Comments on Hypothesis Selection

John Skilling

July 30, 2017

1 Inference

The sum and product rules of probability follow from simple symmetries, and cannot be broken without disobeying at least one of those elementary properties, thereby exposing the proposal to compelling counter-example. I start from Bayes' theorem (really just the product law) which is to be used to infer parameter(s) θ from data d.

$$\begin{array}{ccccc} \underbrace{\Pr(\theta)}_{\text{Prior}} & \underbrace{\Pr(d \mid \theta)}_{\text{Likelihood}} & = & \underbrace{\Pr(d)}_{\text{Evidence}} & \underbrace{\Pr(\theta \mid d)}_{\text{Posterior}} \\ \underbrace{\pi(\theta)}_{\text{measure}} & \underbrace{L(\theta)}_{\text{function}} & = & \underbrace{Z}_{\text{scalar}} & \underbrace{P(\theta)}_{\text{measure}} \end{array} \tag{1}$$

Marginalising over the unknown θ (*i.e.* using the sum rule) gives

$$\text{Evidence} \quad Z = \int L(\theta)\, \pi(\theta) d\theta \tag{2}$$

which — plausibly — is simply the mean (with respect to the prior) likelihood value $\langle L \rangle_\pi$, and thence

$$\text{Posterior} \quad P(\theta) = \frac{L(\theta)\, \pi(\theta)}{Z} \tag{3}$$

Although it's neater to write measure elements $\pi(\theta)d\theta$ and $P(\theta)d\theta$ as $d\pi$ and dP respectively, I will use the less sophisticated notation with π and P being densities over "volume" $d\theta$.

2 Hypothesis comparison

Bayes' theorem is the unique tool for inference generally. Here, we seek to compare rival hypotheses. Suppose that our prior π contains a switch which can be 0 (with probability $\frac{1}{2}$) or 1 (with complementary probability $\frac{1}{2}$), so that it can be written as

$$\pi(\theta) = \tfrac{1}{2}\pi_0(\theta) + \tfrac{1}{2}\pi_1(\theta) \tag{4}$$

where π_0 and π_1 are subsidiary rival priors (i.e. normalised measures).

The evidence decomposes as

$$Z = \tfrac{1}{2}Z_0 + \tfrac{1}{2}Z_1 \tag{5}$$

where

$$Z_0 = \int L(\theta)\, \pi_0(\theta) d\theta \quad \text{and} \quad Z_1 = \int L(\theta)\, \pi_1(\theta) d\theta \tag{6}$$

and then the posterior decomposes as

$$P(\theta) = \frac{L(\theta)\, (\frac{1}{2}\pi_0(\theta) + \frac{1}{2}\pi_1(\theta))}{\frac{1}{2}Z_0 + \frac{1}{2}Z_1} = \frac{Z_0}{Z_0 + Z_1} P_0(\theta) + \frac{Z_1}{Z_0 + Z_1} P_1(\theta) \tag{7}$$

where

$$P_0(\theta) = \frac{L(\theta)\pi_0(\theta)}{Z_0} \quad \text{and} \quad P_1(\theta) = \frac{L(\theta)\pi_1(\theta)}{Z_1} \tag{8}$$

are the subsidiary rival posteriors that would be inferred from π_0 and π_1 taken individually.

The above analysis follows the sum and product rules and is thereby incontrovertible. It allows us to compare hypotheses. Suppose that "Agent 0" proposes some hypothesis in the form of a prior π_0, while "Agent 1" likewise proposes π_1. If I choose to treat the agents symmetrically, I will assign prior $\frac{1}{2}$ to each,

$$\Pr(\text{Agent } 0) = \tfrac{1}{2} \quad \text{and} \quad \Pr(\text{Agent } 1) = \tfrac{1}{2} \tag{9}$$

and I will then be forced to use the above analysis. That amounts to taking the individual posteriors P_0 and P_1, and weighting them by posterior assignments

$$\Pr(\text{Agent } 0 \mid d) = \frac{Z_0}{Z_0 + Z_1} \quad \text{and} \quad \Pr(\text{Agent } 1 \mid d) = \frac{Z_1}{Z_0 + Z_1} \tag{10}$$

In this way, different agents acquire differing credibilities as their various hypotheses are tested against the data. The ratio

$$\frac{\Pr(\text{Agent } 0 \mid d)}{\Pr(\text{Agent } 1 \mid d)} = \frac{Z_0}{Z_1} \tag{11}$$

is known as the *Bayes factor*.

