LR material removed from main higher-oder paper

Rafal Urbaniak and Marcello Di Bello

2024-04-12

Table of contents

Abstract. Rational agents are often uncertain about the truth of many propositions. To represent this uncertainty, it is natural to rely on probability theory. Two options are typically on the table, precise and imprecise probabilism, but both fall short in some respect. Precise probabilism is not expressive enough, while imprecise probabilism suffers from belief inertia and the impossibility of proper scoring rules. We put forward a novel version of probabilism, higher-order probabilism, and we show that it outperforms existing alternatives.

1 LR Evidence aggregation: the simple case - SET ASIDE FOR BOOK

Rational agents are often tasked with aggregating pieces of evidence and assessing their value relative to a hypothesis. In this and the next section, we examine the question of how multiple items of evidence should be evaluated together. This question raises novel difficulties for both precise and imprecise probabilism. We show how higher-order probabilism can handle them.

For the precise probabilist, a natural measure of the value of the evidence is the likelihood ratio. This ratio is relative to a pair of competing hypotheses, say H and its negation $\neg H$ (though the two hypotheses need not be one the negation of the other). Relative to these hypotheses, the likelihood ratio of a single piece of evidence E is the probability of E given H divided by the probability of E given $\neg H$, or in short, $\frac{\mathrm{P}(E|H)}{\mathrm{P}(E|\neg H)}$. Degrees of evidential value (or support, strength) can be expressed as follows:

> the higher $\frac{\mathrm{P}(E|H)}{\mathrm{P}(E|\neg H)}$ (if greater than one), the more strongly E supports H.

The value of the evidence increases whenever $\mathrm{P}(E|H)$ increases or whenever $\mathrm{P}(E|\neg H)$ decreases. The higher $\mathrm{P}(E|H)$, the better the evidence at tracking H (a true positive); the lower $\mathrm{P}(E|\neg H)$, the better the evidence at avoiding $\neg H$ (a true negative). If the probability of E is the same given hypothesis H as given its negation, that is, the likelihood ratio equals one, the evidence would have no value for H.

Likelihood ratios can also be used for assessing the value of multiple pieces of evidence in the aggregate, again relative to a pair of hypotheses of interest. In the simplest case (for more complex cases, see the next section), multiple items of evidence all bear on the same hypothesis. Then, to obtain their combined evidential value, it is enough to multiply their individual likelihood ratios.

$$\frac{\mathrm{P}(E_1 \wedge E_1 \ldots E_k|H)}{\mathrm{P}(E_1 \wedge E_2 \ldots E_k|\neg H)} = \frac{\mathrm{P}(E_1|H)}{\mathrm{P}(E_1|\neg H)} \times \frac{\mathrm{P}(E_2|H)}{\mathrm{P}(E_2|\neg H)} \times \cdots \times \frac{\mathrm{P}(E_k|H)}{\mathrm{P}(E_k|\neg H)}$$

The equality holds provided $E_1, E_2, \ldots, E_k$ are probabilistically independent conditional on hypothesis H and its negation. Think, for example, at several diagnostic tests performed by independent laboratories or independent witnesses in a trial testifying about the same issue.

To see how likelihood ratios can be deployed, it is worth working through a specific case. In a murder case, the police recover two items of trace evidence, both against the defendant. First, hair found at the crime scene matches the defendant's hair; call this evidence 'hair.' Second, the fur of the defendant's dog matches the fur found in a carpet wrapped around one of the bodies; call this evidence 'fur.'[1] The two matches favor the hypothesis that the defendant (and the defendant's dog) must be the source of the crime traces; call this hypothesis 'source'. If the two matches are independent lines of evidence (conditional on the source hypothesis and its negation), their likelihood ratios can be multiplied:

[1] The hair evidence and the dog fur evidence are stylized after two items of evidence in the notorious 1981 Wayne Williams case (Deadman, 1984b, 1984a).

$$\frac{\mathrm{P}(\text{fur} \wedge \text{hair}|\text{source})}{\mathrm{P}(\text{fur} \wedge \text{hair}|\neg\text{source})} = \frac{\mathrm{P}(\text{fur}|\text{source})}{\mathrm{P}(\text{fur}|\neg\text{source})} \times \frac{\mathrm{P}(\text{hair}|\text{source})}{\mathrm{P}(\text{hair}|\neg\text{source})}$$

