexity of the model and the number of assessments. For lower previsions, the computation of inferences and decisions by natural extension involves linear programming. More work is needed to develop computationally efficient methods and tractable models, particularly based on conditional independence. Bayesian probabilities, belief functions and first-order possibility measures, as special types of lower or upper previsions, may be computationally simpler in some cases. (They are tractable in singly-connected belief networks, for example.) Computational methods are most highly developed for Bayesian models. Second-order possibility measures are less tractable than the other measures—computations involve nonlinear programming.

8. Conclusion

So what measures of uncertainty should be used in expert systems? I believe that Bayesian probabilities, upper and lower probabilities, belief functions and first-order possibility measures can all be useful in different types of problems, for instance when the information is in the form of extensive statistical data, bounds on probabilities, multivalued mappings or natural-language judgements respectively.

All these measures can be useful in special types of problems, but none of them is adequate as a general model of uncertainty. For example, none of them can adequately model the three qualitative judgements in the football example. Bayesian probabilities are not sufficiently general because, in many problems, information is scarce and judgements are imprecise. Belief functions and possibility measures are not sufficiently general because many examples involve lower probabilities that are not even 2-monotone. Upper and lower probabilities are not sufficiently general because they do not uniquely determine upper and lower previsions and conditional probabilities; upper and lower previsions produce greater precision in inferences and decisions. I suggest that upper and lower previsions, which include the other measures as special cases, are sufficiently general to model the most important types of uncertainty.

This raises the question: to what extent is the theory of coherent lower previsions compatible with the other theories? It is compatible with the Bayesian theory as the two theories have a similar behavioural interpretation and the rules of the Bayesian theory are special types of natural extension. So the theory of lower previsions can be regarded as a generalisation of the Bayesian theory; the two theories agree in the special case where all probability models are precise. The theory of Bayesian sensitivity analysis or "probability intervals" is also, to a large extent, compatible with the theory of lower previsions.

Possibility theory and Dempster–Shafer theory appear to be less compatible with the theory of lower previsions. However I believe that the theory of lower previsions can incorporate some of the most useful features of these theories. In particular, one of the main contributions of the two theories has been to suggest some flexible and powerful methods for modelling particular types of partial information, notably through multivalued mappings and natural-language judgements. These can be used as methods for assessing lower previsions.

The theories differ in their interpretation of uncertainty measures. The interpretation of belief functions and possibility measures is unsettled and controversial, but it seems to me that at least some of the interpretations that have been proposed for these measures are consistent with the behavioural interpretation of lower and upper previsions. Two advantages of the behavioural interpretation are that it relates uncertainty measures to decisions and thereby explains how they can be used, and it leads to coherence principles which can be used to check consistency of models with conclusions. (Both features are lacking in Dempster–Shafer and possibility theory.) The behavioural interpretation is sufficiently general to encompass multivalued mappings, inexact judgements and lower envelopes of Bayesian probability measures as sources of lower previsions, yet specific enough to support the principles of coherence and the rules of natural extension.

The theories have quite different rules for combining and updating uncertainties, and in this respect they do appear to be incompatible. Dempster's rule of combination and the minimum rule for combining possibility distributions can, in some problems, produce upper and lower probability models that are incoherent, and in these cases the rules are not compatible with the theory of lower previsions. These rules are controversial. They can lead to intuitive inconsistencies as well as mathematical incoherence and they seem to be applicable only in a limited range of problems. If these rules were given a more restricted role in the Dempster–Shafer theory and in possibility theory then the three theories would be considerably more compatible.

Natural extension is an alternative to Dempster's rule and the minimum rule, and it is worth investigating other rules that preserve coherence but produce stronger conclusions than natural extension. This may be a way of developing the calculus of belief functions and possibility measures. The behavioural interpretation and principles of coherence can impose some much needed discipline on the theories of belief functions and possibility measures, without which these theories can produce inconsistencies.

Of course the theory of lower previsions is itself undeveloped in some important repects. Further investigation is needed into the practical problems of modelling, assessment and computation. Particular problems are to determine how best to model independence judgements, and how to compute natural extensions and propagate lower previsions efficiently. It is also important to compare the four approaches in some realistically complex expert systems. I hope that this paper will persuade some people to consider these problems.

