Robust Bayesianism: Imprecise and Paradoxical Reasoning

Stefan Arnborg
Kungl Tekniska Högskolan
Stockholm
SE-100 44
Sweden
stefan@nada.kth.se

Abstract – *We are interested in understanding the relationship between Bayesian inference and evidence theory, in particular imprecise and paradoxical reasoning. The concept of a set of probability distributions is central both in robust Bayesian analysis and in some versions of Dempster-Shafer theory. Most of the literature regards these two theories as incomparable. We interpret imprecise probabilities as imprecise posteriors obtainable from imprecise likelihoods and priors, both of which can be considered as evidence and represented with, e.g., DS-structures. The natural and simple robust combination operator makes all pairwise combinations of elements from the two sets. The DS-structures can represent one particular family of imprecise distributions, Choquet capacities. These are not closed under our combination rule, but can be made so by rounding. The proposed combination operator is unique, and has interesting normative and factual properties. We compare its behavior on Zadeh's example with other proposed fusion rules. We also show how the paradoxical reasoning method appears in the robust framework.*

Keywords: DS-structures, Modified Dempster-Shafer rule, Capacities, Evidence theory, Likelihood, imprecise probability

1 Introduction

Several apparently incomparable approaches exist for uncertainty management. It has been a goal in research to encompass all aspects of uncertainty management in a single framework. Attaining this goal should make the topic teachable in undergraduate and graduate engineering curricula and facilitate engineering applications development. We approach the problem by asking if robust Bayesian analysis could be such a framework. The DS theory originated within Bayesian statistical analysis[1], but when developed by Shafer[2] took the concept of belief assignment as primitive. The assumption being that bodies of evidence - probabilistic statements about the possible worlds of interest - can be taken as primitives rather than sampling functions and priors. When the connection to Bayes method and Dempster's application is broken, it is no longer necessary to use the Dempster combination rule, and evidence theory abounds with proposals on how bodies of evidence should be interpreted and combined. But there seems not to exist other bases for obtaining bodies of evidence than likelihoods and priors, and therefore an analysis of a hypothetical Bayesian obtainment of bodies of evidence can bring light to problems in evidence and aggregation theory. Particularly, a body of evidence represented by a DS-structure (bpa) has an interpretation as a set of possible probability distributions, and combining or aggregating two such structures can be done in robust Bayesian analysis. The resulting combination operator is trivial, but compared to other similar operators it has interesting behavior and normative advantages. It appears to be missing in recent overviews of evidence and imprecise probability theory. Our ideas are closely related to problems discussed in [3] and in the recent and voluminous report[4], which also contains a quite comprehensive bibliography. In section 2 we review Bayesian and robust Bayesian analysis and some of its relations to DS theory; in section 3 we discuss Zadeh's example. In section 4 we derive the robust combination operator and we apply it in section 5 to Zadeh's problem and in section 6 to the paradoxical fusion principle.

2 Bayesian analysis

Bayesian analysis is usually explained[5, 6, 7] using the formula

$$f(\lambda|x) \propto f(x|\lambda)f(\lambda), \tag{1}$$

where $\lambda \in \Lambda$ is the world of interest among $n = |\Lambda|$ possible worlds (sometimes called parameter space), and $x \in X$ is an observation among possible observations. The distinction between observation and world space is not necessary but is convenient - it indicates what our inputs are (observations) and what our outputs are (belief about possible worlds). The functions in the formula are probability distributions, discrete or continuous. The sign $\propto$ indicates that the left side is proportional to the right side (as a function of λ), with the normalization constant left out. In equation (1), $f(x|\lambda)$ is a sampling function which connects observation space and possible world space by giving a probability distribution of observed value for each possible world, and $f(\lambda)$ is a prior describing our expectation on what the world might be. The rule (1) gives the distribution $f(\lambda|x)$ over possible worlds λ conditional on observations x. A paradox arises if the supports of $f(\lambda)$ and $f(x|\lambda)$ are disjoint(since each possible world is ruled out either by the prior or by the likelihood), a possibility we will ignore throughout this paper. Equation (1) is free of technical complication and easily explainable. It generalizes however to surprisingly complex settings, as required of any

device helpful in design of complex technical systems. Ed Jaynes made (1) the basis for teaching science and interpretation of measurements[5], an idea that caught on well by students devouring his unfinished lecture notes on the web. In general, for infinite (compact metric) observation spaces or possible world sets, some measure-theoretic caution is called for, but it is also possible to base the analysis on well-defined limit processes in each case as pointed out by, among others, Jaynes[5]. We will here assume Jaynes' approach and discuss thus only the finite case. Equation (1) is valid under the assumption that observations are selected and missing 'at random', i.e., not dependent on world state except through recorded observations. We will assume this throughout. When selection is made based on unrecorded circumstances, we have *selection bias* which can and should be entered into the statistical model. Ways of handling data selection biases are discussed thoroughly in [7]. In sensor management, selection bias occurs when the 'reason' for directing sensors or excluding observations is not recorded in the probabilistic model – it is however normally assumed that this problem does not exist, and maybe it doesn't for the more rational methods of sensor management.

