A personal essay on Bayes factors

Danielle J. Navarro

Original: September 2018; PsyArXiv: December 2020

Abstract:

This is an archived version of a blog post on Bayes factors. It is a personal reflection on some of the practical issues that one encounters when attempting to apply Bayes factors to difficult inference problems. The main message of the piece is real world inference is hard and that being too prescriptive about how statistics *must* be done is generally a recipe for disaster.

GitHub: https://github.com/djnavarro/bayes-factor-essay

PsyArXiv: https://psyarxiv.com/

Important note:

This essay is not a traditional academic output. It has not been peer-reviewed or submitted to a journal, nor indeed will it be. It was originally published as a blog post that no longer exists (for reasons that the author chooses not to go into). However I have received a few requests for the post to be made available again, and it seems like the post was valuable to a few people, so here it is. Correspondence concerning my ramblings, should such be necessary for some unfathomable reason, can be addressed to Danielle Navarro (School of Psychology, University of New South Wales, d.navarro@unsw.edu.au). Not surprisingly, as I wrote it originally as a personal reflection on Bayesian inference, the author asserts no relevant conflicts of interest and did not receive funding for this work. Duh.

'Cause you're hot then you're cold
You're yes then you're no
You're in then you're out
You're up then you're down
You're wrong when it's right
It's black and it's white
– Katy Perry

I have mixed feelings about Bayes factors. As Katy Perry once observed, it's *extremely* hard to make valid inferences when you aren't even sure what the hell it is you're trying to make inferences about. Oh sure, it's easy enough to tell pretty stories about rational reasoning with respect to a prior, but if we're going to have a serious statistical relationship you need to bring more than a Dutch book argument to the table.

Love at first sight

I first discovered Bayes factors in 1999, as a new Ph.D. student working on problems in similarity judgment and categorisation. I read Kass and Raftery (1995)[1] and was head over heels. As long as one accepts the principle of inverse probability, a Bayesian reasoner can evaluate a hypothesis h in light of data d in a simple fashion:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

I suspect most people reading this post already knows what Bayes rule says, but not everyone who follows me cares that much so let's break this down:[2]

- $P(h)$ is my *prior belief*: how plausible was h before the data arrived?
- $P(h|d)$ is my *posterior belief*: how plausible is h now that I've seen the data?
- $P(d|h)$ is the *likelihood*: the probability that we would have observed data d if the hypothesis h describes the true data generating mechanism [3] [4]

When comparing two competing hypotheses h_0 and h_1, I can compute the posterior odds favouring one over the other simply by dividing the two posterior probabilities,

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)}{P(d|h_0)} \times \frac{P(h_1)}{P(h_0)}$$

Or, in something closer to every day language:

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

Thus the Bayes factor (BF) is defined by the ratio of the two likelihoods, and it has a natural interpretation as a *weight of evidence*. It tells me how I need to adjust my beliefs in light of data. And it's so simple...

[1] https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572

[2] Throughout this post I will follow my usual tendency to ignore things like the difference between densities and probabilities, and I will absolutely *not* waste everyone's time by introducing σ-algebras, because this is a blog post not a bloody measure theory textbook. If such things unnerve you too greatly, I refer you whichever section of Mark Schervish's very excellent *Theory of Statistics* book will allow you to feel love again.

[3] Again... if you feel inclined to instruct me on the difference between $P(x|\theta)$ and $\mathcal{L}(\theta|x)$... don't. Go take it up with Fisher's ashes. He's the one who tried to misuse the ordinary natural language meaning of the word "likelihood" by inappropriately attaching probabilistic connotations to a score function that frequentists are explicitly forbidden to interpret as a probability

[4] No seriously. Go visit his ashes. Fisher retired to St Mark's college in Adelaide, and his ashes are kept in St Peter's Cathedral in North Adelaide, a short walk from the University. The staff there are very friendly and will gladly show you to them.

$$\text{BF} = \frac{P(d|h_1)}{P(d|h_0)}$$

What's not to love?