Acknowledgements

I am grateful to Dr. Hitoshi Furuta for inviting me to lecture on this subject in Osaka in April 1990 and for suggesting this paper. Many people commented on an earlier version of the paper [100] which was circulated in April 1991. In particular I must thank Terrence Fine, Frank Hampel and Philippe Smets for stimulating discussions, and Smets, Nic Wilson and George Klir for sending me copies of relevant papers. Nic Wilson contributed some exceptionally detailed and insightful comments on an earlier version and the final version has benefited greatly from his suggestions.

References

[1] K.-P. Adlassnig and G. Kolarz, CADIAG-2: computer-assisted medical diagnosis using fuzzy subsets, in: M.M. Gupta and E. Sanchez, eds., *Approximate Reasoning in Decision Analysis* (North-Holland, Amsterdam, 1982) 219–247.

[2] S.K. Andersen, K.G. Olesen, F.V. Jensen and F. Jensen, HUGIN—a shell for building Bayesian belief universes for expert systems, in: *Proceedings IJCAI-89*, Detroit, MI (Morgan Kaufmann, San Mateo, CA, 1989) 1080–1085.

[3] R. Beyth-Marom, How probable is probable? Numerical translations of verbal probability expressions, *J. Forecasting* **1** (1982) 257–269.

[4] G. Biswas and T.S. Anand, Using the Dempster–Shafer scheme in a mixed-initiative expert system shell, in: L.N. Kanal, T.S. Levitt and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **3** (North-Holland, Amsterdam, 1989) 223–239.

[5] G. Biswas, M. Oliff and R. Abramczyk, OASES (Operations Analysis Expert System): an application in fiberglass manufacturing, *Int. J. Expert Syst. Res. Appl.* **1** (1988) 193–216.

[6] B.G. Buchanan and E.H. Shortliffe, eds., *Rule-Based Expert Systems* (Addison-Wesley, Reading, MA, 1984).

[7] J.-C. Buisson, H. Farreny and H. Prade, The development of a medical expert system and the treatment of imprecision in the framework of possibility theory, *Inf. Sci.* **37** (1985) 211–226.

[8] R. Buxton, Modelling uncertainty in expert systems, *Int. J. Man-Mach. Stud.* **31** (1989) 415–476.

[9] L.M. De Campos, M.T. Lamata and S. Moral, The concept of conditional fuzzy measure, *Int. J. Intell. Syst.* **5** (1990) 237–246.

[10] J.E. Cano, S. Moral and J.F. Verdegay, Propagation of convex sets of probabilities in directed acyclic networks, in: *Proceedings Fourth IPMU Conference* (1992) 289–292.

[11] P. Cheeseman, Probabilistic versus fuzzy reasoning, in: L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1986) 85–102.

[12] L.J. Cohen, *The Probable and the Provable* (Clarendon Press, Oxford, 1977).

[13] P.R. Cohen, *Heuristic Reasoning about Uncertainty: An Artificial Intelligence Approach* (Pitman, London, 1985).

[14] R.G. Cowell, BAIES—a probabilistic expert system shell with qualitative and quantitative learning, in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., *Bayesian Statistics* **4** (Clarendon Press, Oxford, 1992) 595–600.

[15] A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, *Ann. Math. Stat.* **38** (1967) 325–339.

[16] D. Dubois and H. Prade, An introduction to possibilistic and fuzzy logics, in: P. Smets, A. Mamdani, D. Dubois and H. Prade, eds., *Non-Standard Logics for Automated Reasoning* (Academic Press, London, 1988) 287–326.

[17] D. Dubois and H. Prade, *Possibility Theory* (Plenum Press, New York, 1988).

[18] D. Dubois and H. Prade, Modelling uncertain and vague knowledge in possibility and evidence theories, in: R.D. Shachter, T.S. Levitt, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **4** (North-Holland, Amsterdam, 1990) 303–318.

[19] D. Dubois and H. Prade, Evidence, knowledge, and belief functions, *Int. J. Approx. Reasoning* **6** (1992) 295–319.

