Nonparametric Protective Weakening: an Illustration of Dempster-Shafer Inference

A. P. Dempster

Harvard University

Slides for talk at DASFIII, Fields Institute, Toronto
April 30, 2010.

Paul Edlefsen, James Ong, and I are developing a paper with further details and illustrations.

Abstract. Assume observations $X_1, X_2, \ldots, X_n$ from a univariate population having unknown continuous cdf $F(x)$. What can be said about $F(b) - F(a)$, for prescribed a and b that may or may not depend on the data? Standard sampling theory provides nonparametric tolerance regions with predictable long run properties. Associated long run frequencies may be reinterpreted as Dempster-Shafer (DS) posterior probabilities, including familiar $p \geq 0$ and $q \geq 0$ "for" and "against" the truth of an assertion about $F(b) - F(a)$, but possibly allowing a third probability $r \geq 0$ of "don't know", so that now $p + q + r = 1$, generalizing the more familiar $p + q = 1$. Such "fiducial-like" probabilities are not typically protected as long run betting probabilities, when the counterparty is allowed to choose the side of each bet. DS theory provides a menu of weakened models whose resulting DS inferences are more conservative, and hence more protected. In my talk I will define and begin to explore one such weakened DS analysis.

Goals of Talk

- Introduce (or review) DS modeling and analysis

- Illustrate DS methodology via the example of nonparametric inference about a continuous univariate population distribution

- Introduce the concept of "robustifying" a DS analysis through model "weakening", using the nonparametric example as an illustration

Elements of the DS Calculus

- Establishing the "state space model" or SSM

- Establishing a collection of independent "DS probability models" or DSPMs over the SSM

- Implementing the numerical steps required to carry out the DS operations leading to desired inferences implied by the assumed SSM and DSPMs

Examples of SSMs

• Textbook examples for hypothetical statistical situations, such as univariate nonparametric or parametric modeling and inference, or multiple testing situations, or assessing covariance structures in high dimensions

• In the nonparametric example, the SSM consists of real-valued sample observables $X_1, X_2, \ldots, X_n$ together with the population distribution of X represented by a continuous cdf F(x).

Examples of SSMs (continued)

• While statisticians traditionally focus on “the data”, and on associated sampling distributions, in almost any actual situation the underlying reality is much more complex. The SSM should be big enough to capture what matters.

• Topical examples include the climate of planet Earth for the past 1000 years, or the world economy for the past 50 years. Important complex subsystems are rapidly becoming accessible to representation, partial observation, modeling, and analysis including forecasting.

What is a DSPM?

- Mathematically, it is a probability distribution across subsets of the set of possible states of the SSM, referred to as the “mass distribution” or “bpa”.

- Interpretation-wise, it represents “your” allocation of formal uncertainty as to which state of the SSM is the “true” state, based on the evidence justifying the DSPM.

Examples of DSPMs

• Included in a given DS model and analysis, there are typically many DSPMs defined over the SSM.

• In the example of this talk, the SSM can be represented by the vector $(X_1, X_2, \ldots, X_n, F(x))$.

• If you observe a specific sample true value, say x_i for X_i, this is a DSPM that assigns mass one to the subset of possible values of the SSM where X_i is limited to x_i. And so on for the remaining elements of the vector representation of the SSM.

A More Complex DSPM

• Functions of the variables defining an SSM are also implicitly part of the SSM. So, for example, the variables

$$U_1 = F(X_i) \text{ for } i = 1, 2, \ldots, n$$

are regarded as included in the SSM.

• In DS terms, standard nonparametric statistics assumes that the unknowns $U_1, U_2, \ldots, U_n$ have mass distribution determined by i.i.d. uniformly distributed variables $u_1, u_2, \ldots, u_n$ on (0, 1). Note: I use lower case letters to represent DSPMs.

Operations with DSPMs : Projection

• A DSPM can be down-projected to a partition or margin of its SSM, yielding the implied DSPM on the marginal SSM.

• A DSPM defined on a marginal SSM can be up-projected to yield the implied DSPM on the full SSM.

• Definitions and illustrations via the nonparametric sampling model are pretty obvious.