Better yet, it even extends naturally from simple hypotheses to full fledged models. Suppose I have a theoretically meaningful computational model $\mathcal{M}$ for some psychological phenomenon, with parameter(s) θ. For any choice of parameter values θ my model provides me with a likelihood function for the data $P(d|\theta)$, and my researcher knowledge of the world provides a prior $P(\theta|\mathcal{M})$ belief about the relative plausibility of different parameters. So the *a priori* prediction that my model makes about the probability of observing data d in my experiment is calculated with the *marginal likelihood* [5]

$$P(d|\mathcal{M}) = \sum_{\theta} P(d|\theta)P(\theta|\mathcal{M})$$

The intuition is dead simple (or so I thought at the time).... if I don't know which parameter θ is the right one, I should hedge my bets by constructing an appropriate weighted average. Easy-peasy. This gives me a Bayes factor that I can use to compare two computational models like so:

$$\text{BF} = \frac{P(d|\mathcal{M}_1)}{P(d|\mathcal{M}_0)}$$

Honestly, I don't see why this "statistics business" is so hard, I thought. All you have to do to scale up from simple hypotheses to serious theory evaluation is turn an italicised h into a squiggly $\mathcal{M}$ and you're done! I read the Myung and Pitt (1997) paper on model selection with Bayes factors[6] and thought yep, this is it. Problem solved. Easy!

Oh, you sweet summer child.

Seeds of doubt

> *I'm feelin' electric tonight*
> *Cruisin' down the coast, goin' about 99*
> *Got my bad baby by my heavenly side*
> *I know if I go, I'll die happy tonight*
> – Lana Del Rey

During my PhD I used Bayes factors (or similar tools) a *lot*. One of my very first papers[7] sought to combine multidimensional scaling methods with overlapping clustering methods in a way that would allow someone to estimate stimulus representations that have both continuous and discrete parts (e.g., our intuitions about number are partly continuous insofar as they pertain to magnitude, but also discrete when they pertain to other arithmetic properties), using Laplace approximations to the Bayes factor to automatically determine the appropriate number of clusters and dimensions. The technique had some problems. Collections of weighted binary features (as used in featural reprentations; Tversky (1977)[8], Shepard and Arabie (1979)[9])

[5] IF YOU EMAIL ME TO TALK ABOUT UNCOUNTABLY INFINITE SETS OR TRY TO DISCUSS LEBESGUE MEASURABLE FUNCTIONS IN MY PRESENCE I WILL HUNT YOU DOWN, CUT YOU INTO INFINITESMALLY THIN HORIZONTAL SLICES AND FEED THE SLICES TO MY CHILDREN.

[6] https://link.springer.com/article/10.3758/BF03210778

[7] https://psyarxiv.com/qejyb/

[8] https://doi.org/10.1037/0033-295X.84.4.327

[9] https://doi.org/10.1037/0033-295X.86.2.87

induce a *qualitatively different parameter space* than co-ordinates in a Minkowski space[10] (Shepard 1974)[11], and so when you try to mix them together into a hybrid similarity representation you get... weirdness.

Any time you compute the marginal likelihood $P(d|\mathcal{M})$ you are implicitly introducing a penalty for excess complexity, so the Bayes factor incorporates a form of automatic Ockham's razor. But when I built the hybrid model I found that the (implied) penalty term for *"adding one more continuous dimension for an MDS solution"* doesn't seem to be commensurate with the (implied) penalty term for *"adding one more discrete feature"* and while I could get some decent solutions in some cases (the numbers example worked pretty well...) I never did find a general version that would "just work".

I put it down to the fact that the priors $P(\theta|\mathcal{M})$ were kind of ad hoc... after all, I didn't know what would make sense as a plausible prior that would render continuous things and discrete things commensurate with one another in a way that made sense for the psychological problems I wanted to solve. I assumed the right answer would come to me one day.

It hasn't yet, but I'm still hoping it will.