[20] R.O. Duda, P.E. Hart and N.J. Nilsson, Subjective Bayesian methods for rule-based inference systems, in: *Proceedings 1976 National Computer Conference, AFIPS* **45** (1976) 1075–1082; also in: B.W. Webber and N.J. Nilsson, eds., *Readings in Artificial Intelligence* (Morgan Kaufmann, Los Altos, CA, 1981) 192–199.

[21] R.O. Duda and R. Reboh, AI and decision making: the PROSPECTOR experience, in: W. Reitman, ed., *Artificial Intelligence Applications for Business* (Ablex, Norwood, NJ, 1984) 111–147.

[22] A. Dutta, Reasoning with imprecise knowledge in expert systems, *Inf. Sci.* **37** (1985) 3–24.

[23] R. Fagin and J.Y. Halpern, A new approach to updating beliefs, in: P.P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **6** (North-Holland, Amsterdam, 1991) 347–374.

[24] H. Farreny, H. Prade and E. Wyss, Approximate reasoning in a rule-based expert system using possibility theory: a case study, in: H.J. Kugler, ed., *Information Processing '86* (North-Holland, Amsterdam, 1986) 407–413.

[25] K.W. Fertig and J.S. Breese, Interval influence diagrams, in: M. Henrion, R.D. Shachter, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **5** (North-Holland, Amsterdam, 1990) 149–161.

[26] M. Fieschi, M. Joubert, D. Fieschi, G. Botti and M. Roux, A program for expert diagnosis and therapeutic decision, *Med. Inf.* **8** (1983) 127–135.

[27] T.L. Fine, *Theories of Probability* (Academic Press, New York, 1973).

[28] B. de Finetti, *Theory of Probability* Vol. 1 (Wiley, London, 1974).

[29] B. de Finetti, *Theory of Probability* Vol. 2 (Wiley, London, 1975).

[30] L.C. van der Gaag, Computing probability intervals under independency constraints, in: P.P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **6** (North-Holland, Amsterdam, 1991) 457–466.

[31] W.A. Gale, ed., *Artificial Intelligence and Statistics* (Addison-Wesley, Reading, MA, 1986).

[32] R. Giles, Foundations for a theory of possibility, in: M.M. Gupta and E. Sanchez, eds., *Fuzzy Information and Decision Processes* (North-Holland, Amsterdam, 1982) 183–195.

[33] R. Giles, Semantics for fuzzy reasoning, *Int. J. Man-Mach. Stud.* **17** (1982) 401–415.

[34] J. Gordon and E.H. Shortliffe, A method for managing evidential reasoning in a hierarchical hypothesis space, *Artif. Intell.* **26** (1985) 323–357.

[35] B.N. Grosof, An inequality paradigm for probabilistic knowledge, in: L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1986) 259–275.

[36] J.Y. Halpern and R. Fagin, Two views of belief: belief as generalized probability and belief as evidence, *Artif. Intell.* **54** (1992) 275–317.

[37] D. Heckerman, E. Horvitz and B. Nathwani, Toward normative expert systems I: the PATHFINDER project, *Methods Inf. Med.* **31** (1992) 90–105.

[38] M. Henrion, J.S. Breese and E.J. Horvitz, Decision analysis and expert systems, *AI Magazine* **12** (1991) 64–91.

[39] Y.-T. Hsia and P.P. Shenoy, An evidential language for expert systems, in: Z. Ras, ed., *Methodologies for Intelligent Systems* **4** (North-Holland, New York, 1989) 9–16.

[40] P.J. Huber, *Robust Statistics* (Wiley, New York, 1981).

[41] J.-Y. Jaffray, Bayesian updating and belief functions, *IEEE Trans. Syst. Man Cybern.* **22** (1992) 1144–1152.

[42] E.T. Jaynes, *Papers on Probability, Statistics and Statistical Physics* (Reidel, Dordrecht, 1983).

[43] F.V. Jensen, S.K. Andersen, U. Kjaerulff and S. Andreassen, MUNIN—on the case for probabilities in medical expert systems, *Lecture Notes in Medical Informatics* **33** (Springer, Berlin, 1987).