I could, of course, have assigned different prior weights in the first place, based on previous history or crude prejudice or whatever. As always, my prior reflects my personal judgement. That's Bayes. But the methodology is unambiguous. Unless it's isomorphic, any other proposal must conflict with the Bayesian analysis, thereby being in conflict with the basic symmetries underlying the sum and product rules.

Bayesian analysis is general. Rival hypotheses can even have different variables θ_0 and θ_1, in which case the analysis as formulated in the joint space $\Theta = \Theta_0 \times \Theta_1$ of all the unknowns factorises. Alternatively, in an application that has historically attracted attention, one hypothesis (the "null hypothesis" proposed by Agent 0) represents a selected subset of the wider hypotheses proposed by Agent 1.

3 The FBST

To test a null hypothesis $\theta \in \Theta_0 \subset \Theta_1$ against the wider alternative $\theta \in \Theta_1$, Pereira and Stern (PS) have proposed the "Full Bayesian Significance Test" (FBST) [ref: *Entropy*, **1**, 99–110, 1999]. The FBST selects a critical location $\theta^\star$ by maximisation. However, neither maximisation nor minimisation appear in the sum and product rules of probability, so the appearance necessarily involves conflict.[1]

It only remains to explain the FBST and demonstrate its failure. PS start by determining the greatest posterior density $P^\star$ (their symbol $f^\star$) that is attainable under the null hypothesis (at location $\theta^\star \in \Theta_0$). Now, it is methodologically unsound to maximise a density, because such selection is not coordinate invariant. As coordinates become (say) squeezed locally, the density of a conserved quantity must amplify in proportional compensation, thereby shifting the maximum. Consider, for example, a standard normal density in two dimensions (polar coordinates $x + iy = re^{i\phi}$):

$$f(r, \phi) = \tfrac{1}{2\pi} e^{-r^2/2}$$

Transformation of radius r to $u = \frac{1}{2\pi}(1 - e^{-r^2/2})^{1/2}$ flattens the density distribution to a constant

$$\tilde{f}(u, \phi) = \pi^{-1}$$

in the unit disk $u < 1$, which annihilates any procedure (such as FBST) that relies on maximisation. Furthermore, tiny subsequent transformations can generate a formal maximum anywhere, making such procedure completely unstable. The only way of avoiding this is to use the *ratio* of densities. On

[1] The variational principle known as maximum entropy does involve maximisation, but of the entropy functional to produce a probability distribution. It does not maximise within a probability distribution.

squeezing coordinates, both densities respond in the same inverse proportion, so their ratio is unaffected. Mathematicians know this invariance as the Radon-Nikodym theorem.

In their examples, PS use uniform priors, so in effect they are using the posterior/prior ratio, which is legitimate. Accordingly, we proceed on the understanding that their practical intention was to legitimately maximise value of likelihood instead of illegitimately maximising density of posterior. Having located the maximum likelihood value $L^\star$ consistent with the null hypothesis Θ_0, PS use it to define the "credible" domain $\{\theta : L(\theta) > L^\star\}$ within the wider hypothesis Θ_1, where the likelihood value exceeds $L^\star$.

Again, this is methodologically unsound. Probability calculus is concerned with prior and probability measures, not pointwise values. All one needs to do is add the immediate neighbourhood of global maximum likelihood to Θ_0 (which can be done within PS's general framework of equality and inequality constraints), and the credible domain is destroyed because there's none of it left.

