Bayesian Analysis (2006) **1**, Number 3, pp. 423–428

Coherence and calibration: comments on subjectivity and "objectivity" in Bayesian analysis (Comment on Articles by Berger and by Goldstein)

David Draper*

Abstract. In this contribution to the discussion of "The case for objective Bayesian analysis" by James Berger and "Subjective Bayesian analysis: principles and practice" by Michael Goldstein, I argue that (a) all Bayesian work is inherently subjective and needs to be guided simultaneously by considerations of both coherence and calibration, and (b) "objective" (diffuse) prior distributions are sometimes, but not always, useful in attaining good calibrative performance—it depends (as usual) on your judgment about how knowns (e.g., past observables) and unknowns (e.g., future observables) are related.

Keywords: Meta-analysis, out-of-sample predictive calibration.

Subjective or "objective" Bayes? As Berger notes, the distinction is false: all Bayesian analysis is subjective, as indeed is all statistical analysis. Description, inference, prediction, and decision-making: these are the four components of statistical work, and all of them centrally involve assumptions and judgments, rendering them subjective at their core. In description, I have to make a judgment about what summaries to retain and what to de-emphasize; in inference, for example from a sample to a population with data gathered observationally (a "sample of convenience"), I need to make a judgment about how the sampled and unsampled units in the population are related to each other; in prediction I have to make a judgment about how the past and future are related; and in decision-making, judgment must guide the choice of utility function and must play a part in assessing the probabilities defining the expected utility to be maximized. There is no need to apologize for the role of subjectivity in statistical analysis: as Goldstein beautifully illustrates with his examples on software reliability and oil reservoirs, all scientific activity that has an inferential character inescapably involves judgment (the data never really "speak for themselves" when you look closely at how scientific inferences are made), and indeed to be human is to make choices based on assumptions and judgments every waking moment.

And yet there is a powerful impulse toward the "objective." To me this stems from a healthy—and crucial—desire: I want to be *coherent* (internally consistent) in my implementation of Bayes, but coherence by itself is not enough to guarantee that my Bayesian answer is a good answer to a real-world question (I am always free in the coherent Bayesian paradigm to insert extremely strong prior information that is, after

*Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA, http://www.ams.ucsc.edu/~draper

the fact, seen to be out of step with the world, and if I do so my Bayesian solution will be poor indeed). This forces me to be guided, not only by coherence, but also by *calibration* (external consistency): as a consultant I want to use Bayesian methods, because Bayes is the best paradigm so far invented for (a) quantifying all relevant sources of uncertainty in real-world problems and (b) effectively propagating that uncertainty through to the final solution to the problem, but if I want to get invited back to consult again (and again) I had better pay attention to how often I get the right answer (e.g., meteorologists who consistently get it wrong about when it will rain will quickly be ignored, or fired, or both), and this is a fundamentally calibrative activity. "Objective" Bayes is a special case of this general desire to be well-calibrated, as I will now attempt to demonstrate.

Consider the problem of prediction of future observables y_i, and to keep things simple suppose that I judge the standard parametric framework $\theta \sim p(\theta), (y_i|\theta) \overset{\text{IID}}{\sim} p(y_i|\theta)$ for $i = 1, \ldots, n$ appropriate for quantifying my uncertainty about the data vector $y = (y_1, \ldots, y_n)$ I'm about to observe, for some choice of prior distribution $p(\theta)$ on a parameter vector θ and some sampling distribution $p(y_i|\theta)$. Let's say, before any data have arrived, that I'm thinking about the moment when y_1 through y_n will have been observed and the next task will be to predict y_{n+1}. If I judge (my uncertainty about) past and future observables to be exchangeable, then I may well wish to use one of Berger's "objective" (diffuse) prior distributions (or something like it) in constructing my predictive distribution $p(y_{n+1}|y)$, because in this case the likelihood function $l(\theta|y) = c\prod_{i=1}^{n} p(y_i|\theta)$ (for some positive c) will contain accurate and well-calibrated information about what I should expect to see when y_{n+1} arrives, and should therefore be allowed to dominate the posterior predictive distribution. But if I have reason to believe that something about "the process giving rise to" y_{n+1} will be different from "the process that gave rise to" $(y_1, \ldots, y_n)$ (the quotes emphasizing the usual and often useful fiction about the existence of an underlying data-generating mechanism), then—in creating what will be seen, after the fact, to be a well-calibrated predictive distribution for y_{n+1}—something other than straightforward use of an "objective" prior will be needed. Here is an example.