So far so good. But how do we fill in the precise probabilies? The numerators can be equated to one: if the defendant is a contributor, the laboratory will declare a match for sure. This is a simplication, but it will do for our purposes. To fill in the denominators, a trial expert will provide so-called match probabilities. They express the likelihood that, by coincidence, a random person (or a random dog) who is not a contributor would still match. The match probabilities are approximated by counting how many matches are found in a representative sample of the human population (or the canine population). Suppose the matching hair type occurs 0.0253 times in a reference database, and the matching dog fur type occurs 0.0256 times in a reference database (more on how these numbers are calculated soon). These frequencies can fill in the match probabilities. Putting everything together:

$$\frac{\mathrm{P}(\text{dog}|\text{source})}{\mathrm{P}(\text{dog}|\neg\text{source})} \times \frac{\mathrm{P}(\text{hair}|\text{source})}{\mathrm{P}(\text{hair}|\neg\text{source})} = \frac{1}{0.0252613} \times \frac{1}{0.025641} = 1543.862069$$

The resulting ratio is large. The two matches, combined, strongly favor the source hypothesis.

This is the story about evidence aggregation told by the precise probabilist. But this story misses something crucial. As it happens, the match probability for hair evidence is based on 29 matches found in a sample database of size 1148, while the match probability for the dog evidence is based on finding two matches in a smaller database of size 78. The relative frequencies are about .025 in both cases, but the two samples differ in size. The smaller the sample, the greater the uncertainty about the match probabilities. So, for individual pieces of evidence, simply reporting the exact numbers makes it seem as though the evidential value of the matches is the same, but actually, it is not.[2] In the aggregate, multiplying the individual likelihood ratios further washes away this difference.

A better alternative is easily available: the evaluation of multiple items of evidence should take into account higher-order uncertainty. **?@fig-densities** (upper part) depicts higher-order probability distributions of different match probabilities given the sample data—the actual number of matches found in the sample databases. As expected, some random match probabilities are more likely than others, and since the sizes of the two databases are different, the distributions have different spreads: the smaller the database the greater the spread, the greater the uncertainty about the match probability. In light of this, **?@fig-densities** (lower part) depicts the probability distribution for the joint match probability associated with both items of match evidence, hair and fur evidence. The aggregate value of the two pieces of match evidence, then, is given by a distribution over possible likelihood ratios. The shape of this distribution conveys the degree of higher-order uncertainty about the value of the aggregate evidence. **Marcello (Rafal/Nikodem to add): Can we have a formula for how the two matches are combined in the higher-order approach? In precise probabilism, you multiply the individual LRs, in higher-order probabilism, what do we do formally? Can we also have a distribution of likelihood ratios? What happens if both the numerator and denominator in the LR are distributions?**

add distributions of LRs figure

[2] The match probabilities in the Wayne Williams case on which our running example is based were 1/100 for the dog fur, and 29/1148 for Wayne Williams' hair. Match probabilities have been slightly but not unrealistically modified to be closer to each other in order to make a conceptual point: the same first-order probabilities, even when they sound precise, may come with different degrees of second-order uncertainty.

