This Physicist's view of Gelman's Bayes

John Skilling
Maximum Entropy Data Consultants Ltd.
Killaha East, Kenmare, County Kerry, Ireland
skilling@eircom.net — August 2008

Abstract.   The author offers fundamentalist commentary on Andrew Gelman's brilliantly provocative comments on Bayes, and the associated discussion.

Keywords: evidence, fundamentals, semi-Bayesian, Emperor's Clothes.

1   Introduction

In publishing Gelman (2008) with commentaries, the Editor is to be congratulated on allowing an exhilirating relaxation of the orthodox norms of professional presentation. Consequently, each author's contribution is seen with unusual clarity. There's no fussy detail. There's no intricate symbolism designed to impress. There is just the natural language of personal communication, so well suited to discussion of basic outlooks.

Yet that very clarity exposes what is oddly missing. The discussions lack any serious account of why we MUST use Bayes or of how I think we SHOULD use Bayes. Readers would think there was a choice. There isn't. Here in complimentary response to Andrew's wonderfully successful provocation is my own polemical rant on the subject.

2   Why we MUST use Bayes

Probability calculus, often called "Bayesian", is not an option to be accepted, modified or rejected at whim. It has a firm logical basis as the unique calculus of rationality. Over sixty years ago, Richard Cox wrote a remarkable paper (Cox (1946)) which Jaynes (2003) considered to be "the most important advance in the conceptual (as opposed to the purely mathematical) formulation of probability theory since Laplace". I have long concurred with that view, except that I omit the bracketed qualification. Although some of us continue to polish and refine the approach, I hold that Cox (1946) remains the foundation authority.

It's not a difficult paper. Cox simply pointed out that any acceptable calculus of inference ought to conform to the obvious basic properties of inference. He required an ordering, represented by transitivity $\geq$ which I'll informally call "more". He presented "more" as a required relationship between degrees of belief, with degree of belief being the underlying idea. Myself, I think it's cleaner to think of "more" as the underlying idea, and degree of belief as merely one application among many. Either way, it's a basic property. Cox also required "or", as in logical combination of possibilities, and "not", negation.

He then sought a calculus for a bi-valuation, with particular application to context-dependent degrees of belief. By considering elementary situations with no more than three *yes/no* propositions, he proved that conformity to these three properties (`more`, `or`, `not`) required the standard sum and product rules of ordinary probability calculus. There was no subtle measure theory or questionable digression into infinity. There was just the theorem, as simple as ABC:

$$\text{Assert}(\texttt{more, or, not}) \implies \text{Bayes} \qquad \text{(Cox theorem)}$$

The derivation is solid: although Cox made some minor assumptions such as differentiability, they aren't necessary. We don't need any other justification, and we certainly don't need weaker ones like Dutch books.

Any other calculus is either equivalent (like percentages or log-probs) so un-necessary, or different. Any different procedure that does not conform to probability calculus thereby contradicts the assumed properties:

$$\text{Assert NOT}(\texttt{more, or, not}) \impliedby \text{NOT Bayes} \qquad \text{(Cox converse)}$$

Which of (`more`, `or`, `not`) would *you* be prepared to give up? If, like me, you want all of them in inference, then, like me, you have no option but to be a Bayesian. If, regardless, you wish to stand by a non-Bayesian procedure, then be prepared for counter-examples that expose contradiction with (`more`, `or`, `not`). Less politely, I will just say that you are wrong.

If despite this you ever manage to produce a non-Bayesian procedure that's convincingly superior to sensible Bayesian analysis, there will be two consequences.

1: I will build upon this contradiction to prove 0=1, give up the intellectual activity that's become pointless and take to gardening.

2: You will be showered with praise and prizes from the sub-logical barbarians who lie (in both senses) all around us.

The barbarians will be delighted with the message that they are no longer asked to think straight. To avoid this we, the guardians of rational inference, have a duty to think straight ourselves.

