Robust Bayesianism: Relation to Evidence Theory

STEFAN ARNBORG
Kungliga Tekniska Högskolan

We are interested in understanding the relationship between Bayesian inference and evidence theory. The concept of a set of probability distributions is central both in robust Bayesian analysis and in some versions of Dempster-Shafer's evidence theory. We interpret imprecise probabilities as imprecise posteriors obtainable from imprecise likelihoods and priors, both of which are convex sets that can be considered as evidence and represented with, e.g., DS-structures. Likelihoods and prior are in Bayesian analysis combined with Laplace's parallel composition. The natural and simple robust combination operator makes all pairwise combinations of elements from the two sets representing prior and likelihood. Our proposed combination operator is unique, and it has interesting normative and factual properties. We compare its behavior with other proposed fusion rules, and earlier efforts to reconcile Bayesian analysis and evidence theory. The behavior of the robust rule is consistent with the behavior of Fixsen/Mahler's modified Dempster's (MDS) rule, but not with Dempster's rule. The Bayesian framework is liberal in allowing all significant uncertainty concepts to be modeled and taken care of and is therefore a viable, but probably not the only, unifying structure that can be economically taught and in which alternative solutions can be modeled, compared and explained.

Manuscript received April 20, 2006; released for publication April 21, 2006.

Refereeing of this contribution was handled by Neil Gordon.

Author's address: Kungliga Tekniska Högskolan, Stockholm, SE-100 44, Sweden, E-mail: (stefan@nada.kth.se).

1. INTRODUCTION

Several, apparently incomparable, approaches exist for uncertainty management. Uncertainty management is a broad area applied in many different fields, where information about some underlying, not directly observable, truth—the state of the world—is sought from a set of observations that are more or less reliable. These observations can be, for example, measurements with random and/or systematic errors, sensor readings, or reports submitted by observers. In order that conclusions about the conditions of interest be possible, there must be some assumptions made on how the observations relate to the underlying state about which information is sought. Most such assumptions are numerical in nature, giving a measure that indicates how plausible different underlying states are. Such measures can usually be normalized so that the end result looks very much like a probability distribution over the possible states of the world, or over sets of possible world states. However, uncertainty management and information fusion is often concerned with complex technical, social or biological systems that are incompletely understood, and it would be naive to think that the relationship between observation and state can be completely captured. At the same time, such systems must have at least some approximate ways to relate observation with state in order to make uncertainty management at all possible.

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 propose here that robust Bayesian analysis is such a framework. The Dempster-Shafer or evidence theory originated within Bayesian statistical analysis [19], but when developed by Shafer [51] took the concept of belief assignment rather than probability distribution as primitive. The assumption being that bodies of evidence—beliefs about the possible worlds of interest—can be taken as primitives rather than sampling functions and priors. Although this idea has had considerable popularity, it is inherently dangerous since it seems to move application away from foundational justification. When the connection to Bayes' method and Dempster's application model 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, as a rule with convincing but disparate argumentation. 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 theory. Particularly, a body of evidence represented by a DS-structure 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, even surprising, behavior and normative advantages. Some concrete progress in working with convex sets of probability vectors has been described in [41, 57, 29]. It appears that the robust combination operator we discuss has not been analyzed in detail and compared to its alternatives, and is missing in recent overviews of evidence and imprecise probability theory. Our ideas are closely related to problems discussed in [32] and in the recent and voluminous report [21], which also contains a quite comprehensive bibliography. The Workshop hosted by the SANDIA lab has resulted in an overview of current probabilistic uncertainty management methods [34]. A current overview of alternative fusion and estimation operators for tracking and classification is given in [45].

The main objective of this paper is to propose that precise and robust Bayesian analysis are unifying, simple and viable methods for information fusion, and that the large number of methods possible can and should be evaluated by taking into account the appropriateness of statistical models chosen in the particular application where it is used. We are aware, however, that the construction of Bayesian analysis as a unifying concept has no objective truth. It is meant as a post-modernistic project facilitating teaching and returning artistic freedom to objective science. The Bayesian method is so liberal that it almost never provides unique exact solutions to inference and fusion problems, but is completely dependent on insightful modeling. The main obstacle to achieving acceptance of the main objective seems to be the somewhat antagonistic relationship between the different schools where sometimes sweeping arguments have been made that seem rather unfair whoever launched them, typical examples being [42, 51] and the discussions following them.

Another objective is to investigate the appropriateness of particular fusion and estimation operations, and their relationships to the robust as well as the precise Bayesian concept. Specifically, we show that the choice between different fusion and estimation operations can be guided by a Bayesian investigation of the application.

We also want to connect the analysis to practical concerns in information fusion and keep the mathematical/theoretical level of the presentation as simple as possible, while also examining the problem to its full depth. A quite related paper promoting similar ideas is Mahler [43], which however is terser and uses somewhat heavier mathematical machinery.

Quite many comparisons have been made of Bayesian and evidential reasoning with the objective of guiding practice, among others [47, 10, 11, 50]. It is generally found that the methods are different and therefore one should choose a method that matches the application in terms of quantities available (evidence or likelihoods and priors), or the prevailing culture and construction of the application. Although the easiest way forward, this advice seems somewhat short-sighted given the quite large lifespan of typical advanced applications and the significant changes in understanding and availability of all kinds of data during this life-span.

In Section 2 we review Bayesian analysis and in Section 3 dynamic Bayesian (Chapman Kolmogorov/Kalman) analysis. In Section 4 we describe robust Bayesian analysis analysis and some of its relations to DS theory; in Section 5 we discuss decisions under uncertainty and imprecision and in Section 6 Zadeh's well-known example. In Section 7 we derive some evidence fusion operations and the robust combination operator. We illustrate their performance on a paradoxical example related to Zadeh's in Section 8, and wrap up with conclusions in Section 9.

2. BAYESIAN ANALYSIS

Bayesian analysis is usually explained [7, 38, 52, 24] using the formula

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

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. We use a generic function notation common in statistics, so the different occurrences of f denote different functions suggested by their arguments. 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 (1), $f(x \mid \lambda)$ is a sampling distribution, or likelihood when regarded as a function of λ for a given x, 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 posterior distribution $f(\lambda \mid x)$ over possible worlds λ conditional on observations x. A paradox arises if the supports of $f(\lambda)$ and $f(x \mid \lambda)$ 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. In such systems, it is possible that x represents a quantity which is not immediately observable, but instead our information about x is given by a probability distribution $f(x)$, typically obtained as a posterior from (1). Such observations are sometimes called fuzzy observations. In this case, instead of using (1) we apply:

$$f(\lambda \mid f(x)) \propto \int f(x \mid \lambda) f(x) f(\lambda) dx. \tag{2}$$

Ed Jaynes made (1) the basis for teaching science and interpretation of measurements [38]. 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-behaved limit processes in each case as pointed out by, among others, Jaynes [38]. We will here follow Jaynes' approach and thus discuss only the finite case. That generalization to infinite and/or complexly structured unions of spaces of different dimensions and quotiented over symmetry relations is possible is known although maybe not obvious. Mahler claims that such applications are not Bayesian in [43], but they can apparently be described by (1) and similar problems are investigated within the Bayesian framework, for example by Green [26]. Needless to say, since the observation and world spaces can be high-dimensional and the prior and likelihood can be arbitrarily complex, practical work with (1) is full of pitfalls and one often encounters what looks like counterintuitive behaviors. On closer investigation, such problems can lead to finding a modeling error, but more often it shows that (1) is indeed better than one's first intuitive attitude.