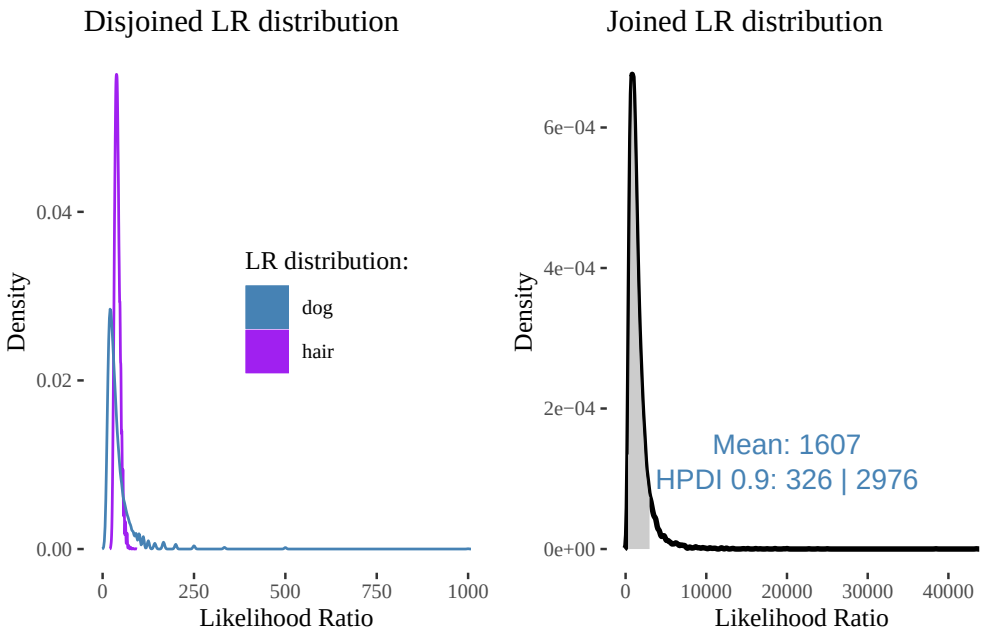

Figure 1: Distributions of dog and hair likelihood ratios and the resulting joint likelihood ratio. Created with the samples from beta distributions. Shaded area on the second one represents HPDI with 0.9 credibility.

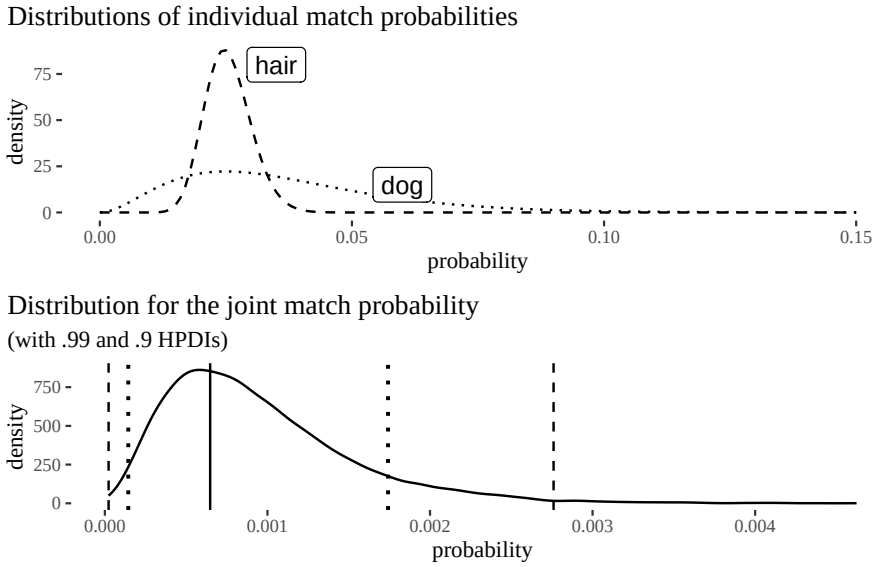

Figure 2: Beta densities for individual items of evidence and the resulting joint density with .99 and .9 highest posterior density intervals, assuming the sample sizes as discussed and independence, with uniform priors.

The precise probabilist might insist that the value of the evidence—one item or multiple items of evidence—is naturally captured by a precise likelihood ratio. In our running example, this ratio equals one over the first-order match probability, and our best assessment of this first-order probability is still the relative frequency of matches found in the database, whether large or small. If this is right, our best assessment of the match probabilities for both fur and hair evidence should be about .025, based on the relative frequencies 2/78 and 29/1148. If we were to bet whether a dog or a human picked at random would have the matching fur

or hair type, our odds should be .025 no matter the size of the database. This argument has some bite when evaluating single items of evidence. In fact, the expected values of the match probabilities for hair and match evidence—based on the higher-order distributions in **?@fig-densities** (upper part)—still end up being about .025. If, as the precise probabilist assumes, first-order probabilities is all we should care about, going higher-order would seem a needless complication.

This line of reasoning, however, breaks down when evaluating two or more items of evidence. What should our betting odds be for the proposition that a human and a dog, both picked at random, would have the matching fur and hair type in question? For the precise probabilist, the answer is straightforward: on the assumption of independence, it is enough to multiply the .025 individual match probabilities and obtain a joint match probability of 6.4772626×10^{-4}. The higher-order probabilist will proceed differently. In assessing first-order match probabilities, they will retain information about higher-order uncertainty as much as possible. This can done in two steps: first, aggregate the higher-order distributions for the two-match probabilities and obtain a higher-order probability distribution for the joint match probability (see **?@fig-densities**); next, to obtain our best assessment of the first-order joint match probability, take the expected value of this latter distribution. The higher-order probabilist will assign 9.381365×10^{-4} to the joint match probability, a value greater than what the precise probabilist would assign.