It has been an important philosophical question to characterize the scope of applicability of (1), which lead to the distinction between objective and subjective probability, among other things. Several books and papers, among others [8, 9, 10, 11], claim that, under reasonable assumptions, (1) is the only consistent basis for uncertainty management. However, the minimal assumptions truly required to obtain this result turn out on closer inspection to be rather complex, for a condensed overview see[12]. One simple assumption usually made in those studies is that uncertainty is measured by a real number or on an ordered scale. Many established uncertainty management methods however measure uncertainty on a partially ordered scale and do apparently not use (1) and the accompanying philosophy. Among probability based alternatives to Bayesian analysis with partially ordered uncertainty concepts are imprecise probabilities or lower/upper prevision theory[13], the Dempster-Shafer(DS)[2], the Fixsen/Mahler(MDS)[14] and Dezert-Smarandache(DSmT)[15] theories. In these schools, it is considered important to develop the theory without reference to classical Bayesian thinking. In particular, the assumption of precise prior and sampling distributions is considered indefensible. Those assumptions are referred to as the dogma of precision in Bayesian analysis[16].

In robust Bayesian analysis[17], one acknowledges that there can be ambiguity about the prior and sampling distributions, and it is accepted that a convex set of such distributions is used in inference. It is possible that all consistent interval-based uncertainty management schemes (where uncertainty is described by an interval of real numbers) can be explained as robust Bayesian analysis, but as of now there appears to be no truly convincing argument for this. The idea of robust Bayesian analysis goes back to the pioneers of Bayesian analysis[8, 18], but the computational and conceptual complexities involved meant that it could not be fully developed in those days. Instead, a lot of effort went into the idea of finding a canonical and unique prior, an idea that seems to have failed except for finite problems with some kind of symmetry, where a natural generalization of Bernoulli's indifference principle has become accepted. The problem is that no proposed priors are invariant under arbitrary rescaling of numerical quantities or non-uniform coarsening or refinement of the current frame of discernment.

Convex sets of probability distributions can be arbitrarily complex. Such a set can be generated by mixing of a set of 'corners' (called simplices in linear programming theory) and the set of corners can be arbitrarily large already for sets of probability distributions over three elements (the family is representable by the set of convex regions in the lower left half of the unit square). In evidence theory, the concept of DS-structure is a representation of a belief over a frame of discernment (possible worlds) by a probability distribution over its powerset, a basic probability assignment bpa, bba or DS-structure (terminology is not stable). Even if it is considered important in many versions of DS theory not to equate a DS-structure with a set of possible distributions, such a perspective is prevalent in tutorials and almost unavoidable in a teaching situation. The DS-structure thus represents all distributions over the set obtainable by reallocation the probability of each non-singleton set A to the singleton members of A. Such a set of distributions is a type of *Choquet capacity*, and these capacities form a particularly concise and flexible family of sets of distributions. These sets will be spanned by at most $\prod_{k=1}^{n} k^{\binom{n}{k}}$ distributions. They can be represented by $n-1$ real numbers - the corresponding DS-structure (whereas an arbitrary convex set can need any number of distributions to span it and needs an arbitrary number of reals to represent it - thus Choquet capacities form a proper and really small subset of all convex sets of distributions). It is definitely possible to introduce more complex but still consistent uncertainty management by going beyond robust Bayesianism, grading the families of distributions and introducing rules on how the grade of combined distributions are obtained from the grades of their constituents. The grade would in some sense indicate how plausible a distribution in the set is. But if the grade is interpreted as a probability distribution over probability distributions, no expressive power is gained. This results in hierarchical Bayesian analysis[7]. Nevertheless, instead of using possibly unnecessarily complex uncertainty methodology, it appears more promising to put efforts into understanding complexly structured observation and possible world spaces, as brought home convincingly in, e.g., [19], where – among other things –the notoriously difficult multiple tracking problem was captured as an inference problem using a dynamic version of (1) with rather complex observation and possible world spaces. A similar development has taken place in genetics, where an unknown number of significant genetic loci are assumed involved as causes of a phenotype like a hereditary disease – and inference aims at finding the number of loci. Finally, in multi-agent systems we must consider the possibility of a gaming component, where an agent must be aware of the possible reasoning processes of other agents, and use information

about their actions and goals to decide its own actions. In this case there appears to be no simple way to separate – as there is in a single agent setting – the uncertainty domain (what is happening?) from the decision domain (what shall I do?) because these get entangled by the uncertainties of what other agents will believe, desire and do. This problem can be approached by game-theoretic analyses[20].