Proceeding nonetheless, PS define the credibility (with regard to the wider hypothesis Θ_1) of the credible set as its integrated probability

$$\kappa^\star = \int_{L(\theta)>L^\star} P(\theta)d\theta \tag{12}$$

Finally, PS define their "evidence" $\texttt{Ev}$ (not the standard Bayesian evidence Z) for the null hypothesis as the complement

$$\texttt{Ev} = 1 - \kappa^\star = \int_{L(\theta)\le L^\star} P(\theta)d\theta \tag{13}$$

This is the probability that a random (with respect to the posterior) θ has a likelihood $L(\theta)$ small enough to be compatible with the null hypothesis.

That usage fails because a common likelihood bound does not require the wider set of low-likelihood points to faithfully represent the null-hypothesis domain.

4 Counter-example to FBST

Consider the simple one-dimensional likelihood function

$$L(\theta) = 1 + \epsilon \sin(8\pi\theta), \qquad 0 < \theta < 1, \quad \epsilon \text{ small.} \tag{14}$$

Take the null hypothesis to be $\theta = \theta^\star$ at some selected location $\theta^\star$. The Bayesian evidence values for and against the null hypothesis are $Z_0 = L(\theta^\star)$ for $\theta = \theta^\star$ and $Z_1 = 1$ for $\theta \neq \theta^\star$, so that the Bayes factor is

$$\frac{Z_0}{Z_1} = L(\theta^\star) \approx 1 \tag{15}$$

This is always close to 1 because ϵ is assumed small. Unsurprisingly, with the likelihood being almost constant, Bayes has no strong preference for one set of locations over any other.

The FBST behaves oppositely. If $\theta^\star$ is chosen at any of the four peaks $\theta = \frac{1}{16}$ or $\frac{5}{16}$ or $\frac{9}{16}$ or $\frac{13}{16}$, then $\texttt{Ev} = 1$ because *all* the posterior lies below. The test overwhelmingly favours the null hypothesis. That was if $\epsilon > 0$, but the situation is reversed if $\epsilon < 0$ because *none* of the posterior then lies below. The test then overwhelmingly denies the null hypothesis.

$$\texttt{Ev} = \begin{cases} 1 \text{ (accept 100\%)} & \text{if } \epsilon \to 0+ \\ \text{undefined} & \text{if } \epsilon = 0 \\ 0 \text{ (reject 100\%)} & \text{if } \epsilon \to 0- \end{cases} \tag{16}$$

The recommendations are also reversed if the any of the valleys at $\theta = \frac{3}{16}$ or $\frac{7}{16}$ or $\frac{11}{16}$ or $\frac{15}{16}$ is selected instead. The FBST can thus display unstable and overwhelming preferences based on negligible variations in likelihood. Such performance cannot be recommended.

5 Commentary

In their final remarks, PS offer various weak arguments in favour of the FBST, none of which address the potential for catastrophic failure.

They argue against traditional p-values, where skepticism is indeed justified. However, they also argue against Bayes, writing "*The Bayes factor is indeed formulated directly in the parameter space, but needs an ad hoc positive prior probability on the precise hypothesis. First we had no criterion to assess the required positive prior probability. Second we would be subject to Lindley's paradox, that would privilege the null hypothesis.*"

That criticism is a classic misrepresentation of Bayesian methodology. Centrally, there *is and always has been* a perfectly good criterion for assessing priors. It's called the *evidence* Z and it's the very first quantity (2) that the calculus produces, even before the posterior (3).

PS argue against the requirement, but *of course* a prior probability distribution is always needed. Here as everywhere else, a user has to show some understanding of the problem by supplying a prior. Personally, I elicit my prior by asking myself where I would place a dozen or more samples of θ in order to illustrate the plausible range, and then I invent some reasonably simple mathematical model with that behaviour. Thus, I find that setting $\theta(p)$ is more intuitive than looking for $p(\theta)$ first. For a start, improper priors become impossible, and in practice a reasonable balance is quickly achieved.