So, the higher-order and precise probabilist will disagree about the betting odds for the proposition that a human and a dog, both picked at random, would have matching fur and hair type. The disagreement will become even starker as a larger number of independent items of evidence are evaluated.[3] Who should be trusted? Since the higher-order probabilist takes into account more information—that is, the higher-distributions—there is good reason to think that the higher-order probabilist should be trusted more than the precise probabilist.[4]

Imprecise probabilism will also run into its own problems when assessing the value of aggregate evidence. Recall that the probability measures in the representor set are those compatible with the evidence. The problem is that almost any random match probability will be compatible with any sample data—with any number of matches found in a reference database. This point should be familiar from the earlier discussion. Think by analogy to coin tossing: even a coin that has a .99 bias toward tails could come up heads on every toss. This series of outcomes is unlikely but possible. Similarly, a hair type that has a match probability extremely small could still be found several times in a sample population. So, it is not clear how to proceed if one takes seriously the binary notion of compatibility. Imprecise probabilism is too permissive because almost any match probability will be compatible with the data.

Another option for the imprecise probabilist is to rely on reasonable ranges of match prob-

[3] Consider the simple case of independent items of evidence whose individual match probabilities are .025. For three, five and seven items of evidence, the joint match probabilities would be: 1.25×10^{-4},3.125×10^{-7} and 7.8125×10^{-10} (for the precise probabilist) and 5.3363868×10^{-4}, 1.6754185×10^{-5} and 9.9986742×10^{-7} (for the higher-order probabilist, based on small databases of size 20).

[4] As a further illustration of this point, consider a couple of variations of our running example. First, suppose the match probabilities associated with two matches are both set to .05 since they are based on the following relative frequencies: one match occurs in a dog fur database and one match occurs in a human hair database, where both databases are small, say of size 20. By multiplying the individual $1/.05$ likelihood ratios associated with the two matches, their evidential value against the defendant would seem quite strong: $1/.05 \times 1/.05 = 400$. However the match probabilities are based on frequencies resulting from small databases, so their evidential value should be rather weak. Precise probability here seems to exaggerate the aggregate value of the evidence. Following higher-order probabilism, the joint likelihood ratio would be 237.8675988, a significantly smaller value. On the other hand, if the same .05 match probabilities were based on larger databases, the evidential value of the two matches should be correspondingly greater, but precise probabilism would make no difference. If, for example, 1,000 hair and fur matches are found in databases of size 20,000, the higher-order probabilist would assign 441.2059925 to the joint likelihood ratio, a much greater value than before. This outcome agrees with our intuitions.

abilities. Suppose these ranges are (.015,.037) (.002, .103), for hair and fur evidence respectively in our original case.[5] As expected, the range is wider for dog fur match evidence than hair match evidence: the uncertainty about the dog fur match probability is greater since the sample database was smaller. This is a desirable feature of the interval approach. Now, to assess the joint uncertainty, it is enough to focus on what happens at the edges of the two intervals. Reasoning with representor members at the edges of the intervals will yield the most extreme probability measure the impreciser is committed to, the worst-case and best-case scenarios. We end up with a new range for the joint match probabilities, (.00003, .003811).[6] The corresponding likelihood ratios could be as high as 3.3333333×10^4 or as low as 262.3983. The two matches could be much stronger or much weaker evidence than previously thought.

Using plausible ranges for the match probabilities leaves the impression that any value in the interval is just as good as any other. Perhaps we should pick the middle value as representative of the interval. However, relying on the entire interval or the middle value will misrepresent the evidence. To see why, consider again **?@fig-densities** (lower part) which depicts the probability distribution for the joint match probability. Interestingly, this distribution is not symmetric. So the most likely value (and the bulk of the distribution, really) does not lie in the middle between the edges. Therefore, only relying on the edges—or taking central values as representative of the interval—can lead to overestimating or underestimating the probabilities at play.[7]

Another problem in taking intervals as representative of the value of the evidence is that they will tend to widen as more items of evidence are evaluated. The size of the likelihood ratio interval was initially 39.64 (hair evidence) and 490.29 (fur evidence). After aggregating the two items of evidence, the likelihood ratio interval widened to 3.3070935×10^4. The size of the match probability interval was initially -0.022 (hair evidence) and -0.101 (fur evidence). After aggregating the two items of evidence, the match probability interval narrowed to -0.003781. Posterior interval (starting with 1:1 prior odds) was initially 0.0209015 (hair evidence) and 0.0913857 (fur evidence). After aggregating the two items of evidence, the posterior interval narrowed to 0.0037665.