A Bayesian data fusion system or subsystem can thus use any level in a ladder with increasing complexity, where each level could be augmented by a gaming component:

- Logic - no quantified uncertainty
- Precise Bayesian fusion
- Robust Bayesianism with Choquet capacities
- General robust Bayesianism (or lower/upper previsions)
- Robust Bayesianism with graded sets of distributions

The ultimate use of data fusion is usually decision making. Precise Bayesianism results in quantities that can be used immediately for expected utility decision making[21]. For the more complex uncertainty representations one uses either minimax criteria or estimates a precise probability distribution to decide from. The latter is a core idea in the transferable belief model, with so-called pignistic transforms[22]. In robust Bayesian analysis, the maximum entropy distribution in a set is often used as an estimate[5]. This choice can be given a decision-theoretic motivation since it minimizes a game-theoretic loss function, and can also be generalized to a range of loss functions[23].

Whether or not this simplistic view (ladder of Bayesianisms) on uncertainty management is tenable in the long run in an educational or philosophical sense is currently not settled.

3 Zadeh's example

We will discuss our problem in the context of Zadeh's example, described and discussed, for example, in[15], of two physicians who investigated a patient independently. The two physicians agree that the problem (the diagnosis of the patient) is within the set $\{M, C, T\}$, where M is Meningitis, C is Concussion and T is brain Tumor. However, they express their beliefs differently, as a probability distribution which is $(0.99, 0, 0.01)$ for the first physician and $(0, 0.99, 0.01)$ for the second. The question is what a third party can say about the patients condition with no more information than that given. This example has been discussed a lot in the literature, see e.g. [15]. It is a classical example on how two independent sets of observations can together eliminate cases to end up with a case not really indicated by any of the two sets in separation. Several such examples have been brought up as good and prototypical in the Bayesian literature, e.g., in [5]. However, in the evidence theory literature the Bayesian solution (also obtained from using Dempster's rule) has been considered inadequate and this particular example has been the starting point for several proposals of alternative fusion rules.

The following are reactions I have met from professionals – physicians, psychiatrists, teachers and military commanders – confronted with similar problems. They are also prototypical for current discussions on evidence theory.

- One of the experts probably made a serious mistake.
- These young men seem not to know what probability zero means, and should be sent back to school.
- It is completely plausible that one eliminated M and the other C in a sound way. So T is the main alternative, or rather T or something else, since there are most likely more possibilities left.
- The assessment for T is probably based mostly on prior information (rareness), so the combined judgment should not make T less likely, rather the opposite.
- An investigation is always guided by the patients subjective beliefs, and an investigation affects those beliefs. This is a possible explanation for the Ulysses syndrome, where persons are seen to embark on endless journeys through the health care system. This view would call for a game-theoretic approach (with parameters difficult to assess).

What the example reactions learn us is that subjects confronted with paradoxical information typically start building their own mental models about the case and insist on bringing in more information, in the form of information about the problem area, the observation protocols underlying the assessments, or a new investigation. The professionals handling of the information problem is usually rational enough, but very different conclusions arise from small differences in mental models.

Similar reactions were observed already by C.S. Pierce in his studies of the human inference process and its relation to logic and emotions[24]. Entertaining essays on this theme can be found in[25, 26]. The person who gets the fusion problem regards the two beliefs expressed as two abstracted observation sets, and tries to understand their combined bearing on patient state. If she feels that the observation sets ought to be similar because of the professional training and standardized operating procedures of the experts she gets worried, otherwise not.

4 Fusion in the Bayesian framework

From a Bayesian point of view, one would analyze Zadeh's and similar problems using an observation space O and a possible world space Λ. The observations are actually sets of observations, test results and interview responses, so O is the powerset of another set X of possible observations. In the case of the example, the world state λ would include all factors that determine the distribution of observation results for the patient. So if physician 1 obtained observation set $X_1 \subseteq X$ and physician 2 obtained observation set $X_2 \subseteq X$, they would obtain a posterior belief of the patients condition expressible as $f_i(\lambda_i|X_i) \propto f_i(X_i|\lambda_i)f_i(\lambda_i)$, for

$i = 1, 2$. Here we have not assumed that the two physicians used the same sampling and prior distributions. Even if training aims at giving the two physicians the same 'knowledge' in the form of sampling function and prior, this ideal cannot be achieved completely in practice. We do assume that the two physicians share the possible world set, since otherwise we would have to make at least some assumptions on their correspondence in order to obtain any type of interesting fusion. In any case, the inference stipulated by the Bayesian method is that physician i states the probability distribution $f_i(\lambda_i|X_i)$ as his belief about the patient. If they use the same sampling function and prior, the Bayesian method also allows them to combine their findings to obtain:

$$f(\lambda|\{X_1, X_2\}) \propto f(\{X_1, X_2\}|\lambda)f(\lambda) = f(X_1|\lambda)f(X_2|\lambda)f(\lambda), \quad (2)$$

under the assumption:

$$f(\{X_1, X_2\}|\lambda)f(\lambda) = f(X_1|\lambda)f(X_2|\lambda). \quad (3)$$

The assumption appears reasonable in many cases under the assumption of no selection bias or other interference, and an adequately fine-grained possible world set Λ.