Take confidence intervals, for example. Yes, a confidence interval may cover the truth 95% of the time. So what? It's easy to find counter-examples in which the confidence interval does *not* cover the truth in the *particular* case that we have. The entire confidence interval can even be logically contradicted by the data. OK, those data were a little unusual. So what? It can happen, and it does. Frequentist confidence intervals are dead. Frequentism was killed in 1946. Let's just leave the twitching corpse behind, and focus on building the rational future, where big and important problems await. Those parts of frequentism that are consistent with Bayesian thinking are un-necessary. Those parts that are in opposition are wrong.

We MUST use Bayes. All alternatives are disproved. It's as simple as that.

3 How we SHOULD use Bayes

As applied to inference, the product law of probability calculus states

$$\underbrace{\Pr(\theta) \quad \times \quad \Pr(x \mid \theta)}_{\text{Inputs}} \quad = \quad \underbrace{\Pr(x) \quad \times \quad \Pr(\theta \mid x)}_{\text{Outputs}} \quad \| \, I$$

$$\text{Prior} \times \text{Likelihood} = \text{Evidence} \times \text{Posterior} \qquad \text{(Bayes)}$$

with the usual identifications θ = unknown, x = data, I = background assumptions.

I use the physicists' name "evidence" for $\Pr(x)$ because a central term in the principal equation of any discipline ought to have an unqualified moniker for what it *is*. Statisticians' names include "prior predictive" which describes how it's often *used*, "marginal likelihood" for how it's often *found*, and others. Double-barrelled monikers like these suggest that $\Pr(x)$ is merely a qualified version of something deeper. That's not the case, and it's liable to confuse outsiders, as well as deflect our own attention.

Probability calculus is a *calculus* for manipulating numbers (ratios, actually). It does not tell us what those numbers are. It is a language in which we can express our opinions. It does not tell us what those opinions, expressed through our prior $\Pr(\theta)$, are. Correctly, I am *allowed* to propose any value, or weighted choice of values, of θ. Nothing is prohibited. The language is fixed, and the content is free.

This is exactly how it should be. I want, and indeed would insist on, the freedom to disagree with you and analyse a problem my own way, according to whatever intellectual skills and experience I may possess. And I grant you the ability and freedom to do likewise. Inference is not an automatic procedure. The universe has no mechanism for telling us "objective" truth. We have to actively seek answers for ourselves. That's how inference actually is. It's no use complaining about it, or trying to evade it. We must live with it, whether we like it or not. As it happens, I do like it. We Bayesians can talk to each other because we all use the same language, and can approach objectivity by aligning our opinions.

My answer to a fellow-physicist or anybody else who wants a non-Bayesian procedure, say "an orthodox confidence interval in which the data speak for themselves", is to recommend them to put such childish and naïve intuitive demands aside, and educate the mind to an adult acceptance of the facts of life. If one's intuition is in conflict with the facts of a matter, then one should educate the intuition. For example in physics, quantum mechanics is supremely counter-intuitive at first, second and even third glance. Yet, with sufficient exposure, one's intuition falls into line, to the point that the preceding classical mechanics becomes no more than a simple approximation, valid in appropriate circumstances. Similarly with relativity. Similarly in statistics.

We do great dis-service to the scientific community and the larger society beyond when statistical packages and recommendations continue to propound counter-rational techniques. I would much prefer that, as a matter of elementary professional standards, we all refuse to provide non-Bayesian procedures. Other sciences speak the truth as best they can, as an uncontroversial matter of course, and it works for them. In inference, that strategy works for me, and my clients find no difficulty with it as they enjoy the

power of the logical and consistent Bayesian approach. (It's more fun too.)

But, in promoting an attitude of unrestricted freedom, has the Bayesian paradigm opened the flood-gates of a fearsome world of uncontrolled anarchy in which anything goes? Not at all! Remember the "evidence", that too-often-ignored half of Bayesian inference. That's how the universe, in the form of data, assesses our prior. Yes, I give my clients keys to the Bayesian F-16 fighter plane, so they can fly anywhere they want — but *my* F-16s come equipped with an altimeter in the form of an algorithm to compute the evidence. They all should. An altimeter helps to avoid crashes. It gives control.