All in all, precise and imprecise probabilism does not fare well in modeling the value of evidence in the aggregate. Instead, the evaluation of multiple items of evidence should take into account higher-order uncertainty (as illustrated in **?@fig-densities**). Whenever probability distributions for the probabilities of interest are available (and they should be available for match

LR intervals widen but match and posterior probability intervals do not? How does that work? How can we claim that uncertainty increases?

[5] These are 99% credible intervals starting with uniform priors. A 99% credible interval is the narrowest interval to which the expert thinks the true parameter belongs with probability .99. For a discussion of what credible intervals are, how they differ from confidence intervals, and why confidence intervals should not be used, see Kruschke (2015).

[6] Redoing the calculations using the upper bounds of the two intervals, .037 and .103, yields the following:

$$\text{P}(\text{dog} \wedge \text{hair}|\neg\text{source}) = .037 \times .103 = .003811.$$

This number is around 5.88 times greater than the original estimate. Given this number, the two matches are much weaker evidence for the source hypothesis than previously thought. The calculation for the lower bounds, .015 and .002, yields the following:

$$\text{P}(\text{dog} \wedge \text{hair}|\neg\text{source}) = .015 \times .002 = .00003$$

This number is around 0.46 times lower than the original estimate. Given this number, the two matchs are much stronger evidence for the source hypothesis than previously thought.

[7] The calculations for the joint interval assume that because the worst- or best-case probability for one event is x and the worst- or best-case probability for another independent event is y, the worst- or best-case probability for their conjunction is xy. However, this conclusion does not follow if the margin of error (credible interval) is fixed. Just because the probability of an extreme value x for one variable X is .01, and so it is for the value y of another independent variable Y, it does not follow that the probability that those two independent variables take values x and y simultaneously is the same. In general, it is impossible to calculate the credible interval for the joint distribution based solely on the individual credible intervals corresponding to the individual events.

evidence and many forms of scientific evidence whose reliability has been studied), those distributions should be reported for assessing the value of the evidence. This approach avoids hiding actual aleatory uncertainties under the carpet. It also allows for a more balanced assessment of the evidence, whereas using point values or intervals may exaggerate or underestimate the value of the evidence.

A couple of clarifications are in order. First, the problem we are highlighting is not confined to match evidence. Say an eyewitness testifies against the defendant: they saw the defendant near the crime scene at the relevant time. To assess the value of this testimony, one should know something analogous to the match probability: if the defendant was not there, how probable is it that the witness would still say the defendant was there? Or suppose a medical test for a disease turns out positive. Here again, to assess the evidential value of the positive test, one should know how probable it is that the test would still turn out positive even when a patient is actually negative. And so on. These false positive probabilities are usually derived from sample-based frequencies in surveys or experiments: how often witnesses misidentify people; how often tests misdiagnose; etc. So, depending on the sample size, the false positive probabilities will have different degrees of uncertainty, and the latter should be taken into account when evaluating eyewitness testimonies, diagnostic test results, and many other forms of evidence. At the same time—and this is the second clarification—this discussion is not meant to suggest that the problem we are highlighting is confined to differences in sample size; it is broader than that. Probabilities can be subject to uncertainty for other reasons, for example, when they are derived from a probability model for which there is little support, or when the sample size is large but unrepresentative. So, in short, the problem of higher-order uncertainty is widespread and goes beyond match evidence and questions of sample size.

Deadman, H. A. (1984a). Fiber evidence and the wayne williams trial (conclusion). *FBI L. Enforcement Bull.*, *53*, 10–19.

Deadman, H. A. (1984b). Fiber evidence and the wayne williams trial (part i). *FBI L. Enforcement Bull.*, *53*, 12–20.

Kruschke, J. (2015). *Doing bayesian data analysis (second edition)*. Boston: Academic Press.