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 [17, 49, 42, 15], 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, as discussed in [7, 64, 33, 31, 46, 35, 2]. One simple assumption usually made in those studies that conclude in favor of (1) 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 [62], the Dempster-Shafer (DS) [51], the Fixsen/Mahler (MDS) [22] and Dezert-Smarandache (DSmT) [53] 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 [63].

Indeed, when the inference process is widened from an individual to a social or multi-agent context, there must be ways to accommodate different assessments of priors and likelihoods. Thus, there is a possibility that two experts make the same inference using different likelihoods and priors. If expert 1 obtained observation set $X_1 \subseteq X$ and expert 2 obtained observation set $X_2 \subseteq X$, they would obtain a posterior belief of, e.g., a patient's condition expressible as $f_i(\lambda_i \mid X_i) \propto f_i(X_i \mid \lambda_i) f_i(\lambda_i)$, for $i = 1,2$. Here we have not assumed that the two experts used the same sampling and prior distributions. Even if training aims at giving the two experts the same "knowledge" in the form of sampling function and prior, this ideal cannot be achieved completely in practice. The Bayesian method prescribes that expert i states the probability distribution $f_i(\lambda_i \mid X_i)$ 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:

$$\begin{aligned} f(\lambda \mid \{X_1, X_2\}) &\propto f(\{X_1, X_2\} \mid \lambda) f(\lambda) \\ &= f(X_1 \mid \lambda) f(X_2 \mid \lambda) f(\lambda) \end{aligned} \tag{3}$$

under the assumption:

$$f(\{X_1, X_2\} \mid \lambda) = f(X_1 \mid \lambda) f(X_2 \mid \lambda).$$

The assumption appears reasonable in many cases. In cases where it is not, the discrepancy should be entered in the statistical model. This is particularly important in information fusion for those cases where the first set of observations was used to define the second investigation, as in sensor management. This is an instance of selection bias. Ways of handling data selection biases are discussed thoroughly in [24]. Data selection bias is naturally and closely related to the missing data problem that has profound importance in statistics [48] and has also been examined in depth in the context of imprecise probability fusion [16].

It is important to observe that it is the two experts likelihood functions, not their posterior beliefs, that can be combined, otherwise we would replace the prior by its normalized square and the real uncertainty would be underestimated. This is at least the case if the experts obtained their training from a common body of medical experience coded in textbooks. If the posterior is reported and we happen to know the prior, the likelihood can be obtained by $f(X \mid \lambda) \propto f(\lambda \mid X)/f(\lambda)$ and the fusion rule becomes

$$f(\lambda \mid X_1, X_2) \propto f(\lambda \mid X_1) f(\lambda \mid X_2)/f(\lambda). \tag{4}$$

The existence of different agents with different priors and likelihoods is maybe the most compelling argument to open the possibility for robust Bayesian analysis, where the likelihood and prior sets would in the first approximation be the convex closure of the likelihoods and prior of different experts.

3. WHAT IS REQUIRED FOR SUCCESSFUL APPLICATION OF BAYES METHOD?

The formula (1) is deceptively simple, and hides the complexity of a real world application where many engineering compromises are inevitable. Nevertheless, any method claimed to be Bayesian must relate to (1) and include all substantive application knowledge in the parameter and observation spaces, the likelihood and the prior. It is in general quite easy to show the Bayesian

method to be better or worse than an alternative by not including relevant and necessary application knowledge in (1) or in the alternative method. Let us illustrate this by an analysis of the comparison made in [56]. The problem is to track and classify a single target. The tracking problem is solved with a dynamic version of Bayes method, known as the Bayesian Chapman-Kolmogorov relationship:

$$f(\lambda_t \mid D_t) \propto f(d_t \mid \lambda_t) \int f(\lambda_t \mid \lambda_{t-1}) f(\lambda_{t-1} \mid D_{t-1}) d\lambda_{t-1}$$

$$f(\lambda_0 \mid D_0) = f(\lambda_0). \tag{5}$$

Here $D_t = (d_1, \ldots, d_t)$ is the sequence of observations obtained at different times, and $f(\lambda_t \mid \lambda_{t-1})$ is the maneuvering (process innovation) noise assumed. The latter is a probability distribution function (pdf) over state λ_t dependent on the state at the previous time-step, λ_{t-1}. When tracking targets that display different levels of maneuvering like transportation, attack and dog-fight for a fighter airplane, it has been found appropriate to apply (5) with different filters with levels of innovation noise corresponding to the maneuvering states, and to declare the maneuvering state that corresponds to the best matching filter. In the paper [56] the same method is proposed for a different purpose, namely the classification of aircraft (civilian, bomber, fighter) based on their acceleration capabilities. This is done by ad hoc modifications of (5) that do not seem to reflect substantive application knowledge, namely that the true target class is unlikely to change, and hence does not work well. The Bayesian solution to this problem would involve looking at (5) with a critical mind. Since we want to jointly track and classify, the state space should be, e.g., $P \times V \times C$, where P and V are position and velocity spaces and C is the class set, $\{c, b, f\}$. The innovation process should take account of the facts that the target class in this case does not change, and that the civilian and bomber aircraft have bounded acceleration capacities. This translates to two requirements on the process innovation component $f(\lambda_t \mid \lambda_{t-1})$ that (assuming unit time sampling):

$$f((p_t, v_t, c_t) \mid (p_{t-1}, v_{t-1}, c_{t-1})) = 0 \quad \text{if} \quad c_t \neq c_{t-1}$$

$$f((p_t, v_t, k) \mid (p_{t-1}, v_{t-1}, k)) = 0 \quad \text{if} \quad |v_t - v_{t-1}| > a_k$$

where a_k is the highest possible acceleration of target class k. Such an innovation term can be (and often is) described by a Gaussian with variance tuned to a_k, or by a bank of Gaussians. With this innovation term, the observation of a high acceleration dampens permanently the marginal probability of having a target class incapable of such acceleration. This is the natural Bayesian approach to the joint tracking and classification problems. Similar effects can be obtained in the robust Bayes and TBM [56] frameworks. As a contrast, the experiments reported by Oxenham et al. [44] use an appropriate innovation term and also give more reasonable results, both for the TBM and the Bayesian Chapman Kolmogorov approaches. The above is not meant as an argument that one of the two approaches compared in [56] is the preferred one. Our intention is rather to suggest that appropriate modeling may be beneficial for both approaches.

The range of applications where an uncertainty management problem is approached using (1) or (5) is extremely broad. In the above example, the parameter λ consists of one state vector (position and velocity vectors of a target) and its target label, thus the parameter space is (for 3D tracking) $R^6 \times C$ where C is a finite set of targets labels. In our main example, λ is just an indicator with three possible values. In many image processing applications, the parameter λ is the scene to be reconstructed from the data x, which is commonly called the film even if it is nowadays not registered on photographic film and is not even necessarily represented as a 2D image. This approach has been found excellent both for ordinary camera reconstruction problems and for special types of cameras as exemplified by Positron Emission Tomography and functional Magnetic Resonance Imaging, the type of camera and reconstruction objective having a profound influence on the choice of likelihood and priors, see [3, 27]. In genetic investigations, complex Bayesian models are also used a lot, and here the parameter λ could be a description of how reproduction in a set of individuals in a family has been produced by selection of chromosomes from parents, the positions of crossovers and the position of one or more hypothesized disease-causing gene(s), whereas the data are the genotypes and disease status of individuals, plus individual covariates that may environmentally influence development of disease. For a unified treatment of this problem family, see [14]. Another fascinating example is Bayesian identification of state space dynamics in time series, where the parameter is the time series of invisible underlying states, a signaling distribution (output distribution as a function of latent state) and the state change probability distributions [59].