Operations with DSPMs : Combination

• Two or more DSPMs on the same SSM can be combined to pool the evidence underlying each, assuming independence of the items of evidence being pooled.

• In words, the DS combination rule is defined mathematically by intersecting the subsets ("focal elements") that carry the mass of each component DSPM, then multiplying masses to find the pooled mass, and lastly accumulating and renormalizing as needed.

Examples of Independent DSPMs

- DSPMs come in several forms. In the nonparametric model, there is explicit DS independence of successive sample u_i, and implicit independence assumed when Boolean combination is applied to the observed sample x_i. More broadly, common examples from statistical practice are marginal independence in contingency tables, and independence of prior and likelihood in Bayesian inference. And of course Boolean combination is all over the place.

A “Standard DS Protocol”

• Set up the SSM, define a set of independent DSPMs on margins of the SSM, up-project to the full SSM, combine at the level of the full SSM, and down-project to margins of interest.

• DS is basically a personalist theory. It is “your” choice to adopt a formal uncertainty model and analysis for reporting or decision-making. Scientific applications of the formal theory of probability typically make many implicit and explicit independence assumptions.

From a Mass Distribution to (p, q, r) Inferences

• Given any subset of the SSM, or equivalently an assertion that the true state of the system lies in the subset, the total mass that must belong to the subset defines the probability "for" the assertion, denoted by p, while the total mass that must belong to the complementary set defines the probability "against" the assertion, and is denoted by q. The remaining mass straddling the subset and its complement is denoted by r, and is called the probability of "don't know". DS replaces the familiar pair (p, q) with $p + q = 1$ with a triple (p, q, r) with $p + q + r = 1$.

Alternative Terminology

• In my 1960s papers, p was called <u>lower probability</u>, while p + r was called <u>upper probability</u>. Shafer's 1976 monograph called these quantities <u>belief</u> and <u>plausibility</u>, respectively. Different terms are used interchangeably.

• I now prefer (p, q, r) because it puts a focus on r, hence points to a species of uncertainty that Bayesian inference is unable to address (a species identified by economists Knight and Keynes long ago).

Fiducial as a Special Type of DS

• The DS scheme outlined in preceding slides grew out of R. A. Fisher's fiducial reasoning. In effect, the concept of probabilities of "don't know" created a broad foundation that includes examples of fiducial inference as a special DS type restricted to p and q with p + q = 1, while providing precise concepts and rules that Fisher was never able to formulate.

Back to the Nonparametric Example

- Included in his many basic contributions to 20th Century inferential statistics, R. A. Fisher invented what later came to be called nonparametric statistics. His treatment involved fiducial probability, a concept regarded as suspect by most later statisticians. When fiducial inference is regarded as a specialization of DS inference, however, a clearly articulated framework for Fisher’s logic is created. This in turn may serve to rebut some of the criticisms of fiducial reasoning.

How Fisher Introduced Nonparametric

• In his 1939 *Annals of Eugenic* obituary of W. S. Gosset, simply entitled ‘Student’, Fisher referred to a 1908 remark by Student to the effect that in a sample of size 2 from a continuous univariate population, the probabilities are ¼, ½, ¼ that the population median lies between the two sample observations. Fisher specifically referred to Gosset’s probabilities as “fiducial” probabilities.

Fisher's Extension of Gosset's Example

• Fisher used the obituary article to extend Gosset's theory from n = 2 to general n, in effect introducing the modern nonparametric model, where the $U_i = F(X_i)$ are independently uniformly distributed on (0, 1). In fiducial terms, the U_i are called pivotal quantities, from which Fisher infers that the $F(X_i)$ are i.i.d. uniform on (0, 1) given the observed data. According to Fisher, these can be used by scientists to associate inferential probabilities with statements about any population quantiles.

How DS Extends Fiducial

• The SSM defined for the nonparametric model includes all possible continuous cdfs, not just the values of the cdf at the observed X_i.

