Marcello's Comments on the Chapter about Weight

9/1/2022

1 Purpose of this document

Marcello's comments on the chapter about weight.

2 Weight and completeness (Sec 5)

Rafal writes:

> So the second difficulty is that on this approach the weight of evidence becomes very sensitive not only to what the actual evidence is, but also to what an ideal evidence in a given case should be. unless as clear and epistemologically principled guidance as to how to formulate such ideal lists is available, this seems to open a gate to arbitrariness. Change of awareness of one's own ignorance, without any major chance to the actual evidence obtained, might lead to overconfidence or under-confidence in one's judgment. Moreover, it is not clear how disagreement about weight arising between agents not due to evidential differences, but rather due to differences in their list of ideal items of evidence should be adjudicated.

I am not sure. What makes a body of evidence complete is subjective, since it depends on what one knows about a given situation. But this fact cannot be used as a criticism for a theory of weight based on completeness. If I did not know the defendant had a huge archive of documents in his office, I might think that all the evidence I have (without the documents in the archive) is complete. But when I learn about the existence of that archive and realize that the evidence lacks the documents in the archive, then my evidence is clearly incomplete. This does not makes the assessment of weight-as-completeness subjective. This is just how things should be.

Perhaps there are different levels of analysis here:

- assess completeness of evidence based on an ideal list that would apply universally to any case like the one under consideration (script approach)
- assess completeness based on a specific recounting of what happened that is agreed by both parties (shared narrative approach).
- assess completeness based on a specific recounting of what happened pout forward by one of the parties (partisan narrative approach).

Arbitrariness might exist in the script approach (we might disagree about the right script to apply to determine the ideal list of items of evidence), but does not exist in the narrative approach.

3 Imprecise probabilities (Sec 8 and 9)

These two sections are very interesting, but we need to think about they can fit in the chapter as such. Most of the examples (and counterexamples) in these sections are about coin tossing and sample size. I think we might need to consider examples of quantitative evidence in the law, DNA evidence, multiple reference populations or different sample sizes. To warrant a discussion of imprecise probabilities here, we need to show that imprecise probability measures (and also Joyce's notion of weight) have a prima facie applicability to the law and then we can show that they are inadequate for various reasons.

The bit about proper scoring rules and teh Brier score is particularly interesting. This can perhaps belong to the section on accuracy. But it does seem a requirement of any weight measure, juyst like any probability measures, that we can connect it up to accuracy in some way.

3.1 Discussion of Joyce measure

The discussion of Joyce is very cool, but I am not sure I can completely follow it. I wonder whether this could be a separate, short Analysis paper in which to criticize Joyce's approach. There seems to be enough materials for a cool short paper here.

It seems to me that Joyce's measure is something like the difference of the variances — the variance before you update by the evidence and the variance (something like a spread) after you update by the weight. Essentially, a credence for him is an expectation—the weighted average of the possible chances. Thus, weight is the difference between the variance (spread) of the credence before and after updating. Is this right?

If so, this characterization is not completely different from what you are suggesting, but as I said, I do not completely follow the discussion here.

Another question, to what extent is Joyce's guiding idea correct and where does it go wrong and why does it go wrong in the way you illustrate?

The next question is, how does the higher order approach you propose get around the problems that Joyce's measure runs into? And why is it, exactly, that your measure fares better than Joyce's?

3.2 More general "big picture" observations

It seems to me that there are two motivations for talking about weight (which is the topic of the chapter):

- The first motivation has to do with Kaynes considerations about the weight of arguments as opposed to balance. That is one bucket of considerations and the stuff about completeness goes in the same bucket. All this stuff has loosely to do with the "quantity" of evidence as opposed to balance.
- The second motivation has to do with coin tossing examples that show we can be uncertainty about the bias of the coin, not only uncertain about the outcome of the coin toss. And these examples motivate the imprecise probability approach.
- These two motivations – (a) weight/balance and (b) imprecise probabilities – seem to be quite independent, but also related in some ways. We can say that these are the two theoretical motivations for this work. Do you agree with this distinction of the two main motivations?
- Perhaps, your higher order approach is unique in that it combined these two strands of literature. At the same time, one recurrent question some people might have is, in what way is your higher order approach novel? In some sense, it is not, because everybody agrees that there can be uncertainty about the bias of a coin, and if the bias is a parameter, we can then represent this uncertainty with a distribution. This is standard statistics, right? But the novelty of your approach consists in the details, and specifically, taking these standard ideas about the uncertainty about a model parameter and apply them to give account of weight, is this right?
- In general, I think it'd be good to trace the intellectual lineage of the approach to allow the reader to place it somewhere in the conceptual landscape. For example, you are adamant that it is a Bayesian approach (at least you told me it is), but then what about the remarks that Meester makes – see my comments later about your Section 12?