Characteristic of cases where (1) and (5) are not as easily accepted is the presence of two different kinds of uncertainty, often called aleatory and epistemic uncertainty, where the former can be called "pure randomness" as one perceives dice (Latin: alea) throwing, while the latter is caused by "lack of knowledge" (from the Greek word for knowledge, episteme). Although one can argue about the relevance of this distinction, application owners have typically a strong sense of the distinction, particularly in risk assessment. The consequence is that the concepts of well-defined priors and likelihoods can be, and have been, questioned. The Bayesian answer to this critique is robust Bayesian analysis.

4. ROBUST BAYES AND EVIDENCE THEORY

In (global) robust Bayesian analysis [5, 36], 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. The idea of robust Bayesian analysis goes back to the pioneers of Bayesian analysis [17, 39], 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. The difficulty of finding precise and unique priors has been taken as an argument to use some other methods, like evidence theory. However, as we shall see, this is an illusion, and avoiding use of an explicit prior usually means implicit reliance on Bernoulli's principle of indifference anyway. Likewise, should there be an acceptable prior, it can and should be used both in evidence theory and in Bayesian theory. This was pointed out, e.g., in [6, ch. 3.4].

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.

In evidence theory, the DS-structure is a representation of a belief over a frame of discernment (set of possible worlds) Λ (commonly called the frame of discernment Θ in evidence theory) by a probability distribution m over its power-set (excluding the empty set), a basic probability assignment bpa, basic belief assignment bba, bma, or DS-structure (terminology is not stable, we will use DS-structure). The sets assigned non-zero probability in a DS-structure are called its focal elements, and those that are singletons are called atoms. A DS-structure with no mass assigned to non-atoms is a precise (sometimes called Bayesian) DS-structure. 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 (e.g., [30, ch. 7] and [8, ch. 8]), explicit in Dempster's work [18], and almost unavoidable in a teaching situation. It is also compellingly suggested by the common phrase that the belief assigned to a non-singleton can flow freely to its singleton members, and the equivalence between a DS-structure with no mass assigned to non-singletons and the corresponding probability distribution [55]. Among publications elaborating on the possible difference between probability and other numerical uncertainty measures are [32, 55, 20].

A DS-structure seen as a set of distributions is a type of Choquet capacity, and these capacities form a particularly concise and flexible family of sets of distributions (the full theory of Choquet capacities is rich and of no immediate importance for us—we use the term capacity interpretation only to indicate a set of distributions obtained from a DS-structure in a way we will define precisely). Interpreting DS-structures as sets of probability distributions entails saying that the probability of a union of outcomes $e \subset \Lambda$ lies between the belief of e ($\sum_{w \subset e} m(w)$) and the plausibility of e ($\sum_{w \cap e \neq \emptyset} m(w)$). The parametric representation of the family of distributions it can represent, with parameters α_{ew}, $e \in 2^\Lambda$, $w \in \Lambda$, is $P(w) = \sum_e \alpha_{ew} m(e)$, all $w \in \Lambda$, where $\alpha_{ew} = 0$ if $w \notin e$, $\sum_{w \in e} \alpha_{ew} = 1$, and all α_{ew} are non-negative. This representation is used in Blackman and Popoli [8, ch. 8.5.3]. The pignistic transformation used in evidence theory to estimate a precise probability distribution from a DS-structure is obtained by making the α_{ew} equal for each e, $\alpha_{ew} = 1/|e|$ if $w \in e$. The relative plausibility transformation proposed by, among others, Voorbraak [60], Cobb and Shenoy [12, 13], on the other hand, is the result of normalizing the plausibilities of the atoms in Λ. It is also possible to translate a pdf over Λ to a DS-structure. Indeed, a pdf is already a (precise) DS-structure, but Sudano [58] studied inverse pignistic transformations that result in non-precise DS-structures by coarsening. They have considerable appeal but are not in the main line of argumentation in this paper [58].

It is illuminating to see how the pignistic and relative plausibility transformations emerge from a precise Bayesian inference: The observation space can in this case be considered to be 2^Λ, since this represents the only distinction among observation sets surviving from the likelihoods. The likelihood will be a function $l : 2^\Lambda \times \Lambda \to R$, the probability of seeing evidence e given world state λ. Given a precise $e \in 2^\Lambda$ as observation and a uniform prior, the inference over Λ would be $f(\lambda \mid e) \propto l(e, \lambda)$, but since we in this case have a probability distribution over the observation space, we should use (2), weighting the likelihoods by the masses of the DS-structures. Applying the indifference principle, $l(e, \lambda)$ should be constant for λ varying over the members of e, for each e. The other likelihood values ($\lambda \notin e$) will be zero. Two natural choices of likelihood are $l_1(e, \lambda) \propto 1$ and $l_2(e, \lambda) \propto 1/|e|$, for $\lambda \in e$. Amazingly, these two choices lead to the relative plausibility transformation and to the pignistic transformation, respectively:

$$f_i(\lambda \mid m) \propto \sum_{\{e:\lambda \in e\}} m(e) l_i(e, \lambda) = \begin{cases} \sum_{\{e:\lambda \in e\}} m(e) \Big/ \sum_e |e| m(e), & i = 1 \\ \sum_{\{e:\lambda \in e\}} m(e)/|e|, & i = 2. \end{cases} \tag{6}$$

Despite a lot of discussion, there seems thus to exist no fundamental reason to prefer one to the other, since

they result from two different and completely plausible statistical models and a common application of an indifference principle. The choice between the models (i.e., the two proposed likelihoods) can in principle be determined by (statistical) testing on the application's historic data.

The capacity corresponding to a DS-structure can be represented by $2^n - 2$ real numbers—the corresponding DS-structure is a normalized distribution over $2^n - 1$ elements (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 capacities form a proper and really small subset of all convex sets of distributions).

It is definitely possible—although we will not elaborate it here—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. It seems however important to caution against unnecessarily diving into the more sophisticated robust and graded set approaches to Bayesian uncertainty management.

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 is not addressed here, but can be approached by game-theoretic analyses, see, e.g., [9].

A Bayesian data fusion system or subsystem can thus use any level in a ladder with increasing complexity:

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

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. We will not further consider the first and the last rungs of the ladder.

4.1. Rounding

A set of distributions which is not a capacity can be approximated by rounding it to a minimal capacity that contains it (see Fig. 1), and this rounded set can be represented by a DS-structure. This rounding "upwards" is accomplished by means of lower probabilities (beliefs) of subsets of Λ. Specifically, in this example we list the minimum probabilities of all subsets of $\Lambda = \{A,B,C\}$ over the four corners of the polytope, to get lower bounds for the beliefs. These can be converted to masses using the Möbius inversion, or, in this simple example, manually from small to large events. For example, $m(A) = \text{bel}(A)$, $m(\{A,B\}) = \text{bel}(\{A,B\}) - m(A) - m(B)$, and $m(\{A,B,C\}) = \text{bel}(\{A,B,C\}) - m(\{A,B\}) - m(\{A,C\}) - m(\{B,C\}) - m(A) - m(B) - m(C)$. Since we have not necessarily started with a capacity, this may give negative masses to some elements. In that case, some mass must be moved up in the lattice to make all masses non-negative, and this can in the general case be done in several ways, but each way gives a minimal enclosing polytope. In the example, we have four corners, and the computation is shown in Table I. In this example we immediately obtain non-negative masses, and the rounded polytope is thus unique.