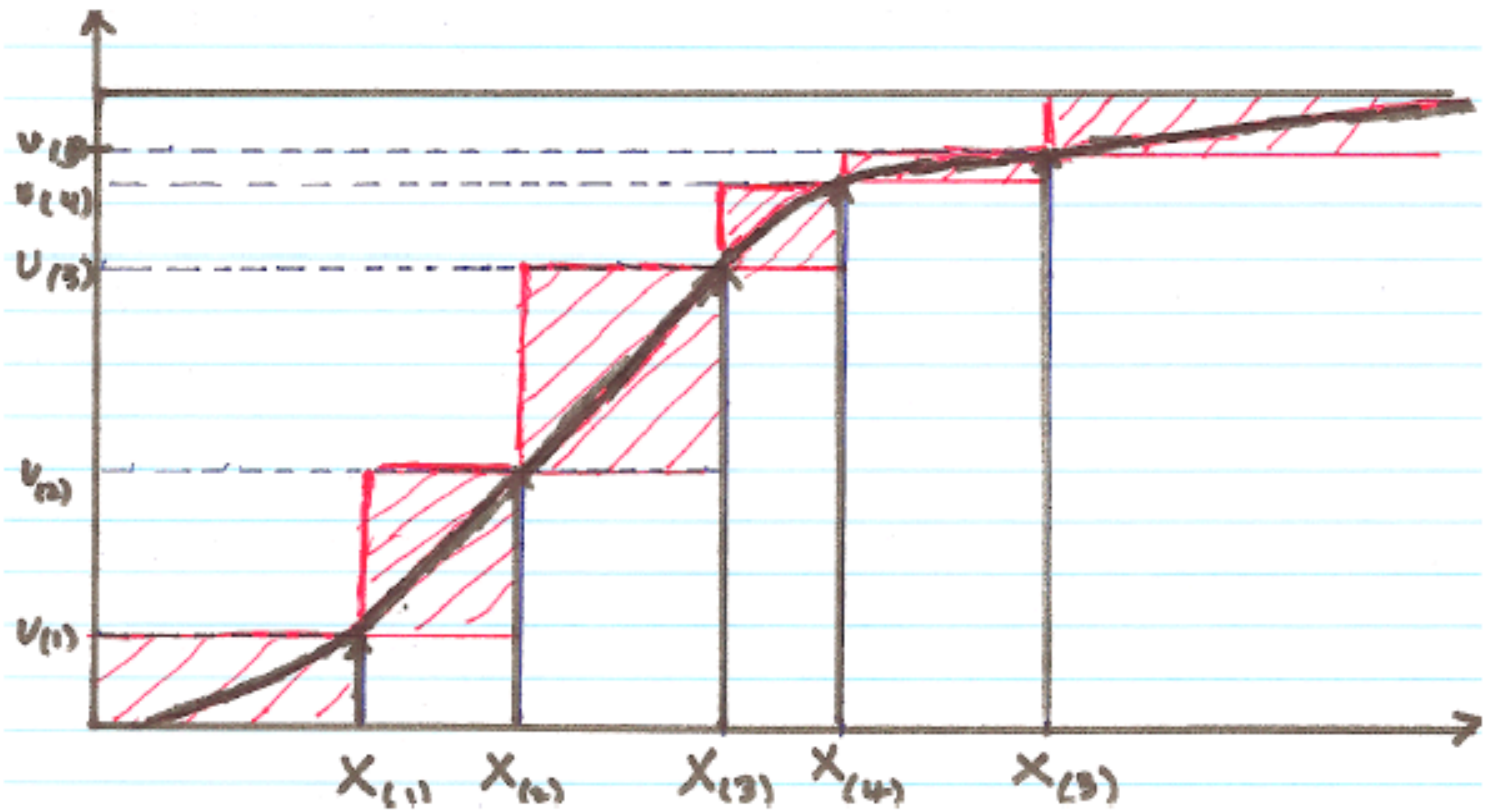
• The concept of DSPM can accommodate all values of F(x) because it assumes a mass distribution on subsets of the SSM. It does this through up-projection from margins, such as from individual observed x, and from assumed uniformly distributed u_i on marginal $U_i = F(X_i)$.

• The resulting inferences generally involve (p, q, r) with $r > 0$, as illustrated in two ways in the rest of my talk.

The First Illustration with $r > 0$

• Essentially any assertion about the unknown $F(x)$ that is either true or false has an associated (p, q, r). A simple example asks questions about $F(b) - F(a)$ where the interval (a, b) is specified without regard to the placement of the observations x_i on the X-axis.