[44] A.C. Kak, K.M. Andress, C. Lopez-Abadia, M.S. Caroll and J.R. Lewis, Hierarchical evidence accumulation in the PSEIKI system, in: M. Henrion, R.D. Shachter, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **5** (North-Holland, Amsterdam, 1990).

[45] S. Kent, Words of estimated probability, *Stud. Intell.* **8** (1964) 49–65.

[46] G.J. Klir and T.A. Folger, *Fuzzy Sets, Uncertainty, and Information* (Prentice-Hall, Englewood Cliffs, NJ, 1988).

[47] P. Krause and D. Clark, *Representing Uncertain Knowledge* (Kluwer, Dordrecht, 1993).

[48] H.E. Kyburg Jr, Bayesian and non-Bayesian evidential updating, *Artif. Intell.* **31** (1987) 271–293.

[49] K.B. Laskey, M.S. Cohen and A.W. Martin, Representing and eliciting knowledge about uncertain evidence and its implications, *IEEE Trans. Syst. Man Cybern.* **19** (1989) 536–557.

[50] S.L. Lauritzen and D.J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *J. Roy. Stat. Soc. Ser. B* **50** (1988) 157–224.

[51] J. Lebailly, R. Martin-Clouaire and H. Prade, Use of fuzzy logic in rule-based systems in petroleum geology, in: E. Sanchez and L.A. Zadeh, eds., *Approximate Reasoning in Intelligent Systems, Decision and Control* (Pergamon Press, Oxford, 1987) 125–144.

[52] J.F. Lemmer, Generalised Bayesian updating of incompletely specified distributions, *Large Scale Syst.* **5** (1983) 51–68.

[53] K.S. Leung, W.S.F. Wong and W. Lam, Applications of a novel fuzzy expert system shell, *Expert Syst.* **6** (1989) 2–10.

[54] I. Levi, Potential surprise: its role in inference and decision making, in: L.J. Cohen and M. Hesse, eds., *Applications of Inductive Logic* (Oxford University Press, Oxford, 1980) 1–27.

[55] I. Levi, Consonance, dissonance and evidentiary mechanisms, in: P. Gärdenfors, B. Hansson and N.E. Sahlin, eds., *Evidentiary Value*, Library of Theoria **15** (Gleerups, Lund, 1983) 27–43.

[56] D.V. Lindley, *Making Decisions* (Wiley, London, 1971).

[57] R.P. Loui, Interval-based decisions for reasoning systems, in: L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1986) 459–472.

[58] J.D. Lowrance, Automated argument construction, *J. Stat. Planning Inference* **20** (1988) 369–387.

[59] R.C. Moore, Semantical considerations on nonmonotonic logic, *Artif. Intell.* **25** (1985) 75–94.

[60] N.J. Nilsson, Probabilistic logic, *Artif. Intell.* **28** (1986) 71–87.