In the resulting up-rounded bba, when transforming it to a capacity, we must consider $2 * 2 * 3 = 12$ possible corner points. However, only five of these are actually corners of the convex hull in this case, and those are the corners visible in the enclosing capacity of Fig. 1. The other possible corner points turn out to lie inside, or inside the facets of, the convex hull. As an example, consider the lowest horizontal blue-dashed line; this is a facet of the polytope characterized by no mass flowing to B from the focal elements $\{A,C\}$, $\{B,C\}$ and $\{A,B,C\}$. The masses of $\{A,C\}$ and $\{A,B,C\}$ can thus be assigned either to A or to C. Assigning both to C gives the left end-point of the facet, both to A gives the right end-point, and assigning one to A and the other to C gives two interior points on the line.

It is also possible, using linear programming, to round downwards to a maximal capacity contained in a set. Neither type of rounding is unique, i.e., in general there may be several incomparable (by set inclusion) up- or down-rounded capacities for a set of distributions.

5. DECISIONS UNDER UNCERTAINTY AND IMPRECISION

The ultimate use of data fusion is usually decision making. Precise Bayesianism results in quantities—probabilities of possible worlds—that can be used immediately for expected utility decision making [49, 4]. Suppose the profit in choosing a from a set $\mathcal{A}$ of possible actions when the world state is λ is given by the utility function $u(a,\lambda)$ mapping action a and world state λ to a real valued utility (e.g., dollars). Then the action maximizing expected profit is $\arg\max_a \int u(a,\lambda) f(\lambda \mid x) d\lambda$. In robust Bayesian analysis one uses either minimax criteria or estimates a precise probability distribution to decide from. Examples of the latter are the pignistic and relative plausibility transformations. An example of

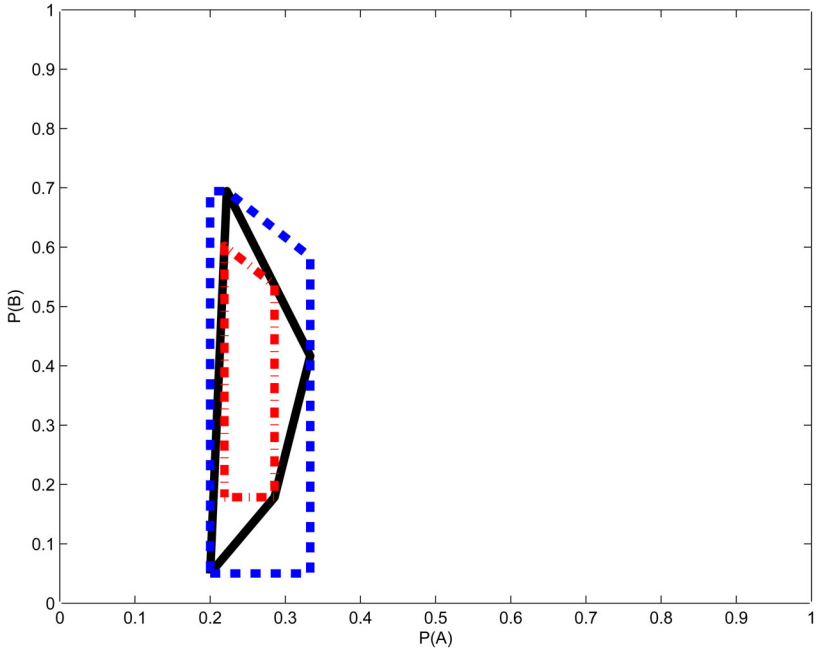

Fig. 1. Rounding a set of distributions over $\{A,B,C\}$. The coordinates are the probabilities of A and B. A set spanned by four corner distributions (black solid), its minimal enclosing (blue dashed), and one of its maximal enclosed (red dash-dotted), capacities.

TABLE I
Rounding a Convex Set of Distributions Given by its Corners*

Focal	Corners				min	m
A	0.200	0.222	0.333	0.286	0.200	0.200
B	0.050	0.694	0.417	0.179	0.050	0.050
C	0.750	0.083	0.250	0.536	0.083	0.083
$\{A,B\}$	0.250	0.916	0.750	0.465	0.250	0
$\{A,C\}$	0.950	0.305	0.583	0.822	0.305	0.022
$\{B,C\}$	0.800	0.777	0.667	0.715	0.667	0.534
$\{A,B,C\}$	1.000	1.000	1.000	1.000	1.000	0.111

*Corners of the black polygon of Fig. 1 are listed clockwise, starting at bottom left.

a decision-theoretically motivated estimate is the maximum entropy estimate, often used in robust probability applications [38]. 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 [28]. Specifically, a Decision maker must select a distribution q while Nature selects a distribution p from a convex set Γ. Nature selects an outcome x according to its chosen distribution p, and the decision makers loss is $-\log q(x)$. This makes the Decision maker's expected loss equal to $E_p\{-\log q(X)\}$. The minimum (over q) of the maximum (over p) expected loss is then obtained when q is chosen to be the maximum entropy distribution in Γ. Thus, if this loss function is accepted, it is optimal to use the maximum entropy transformation for decision making.

The maximum entropy principle differs significantly from the relative plausibility and pignistic transformations, since it tends to select a point on the boundary of a set of distributions (if the set does not contain the uniform distribution), whereas the pignistic transformation selects an interior point.

The pignistic and relative plausibility transformations are linear estimators, by which we mean that they are obtained by normalization of a linear function of the masses in the DS-structure. If we buy the concept of a DS-structure as a set of possible probability distributions, it would be natural to require that as estimate we choose a possible distribution, and then the pignistic transformation of Smets gets the edge—it is not difficult to prove the following:

PROPOSITION 1 *The pignistic transformation is the only linear estimator of a probability distribution from a DS-structure that is symmetric over Λ and always returns a distribution in the capacity represented by the DS-structure.*

Although we have no theorem to this effect, it seems as if the pignistic transformation is also a reasonable decision-oriented estimator approximately minimizing the maximum Euclidean norm of difference between the chosen distribution and the possible distributions, and better than the relative plausibility transformation as well as the maximum entropy estimate for this objective function. The estimator minimizing this maximum norm is the center of the smallest enclosing sphere. It will not be linear in m, but can be computed with some effort using methods presented, e.g., in [23]. The centroid is sometimes proposed as an estimator, but it does not correspond exactly to any known robust loss function—rather it is based on the assumption that the probability vector is uniformly distributed over the imprecision polytope.

The standard expected utility decision rule in precise probability translates in imprecise probability to producing an expected utility interval for each decision alternative, the utility of an action a being given by the interval $I_a = \bigcup_{f \in F} \int u(a,\lambda) f(\lambda \mid x) d\lambda$. In a refinement proposed by Voorbraak [61], decision alternatives are compared for each pdf in the set of possible pdfs: $I_{af} = \int u(a,\lambda) f(\lambda \mid x) d\lambda$, for $f \in F$. Decision a is now better than decision b if $I_{af} > I_{bf}$ for all $f \in F$.

Some decision alternatives will fall out because they are dominated in utility by others, but in general several possible decisions with overlapping utility intervals will remain. In principle, if no more information exists, any of these decisions can be considered right. But they are characterized by larger or smaller risk and opportunity.