Seeing other statistics (that's what I said I was doing)

We lay on the bed there
Kissing just for practice
Could we please be objective?
'Cause the other boys are queuing up behind us
– Belle & Sebastian

At about this point in time, I became fascinated with some of Jay Myung and Mark Pitt's other papers on alternative ways to do model selection. For instance, in 2003 they advocated the use of model selection by minimum description length (MDL).[12] The MDL approach to statistical inference comes out of algorithmic information theory and can be viewed as a stripped down form of Kolmogorov complexity (KC). In KC we would say something like this...

> The Kolmogorov complexity of a string S with respect to programming language L is the length (in bits) of the shortest program P that prints S and then halts.

... so the idea would be to think of a model $\mathcal{M}$ as a program and use it as a tool to compress the data d. Whichever model compresses the data the most is the winner. Strictly speaking, KC is useless in real life because it's uncomputable[13] but there are many ways of taking the idea and transforming it to something that you can use. The best known (I think?) is Jorma Rissanen's[14] stochastic complexity approach (borrowing from work by Shtarkov) but I've always had a soft spot for Wallace and Dowe's[15] explicitly Bayesian formulation of the problem.

As you can probably tell, during the early 2000s I read a lot of statistics papers that I didn't understand all that well.

What I did notice though is that many of these techniques end up constructing some version of the marginal likelihood $P(d|\mathcal{M})$. They all have different motivations and not all of them allow a clear probabilistic interpretation (Rissanen doesn't endorse a Bayesian interpretation of MDL, for instance), but they have more in common with one another than I'd originally thought. I even started reading some information geometry[16] and found roughly the same thing. A large number of these model selection criteria can be viewed as series expansions of $\ln P(d|\mathcal{M})$, with "small" terms omitted (usually $O(1)$). Yay, I thought! This

[10] Something about metric MDS rather than nonmetric MDS... don't @ me
[11] https://doi.org/10.1007/BF02291665
[12] https://doi.org/10.1037/0033-295X.109.3.472
[13] FOR REASONS
[14] https://doi.org/10.1109/18.481776
[15] https://academic.oup.com/comjnl/article-abstract/42/4/270/558949
[16] https://doi.org/10.1073/pnas.170283897

is fantastic. Particulars notwithstanding, there is a strong theoretical justification for basing my inferences on the marginal likelihood.

It didn't take long for my enthusiasm to fade again. The first time I tried to use this for model selection in the wild (selecting between different retention functions in recall memory tasks) I broke it pretty badly[17]. It turns out that $O(1)$ terms can be *very fucking large* in practice, and you can get all sorts of absurd results (e.g., a nested model that is judged to be more complex than the full one) when you use these model selection criteria with "small" (say, a mere 1000 or so observations) samples.

I expanded my dating pool further. I had an on again off again thing with Bayesian nonparametrics [18] [19] [20], I dated normalised maximum likelihood[21], and various other things besides. They all let me down somehow. It turns out that NML is mostly useless in real life, Bayesian nonparametric models don't converge to anything sensible in some situations, and so on.

I never dated a p-value though. I do have standards.

What problems do we study?

> *But I got smarter, I got harder in the nick of time*
> *Honey, I rose up from the dead, I do it all the time*
> *I've got a list of names and yours is in red, underlined*
> – Taylor Swift

Just lately I've been wondering how many of the practical problems I've encountered stem from the fact that almost no statistical problems worth caring about are $\mathcal{M}$-closed (I blame Dan Simpson for this) I'm thinking about this, but it's been a recurring theme in my thoughts for a long time). At the moment I'm reading a paper[22] by Clarke, Clarke and Yu (2013), and I'll steal their words. The first paragraph of the paper starts with this

> Prediction problems naturally fall into three classes, namely M-closed, M-complete, and M-open, based on the properties of the data generator (DG) (Bernardo and Smith 2000). Briefly, M-closed problems are those where it is reasonable to assume that the true model is one of the models under consideration, i.e., the true model is actually on the model list (at least in the sense that error due to model mis-specification is negligible compared to any other source of error). This class of problems is comparatively simple and well studied.