4 Sec 10 (highere order approach)

The general idea here is clear, but I wonder if a simple example or two for legal application is helpful. Perhaps an example with DNA evidence and sample size or something to that effect.

5 Weight of a distribution (Section 11)

- This and the next one are the crucial sections. This is the map I have now in my mind to follow what is going on in this section:
  i. Notion of information/entropy in general
  ii. Entropy of distributions
  iii. Difference of entropy between distributions (cross-entropy, KL divergence)
  iv. weight of a distribution D is the difference (in some sense to be specificed) between the entropy of the distribution D compared to the uniform distribution (which by default has maximal entropy). But, crucially, this measure iof weight is NOT the KL divergence.
  v. weight of evidence (in the next section)
- The move from the example with the three forks to a distribution of parameter values (each associated with a different probability) is not completely clear.Basically this is the step from (i) "entropy in general" to (ii) entropy of distribution. I can sort of see the connection, but it is not spelled out clearly. This is the part that says "A measure of (lack of) information contained in a whole distribution, is entropy, which is the average Shannon information:..."
- Again, in the discussion of entropy it would be good to have a clear running example, possibly legal in nature.
- Using the grip approximation for continuous distributions is fine, but what is the reason? You seem to say it is because we will compare continuous and discrete distribution. That seems sensible. Can you give an example?
- KL divergence. This is the difference between the entropy associated between the two distribution and because of the properties of logarithm it is the log of the ratio. Right?
- After that, you write "The idea is that the more informative a piece of evidence is, as compared to the uniform distribution, the more weight it has, on scale 0 to 1:" Are you talking about a piece of *evidence* or a *distribution*?
- Footnote 14 seems crucial here ("The reader might ask: why not to use the Kullback-Leibler divergence from the uniform distribution instead? Because this divergence does not measure the difference in how informed the distribution is. "). This might need more explanation. But if the KL divergence does not measure how informative a distribution is, then what does it measure? What are the sues for KL measure as opposed to the weight measure you propose?Also, I think one can understand your measure of weight without the KL divergence measure, which seems to be more complicated and meant to do something else.
- Another question is why the weight of a distribution P isn't simply 1-H(P). What are the advantages of comparing with the uniform distribution and using 1-(H(P)/H(uniform)? Another option for weight of a distribution would be the difference between the entropy of the uniform and the entropy of P, that is, H(uniforms)-H(P). Is this equivalent to something like the log of the ratio of the distribution, log(P/uniform)? Why then not take 1-log(P/uniform)? At any rate, more explanation why you picked that measure and excluding others in the vicinity would be useful.
- The graphs used to compare Urbaniak-weight with Joyce-weight need additional explanation. You earlier complained that Joyce-weight display strange patterns. But O noticed that the graph of Urbaniak-weight are similar to those of Joyce-weight. So do you see strange patterns here too, or not? For example, why does the weight of a beta distribution drop as the ratio heads/tails become less sweked, but suddenly increases when teh ratio is close to 50/50? Is this a normal behavior? You might to spend mroe time, suggesting that this behavior is intuitive and to be expected.
- You say: "the entropy decreases with the number of observations". A beta(50,50) will have greater weight than beta(5, 5). I can see why that is because beta(50,50) is more sharply concentrated around .5 than beta(5,5). But some explanatio might be helpful.
- You also say: "it [=entropy] decreases faster if the proportions are closer to the extremes. " Beta(80,20) should presumably have more weight than beta(50,50) since it is more extreme. Why is that exactly? Why is it that a more skewed distribution is weightier? I can sort of see why, but more explanation would be needed. The problem here is that—looking the general graph of beta distribution and also from the comparison with Joyce— it seems that beta(50,50) is weightier than beta (55,45). There is slight increase in weight when we get to 50/50. Why is that? is this intended? Isn't that alike one of the strange patterns that we see in Joyce-weight?