6. ZADEH'S EXAMPLE

We will now discuss our problem in the context of Zadeh's example of two physicians who investigated a patient independently—a case prototypical, e.g., for the important fusion for target classification problem. 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. If the two expert opinions are taken as likelihoods, or as posteriors with a common uniform prior, this problem is solved by taking Laplace's parallel composition (1) of the two probability vectors, giving the result $(0, 0, 1)$, i.e., the case T is certain. This example has been discussed a lot in the literature, see e.g. [53]. 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 [38]. However, in the evidence theory literature the Bayesian solution (which is also obtained from using Dempster's and the Modified 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 experts 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.
- It seems as if estimates are combined at a too coarse level: it is in this case necessary to distinguish in Λ between different cases of the three conditions that are most likely to effect the likelihoods from observations: type, size and position of tumor, bacterial, viral or purely inflammatory meningitis, position of concussion. The frame of discernment should thus not be determined solely from the frame of interest, but also on what one could call homogeneity of likelihoods or evidence.
- The assessments for T are probably based mostly on prior information (rareness) or invisibility in a standard MR scan, so the combined judgment should not make T less likely, rather the opposite.
- An investigation is always guided by the patient's subjective beliefs, and an investigation affects those beliefs. So it is implausible that the two investigations of the same patient are "really" independent. 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 teach 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, a new investigation, or pure speculation. The professionals handling of the information problem is usually rational enough, but very different conclusions arise from small differences in mental models. This is a possible interpretation of the prospect theory of Kahneman and Tversky [40].

To sum things up, if we are sure that the experts are reliable and have the same definitions of the three neurological conditions, the result given by Bayes' and Dempster's rules are appropriate. If not, the assumptions and hence the statistical model must be modified. It seems obvious that the decision makers belief in the experts reliability must be explicitly elicited in similar situations.

7. FUSION IN EVIDENCE AND ROBUST BAYESIAN THEORY

The Dempster-Shafer combination rule [51] is a straightforward generalization of Laplace's parallel composition rule. 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 [18] 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 multi-valued mapping, ambiguity in how the information of

the sources should be translated to a common frame of discernment. It is fairly plausible that the information given by the source is well representable as a DS structure interpreted as a capacity. What is much less plausible is that the information combined from several sources is well captured by Dempster's rule rather than by the Fixsen/Mahler combination rule or the robust combination rule to be described shortly. The precise assumptions behind Dempster's rule are seldom explained in tutorials and seem not well known, so we recapitulate them tersely: It is assumed that evidence comes from a set of sources, where source i has obtained a precise probability estimate p_i over its private frame X_i. This information is to be translated into a common frame Λ, but only a multi-valued mapping Γ_i is available, mapping elements of X_i to subsets of Λ. For the tuple of elements $x_1,\ldots,x_n$, their joint probability could be guessed to be $p_1(x_1)\cdots p_n(x_n)$, but we have made assumptions such that we know that this tuple is only possible if $\Gamma_1(x_1)\cap\cdots\cap\Gamma_n(x_n)$ is non-empty. So the probabilities of tuples should be added to the corresponding subset of Λ probabilities, and then conditioning on non-emptiness should be performed and the remaining subset probabilities normalized, a simple application of (1). From these assumptions Dempster's rule follows.

This is postulated by Dempster as the model required. One can note that it is not based on inference, but derived from an explicit and exact probability model. It was claimed incoherent (i.e. violating the consistent betting paradigm) by Lindley [42], but Goodman, Nguyen and Rogers showed that it is not incoherent [25]. Indeed, the assumption of multi-valued mappings seems completely innocent, if somewhat arbitrary, and it would be unlikely to lead to inconsistencies. The recently introduced Fixsen/Mahler MDS combination rule [22] involves a re-weighting of the terms involved in the set intersection operation: whereas Dempster's combination rule can be expressed as

$$m_{\mathrm{DS}}(e)\propto\sum_{e=e_1\cap e_2} m_1(e_1)m_2(e_2),\qquad e\neq\emptyset \tag{7}$$

the MDS rule is

$$m_{\mathrm{MDS}}(e)\propto\sum_{e=e_1\cap e_2} m_1(e_1)m_2(e_2)\frac{|e|}{|e_1||e_2|},\qquad e\neq\emptyset. \tag{8}$$

The MDS rule was introduced to account for non-uniform prior information about the world and evidence that contains prior information common to all sources. In this case $|e|$, etc, in the formula are replaced by the prior probabilities of the respective sets. The rule (8) is completely analogous to (4): the denominator of the correction term takes the priors out of the posteriors of both operands, and the numerator $|e|$ reinserts it once in the result. But as we now will see, the MDS rule can also be considered a natural result of fusing likelihood describing information with a different likelihood function.

It is possible to analyze the source fusion problem in a (precise) Bayesian setting. If we model the situation with the likelihoods on $2^\Lambda\times\Lambda$ of (6), Section 4, we find the task of combining the two likelihoods $\sum_e m_1(e)l(e,\lambda)$ and $\sum_e m_2(e)l(e,\lambda)$ using Laplace's parallel composition as in (2) over Λ, giving

$$f(\lambda)\propto\sum_{e_1,e_2} m_1(e_1)m_2(e_2)l_i(e_1,\lambda)l_i(e_2,\lambda).$$

For the choice $i=1$, this gives the relative plausibility of the result of fusing the evidences with Dempster's rule; for the likelihood l_2 associated with the pignistic transformation, we get $\sum_{e_1,e_2} m_1(e_1)m_2(e_2)l(e_1,\lambda)l(e_2,\lambda)/(|e_1||e_2|)$. This is the pignistic transformation of the result of combining m_1 and m_2 using the MDS rule. In the discussions for and against different combination and estimation operators, it has sometimes been claimed that the estimation operator should propagate through the combination operator. This claim is only valid if the above indicated precise Bayesian approach is bought, which would render DS-structures and convex sets of distributions unnecessary. In the robust Bayesian framework, the maximum entropy estimate is completely kosher, but it does not propagate through any well known combination operation. The combination of Dempster's rule and the pignistic transformation cannot easily be defended in a precise Bayesian framework, but Dempster's rule can be defended under the assumption of multi-valued mappings and reliable sources, whereas the pignistic transformation can be defended in three ways: (1) It can be seen as "natural" since it results, e.g., from an indifference principle applied to the parametric representation of Blackman and Popoli; (2) Smets argument [54] is that the estimation operator (e.g., the pignistic transformation) should propagate, not through the combination operator, but through linear mixing; (3) An even more convincing argument would relate to decisions made, e.g., it seems as if the pignistic transformation is, not exactly but approximately, minimizing the norm of the maximum (over Nature's choice) error made measured as the Euclidean norm of the difference between the selected distribution and Nature's choice.

7.1 The Robust Combination Rule

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 [41, 57]. 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-structures. The

most obvious combination rule would yield the set of probability functions that can be obtained by taking one member from each set and combining 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 (3) would thus take the form (where F denotes convex families of functions):

$$F(\lambda \mid \{X_1, X_2\}) \propto F(\{X_1, X_2\} \mid \lambda) \times F(\lambda) = F(X_1 \mid \lambda) \times F(X_2 \mid \lambda) \times F(\lambda). \quad (9)$$

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, for reasons similar to the corresponding problem in interval arithmetic: 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 the state (distance, inclination, etc) of the 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.

We must now explain how to compute the operator of Definition 1. The definition given of the robust Bayesian combination operator involves infinite sets in general and is not computable directly. For singleton sets it is easily computed, though, with Laplace's parallel composition rule. It is also the case that every corner in the resulting set can be generated by combining two corners, one from each of the operands. This observation 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 tessellate the boundary and remove those points falling in the interior of the polytope. The figures of this paper were produced by a Matlab implementation of robust combination, Dempster's and the MDS rule, maximum entropy estimation, and rounding. The state of the art in computational geometry software thus allows easy and efficient solutions, but of course as the state space and/or the number of facets of the imprecision polytopes become very large, some tailored approximation methods will be called for. The DS and MDS rules have exponential complexity in the worst case. The robust rule will have a complexity quadratic in the number of corners of the operands, and will thus depend on rounding for feasibility. For very high-dimensional problems additional pruning of the corner set will be necessary (as is also the case with the DS and MDS operators).