A user with no domain knowledge who flaunts his ignorance by supplying an uninformative prior over an infinite range will be rewarded by an evidence value of zero, $Z = 0$. He was totally unable to predict what the data eventually turned out to be. That means that he is infinitely out-classed by any rival who has even the remotest idea of how the application should actually behave. In short, he's useless. Bayes can be used by by ill-informed and by well-informed users alike. However, the evidence (*a.k.a.* prior predictive) Z is available to assess their predictive strengths.

Logic dictates to us the standard sum and product rules that we call Bayesian calculus. Logic does *not* tell us what to put into those rules. We have to ask a question (set a prior) before we can get an answer (evidence and posterior) from our observations (likelihood). That's how the world *is*. Yes, our modelling is subjective. It has to be. Get used to it.

As for Lindley's paradox (from which we have moved on in the intervening 60 years), you might equally wrongly call (14) Skilling's paradox. It's not a paradox at all. Such so-called paradoxes merely show the advantage of rationality in the analysis of sensible hypotheses. In our game with the world, there are no paradoxes. If there were, the number of admissible calculi of inference, already reduced to just one (Bayes) would be further reduced to zero. Is that what proponents want?

6 Recommendation

Proposals should be tested "*hard*".

Most proposals that are even vaguely plausible will not perform too badly when presented with applications having sufficient dynamic range to overcome the failings. If constraints are tough, meaning in the context of inference that likelihoods have high dynamic range, it will be difficult to disobey. Such testing would be "soft", being forgiving of imperfections. PS's tests of the FBST are soft in this sense. They use likelihoods involving polynomials and exponentials of sufficiently high power to dominate, so that differences between methods are relatively minor (though there are places where Bayes and FBST differ by more than a factor 10, which might have been worthy of comment). These soft tests in a forgiving environment have misled PS into thinking that their FBST is acceptable, which it is not.

Users seek methods that will not fail when challenged. In inference as in engineering, hard testing is required. Here, the FBST fails catastrophically, while Bayes inevitably continues — as always — to give faithful analysis.

Comments on selection

John Skilling

July 24, 2017

1 Basic Bayes

Suppose I give you two functions u and v

$$u(\theta) \geq 0\,, \qquad v(\theta) \geq 0\,, \qquad \int u(\theta)d\theta = \int v(\theta)d\theta = 1\,.$$

I now tell you that one is a prior distribution $\Pr(\theta)$, normalised by definition. The other is a likelihood function, $\Pr(\texttt{data} \mid \theta)$, normalised over θ by happy accident. Can you tell which is which?

Of course you can't. What I told you is symmetric between u and v. So, to represent the situation faithfully, your analysis must respect that symmetry. And Bayes does respect the symmetry. Whichever choise you make, you get the same joint distribution

$$\Pr(\theta, \texttt{data}) = u(\theta)\, v(\theta)$$

which is the central construction from which all inference follows. As always, marginalising over θ gives the evidence (*aka* prior predictive)

$$Z = \int u(\theta)\, v(\theta)\, d\theta$$

after which the posterior

$$\Pr(\theta \mid \texttt{data}) = u(\theta)\, v(\theta)\, /\, Z$$

is available.

2 A switch s

Next, I tell you that the parameters θ involve a binary variable s that could be 0 or 1, along with other parameters ϕ, so that $\theta = \{s, \phi\}$. The function u could be either u_0 or u_1, each being individually normalised.

$$u_0(\phi) \geq 0\,, \qquad u_1(\phi) \geq 0\,, \qquad \int u_0(\phi)d\phi = \int u_1(\phi)d\phi = 1\,.$$

$$u(s, \phi) = \begin{cases} u_0(\phi) & (\text{if } s = 0) \\ u_1(\phi) & (\text{if } s = 1) \end{cases}$$

I tell you nothing about the setting of s, so (symmetry again) you will faithfully assign prior probability $\frac{1}{2}$ to $s = 0$ and to $s = 1$. Meanwhile, the function v depends only on the other parameters ϕ, not on s.