- There is a potential ambiguity (or potential source of confusion) in the illustration using beta distributions. The guiding intuitions is that beta(a, b) represents a set of observations with a successes and b failures. So the beta distribution in fact represent a bunch of observations. Under this interpretation, beta(50,50) is weightier than beta (5,5) because it represents more observations. So, in some sense, here we are not thinking about the weight of a distribution, but the weight of the observations (the more observation, the weightier they are). At the same time, your account shows that the mere number of observation does not solely determine weight. What also matters is how the outcomes of he observation are allocated (the proportion of successes versus failures). So, then, it seems that weight tracks (a) how many observations there are and (b) how they are distributed of structured? So, the question is, when we talk about the weight of a distribution, are we talking about anything other than the weight of (a) and (b)—i.e. observation plus how they are structured?
- One interesting thing to know would be, when do two beta distributions (with different levels of proportion) have the same weight? For example, consider beta(80,20), so the proportion here is 80/20. Now let's take a proportion like 60/40? What would be a beta distribution that satisfies this proportion and that has the same weight as beta(80,20)? I suppose that the beta distribution will have to have more sample observations. But how many more?
- Here is another way of thinking about this problem intuitively. Suppose I want to know whether this coin is fair (0.5). Contrast this with suppose I want to whether the coin is biased with bias 0.8 heads. The question is how many observations, and in what proportions, would I need to be equally sure (=to have equally weighty observations) to conclude that the coin is fair versus 0.8 biased toward heads?

6 Higher order and BN (sec 12)

This section is interesting, but only sketches the ideas. We might need to think about the following:

- What are some of the recurrent objections—theoretical and practical—against using probabilities for assessing evidence in complex legal cases?
- Does the higher order approach address some of these objections, and if so, which ones?
- Which objections are instead still outstanding?

We need to be clear about the scope and applicability, but also the limitations of the higher order approach.

In this sense, it is interesting to read a short, 8-page 2019 paper "The Limits of Bayesian Thinking in Court" by Ronald Meester (https://doi.org/10.1111/tops.12478Citations), part of the edited volume "Models of Rational Proof in Criminal Law," Henry Prakken, Floris Bex and Anne Ruth Mackor.

Meester is criticizing Christian Dahlman's model of the Dutch murder case Simsonhaven (interesting case to look at, by the way, perhaps more than Sally Clark). Consider some common criticism Meester makes. Others made similar criticisms (Verheij, Allen, Cohen etc):

Criticism 1: Unfounded numbers for the likelihood ratio

Ronald Meester writes:

> My greatest concern with Dahlman's paper is his way of assigning numerical values to likelihood ratios. It is hard to take these numbers seriously, and for most of his conditional probabilities, it would be hard to imagine that a well-founded number can be assigned to it. Subjectivity does not mean that anything goes. The problem is that Dahlman gives no arguments for his choices. I believe that this is so because for many of his likelihood ratios, it would be impossible to assign a number in a well-founded way. ... Take, for instance, the following quote:
>
> "At the end of the day it must be more likely that we would see a bundle of evidence if EL is guilty than if he is innocent. I estimate it as five times more likely. Likelihood ratio: 5."
>
> Why 5? Why not 25? Why not 2? Such differences potentially matter a lot, especially (as is the case here) when many likelihood ratios must be multiplied.

I wonder if the higher order approach can in part rebut these criticisms. It would be good if it could. If one has no idea whatsoever about what the number should be, then I guess one would simply use the uniform distribution, right?

But what if one at least reasonably believes that the evidence has positive value? This might constraint the distribution somewhat. If so, how? What are some other examples?