We can now make a few statements, most of which are implicitly present in [19, Discussion by Aitchison] and [32], about fusion in the robust Bayesian framework:

- The combination operator is associative 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 in the robust framework 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.
- Total skepticism cannot be expressed with Dempster's rule, since it never introduces a focal element which is a superset of all focal elements in one operand.

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 combination operator is, and the rounded robust operator can be made (note: it is not unique) monotone with respect to imprecision, i.e., if $F_i' \subseteq F_i$, then $F_1' \times F_2' \subseteq F_1 \times F_2$.*

PROPOSITION 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 By contradiction; thus assume 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.

We can also show that the MDS combination rule has the "nice" property of giving a result that always overlaps the robust rule result, under the capacity interpretation of DS-structures:

PROPOSITION 3 *Let m_1 and m_2 be two DS-structures and let F_1 and F_2 be the corresponding capacities. If F is the*

capacity representing $m = m_1 *_{\text{MDS}} m_2$ and F' is $F_1 \times F_2$, then F and F' overlap.

PROOF Since the pignistic transformation propagates through the MDS combination operator, and by Proposition 1 the pignistic transformation is a member of the capacity of the DS-structure, the parallel combination of the pignistic transformations of m_1 and m_2 is a member of F' and equal to the pignistic transformation of m, which for the same reason is a member of F. This concludes the proof.

The argument does not work for the original Dempster's rule, for reasons that will become apparent in the next section. It was proved by Jaffray [37] that Dempster's rule applied with one operand being precise gives a (precise) result inside the robust rule polytope. The same holds of course, by Proposition 3, for the MDS rule. We can also conjecture the following, based on extensive experimentation with our prototype implementation, but have failed in obtaining a short convincing proof:

CONJECTURE 1 *The MDS combination rule always gives a result which is, in the capacity interpretation, a subset of the robust rule result. The MDS combination rule is also a coarsest symmetric bilinear operator on DS-structures with this property.*

8. A PARADOXICAL EXAMPLE

In [1] we analyzed several versions of Zadeh's example with "discounted" evidences to illustrate the differences between robust fusion and the DS and MDS rules, as well as some different methods to summarize a convex set of pdfs as a precise pdf. Typically, the DS and MDS rules give much smaller imprecision in the result than the robust rule, which can be expected from their behavior with one precise and one imprecise operand. One would hope that the operators giving less imprecision would fall inside the robust rule result, in which case one would perhaps easily find some plausible motivation for giving less imprecision than indicated in the result. In practice this would mean that a system using robust fusion would sometimes find that there is not a unique best action while a system based on the DS or MDS rule would pick one of the remaining actions and claim it best, which is not obviously a bad thing. However, the DS, MDS and robust rules do not only give different imprecision in their results, they are also pairwise incompatible (sometimes having an empty intersection) except for the case mentioned in Conjecture 1. Here we will concentrate on a simple, somewhat paradoxical, case of combining two imprecise evidences and decide from the result.

Varying the parameters of discounting a little in Zadeh's example, it is not difficult to find cases where Dempster's rule gives a capacity disjoint (regarded as a geometric polytope) 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 Fig. 2 we show an example where this is not the case. The two combined evidences are imprecise probabilities over three elements A, B and C, the first spanned by the probability distributions $(0.2, 0.2, 0.6)$ and $(0.2, 0.5, 0.3)$, the second by $(0.4, 0.1, 0.5)$ and $(0.4, 0.5, 0.1)$. These operands can be represented as DS structures, as shown in Table II, and they are shown as vertical green lines in Fig. 2. They can be combined with either the DS rule, the MDS rule, or the robust rule, as shown in Table III. The situation is illustrated in Fig. 2, where all sets of pdfs are depicted as lines or polygons projected on the first two probabilities. The figure shows that the robust rule claims the probability of the first event A (horizontal axis) to be between 0.2 and 0.33, whereas Dempster's rule would give it an exact probability around 0.157. The MDS rule gives a result that falls nicely inside the robust rule result, but it claims an exact value for the probability of A, namely 0.25. Asked to bet with odds six to one on the first event (by which we mean that the total gain is six on success and the loss is one on failure), the DS rule says decline, the robust and MDS rules say accept. For odds strictly between four and five to one, the robust rule would hesitate and MDS would still say yes. For odds strictly between three and four to one, DS and MDS would decline whereas the robust rule would not decide for or against. Including the refinement proposed by Voorbraak (see Section 5) would not alter this conclusion unless the imprecisions of the two operands were coupled, e.g., by common dependence on a third quantity.

In an effort to reconcile Bayesian and belief methods, Blackman and Popoli [8, ch. 7] propose that the result of fusion should be given the capacity interpretation as a convex set, whereas the likelihoods should not—an imprecise likelihood should instead be represented as the coarsest enclosing DS-structure having the same pignistic transformation as the original one. When combined with Dempster's rule, the result is again a prior for the next combination whose capacity interpretation shows its imprecision. The theorem proved—at some length—in [8, App. 8A] essentially says that this approach is compatible with our robust rule for precise likelihoods. In our example, if the second operand is coarsened to $\{m_2'(A) \mapsto 0.1, m_2'(\{A,B,C\}) \mapsto 0.9\}$, the fusion result will be a vertical line at 0.217, going from 0.2 to 0.49, just inside the robust rule result. However no mass will be assigned to a non-singleton set containing A, so the rule still gives a precise value to the probability of A. The philosophical justification of this approach appears weak.

The example shows that Dempster's rule is not compatible with the capacity interpretation, whereas the MDS rule is: there is no pair of possible pdfs for the operands that combine to any possible value in the

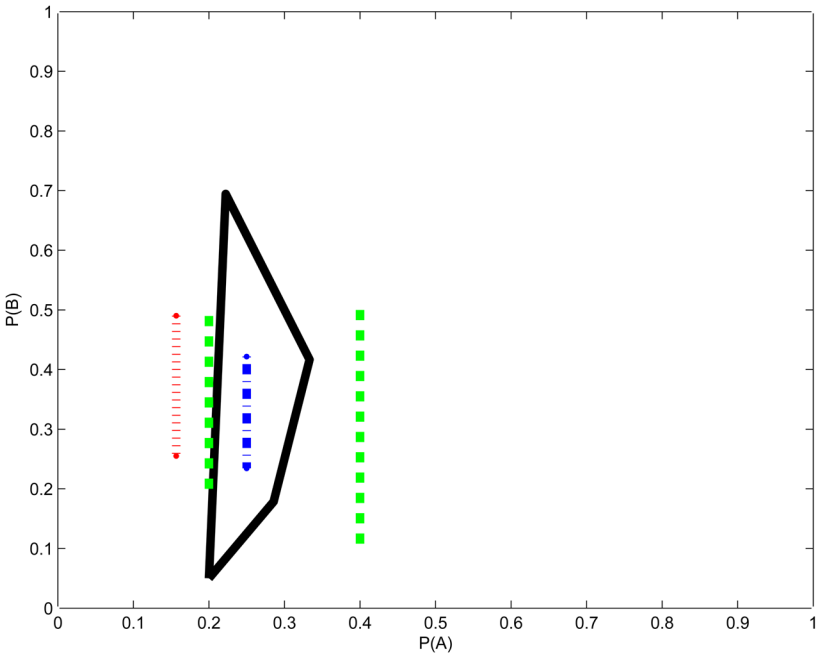

Fig. 2. A case where the robust rule and Dempster's rule give paradoxical results. The coordinates are the probabilities of A and B. The operands are shown in green dashed, the result of the robust combination rule is shown in black solid (same as in Fig. 1), Dempster's rule gives the result shown in red dotted, the Fixsen/Mahler MDS rule shown in blue dash-dotted lines.