$$v(s, \phi) = v(\phi)$$

The joint distribution of everything, from which all else follows, is now

$$\Pr(s, \phi, \texttt{data}) = \begin{cases} \frac{1}{2}u_0(\phi)v(\phi) & (\text{for } s = 0) \\ \frac{1}{2}u_1(\phi)v(\phi) & (\text{for } s = 1) \end{cases}$$

As always, marginalising over the parameters gives the evidence

$$Z = \tfrac{1}{2}Z_0 + \tfrac{1}{2}Z_1 \qquad \text{where} \qquad Z_0 = \int u_0(\theta)\, v(\theta)\, d\theta\,, \quad Z_1 = \int u_1(\theta)\, v(\theta)\, d\theta$$

after which the posterior for s

$$\Pr(s \mid \texttt{data}) = \begin{cases} Z_0\,/(Z_0 + Z_1) & (\text{for } s = 0) \\ Z_1\,/(Z_0 + Z_1) & (\text{for } s = 1) \end{cases}$$

and the conditional posteriors for ϕ

$$\Pr(\phi \mid s, \texttt{data}) = \begin{cases} u_0(\phi)v(\phi)\,/\,Z_0 & (\text{for } s = 0) \\ u_1(\phi)v(\phi)\,/\,Z_1 & (\text{for } s = 1) \end{cases}$$

are available. So, as expected, data will inform us about the parameters s and ϕ in the usual standard way. But you still don't know which of u and v is the prior and which is the likelihood.

3 Ontological interpretation

The likelihood is u and the prior is v. The switch s is a physical switch inside the apparatus which changes its response from u_0 to u_1.

Although we did not originally know how it was set, we can (to the extent that Z_0 and Z_1 differ) find out after acquiring data. Of course we can. It would be crazy if $Z_1 \gg Z_0$ did not favour $s = 1$. In the extreme, if Z_0 was zero because u_0 and v had no common support, then we would *know* as a matter of logical deduction that $s = 0$ was incompatible with our prior knowledge, so that s had to be 1 (unless Z_1 was also zero, in which case our modelling must be wrong). Moral: test hard, not soft.

So it must be possible for the data to inform us about the switch. And the only rational calculus is Bayes.

4 Epistemological interpretation

The likelihood is v and the prior is u. The switch s is our opinion about whether to trust Agent 0 who proposes u_0 or Agent 1 who proposes u_1.

Although we did not originally know which agent to believe, we can (to the extent that Z_0 and Z_1 differ) find out after acquiring data. Of course we can. It would be crazy if if $Z_1 \gg Z_0$ did not favour $s = 1$. In the extreme, if Z_0 was zero because u_0 and v had no common support, then we would *know* as a matter of logical deduction that Agent 0's proposal was incompatible with our data, so that he had to be rejected in favour of Agent 1 (unless Z_1 was also zero, in which case either an agent or the data must be wrong). Moral: test hard, not soft.

So it must be possible for the data to inform us about the agents. And the only rational calculus is Bayes.

5 Synthesis

Ontology and epistemology are interchangeable. Either way round, the evidence Z is the same, so there is no way of deciding which is preferable.

Nothing in probability calculus tells us how to interpret the symbols. Once we have abstracted our problem into symbols, the interpretation has vanished. Despite what philosophers may claim, it's nonsense to suppose that there is some peculiar metaphysical distinction between probabilities of different types. It's all the same, always obeying the same unique calculus of simple proportion that's forced by basic symmetries which apply to ontology and epistemology alike.

If that feels counter-intuitive (and it's common to find this outlook strange at first), then the remedy is to educate one's intuition to align it with rational analysis. And the educated analyst has enhanced

intellectual resources. For example, uninformative priors are often praised for their lack of prejudice, despite their objectively low evidence values. But nobody would seriously propose an uninformative experiment. To the educated Bayesian, an uninformative prior is equally silly.