It is important to observe that it is the two physicians likelihood functions, not their posterior beliefs, that can be combined, otherwise we would replace the prior by its normalized square which means that its mode would get a too large influence and the real uncertainty would be underestimated. This is at least the case if they obtained their training from a common body of medical experience coded in textbooks. To the extent they both obtained their priors in independent practice, they should however be combined. If the posterior is reported and we happen to know the prior, the likelihood can be obtained by $f(X|\lambda) \propto f(\lambda|X)/f(\lambda)$. The posterior, likelihood, and prior can be viewed (after normalization in the case of the likelihood functions) as probability distributions or as random sets having a singleton value. It is interesting to note (as is well known[19]) that the combination rule is that the posterior of the combined evidence can be expressed as the (nonempty) set intersection of the (singleton) random sets describing the prior and the two likelihoods.

The Dempster-Shafer combination rule[2] is computationally equivalent to allowing the operands as well as the result in this combination to be nonempty, not necessarily singleton, random sets. By this statement we do not claim that this is the way DS theory is usually motivated. But the model in which Dempster's rule is motivated[27] is different from ours: there it is assumed that each source has its own possible world set, but precise beliefs about it. The impreciseness results only from a multivalued mapping, ambiguity in how the sources information should be translated to a common frame of discernment. In Dempster's model the random set intersection is the required result of fusion. But on closer inspection of his application example, the impreciseness of sources incurred by the multivalued mapping is easily interpretable as an imprecise probability distribution and it seems not at all clear why the natural robust fusion operator was not chosen. The recently introduced Fixsen/Mahler MDS combination rule[14] involves a re-weighting of the terms involved in the set intersection operation: Whereas Dempster's combination rule can be expressed as $m_{DS}(A) \propto \sum_{A=B\cap C} m_1(B)m_2(C)$ (where $A \neq \Theta$), the MDS rule is $m_{MDS}(A) \propto \sum_{A=B\cap C} m_1(B)m_2(C)|A|/(|B||C|)$. The MDS and DS rules are identical to Bayesian fusion for precise distributions, but when both operands are imprecise, the MDS rule seems to have a fundamental advantage over the DS rule, as we shall see in section 5.

A set of distributions which is not a Choquet capacity can be approximated by *rounding* it to a minimal Choquet capacity that contains it (see Fig. 1), and this rounded set can be represented by a DS-structure. (Figures represent sets of three-item pdf:s, by projection on two items). It is also possible, using linear programming, to round downwards to a maximal Choquet capacity contained in a set. Neither type of rounding is unique. For large frames it will also be necessary to constrain the focal elements to a subset of the powerset.

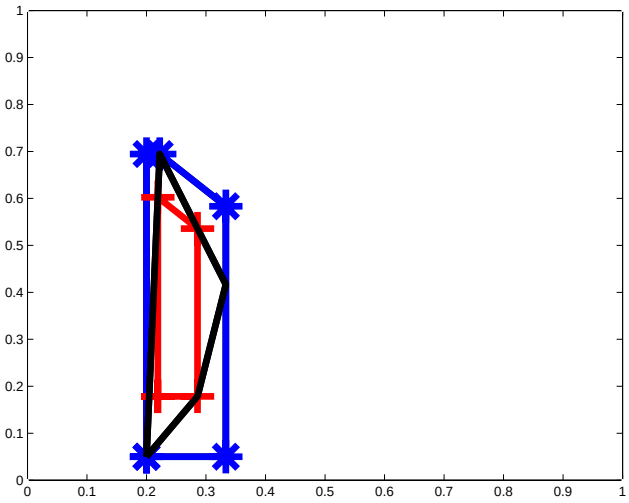

Fig. 1: Rounding a set of distributions over three items. A set spanned by four corner distributions (black), one of its minimal enclosing (blue *), and one of its maximal enclosed (red +), Choquet capacities.

Our approach to impreciseness is that impreciseness in conclusions is caused by impreciseness in sampling functions and priors. In terms of sampling functions used, one can assume that these are imprecise, non-stationary, estimated with bias, or that they vary within atoms of the current frame of discernment which is too coarse for reliably combining precise probability assessments (inhomogeneity). Impreciseness in priors is caused by lack of information about the processes involved. An assessment that the sampling function (actually the product of all sampling functions used in the assessment) is imprecise gives the same effect on the body of evidence, regardless of what the reason is. The difficulty lies in assessing the magnitude of impreciseness, and this difficulty is somewhat unavoidable.

It seems not to be different from the problems of assessing subjective probability or belief.

The imprecise distributions we use can, if constrained by rounding to Choquet capacities, be viewed as random sets. The corresponding random sets can be combined as before: take the intersection of the participating random sets and condition on the result being non-empty. The resulting random set can be regarded as a Choquet capacity, the set of possible distributions for λ. It has been argued in several papers, among others [28], that a random set union is more appropriate than Dempster's rule as a combination rule, and indeed a large number of alternative combination rules have been proposed over the years. Another alternative to Dempster's rule is Yager's rule[29]. For recent surveys see [4, 30].