Criticism 2: The problem of negation

Ronald Meester writes, same paper as above:

> it is possible that one has no belief in a particular statement but at the same time no belief in its negation either. If, for instance, a car caused an accident, and it is confirmed that there were two persons in the car at the moment of the accident, then it may happen that there is no information whatsoever about who was driving. In such a case, one's belief that person 1 was driving is small, but this does not imply that the belief that person 2 was driving is high. Hence, not all expressions of belief can be captured by probability theory. Epistemic uncertainty is, simply, more complicated than what probability theory allows for,

This is a common criticism of probability theory, starting with Cohen. It'd be good if the higher order approach has an answer here, which I think it does. Or perhaps the answer is already contained in using LR as a measure of the value of the evidence?

Criticism 3: LRs are not model parameters to be estimted

This is a technical criticism, but an interesting one, since one might think that in the higher order approach we are treating probabilities essentially as a parameters, and this – according to Meester who says he is a Bayesian– does fit well with the Bayesian approach. Ronald Meester writes, same paper as before:

> ...considering an extreme situation, namely a situation with no relevant data at all. In such a situation, obviously one cannot make a meaningful statistical estimate of any model parameter, and uncertainty remains. On the other hand, any likelihood ratio corresponding to such a situation is well defined and known, namely equal to 1. It is precisely equal to 1, without any uncertainty....So the likelihood ratio must be of a different nature than a model parameter.
>
> ....consider a situation of a full match of a suspect S with a DNA profile found at the scene of the crime, with population frequency p. If this p were know to us, then the likelihood ratio of S being the donor versus an unrelated person being the donor would be 1/p. But in reality we do not know p, and we describe our knowledge about p by a probability distribution. When we do that, the likelihood ratio becomes a function of this probability distribution; that is, it is a function of our knowledge about p, and not of p itself.
>
> If we make more observations, that is, if the data change, then so does the likelihood ratio, since our knowledge changes. The likelihood ratio is thus not a quantity that one could determine, or at least approximate arbitrarily well, if only enough data would be at our disposal. On the contrary, by its very nature, the likelihood ratio is a quantity that depends only on the data that we have seen (and our original conviction since one has to start somewhere). It is not some existing reality. There is simply no such thing as approximating a likelihood ratio "arbitrarily well" by performing more experiments. If we performed more experiments, the likelihood ratio would change, but it would be a different likelihood ratio depending on different data. It would not be a more precise version of the earlier likelihood ratio. ... One may even imagine a situation in which no uncertainty exists at all anymore, and in which the corresponding likelihood ratio is exactly expressible in terms of model parameters, in the above case as 1/p. This will then be the appropriate likelihood ratio in such a situation, but it, simply, expresses something different than a likelihood ratio based on less data. In classical estimation procedures, the estimates always refer to the same underlying parameters, but in the case of likelihood ratios this is not true. They express conditional probabilities based on certain knowledge, and if the knowledge changes, so do the probabilities.

He refers to his book Meester, R., & Slooten, K. (2020). Theory and philosophy of statistical evidence in forensic science. Cambridge, UK: Cambridge University Press – which we should probably have a look at.

Criticism 5: The logic of the case, not the numerical details

Ronald Meester is a Bayesian statistician, but he thinks that assigning precise numbers is too much and possibly misleading. This is his view about how the Bayesian approach can be useful:

> In my own casework, I use a Bayesian approach primarily to aid the reasoning in the case. I

use numbers only when these are both founded and relevant to the case. I help the judiciary with the logic of the case, and in my experience this is typically found very useful.

So, setting numbers aside when they are not available, does the higher order approach help in understanding the "logic of the case" better than the standard Bayesian approach? If so, how?

Criticism 6: People reason informally (and often correctly) without fully specificied probability distributions

This is a criticism that I heard many times from Verhej but also Allen. So the idea is that people are able to muddle through the evidence, and even without a fully specified probability distribution, they get at something correct. For example, what if one reasonably believes that one piece of evidence is stronger than another but does not know exactly how strong and on that basis one can correctly conclude that the evidence isn't strong enough to convict? How would that be represented? I think the key might how to represent reasonable, but somewhat informal, assessments of the evidence without full numerical precision. Can the higher order approach help here?