• I'll explain this without explicit formulas with reference to the following rough picture that illustrates the class of posterior cdfs $F(x)$ that remain possible after the observations x_i are in hand, and a draw from the probabilistic variables $u_1, u_2, \dots, u_n$ is fixed.

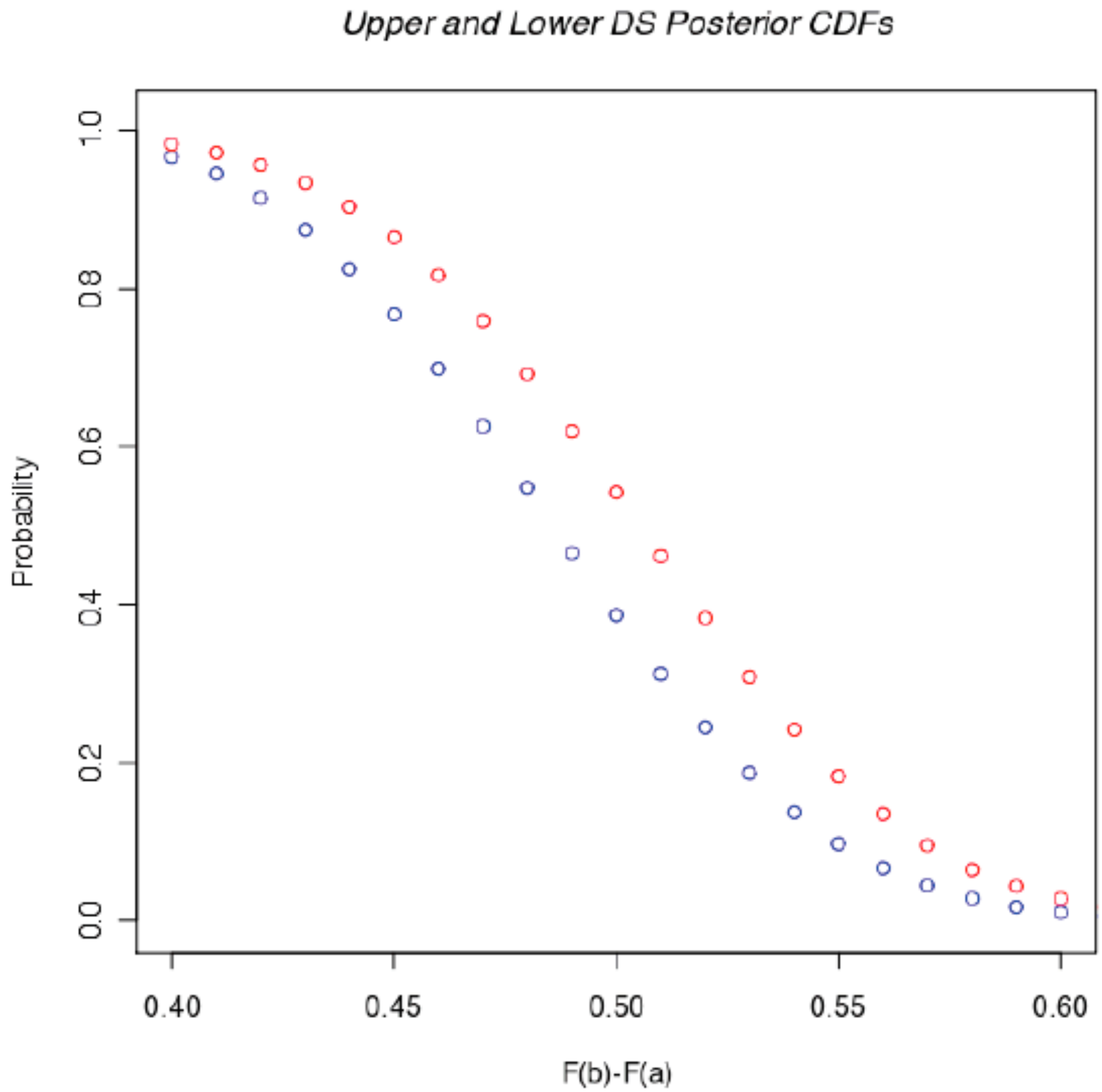


U(5)
U(4)
U(3)
U(2)
U(1)
X(1)
X(2)
X(3)
X(4)
X(5)

Some Details

• I explain here why upper and lower posteriors are provided by beta distributions.