The combination of evidence – likelihood functions normalized so they can be seen as probability distributions – and a prior over a finite space is thus done simply by component-wise multiplication followed by normalization. The resulting combination operation agrees with the DS and the MDS rules for precise beliefs. The robust Bayesian version of this would replace the probability distributions by sets of probability distributions, for example represented as DS beliefs. The most obvious combination rule would yield the set of probability functions that can be obtained by taking one member from each set and combine them. Intuitively, membership means that the distribution can possibly be right, and we would get the final result, a set of distributions that can be obtained by combining a number of distributions each of which could possibly be right. The combination rule (2) would thus take the form (where F denotes convex families of functions):

$$F(\lambda|\{X_1, X_2\}) \propto F(\{X_1, X_2\}|\lambda) \times F(\lambda) = F(X_1|\lambda) \times F(X_2|\lambda) \times F(\lambda). \quad (4)$$

Definition 1 *The* robust Bayesian combination operator $\times$ *combines two sets of probability distributions over a common space* Λ. *The value of* $F_1 \times F_2$ *is* $\{cf_1f_2 : f_1 \in F_1, f_2 \in F_2, c = 1/\sum_{\lambda\in\Lambda} f_1(\lambda)f_2(\lambda)\}$

The operator can easily be applied to give too much impreciseness: The impreciseness of likelihood functions has typically a number of sources, and the proposed technique can give too large uncertainties when these sources do not have their full range of variation within the evidences that will be combined. A most extreme example is the sequence of plots returned by a sensor: variability can have its source in the target, in the sensor itself, and in the environment. But when a particular sensor follows a particular target, the variability of these sources are not fully materialized. The variability has its source only in state (distance, inclination, etc) of target, so it would seem wasteful to assume that each new plot comes from an arbitrarily selected sensor and target. This and similar problems are inherent in system design, and can be addressed by detailed analyses of sources of variation, if such are feasible.

The definition of the robust Bayesian combination operator involves infinite sets in general and is not computable directly. For singleton sets is is easily computed, though. Using this we can immediately combine operands each of which is generated by mixing of a finite set of corners:

Theorem 1 *If* $F_1 = \{\sum_{i\in I} c_i g_i : 0 \le c_i, \sum_{i\in I} c_i = 1\}$ *and* $F_2 = \{\sum_{j\in J} c_j h_j : 0 \le c_j, \sum_{j\in J} c_j = 1\}$, *then* $F_1 \times F_2 = \{\sum_{i\in I, j\in J} c_{ij}\{g_i\} \times \{h_j\} : 0 \le c_{ij}, \sum_{ij} c_{ij} = 1\}$.

Proof hint: Let the *cone* of a pdf set be the set of non-negative scalings of its members. Consider obtaining the cone of the combination by unnormalized combination of the cones of the operands.

This theorem gives the method for implementation of the robust operator. After the potential corners of the result have been obtained, a convex hull computation as found, e.g., in MATLAB and OCTAVE, is used to tesselate the boundary and remove those points falling in the interior of the polytope. We can now make a few statements, most of which are mentioned in [1, Discussion by Atkinson],[3], about fusion in the robust Bayesian framework:

- The combination operator is associate and commutative, since it inherits these properties from the multiplication operator it uses.
- Precise beliefs combined gives the same result as Dempster's rule and yield new precise beliefs.
- A precise belief combined with an imprecise belief will yield an imprecise belief in general - thus Dempster's rule underestimates imprecision compared to the robust operator.
- Ignorance is represented by a uniform precise belief, not by the vacuous assignment of DS-theory.
- The vacuous belief is a belief that represents total skepticism, and will when combined with anything yield a new vacuous belief (it is thus an absorbing element). This belief has limited use in the robust Bayesian context.
- Dempster's rule is clearly inadequate for combining the vacuous belief with anything, but here the union rule gives the 'right' answer.

So it seems that none of the established combination rules captures the idea of robust Bayesian analysis. Why is the robust combination operator not considered an interesting option? One possible answer is that our proposed combination is not closed under restriction to Choquet capacities. The more imprecise evidence we have combined, the more corners will we need to span the result, and Choquet capacities only allow for a bounded number of these. Some type of approximation is required if we want to stay within the belief function framework. The most natural approximation is rounding. In a sense we fit the right answer into our constraints by creating more – possibly too much – impreciseness.

Definition 2 *A* rounded robust Bayesian combination operator *combines two sets of probability distributions over a common space* Λ. *The robust operation is applied to the rounded operands, and the result is then rounded.*

An important and distinguishing property of the robust rule is:

Observation 1 *The robust and rounded robust operators are* monotone with respect to imprecision, *i.e., if* $F_i' \subseteq F_i$, *then* $F_1' \times F_2' \subseteq F_1 \times F_2$.

Theorem 2 *For any combination operator* $\times'$ *that is monotone* wrt *imprecision and is equal to the Bayesian (Dempster's) rule for precise arguments,* $F_1 \times F_2 \subseteq F_1 \times' F_2$, *where* $\times$ *is the robust rule.*

Proof outline: By contradiction; assume thus there is an $f \in F_1 \times F_2$ with $f \notin F_1 \times' F_2$. By the definition of $\times$, $f = \{f_1\} \times \{f_2\}$ for some $f_1 \in F_1$ and $f_2 \in F_2$. But then $f = \{f_1\} \times' \{f_2\}$, and since $\times'$ is monotone *wrt* imprecision, $f \in F_1 \times' F_2$, a contradiction.