Sadly, Gelman ignored the evidence in defining "Bayesian inference" as the generation of the conditional posterior $\Pr(\theta \mid x)$ alone. In doing that he missed half the point. I know no other discipline in which half of the principal equation is so widely ignored, and it should not be ignored here either. I could (and often do) argue that the evidence

$$Z = \Pr(x) = \sum_{\theta} \Pr(\theta, x)$$

is even more important than the posterior

$$\Pr(\theta \mid x) = \frac{\Pr(\theta, x)}{Z}$$

on the grounds that algebraically it has to be evaluated first, and logically there's no need to proceed to the posterior if the evidence is unacceptably weaker than that from some other candidate prior. So it's the posterior that's subordinate to the evidence, and definitely *not* the other way round. Myself, I think of "Bayesian inference" as the generation of the evidence, with the posterior following if needed. Evidence is central. Gelman's definition was only semi-Bayesian.

In summary, we SHOULD use Bayes by assigning subjective priors and inspecting evidence values.

4   Priors

Almost always in practice, there is a choice of prior. Do I make my prior for θ a Gaussian, a Cauchy, or something else entirely? I suppose that I ought to include all those forms that are consistent with my loose subjective opinions, but in practice I don't. Life is too short, and I approximate my opinions with one chosen formulation.

A controversial factor in this choice can be the supposed influence of the likelihood function. As a matter of practical computation, it has often been convenient to have a "conjugate" prior that looks similar to the likelihood. Indeed, the design of instrumentation (a telescope, say) clearly limits what we can see with it, and thereby influences the ideas (on astrophysics) from which we generate priors for the objects which the instruments will later observe. So the functional form of the likelihood could legitimately influence our prior.

However, selecting a prior that's conjugate to a particular likelihood is dubious. Firstly, knowledge of astrophysics derives from a variety of sources, so a prior should not be restricted to the particular instrument that happened to collect the current dataset. Secondly, practical likelihoods tend to be full of drop-outs, saturation effects, hysteresis, correlations, and all sorts of such complications. A prior that attempted to mirror such mundane technical detail would be implausible and impractical for astronomical objects. Thirdly and lastly, conjugacy removes some freedom that I may want.

Hence I judge it better to keep prior assignment separate from the likelihood. Coping with this divorce is a matter for algorithm design, for example Skilling (2004). I'm sympathetic to reference priors as pragmatic default options, but I still want to be free.

A vague prior will predict data badly, thereby producing a low evidence. In the limit, an improper prior will give a zero value, because its value is 0 everywhere (even if its integral over its infinite domain does sum to 1, as any probability must). It will be infinitely outclassed by any other prior that incorporates even the slightest idea of the vague order of magnitude of the parameters being measured. That's why improper priors are silly. At the other extreme, the "sure thing" prior which predicts the data perfectly has the highest evidence of all. But, to play the game fairly, we are supposed to have our prior in place independently of analysing the data, and it might be difficult to propose "sure thing" independently. Try to cheat by using the data twice, and Bayesian inference will come straight back and mislead you. What else *could* it do? Play fair, and the evidence will give you a fair assessment. How could it be otherwise?

Of course, I don't take the evidence as an absolute guide. If you find a larger evidence than mine, then — other considerations being equal — I will favour your prior above my own, quantifying that favour by the ratio of evidences known as the Bayes factor. But if I judge your prior to be intrinsically absurd, then the only effect may be to make me soften my feeling of absurdity. It works the other way round, too. If I applaud the brilliance of your insight, then I'll favour your prior despite some penalty in evidence. More formally, I will use my intellectual skills and experience to assign prior weights to the various candidate priors before me, and use their evidence values to modulate those weights. Naturally, there's nothing absolute about my subjective opinions. Why should there be? As always, the universe does not give us absolute answers. We have to seek answers subjectively. Like it or not, that's the nature of inference.

Incidentally, this construction also shows why the numerical evidence should always be provided in presentation of Bayesian results. Only thus can other people assess the relative values of different analyses. Somebody can quite legitimately believe 100% in creationism, to the exclusion of everything else. Bayesian analysis will still work smoothly and without contradiction. Evolutionary theory, though, will see connections that make the fossil record much more predictable so hugely less surprising. Accordingly, evidence values using fossil data will hugely favour the evolutionist, to the point that any dispassionate observer will reject creationism. We don't bother to work the numbers out because we can judge at once that the disparity is enormous. But we could. The Bayesian outlook works. It tends to demolish dogma, which is a Good Thing because dogma acts against the freedom our civilization needs in order to move forward.