[61] P. Orponen, Dempster's rule is #P-complete, *Artif. Intell.* **44** (1990) 245-253.
[62] G. Paass, Probabilistic logic, in: P. Smets, A. Mamdani, D. Dubois and H. Prade, eds., *Non-Standard Logics for Automated Reasoning* (Academic Press, London, 1988) 213-251.
[63] J. Pearl, *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufmann, San Mateo, CA, 1988).
[64] J. Pearl, Reasoning with belief functions: an analysis of compatibility, *Int. J. Approx. Reasoning* **4** (1990) 363-389.
[65] J. Pearl, Rejoinder to comments on "Reasoning with belief functions: an analysis of compatibility", *Int. J. Approx. Reasoning* **6** (1992) 425-443.
[66] J.R. Quinlan, Inferno: a cautious approach to uncertain inference, *Comput. J.* **26** (1983) 255-269.
[67] R. Reiter, A logic for default reasoning, *Artif. Intell.* **13** (1980) 81-132.
[68] R. Reiter, Nonmonotonic reasoning, *Ann. Rev. Comput. Sci.* **2** (1987) 147-186.
[69] A. Saffiotti and E. Umkehrer, PULCINELLA: a general tool for propagating uncertainty in valuation networks, in: B.D. D'Ambrosio, P. Smets and P.P. Bonissone, eds., *Proceedings Seventh International Conference on Uncertainty in AI*, Los Angeles, CA (Morgan Kaufmann, San Mateo, CA, 1991) 323-331.
[70] G.L.S. Shackle, *Decision, Order, and Time in Human Affairs* (Cambridge University Press, Cambridge, 1969).
[71] G. Shafer, *A Mathematical Theory of Evidence* (Princeton Universtiy Press, Princeton, NJ, 1976).
[72] G. Shafer, Constructive probability, *Synthese* **48** (1981) 1-60.
[73] G. Shafer, Belief functions and parametric models (with discussion), *J. Roy. Stat. Soc. Ser. B* **44** (1982) 322-352.
[74] G. Shafer, Probability judgement in artificial intelligence and expert systems, *Stat. Sci.* **2** (1987) 3-16.
[75] G. Shafer, Perspectives on the theory and practice of belief functions, *Int. J. Approx. Reasoning* **4** (1990) 323-362.
[76] G. Shafer, Rejoinders to comments on "Perspectives on the theory and practice of belief functions", *Int. J. Approx. Reasoning* **6** (1992) 445-480.
[77] G. Shafer and R. Logan, Implementing Dempster's rule for hierarchical evidence, *Artif. Intell.* **33** (1987) 271-298.
[78] P. Shenoy and G. Shafer, Propagating belief functions with local computations, *IEEE Expert* **1** (3) (1986) 43-52.
[79] F.K.J. Sheridan, A survey of techniques for inference under uncertainty, *Artif. Intell. Rev.* **5** (1991) 89-119.
[80] E.H. Shortliffe, *Computer-Based Medical Consultations: MYCIN* (Elsevier, New York, 1976).
[81] E.H. Shortliffe and B.G. Buchanan, A model of inexact reasoning in medicine, *Math. Biosci.* **23** (1975) 351-379.
[82] P. Smets, Belief functions, in: P. Smets, A. Mamdani, D. Dubois and H. Prade, eds., *Non-Standard Logics for Automated Reasoning* (Academic Press, London, 1988) 253-286.
[83] P. Smets, The transferable belief model and other interpretations of Dempster-Shafer's model, in: P.P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **6** (North-Holland, Amsterdam, 1991) 375-383.
[84] P. Smets, Resolving misunderstandings about belief functions, *Int. J. Approx. Reasoning* **6** (1992) 321-344.
[85] P. Smets and R. Kennes, The transferable belief model, *Artif. Intell.* **66** (1994) 191-234.
[86] P. Smets, A. Mamdani, D. Dubois and H. Prade, eds., *Non-Standard Logics for Automated Reasoning* (Academic Press, London, 1988).
[87] C.A.B. Smith, Consistency in statistical inference and decision (with discussion), *J. Roy. Stat. Soc. Ser. B* **23** (1961) 1-37.
[88] M. Smithson, *Ignorance and Uncertainty* (Springer, Berlin, 1989).
[89] D.J. Spiegelhalter, A statistical view of uncertainty in expert systems, in: W.A. Gale, ed., *Artificial Intelligence and Statistics* (Addison-Wesley, Reading, MA, 1986).
[90] D.J. Spiegelhalter, A.P. Dawid, S.L. Lauritzen and R.G. Cowell, Bayesian analysis in expert systems (with discussion), *Stat. Sci.* **8** (1993) 219-283.
[91] D.J. Spiegelhalter and R.P. Knill-Jones, Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology (with discussion), *J. Roy. Stat. Soc. Ser. A* **147** (1984) 35-76.