5 The robust combination operator on Zadeh's example

The example of Zadeh can be seen as a classical inference where the case T is inferred by elimination of all alternatives. This must be possible in any useful uncertainty management scheme. We will illustrate the robust combination rule by comparing it with standard combination operators from the literature, on Zadeh's example and on two versions of it where we discounted the physicians assessments. In order to illustrate the result graphically we change the example so that the T alternative has probability 0.1 instead of 0.01 in the two bodies of evidence.

Assume thus that we have obtained information that physician 2 used a set of tests to eliminate Meningitis which is unreliable in the sense that there are types of this disease – unfortunately with unknown frequency – that will only be eliminated with probability 0.9, whereas other types can be eliminated with probability 1. These tests have no bearing on distinguishing C from T. This means that there are persons with Meningitis of this type that will test negative with probability 0.1. Since we have no prior information on the frequencies of these types, and since physician 2 has reported a precise body of evidence, our conclusion is that his assessment of Meningitis should be the interval $(0, 0.1)$ instead of the value 0. The relationship between T and C should not be altered, since the tests used for Meningitis have no discriminating power here. So the discounted assessment should be $(k, 0.9 * (1-k), 0.1 * (1-k))$, for some $k \in [0, 0.1]$. This set is spanned by the distributions $(0, 0.9, 0.1)$ and $(0.1, 0.81, 0.09)$. It cannot be represented as a DS-structure, but can be rounded to $\{m(T) = 0.09, m(C) = 0.81, m(MT) = 0.01, m(CM) = 0.09\}$.

In figure 2 we have combined the original precise assessments. Dempster's rule and the robust rule give the same result, $(0, 0, 1)$, as we expect. In figure 3 we discounted physician 2. The Dempster combination moved all the way from $(0, 0, 1)$ to $(0.9, 0, 0.1)$, and the MDS rule to $(0.8257, 0, 0.1743)$. The robust rule gives a result spanned by $(0, 0, 1)$ and $(0.9091, 0, 0.0909)$. This reflects real uncertainty correctly in the Bayesian interpretation and also shows that T and M are both completely plausible while C

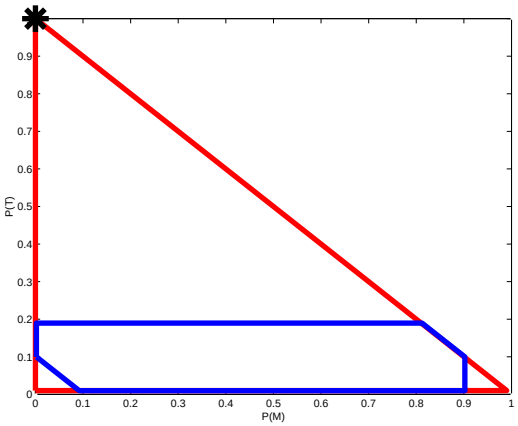

Fig. 2: Combing two precise probability evidences in Zadeh's example. Dempster's rule and the robust combination rule are the same, $P(T) = 1$, $P(M) = 0$ (black *). The disjunctive rule (blue) gives little possibility of T, and Yager's rule (red) is non-informative

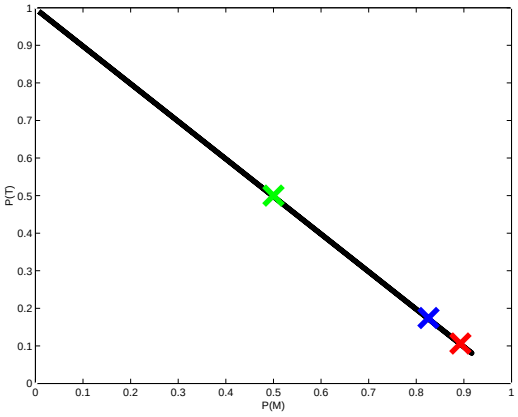

Fig. 3: Combining imprecise and precise evidence in Zadeh's modified example (discounting physician 2). robust rule: black line; Dempster's rule: red; MDS rule: blue; maxent estimate: green.

is not, since it was eliminated by the first physician, who is not yet discounted. Moreover, new information affecting the credibility of T will also affect M, and vice versa.