7 Weight of evidence (Sec 13)

- The start of this section is revealing:

  > "So far we have discussed the weight of a distribution, meant to measure how informed an agent is about an issue. If the agent starts with a uniform prior, this is a good enough approximation of how informed the evidence made them. But in general, how much more information is obtained is context-dependent. We want a prior-relative notion of weight, following the intuition that weight consideration should guide our information gathering also in making us stop collecting further evidence in light of what we already know. But for weight of evidence to have this feature, it has to depend on what we already know."

  So, if I understand this right, when you talk about the "weight of a distribution", you are actually talking about the "weight of the evidence/observations" assuming you started out from uniform prior. This makes sense since the weight of a distribution is measured against the baseline of a uniform distribution. Ss my interpretation right?
- Earlier I thought that "weight of a distribution" and "weight of evidence" were two different things. Now, if I understand this right, looks like "weight of evidence" is a generalization of the idea of weight of a distribution. Weight of a distribution is the weight of the evidence when the prior is uniform. Is this right?
- If the preceding point is right, it might also address the ambiguity I pointed out in my earlier remarks–ambiguity between "weight of distribution" and "weight of the observations" (which underline the distribution).
- The way you suggest to measure the weight of the evidence is as follows:

  > In a given context, consider your distribution for the target hypothesis H given what you already know. Then update on the evidence. This might increase the weight for H, if the evidence confirms your conviction, or decrease it, if it goes against what the previous evidence tells you. Take the difference between the prior weight and the posterior weight (Δw) as your measure of the weight of evidence in that context.

  The idea is to take the difference between W(prior distribution) and W(posterior distribution). But, in light my remarks above, I wonder why "weight of evidence" could not be measured, follows:

$$W(E) = 1 - \frac{H(\textit{posterior distribution (after updating on E)}}{H(\textit{prior distribution})}$$

  This is just a generalization of your earlier measure of weight of distribution. Your measure of the weight of a distribution had in the denominator the uniform distribution. Now, instead of the uniform, the weight of evidence has the prior distribution (which can be any distribution)

$$W(P) = 1 - \frac{H(\textit{distribution P}}{H(\textit{uniform distribution})}$$

Any thought about this approach and your approach in terms of delta-weight? Are the two approaches equivalent?

- The discussion of the rocking and the abused child is interesting, but not completely clear yet.

8 Weight and BN (Sec 14)

This is a very short section. I guess my questions are simply the same as those I formulated for Section 12.

8.1 Expected weight

The notion of expected weight is certainly interesting, but we need to discuss what it does. But here is an **objection**:

Nance (and also Hamer I believe) thinks that the missing evidence could go either way—positive or negative—so overall its expected impact on the existing evidence is null, right? So wouldn't the expected weight also be null?

Is the answer to this objection something like this (see my discussion of Hamer in my notes in the other file): depending on the causal structure of the Bayesian network, the expected weight of a given piece of evidence (which could turn out to have positive or negative probative value) need not always be null overall?

9 Weight and Accuracy (Sec 15)

You spend some time spelling out the correct notion of accuracy that should be used here. This measure of accuracy seems intuitively correct to me. Your "proper score' ' result is also very cool.

This section needs to be expanded, but since we already have so much material, we'll need to think about the what we want to say in the chapter!

Here are some other comments:

- Your measure is a generalization of the Brier score, right? So is it the case that in the special case where we have sharp probability value, your accuracy measure essentially collapses to the Brier score? Just want to make sure I understand.
- A general thought about weight and accuracy is this: a probability assessment that is based on more evidence (or weightier evidence) will tend to be more accurate (in the sense of being closer to the true value) than a probability estimate based on less evidence (less weighty evidence). Hamer show that more evidence reduced uncertainty (probability assessments will tend to be more concentrated toward 0 and 1, assuming true values are 0 and 1). Do you want to establish a result of this sort?
- If I understand the strategy correctly, the idea is this. (a) The first step is to formulate something like the Brier score but for distributions. You suggest to use the KL measure between the indicator distribution (true distribution) and the distribution in question. This seems right as intuitively it measures how distant a distribution is from the true distribution. (b) The second step is to show that a weightier distribution will also have a better KL measure. Is this right? The question is, how does this map onto the general idea that weightier evidence improves accuracy?