Ouch. That's about 99% of the statistical methodology that I was taught (and see in the psychological literature) and they've discarded it as too simplistic to be bothered talking about. It'd hurt less if they weren't entirely correct. Almost all of what we talk about in psychology frames the problem of inference as one of "choosing the true model", and it's implicit that one of the models is presumed to be correct.

This is never accurate in real life. We often hand wave this way by quoting George Box's famous aphorism *all models are wrong but some are useful*, yet we are rarely explicit in psychology in saying what we mean by "useful". At one point I tried formulating what I thought I meant: for many cognitive science experiments that are designed to be "operationalised" versions of a more complex real world situation, I think it makes little sense to bother making predictions about low-level features of the data, and a model is most useful if when makes the correct a priori predictions about theoretically-relevant ordinal patterns in the data[23]. But that's not a very generalisable criterion, it doesn't apply in situations where you actually do have to care about all the features in the data, and so on. I've never seen anyone come up with anything that I found compelling either.

[17] https://www.mitpressjournals.org/doi/10.1162/0899766041336378
[18] http://dx.doi.org/10.1016/j.jmp.2005.11.006
[19] http://dx.doi.org/10.1162/neco.2008.04-07-504
[20] https://doi.org/10.1037/rev0000077
[21] http://dx.doi.org/10.1016/j.jmp.2005.06.008
[22] https://projecteuclid.org/download/pdfview_1/euclid.ba/1378729923
[23] http://dx.doi.org/10.1037/0033-295X.113.1.57

That's the thing about stepping outside of the $\mathcal{M}$-closed world... nothing really works the way it's supposed to. In the most difficult case you have $\mathcal{M}$-open problems:

> M-open problems are those in which the DG does not admit a true model. The DG is so complex (in some sense) that there is no true model that we can even imagine. For instance, one can regard the Collected Works of William Shakespeare as a sequence of letters. Unarguably this data set had a DG (William Shakespeare), but it makes no sense to model the mechanism by which the data was generated. One might try to use the first n letters to predict the n + 1 letter and do better than merely guessing, but one should not expect such a predictor, or any model associated with it, to generate more great literature. The same point applies to the nucleotide sequence in a chromosome, the purchases of a consumer over time, and many other settings. In these cases, we are only able to compare different predictors without reference to a true model.

Oh yes. $\mathcal{M}$-open problems are *nasty*. You have to find some sensible way to discuss what it means to make good prediction that doesn't rely on any notion of "true models", because there is no sense in which the data generating mechanism can possibly be mapped to anything that you or I would *ever* call a "model". I suspect that this is part of the reason why some of the MDL people (e.g. Jorma Rissanen) *don't* want to formulate their model selection procedures with reference to any notion of a "true model". The moment you allow yourself the "crutch" of assuming that a true model exists, you're left unable to justify any claims in an $\mathcal{M}$-open world. Clarke et al comment on that actually...

> From a log-loss point of view, the Shtarkov solution (Shtarkov 1987) has also been extensively studied in the M-open case, see Cesa-Bianchi and Lugosi (2006), but has not caught on partially because the conditions for it to exist are so narrow

... where (assuming it's the paper I'm thinking of) Shtarkov's work is linked to Rissanen's approach to MDL that some folks in psychology (such as myself, once upon a time!) had argued for. But it's like Clarke et al say, this approach is basically useless in real life because there are so few scenarios where you can *do* anything with it.

On the other hand, there's a sense in which the $\mathcal{M}$-open scenario above is more pessimistic than it needs to be. Not every statistical problem is as hard as generating new Shakespeare novels....

> By contrast, M-complete problems are those where the DG has a true model that can be imagined but is not identifiable in any closed form. Inability to write a model explicitly may arise because the model is too complicated or because its constituent pieces are not known. The key point for an M-complete problem is that it is plausible to assume that a true model – also called a 'belief model' – exists because this enables its use in reasoning even if a prior cannot be meaningfully assigned in the usual way. For instance, if a true model exists a bias-variance decomposition can be developed, at least in principle, even when the true model is not explicitly known.