In figure 4 we discounted both physicians, but only by 5% instead of the one physician discounted by 10% in the last example. The Dempster combination moved a long way again, to the line spanned by $(0.4386, 0.4593, 0.1021)$ and $(0.4593, 0.4386, 0.1021)$. This line touches the boundary of the robust combination result, which is now not a Choquet capacity (black in figure) but has a good rounded approximation (cyan in figure). The MDS result lies well inside the robust rule result. The maximum entropy estimate for the robust rule result is the non-informative distribution $(1/3, 1/3, 1/3)$. When interpreting DS-structures

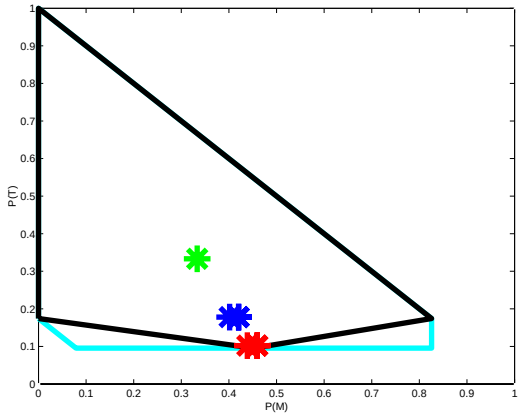

Fig. 4: Combining imprecise and precise evidence in Zadeh's modified example, both experts discounted. red: Dempster's rule; blue: MDS rule; green:maxent estimate; black: robust rule; cyan: rounded robust rule.

as Choquet capacities in the natural way (this interpretation can be found in quite many tutorials of DS theory and is present, somewhat implicitly, in [2]), it is highly desirable that the combination of evidence gives a capacity that is contained in, or at least not disjunct from, the robust rule result. The MDS rule is designed so that the pignistic transform (reallocating the mass of every non-singleton focal element A uniformly over the members of A) of the result is the result of Bayesian fusion of the pignistic transforms of the operands[14]. Therefore, the results of MDS and robust Bayesian fusion always intersect. It is also not difficult to see that the MDS result, viewed as a capacity, is contained in the robust Bayesian fusion result.

Varying the parameters of discounting a little in our example, it is not difficult find cases where Dempster's rule gives a capacity disjoint from the robust rule result. A simple Monte Carlo search indicates that disjointness does indeed happen in general, but infrequently. Typically, Dempster's rule gives an uncertainty polytope that is clearly narrower than that of the robust rule, and enclosed in it. In figure 5 we show an example where this is not the case and the result is somewhat paradoxical. It is paradoxical in the sense that a person viewing DS-structures as capacities would find some bets on the outcome clearly advantageous if he used Dempster's rule but disadvantageous if he used the robust Bayesian rule. This seems to be a quite compelling argument in favor of the MDS rule, where this cannot happen, or against the habit of explaining DS structures with the capacity interpretation.

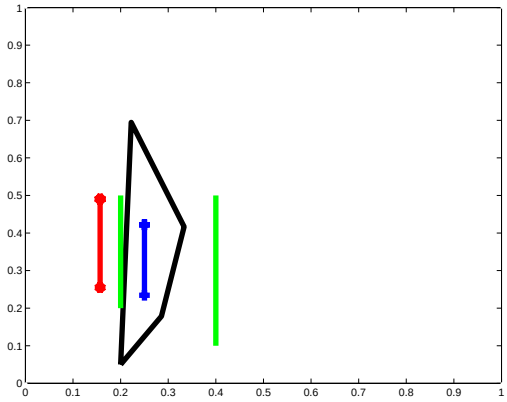

Fig. 5: A case were the robust rule and Dempster's rule give paradoxical results. The operands are shown in green, the result of the robust combination rule is shown in black (same as in figure 1), Dempster's rule gives the result shown in red *, the Fixsen-Mahler MDS rule shown in blue +.

Of course, one question remains: the DS and MDS operators are clearly not monotone *wrt* imprecision. This means that they either underestimate imprecision or eliminate imprecision in a way that can not easily be defended, since it is a by-product of the somewhat arbitrary random set interpretation. The maximum entropy principle can be given a rational game theoretic interpretation, and gives a quite different result in many cases.

6 Paradoxical Bayesian Reasoning

The DSmT theory of Dezert and Smarandache has given a new way to resolve conflicting beliefs[15]. The main idea is that the frame of discernment (possible world set) Θ is expanded to the set D^Θ of symbolic expressions in the original frame using $\cap$ and $\cup$ and with equivalence over the algebraic rules (associativity, commutativity and distributivity) of equivalence. In this frame there are no intersections known to be empty, so Dempster's rule translates to a rule where no normalization is required, a natural random set intersection. Zadeh's original example translates to the combination $\{m(T) = 0.01, m(C) = 0, m(M) = 0, m(C \cap M) = 0.81, m(C \cap T) = 0.09, m(M \cap T) = 0.09\}$. The fused evidence can also be combined with evidence pointing to possible emptiness of atoms in this frame, for example $C \cap M$, which would raise the plausibility of atoms containing T.

Apparently, the DSm theory combines two ideas: the elaboration of the frame of discernment, and the use of the DS combination rule. The two ideas seem orthogonal, one can thus in principle use any combination rule in the extended frame, like the MDS or robust rule. In order to see how the robust rule appears in the extended frame of Zadeh's example, consider a belief on an atom, e.g., M. In an ambiguous context, this is a belief which

can point to M, $M \cap T$, $M \cap C$ or $M \cap C \cap T$, i.e., can be represented by a mass assignment to the focal element $M \cup (M \cap T) \cup (M \cap C) \cup (M \cap C \cap T)$ in 2^Λ.