Dempster's rule result, wheras every possible pdf in the MDS rule results from combining some pair of possible pdfs for the operands. If Conjecture 1 can be proved, the last is true for all pairs of operands, but there are also many particular examples where even Dempster's rule gives a compatible result. It has been noted by Walley that Dempster's rule is not the same as the robust combination rule [62], but I have not seen a demonstration that the two are incompatible in the above sense. There is, of course, a rational explanation of the apparent paradox, namely that the assumptions of private frames of discernment for sources and of a multi-valued mapping for each source is very different from the assumption of imprecise likelihoods, and this means that some informa-

TABLE II
Two Operands of the Paradoxical Example*

Focal	op_1			op_2		
	c_1	c_2	m	c_1	c_2	m
A	0.2	0.2	0.2	0.4	0.4	0.4
B	0.2	0.5	0.2	0.1	0.5	0.1
C	0.6	0.3	0.3	0.5	0.1	0.1
$\{B,C\}$			0.3			0.4

*Columns marked m denote DS-structures and those marked c_1, c_2 denote corners spanning the corresponding capacity. Values are exact.

TABLE III
Fusing the Operands of Table II with the DS, MDS and Robust Rules*

Focal	Fusion Result										
	DS			MDS			Robust				Uprounded
	c_1	c_2	m	c_1	c_2	m	c_{11}	c_{22}	c_{12}	c_{21}	m
A	0.157	0.157	0.157	0.250	0.250	0.250	0.200	0.222	0.333	0.286	0.200
B	0.255	0.490	0.255	0.422	0.234	0.234	0.050	0.694	0.417	0.179	0.050
C	0.588	0.353	0.353	0.328	0.516	0.328	0.750	0.083	0.250	0.536	0.083
$\{A,B\}$			0			0					0
$\{A,C\}$			0			0					0.022
$\{B,C\}$			0.235			0.188					0.534
$\{A,B,C\}$			0			0					0.111

*The result for DS and MDS shown as two corners (c_1 and c_2), and as an equivalent DS-structure (m). For the robust rule result, its four spanning corners are shown, where, e.g., c_{21} was obtained by combining the second corner c_2 of op_1 with c_1 of op_2, etc. These corners are the corners of the black polygon in Fig. 2. The robust rule result is also shown as a DS-structure for the up-rounded result (blue dashed line in Fig. 1). Values are rounded to three decimals.

tion in the private frames is still visible in the end result when Dempster's rule is used. Thus Dempster's rule effectively makes a combination in the frame 2^Λ instead of in Λ as done by the robust rule. It is perhaps more surprising that the paradoxical result is also obtainable in the frame Λ using precise Bayesian analysis and the likelihood $l_1(e, \lambda)$ (see Section 4). The main lesson here, as in other places, is that we should not use Dempster's rule unless we have reason to believe that imprecision is produced by the multi-valued mapping of Dempster's model rather than Fixsen/Mahler's model or incomplete knowledge of sampling functions and prior. If the MDS operator is used to combine likelihoods or a likelihood and a prior, then posteriors should be combined using the MDS rule (8), but with all set cardinalities squared.

Excluding Bayesian thinking from fusion may well lead to inferior designs.

9. CONCLUSIONS

Despite the normative claims of evidence theory and robust Bayesianism, the two have been considered different in their conclusions and general attitude towards uncertainty. The Bayesian framework can however describe most central features of evidence theory, and is thus a useful basis for teaching and comparison of different detailed approaches to information fusion. 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 medical researchers, military commanders, and their teachers in their roles as evaluators, the need for clarity and economy of concepts cannot be exaggerated. The arguments put forward above suggest that an approach based on the precise Bayesian and the robust Bayesian fusion operator is called for, and that choosing decision methods based on imprecise probabilities or DS structures should preferably be based on decision-theoretic arguments. Our example shows how dangerous it can be to apply evidence theory without investigating the validity in an application of its crucial assumption of reliable private frames for all sources of evidence and precise multi-valued mappings from this frame to the frame of interest. The robust rule seems to give a reasonable fit to most fusion rules based on different statistical models, with the notable exception of Dempster's rule. Thus, as long as the capacity interpretation is prevalent in evidence theory applications, there are good reasons to consider if the application would benefit from using the MDS rule (complemented with priors if available) also for combining information in the style of likelihoods. In this case, however, the combination of the MDS rule with pignistic transformation is interpretable as a precise Bayesian analysis. In most applications I expect that the precise Bayesian framework is adequate, and it is mainly in applications with the taste of risk analysis that the robust Bayesian framework will be appropriate.

ACKNOWLEDGMENTS

Discussions with members of the fusion group at the Swedish Defence Research Agency (FOI), students in the decision support group at KTH, and colleagues at Saab AB, Karolinska Institutet and the Swedish National Defense College (FHS) have been important for clarifying ideas presented above. The referees have further made clear the need to clarify the argumentation, and also by their comments made me strengthen my claims somewhat.

REFERENCES

[1] S. Arnborg
Robust Bayesianism: Imprecise and paradoxical reasoning. In P. Svensson and J. Schubert (Eds.), *Proceedings of the Seventh International Conference on Information Fusion*, Vol. I, Stockholm, Sweden, International Society of Information Fusion, June 2004, 407–414.

[2] S. Arnborg and G. Sjödin
Bayes rules in finite models.
In *Proceedings of European Conference on Artificial Intelligence*, Berlin, 2000, 571–575.

[3] R. G. Aykroyd and P. J. Green
Global and local priors, and the location of lesions using gamma camera imagery.
Philosophical Transactions of the Royal Society of London A, **337** (1991), 323–342.

[4] J. O. Berger
Statistical Decision Theory and Bayesian Analysis.
New York: Springer-Verlag, 1985.

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

[6] N. Bergman
Recursive Bayesian Estimation.
Ph.D. thesis, Linköping University, Linköping, 1999.

[7] J. M. Bernardo and A. F. Smith
Bayesian Theory.
New York: Wiley, 1994.

[8] S. Blackman and R. Popoli
Design and Analysis of Modern Tracking Systems.
Boston, London: Artech House, 1999.

[9] J. Brynielsson and S. Arnborg
Bayesian games for threat prediction and situation analysis.
In P. Svensson and J. Schubert (Eds.), *Proceedings of the Seventh International Conference on Information Fusion*, Vol. II, Stockholm, Sweden, International Society of Information Fusion, June 2004, 1125–1132.

[10] S. Challa and D. Koks
Bayesian and Dempster-Shafer fusion.
Sãdhãna, 2004, 145–174.

[11] B. Cobb and P. Shenoy
A comparison of Bayesian and belief function reasoning.
Technical Report, University of Kansas School of Business, 2003.

[12] B. Cobb and P. Shenoy
A comparison of methods for transforming belief function models to probability models.
In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Vol. 2711, *LNCS*, Berlin: Springer, 2004, 255–266.

[13] B. Cobb and P. Shenoy
On the plausibility transformation method for translating belief function models to probability models.
International Journal of Approximate Reasoning, to appear.