[92] T.M. Strat, Decision analysis using belief functions, *Int. J. Approx. Reasoning* **4** (1990) 391–417.
[93] C. Sundberg and C. Wagner, Generalized finite differences and Bayesian conditioning of Choquet capacities, Unpublished manuscript (1991).
[94] B. Tessem, Interval probability propagation, *Int. J. Approx. Reasoning* **7** (1992) 95–120.
[95] F. Voorbraak, On the justification of Dempster's rule of combination, *Artif. Intell.* **48** (1991) 171–197.
[96] P. Walley, Coherent lower (and upper) probabilities, Statistics Research Report, University of Warwick, Coventry (1981).
[97] P. Walley, The elicitation and aggregation of beliefs, Statistics Research Report, University of Warwick, Coventry (1982).
[98] P. Walley, Belief function representations of statistical evidence, *Ann. Stat.* **15** (1987) 1439–1465.
[99] P. Walley, *Statistical Reasoning with Imprecise Probabilities* (Chapman and Hall, London, 1991).
[100] P. Walley, Measures of uncertainty in expert systems, Statistics Research Report, Department of Mathematics, University of Western Australia, Perth (1991).
[101] P. Walley, Inferences from multinomial data: learning about a bag of marbles (with discussion), *J. Roy. Stat. Soc. Ser. B* **58** (1996) 3–57.
[102] P. Walley, Statistical inferences based on a second-order possibility distribution (with discussion), *J. Am. Stat. Assoc.* **91** (1996).
[103] P. Walley and F.M. Campello de Souza, Uncertainty and indeterminacy in assessing the economic viability of energy options: a case study of solar heating systems in Brazil, *Energy Syst. Policy* **14** (1990) 281–304.
[104] P. Walley and T.L. Fine, Varieties of modal (classificatory) and comparative probability, *Synthese* **41** (1979) 321–374.
[105] P. Walley and T.L. Fine, Towards a frequentist theory of upper and lower probability, *Ann. Stat.* **10** (1982) 741–761.
[106] T.S. Wallsten, D.V. Budescu, A. Rapoport, R. Zwick and B. Forsyth, Measuring the vague meanings of probability terms, *J. Exper. Psych. General* **115** (1986) 348–365.
[107] L.A. Wasserman, Comments on Shafer's "Perspectives on the theory and practice of belief functions", *Int. J. Approx. Reasoning* **6** (1992) 367–375.
[108] L.A. Wasserman and J. Kadane, Bayes' theorem for Choquet capacities, *Ann. Stat.* **18** (1990) 1328–1339.
[109] S.R. Watson, J.J. Weiss and M.L. Donnell, Fuzzy decision analysis, *IEEE Trans. Syst. Man Cybern.* **9** (1979) 1–9.
[110] P.M. Williams, Indeterminate probabilities, in: M. Przelecki, K. Szaniawski and R. Wojcicki, eds., *Formal Methods in the Methodology of Empirical Sciences* (Reidel, Dordrecht, 1976) 229–246.
[111] N. Wilson, A Monte-Carlo algorithm for Dempster-Shafer belief, in: B.D. D'Ambrosio, P. Smets and P.P. Bonissone, eds., *Proceedings Seventh International Conference on Uncertainty in AI*, Los Angeles, CA (Morgan Kaufmann, San Mateo, CA, 1991) 414–417.
[112] N. Wilson, The combination of belief: when and how fast?, *Int. J. Approx. Reasoning* **6** (1992) 377–388.
[113] N. Wilson and S. Moral, A logical view of probability, in: A. Cohn, ed., *Proceedings Eleventh European Conference on Artificial Intelligence* (Wiley, London, 1994) 386–390.
[114] H. Xu, An efficient implementation of belief function propagation, in: B.D. D'Ambrosio, P. Smets and P.P. Bonissone, eds., *Proceedings Seventh International Conference on Uncertainty in AI*, Los Angeles, CA (Morgan Kaufmann, San Mateo, CA, 1991) 425–432.
[115] R.R. Yager, An introduction to applications of possibility theory (with discussion), *Human Syst. Manage.* **3** (1983) 246–269.
[116] L.A. Zadeh, Fuzzy sets, *Inf. Control* **8** (1965) 338–353.
[117] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning (Part 3), *Inf. Sci.* **9** (1976) 43–80.
[118] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets Syst.* **1** (1978) 3–28.
[119] L.A. Zadeh, A theory of approximate reasoning, in: J. Hayes, D. Michie and L.I. Mikulich, eds., *Machine Intelligence* **9** (Halstead Press, New York, 1979) 149–194.
[120] L.A. Zadeh, The role of fuzzy logic in the management of uncertainty in expert systems, *Fuzzy Sets Syst.* **11** (1983) 199–227.

[121] L.A. Zadeh, Is probability theory sufficient for dealing with uncertainty in AI: a negative view, in: L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1986) 103–116.