The robust Bayesian analog to DSmT fusion would expand the possible world set Λ to 2^Λ, and interpret an estimate of probability $p : \Lambda \rightarrow \mathbf{R}$ as an imprecise assignment where $p(\lambda)$ is allocated to the union of sets containing λ.

For Zadeh's example, the rounded robust combination rule yields the bpa: $\{m((T \cap C) \cup (M \cap C) \cup (M \cap T) \cup T) = 1\}$, a rather uninformative conclusion. The reason that results become rather vague with the robust rule is apparently that it is significantly more conservative (possibly overstating impreciseness) than the random set intersection rule. In the robust Bayesian framework we can not draw useful conclusions unless we assume that experts, at least to some extent, know what they are talking about. This we must accomplish - as the theory predicts - with prior information somewhat damping the atoms that are intersections of the original atoms. This latter seems to be a key concept in the hybrid DSm theory. A detailed investigation of the behavior of different fusion operators in this framework must unfortunately wait, but seems a promising future project.

7 Conclusions

Despite the normative claims of evidence theory and robust Bayesianism, the two are very different in their conclusions. Further work is required for understanding the basis for assessing uncertainty objectively, so that a given problem will not have incompatible solutions in the two frameworks. The latter is particularly important for obtaining an accepted basis for teaching data fusion, where the robust and MDS rules seem to have pedagogical advantages. The teaching aspect is not limited to persuading engineers to think in certain ways. For higher level uncertainty management, dealing with quantities recognizable to users like military commanders and their teachers in their roles as evaluators, the need for clarity cannot be exaggerated.

Acknowledgments

Discussions with members of the fusion group at FOI, students in the decision support group at KTH, and colleagues at SaabTech, Karolinska Institutet and the Swedish National Defense College have been important for clarifying ideas presented above.

References

[1] A.P. Dempster. A generalization of Bayesian inference (with discussion). *J. of the Royal statistical Society B*, 30:205–247, 1968.

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

[3] J. Y. Halpern and R. Fagin. Two views of belief: belief as generalized probability and belief as evidence. *Artificial Intelligence*, 54:275–318, 1992.

[4] S. Ferson, V. Kreinovich, L. Ginzburg, D. Myers, and K. Sentz. Constructing probability boxes and Dempster-Shafer structures. Technical report, Sandia National Laboratories, 2003.

[5] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[6] D. S. Sivia. *Bayesian Data Analysis, A Bayesian Tutorial*. Clarendon Press: Oxford, 1996.

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis (Second Edition)*. Chapman & Hall, New York, 2003.

[8] B. de Finetti. *Theory of Probability*. London:Wiley, 1974.

[9] L.J. Savage. *Foundations of Statistics*. John Wiley & Sons, New York, 1954.

[10] D. V. Lindley. Scoring rules and the inevitability of probability (with discussion). *Internat. Stat. Rev*, 50:1–26, 1982.

[11] R.T. Cox. Probability, frequency, and reasonable expectation. *Am. Jour. Phys.*, 14:1–13, 1946.

[12] S. Arnborg and G. Sjödin. What is the plausibility of probability? 2002. Manuscript.

[13] P. Walley. *Statistical Reasoning with Imprecise Probability*. Chapman and Hall, 1991.

[14] D. Fixsen and R.P.S. Mahler. The modified Dempster-Shafer approach to classification. *IEEE Trans. SMC-A*, 27(1):96–104, January 1997.

[15] Jean Dezert. Foundations for a new theory of plausible and paradoxical reasoning. *Information and Security*, 9:90–95, 2002.

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

[17] J. O. Berger. An overview of robust Bayesian analysis (with discussion). *Test*, 3:5–124, 1994.

[18] Harold Jeffreys. *Scientific Inference*. Cambridge University Press, 1931.

[19] I.R. Goodman, R. Mahler, and H.T. Nguyen. *The Mathematics of Data Fusion*. Kluwer Academic Publishers, 1997.

[20] J. Brynielsson and S. Arnborg. Bayesian games for threat prediction and situation analysis. In *FUSION 2004*. International Society of Information Fusion, 2004.

[21] J. O. Berger. *Statistical decision Theory and Bayesian Analysis*. Springer-Verlag, 1985.

[22] P. Smets and R Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[23] P.D. Grünwald and A.P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 2004. to appear.

[24] J. Buchler, editor. *Philosophical Writings of Peirce*. Dover, 1955.

[25] U. Eco and T. Sebeok, editors. *The Sign of Three*. Indiana University Press, Bloomington, 1983.

[26] A. Munthe. *The story of San Michele*. London, 1929.

[27] A.P. Dempster. Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[28] H. Sung and M. Farooq. A novel justification of the disjunctive combination rule. In *FUSION 2003*, pages 1244–1251. International Society of Information Fusion, 2003.

[29] R. Yager. On the Dempster-Shafer framework and new combination rules. *Information Sciences*, 41:93–137, 1987.

[30] K. Sentz and S. Ferson. Combination of evidence in Dempster-Shafer theory. Technical report, Sandia National Laboratories, 2003.