I think this is where most of our practical problems in science lie. If we knew enough about a phenomenon to put us in $\mathcal{M}$-closed world we wouldn't bother to study it, and if we knew so little that we couldn't even imagine a true model (putting us in $\mathcal{M}$-open land) it would be foolish to try. So in practice we live in the land of $\mathcal{M}$-complete inference problems. There are interesting results in this situation. I haven't read much about this in a long time, but my recollection from earlier reading was that in this situation a Bayes factor selection procedure will asymptotically converge to the model $\mathcal{M}$ that is closest to the true distribution in Kullback-Leibler divergence.

I used to find this reassuring. I'm less sure now.

It's the little things

Oh, life is bigger
It's bigger
Than you and you are not me
The lengths that I will go to

The distance in your eyes
Oh no, I've said too much
I set it up
– R.E.M.

The worry I have with leaning so heavily on "convergence in KL terms" comes from a few sources. For one thing I'm starting to go beyond the limits of my own skill. You actually have to have a very good grasp of the theory to know what the hell this actually means, and I'm not sure I do. I'm a little unsure about what *practical* conclusions I should draw about a model if all I can say is that it is closer to the truth in the sense of a very specific information distance measure defined over distributions.

The impression I have had when working with KL divergence is that it really does seem to depend on every property of the distributions, but as a researcher I often don't care about every little thing in the data. Worse, to the extent that Bayes factors specifically depend on the *prior* to specify the marginal distribution in question, I have this intuition that even modest mistakes in you specify the prior (especially the tails) could do very strange things. Looking back over the various papers I've written about in this post, I feel like it's been a recurring theme that the details really matter. Just in this little reminiscence...

- When thinking about similarity modelling, I found stimulus features and stimulus dimensions don't seem to have commensurate complexity as judged by the most sensible Bayesian method I could think of
- When doing very simple memory modelling, the best approximations I knew of (Fisher information approximation to MDL) gave absurd predictions because of the weird structure of the models
- In categorisation, when using "infinite dimensional" nonparametric Bayesian models ... oh, don't even get me started.

... and the thing is *these issues have caused my inferences to misbehave every single time I have tried to automate them.*

In real world data analysis, nothing works the way it's supposed to and I have grown deeply skeptical that *any* rule governed approach to automating statistical inference makes much sense.

What to do?

'Cause I'm gonna be free and I'm gonna be fine
(Holding on for your call)
'Cause I'm gonna be free and I'm gonna be fine
(Maybe not tonight)
– Florence and the Machine

Honestly, I don't know. I like Bayesian inference a great deal, and I still find Bayes factors useful in those circumstances where I (1) trust the prior, (2) understand the models and (3) have faith in the (either numerical or analytic) approximations used to estimate it. I don't have a better alternative, and I'm certainly not going to countenance a return to using p-values[24]. More than anything else, the one thing I don't want to see happen is to have the current revival off Bayesian methods in psychology ossify into something like what happened with p-values.

What I think happened there is not necessarily that p-values are inherently useless and that's why our statistics went bad. Rather, it's that introductory methods classes taught students that there was *A RIGHT WAY TO DO THINGS* and those students became professors and taught other students and eventually we ended up with an absurd dogs breakfast of an inference system that (I suspect) even Fisher or Neyman would have found ridiculous. If I've learned nothing else from my research on cultural evolution and iterated learning[25] it's that a collection of perfectly-rational learners can in fact ratchet themselves into believing

[24] That's not to say I think there is no role for orthodox inference, nor that controlling error rates is a thing we should just not think about anymore. I just don't think that this is a sensible idea to build an entire theory of inference around

[25] http://dx.doi.org/10.1111/cogs.12667

foolish things, and that it's the agents with most extreme biases that tend to dominate how the system evolves.

Whatever we do with Bayesian methods, whatever role Bayes factors play, whether we use default or informed priors, the one thing I feel strongly about is this... we should try to avoid anything that resembles a prescriptive approach to inference that instructs scientists *THIS IS HOW WE DO IT* and instills in them the same fear of the Bayesian gods that I was once taught to have for the frequentist deities.

It doesn't help anyone, and it makes science worse.