5   Approximations

My formulation of a specific prior often includes parameters that I don't really know in advance. A client gives me some data x to analyse. He gives me a well-calibrated likelihood function $\Pr(x \mid \theta)$, and we agree on the shape of a prior $\Pr(\theta)$, but he omits to tell me the magnitude of θ that he expects. I certainly don't know it. I'm just the data analyst, faced with a pile of numbers. A complicated hierarchical problem might be controlled by several of these unknown hyperparameters. What do I do?

Well, I cheat of course. I peek at the data to discover what a sensible range $\Delta\theta$ of values seems to be, and use that. Or I might look at the likelihood function to discover the range of θ that it was capable of observing, and use that on the grounds that the instrument was presumably intended to be useful. I half-justify my cheating by saying that my assumed range is what my client would have told me, if he had played fair with me. But it's still a cheat, which I try to remember to apologise for (and blame the client) whenever it might matter. Usually it doesn't matter much, provided I remember to factor $\Delta\theta$ into my prior, and perhaps (if I'm feeling particularly professional) reduce the computed evidence a bit to compensate for the cheat in a rough and ready way.

In short, I'm perfectly prepared to approximate. Fanatical perfection is usually too expensive and I quite often find myself using variants of maximum likelihood, maximum a posteriori probability, empirical Bayes (which we all know isn't Bayesian), cross-validation as a preliminary sanity check, and so on. In a suitably safe environment, I can even envisage using a confidence interval, though I can't recall an occasion. Basically, I'll use whatever common-sense approximation looks appropriate at the time. Usually what to do is so obvious that I just do it without invoking a technical name. I emphatically do not call these approximations "principles", or suggest they have intellectual status beyond their station. They are just approximations to a properly stated Bayesian formulation, removable at the cost of some extra work, perhaps in collaboration with the client. That's all.

It is a mistake to elevate an approximation to the status of a principle (as in MLP = Maximum Likelihood Principle). That just encourages the barbarians to point out with glee that acceptance of a principle that's incompatible with Bayes implies rejection of Bayes. Barbarians, you see, are capable of using logic when it suits them, to our distraction and our clients' confusion. The defence is to think straight in the first place. Everybody knows that approximations admit counter-examples when used inappropriately. Principles, by contrast, are supposed to be reliable.

6   The take-home messages

If you understand your problem, then you know enough to set it up in Bayesian form. Logically, you have no choice about this, because probability is the only calculus conforming to the structure of logical inference. If you don't understand your problem, ask somebody who does.

The "subjective" Bayesian freedom to assign priors is liberating, but don't just be a semi-Bayesian satisfied by the posterior. If you only use half the theory, don't be surprised if you seem unconvincing and attract controversy. Be a full Bayesian and use the evidence too. Play fair with the data, though, otherwise you may be misled.

As for practicalities, approximations are as necessary and as acceptable in inference as anywhere else. Just be honest about what they are. Don't describe an approximation as a principled treatment with impressive title backed up by weighty literature. Remember the fable of the Emperor's Clothes.

Finally, I expect and perhaps hope that this rant will duly attract criticism from Authority. That's fine. I've tried to present my views honestly and if I'm wrong, please show me where. That's how I learn. My main aim, though, is the next generation — students whose careers will be so much more productive if they adopt the full power of a Bayesian outlook. Consider the evidence . . . .

References

Gelman, A. and discussants Bernardo, J. M., Kadane, J. B., Senn, S. and Wassermann, L. 2008. Objections to Bayesian statistics. *Bayesian Analysis* 3: 445–478.

Cox, R. T. 1946. Probability, frequency, and reasonable expectation. *Amer. J. Phys* 14: 1–13.

Jaynes, E. T. 2003. *Probability theory — the logic of science*, page 686. Cambridge Univ. Press.

Skilling, J. 2004. Nested sampling for general Bayesian computation. *Bayesian Analysis* 1: 833–860.