[14] J. Corrander and M. Sillanpää
A unified approach to joint modeling of multiple quantitative and qualitative traits in gene mapping.
Journal of Theor. Biology, **218** (2002), 435–446.

[15] R. T. Cox
Probability, frequency, and reasonable expectation.
Am. Journal of Physics, **14** (1946), 1–13.

[16] G. de Cooman and M. Zaffalon
Updating beliefs with incomplete observations.
Artificial Intelligence, **159**, 1–2 (2004), 75–125.

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

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

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

[20] D. Dubois and H. Prade
Representing partial ignorance.
IEEE Transactions on Systems, Man, and Cybernetics, Pt. A, 1996, 361–377.

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

[22] D. Fixsen and R. P. S. Mahler
The modified Dempster-Shafer approach to classification.
IEEE Transactions on SMC-A, **27**, 1 (Jan. 1997), 96–104.

[23] B. Gärtner
Fast and robust smallest enclosing balls.
In *ESA*, Vol. 1643, *LNCS*, Berlin: Springer, 1999, 325–338.

[24] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin
Bayesian Data Analysis (2nd ed.).
New York: Chapman & Hall, 2003.

[25] I. Goodman, H. T. Nguyen, and G. Rogers
On the scoring approach to admissibility of uncertainty measures in expert systems.
Journal of Math. Analysis Appl., **159** (1991), 550–594.

[26] P. J. Green
Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.
Biometrika, **82** (1995), 711–732.

[27] P. J. Green
Markov Chain Monte Carlo in image analysis.
In W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds.), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall, 1996.

[28] P. D. Grünwald and A. P. Dawid
Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory.
Annals of Statistics, **32**, 4 (2004).

[29] V. Ha, A-H. Doan, V. Vu and P. Haddaway
Geometric foundations for interval-based probabilities.
Annals of Mathematics and Artificial Intelligence, 1998.

[30] D. L. Hall and J. Llinas
Handbook of Multisensor Data Fusion.
Boca Raton: CRC Press, May 2001.

[31] J. Halpern
A counterexample to theorems of Cox and Fine.
Journal of AI Research, **10** (1999), 67–85.

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

[33] D. Heath and W. Sudderth
On finitely additive priors, coherence, and extended admissibility.
Annals of Statistics, **6** (1978), 233–345.

[34] J. Helton and K. Weichselberger (Eds.)
Special issue on alternative representations of epistemic uncertainty.
Reliability Engineering and System Safety, 2004, 1–369.

[35] C. Howson and P. Urbach
Scientific Inference: The Bayesian Approach.
Chicago: Open Court Publishing Company, 1989.

[36] D. Rios Insua and F. Ruggeri (Eds.)
Robust Bayesian Analysis.
La Salle: Springer-Verlag, 2000.

[37] J. Jaffray
Bayesian updating and belief functions.
IEEE Transactions on Systems, Man, and Cybernetics, 1996, 1144–1152.

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

[39] H. Jeffreys
Scientific Inference.
Cambridge: Cambridge University Press, 1931.

[40] D. Kahneman (Ed.)
Judgment Under Uncertainty: Heuristics and Biases.
Cambridge: Cambridge University Press, 1982.

[41] H. Kyburg and M. Pittarelli
Set-based Bayesianism.
IEEE Transactions on Systems, Man, and Cybernetics, 1996, 324–339.

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

[43] R. Mahler
Can the Bayesian and Dempster-Shafer approaches be reconciled? Yes.
In *FUSION 2005*, International Society of Information Fusion, 2005, C7-3.

[44] M. G. Oxenham, S. Challa, and M. R. Morelande
Decentralised fusion of disparate identity estimates for shared situation awareness.
In P. Svensson and J. Schubert (Eds.), *Proceedings of the Seventh International Conference on Information Fusion*, vol. I, Stockholm, Sweden, International Society of Information Fusion, June 2004, 167–174.

[45] M. G. Oxenham, S. Challa, and M. R. Morelande
Fusion of disparate identity estimates for shared situation awareness in a network-centric environment.
Information Fusion, to appear.

[46] J. B. Paris
The Uncertain Reasoner's Companion.
Cambridge: Cambridge University Press, 1994.

[47] S. Parsons and A. Hunter
A review of uncertainty handling formalisms.
In *Applications of Uncertainty Formalisms*, Vol. 1455, *LNCS*, Berlin: Springer, 1998, 266–302.

[48] D. Rubin
Inference and missing data.
Biometrika, **63** (1976), 581–592.

[49] L. J. Savage
Foundations of Statistics.
New York: Wiley, 1954.

[50] J. Schubert and P. Svensson
Methodology for guaranteed and robust high level fusion performance: A literature study.
Technical Report FOI-D-0216-SE, Swedish Defence Research Agency, 2005.

[51] G. Shafer
A Mathematical Theory of Evidence.
Princeton: Princeton University Press, 1976.

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

[53] F. Smarandache and J. Dezert (Eds.)
Advances and Applications of DSmT for Information Fusion.
Rehoboth: American Research Press, 2004.

[54] P. Smets
Decision making in the TBM: The necessity of the pignistic transformation.
International Journal of Approximate Reasoning, **38**, 2 (2005), 133–214.

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

[56] P. Smets and B. Ristic
Kalman filter and joint tracking and classification in the TBM framework.
In P. Svensson and J. Schubert (Eds.), *Proceedings of the Seventh International Conference on Information Fusion*, Vol. I, Stockholm, Sweden, International Society of Information Fusion, June 2004, 46–53.

[57] W. Stirling and A. Morelli
Convex Bayesianism decision theory.
IEEE Transactions on Systems, Man, and Cybernetics, (1991), 173–183.

[58] J. Sudano
Inverse pignistic probability transforms.
In *FUSION 2002*, International Society of Information Fusion, 2002, 763–768.

[59] H. Valpola and J. Karhunen
An unsupervised ensemble learning method for nonlinear dynamic state-space models.
Neural Computation, **14**, 11 (2002), 2647–2692.

[60] F. Voorbraak
A computationally efficient approximation of Dempster-Shafer theory.
International Journal of Man-Machine Studies, (1989), 525–536.

[61] F. Voorbraak
Partial probability: Theory and applications.
In G. de Cooman, F. G. Cozman, S. Moral, and P. Walley (Eds.), *Proceedings of the First International Symposium on Imprecise Probabilities and their Applications*, Gent University, 1999.

[62] P. Walley
Statistical Reasoning with Imprecise Probability.
London: Chapman and Hall, 1991.

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

[64] N. Wilson
Extended probability.
In *Proceedings of the 12th European Conference on Artificial Intelligence*, 1996, 667–671.

Stefan Arnborg took the civilingenjör (M.Sc.E.) exam in engineering physics at KTH in 1968, and the Ph.D. in information processing at KTH in 1972.

He worked as a supercomputer programmer at the Courant Institute (1966), with object-oriented programming systems at the Norwegian Computing Centre (1969–1970) and CAP Sweden (1971), as an OR analyst at the Swedish Defence Research Institute FOA (1972–1979), and as a distributed systems and computer guru at Philips Financial Terminal Systems (PEAB, 1979–1982). He spent sabbatical terms at the University of Oregon (1986, 1989, 1993) and was part-time advisor to the Swedish Institute of Computer Science SICS. He is the director of the theory group and of the M.Sc.E. program in computer engineering. He has been a professor of computer science at KTH since 1982.

His research interests are algorithms and complexity, verification, computer algebra, data engineering and uncertainty management, data mining, human brain informatics, command and control, and aesthetics in engineering education.