• A numerical illustration in the case where n = 100, and a is between the 25^{th} and 26th data point, while b is between the 75^{th} and 76^{th} data point. The resulting (p, q, r) that F(b) – F(a) exceeds .58 is (.0281 , .9354 , .0365). The corresponding upper and lower posteriors are shown on the next slide.



Upper and Lower DS Posterior CDFs
Probability
F(b)-F(a)
0.0
0.2
0.4
0.6
0.8
1.0
0.40
0.45
0.50
0.55
0.60

The Second Illustration

• Here I propose to "weaken" the standard nonparametric model in a way that can protect against possible selection effects associated with a collection of proffered bets. Suppose I am a broker wanting to quote odds for and against F(b) – F(a) exceeding any threshold for any choices of an interval (a, b). You might be reluctant to accept any particular bet with odds quoted by the standard nonparametric DS model, on the grounds that your counterparty may be correctly guessing that clustering of consecutive u_i explains clustering of observed x_i rather than large F(b) – F(a), allowing the offered bet to be biased in favor of the counterparty.

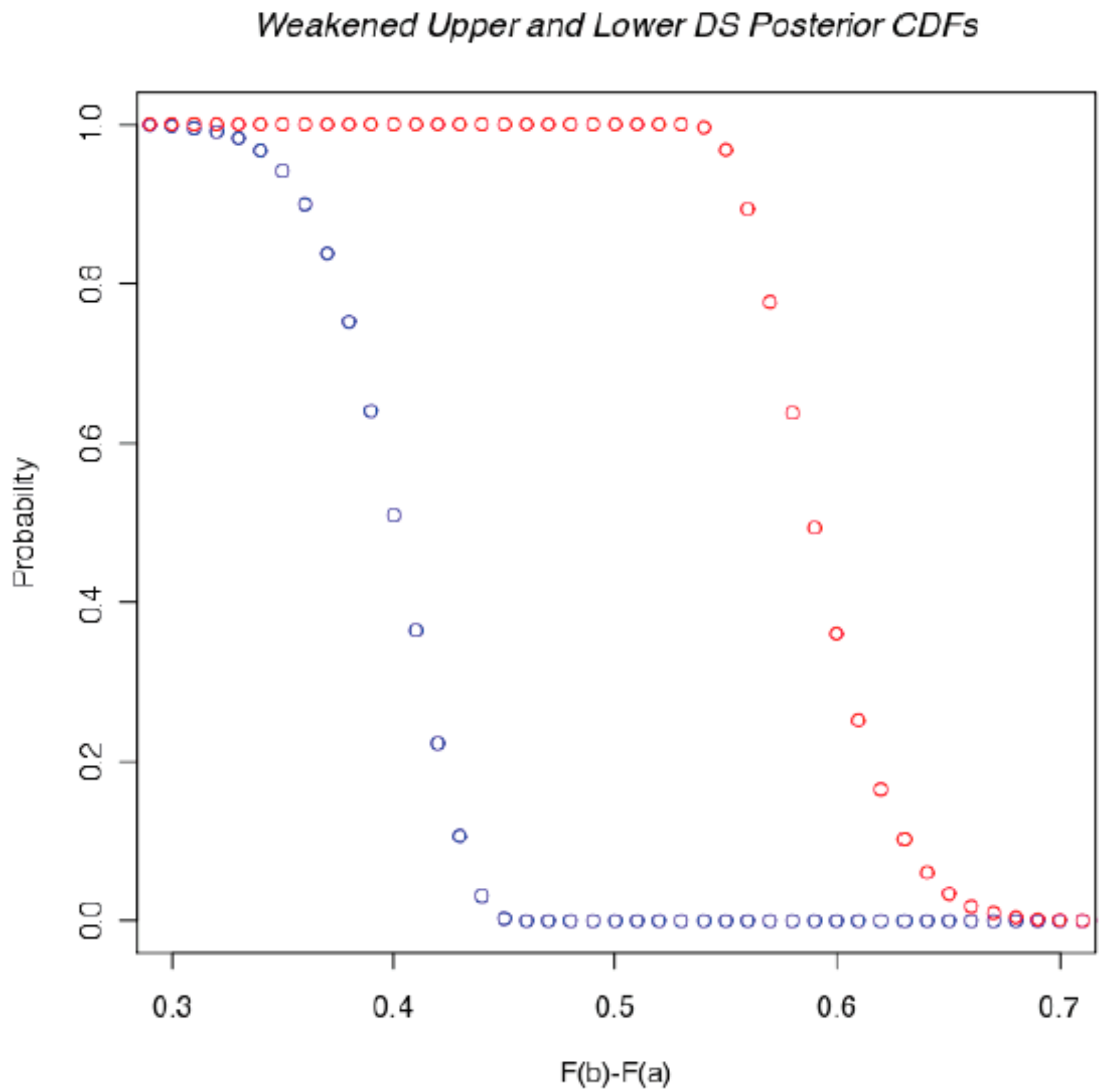
The Proposed Weakening

• A sample of n i.i.d. u_i determines a set of n + 1 ordered intervals on (0, 1). The lengths of these intervals are distributed uniformly on the simplex according to a symmetric D(1, 1, … , 1) Dirichlet distribution. It follows that a set $(u_1, u_2, \dots , u_n)$ after it is ordered can be used to determine any one of n + 1 shifted intervals after (0, 1) is mapped onto a circle of unit circumference. Suppose that for a given draw of x you "don't know" which of the n + 1 choices underlies the observed sample.

Consequences of the Proposed Weakening

- Since conditions “for” and “against” now require that all of the n + 1 possible samples associated with a given set of u_i satisfy a specified condition, it is evident that p and q in standard (p, q, r) inferences are reduced sometimes drastically. For example, (p, q, r) for the sample interquartile range exceeding .58 becomes (0 , .3626 , .6374).

- The upper and lower distributions on the preceding slide become much more separated as illustrated on the following slide.


Weakened Upper and Lower DS Posterior CDFs
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0.3
0.4
0.5
0.6
0.7
F(b)-F(a)

More Examples under the Proposed Weakening

• Fisher (op cit) extended Student's n = 2 example by giving three further examples as follows:
"Thus, for a sample of 6 the chance that [the population median] lies outside the observed range is only 0.03125; while for a sample of 9 the fiducial probability of it lying outside the penultimate pair is only 0.03906. Similarly, for a sample of 12, the chance is only 0.0386 of it lying outside the range of the central 8 observations."

• Under the proposed weakening scheme these precise fiducial probabilities become (p, q, r) triples (0, .25, .75), (0, .89, .11),(0, .8432, .1568), (0, .8287, .1713).

Endnote (1)

• The most common complaint about fiducial inference is that pivotal variables cannot both be independent of the population parameters and independent of the observations. In DS-land, this criticism has no force. For example, in the nonparametric model, if you should acquire prior evidence that specifies F(x), that is one DSPM. If you are comfortable with the u_i being i.i.d. uniformly distributed, this is another representation of evidence. The data constitute yet a third piece of evidence. Any pair, or even all three, of the evidential inputs, assuming you simultaneously adopt them, can be taken to be independent.

Why Might this be Counterintuitive?

• An implicit assumption in standard presentations of statistical inference is what might be called the random world hypothesis, namely, that an agency such as Nature determines a stochastic model, which in turn is used to “generate the data” by a “random” process. The statistician’s job is to perform uncertain inference about properties of the random process. The alleged flaw in Fisher’s reasoning is that the pivotal variables are used to obtain the data from the model, so it is absurd to regard the data as independent of the pivotal variables. The personalist viewpoint rejects the random world hypothesis as not being how real scientific systems operate.

Endnote (2)

- By concentrating on the elementary nonparametric example, I did not come back to my argument that DS modeling and analysis is suited to the analysis of real world complex systems. The claim is based on the separation of the SSM and the DSPMs in the process of formal model construction. The former may need to be very complex if essential aspects of the real world are to be incorporated, while the latter may need to be more circumscribed to fit limitations on “your” evidence. In other words, DS encourages representation and assessment of a realistic degree of “don’t know’.