In 1992 researchers in Scotland (GREATgroup (1992); also see Leonhardt (2001)) published (in the *British Medical Journal* (BMJ)) the results of a randomized clinical trial with 311 patients to test the effectiveness of a thrombolytic ("clot-busting") drug (anistreplase; call it A) in preventing death following heart attack. The BMJ article reported that the drug reduced the death rate in the treatment group by 49% in relation to the control patients, a result which would revolutionize the treatment of heart attack if it (a) held up under replication and (b) generalized to other thrombolytic agents. Pocock and Spiegelhalter (1992) were skeptical that this result would generalize and replicate, because—while the Scottish experiment was the first high-quality clinical study of drug A—by 1992 there were several trials of other thrombolytic drugs with chemical composition similar to that of A, and the effects of these other drugs were by and large considerably less dramatic. Pocock and Spiegelhalter in effect set themselves the task of trying to predict (in 1992) what a good meta-analysis of the literature on the general class of thrombolytic drugs (including A) roughly 10 years later would

conclude about the effectiveness of these drugs, using for likelihood information only the results of the Scottish study but permitting the results for the other clot-busting drugs accumulated by 1992 to specify a prior (on the true mean effect of the class of thrombolytics, in their hierarchical model) that was strongly informative and that differed substantially from the likelihood. They published a letter in the BMJ in 1992 predicting that the future medical literature, once other studies on thrombolytics were available, would conclude that these drugs as a class only reduce mortality by about 27%. Eight years later, in May 2000, a high-quality meta-analysis (Morrison et al. (2000)) of drug *A* and two other thrombolytics was published in the *Journal of the American Medical Association*, and the conclusion was that this class of drugs lowers mortality by about 17% (this is the prevailing medical consensus today). Pocock and Spiegelhalter, using only the Scottish data for likelihood information, were able to come far closer to the right answer than the Scottish study itself, because (i) they judged that the Scottish trial and future thrombolytic studies were not exchangeable and (ii) they were prepared to use a highly informative (non-"objective") prior (based on sound science) to quantify their judgment. The point of this example is that diffuse ("objective") priors will not always lead to well-calibrated Bayesian predictions; it depends on (your judgment about) how the past and future are related.

Thus the twin pillars guiding practical Bayesian statistical analysis—coherence and calibration—are different with respect to subjectivity and "objectivity." There is nothing subjective about coherence (a particular set $\mathcal{S}$ of probability assessments either does or does not obey the usual rules of probability, and Dutch book either can or cannot be made against $\mathcal{S}$), but calibration is a concept with both subjective and "objective" aspects: I can "objectively" assess whether a particular method for prediction of observables has been well calibrated in the past (e.g., by pretending that subsets of past data are "future" data and seeing how well I can predict them from "past" data), but assertions of good calibration of such a method on future data are inevitably and subjectively based on judgments about how the past and (as-yet-unseen) future are related.

As always when Berger and Goldstein write about something, both the topic and the points they make are important (even if you don't agree with everything they say). While acknowledging that all Bayesian work is subjective at its core, Berger's main contribution is—in the language of my discussion here—to emphasize a particular class of prior distributions that is (a) typically at least "ϵ-coherent" (a description of priors which yield posteriors that are limiting approximations to posteriors from coherent priors) and (b) useful in achieving good calibration when (in your judgment) the past and the future are exchangeable; Goldstein's main contribution is to illuminate the crucial and inescapable role of subjectivity in all good uncertainty quantification. I hope that people will keep both coherence and calibration in mind as they think about how these two papers inform their own Bayesian sensibilities.

Some specific comments and questions about the papers, to draw this discussion to a close:

- In his Section 2.1 ("The *appearance* of objectivity is often required"; my italics) it would seem that Berger is trying to have it both ways: he acknowledges that

(a) statistics is inherently subjective ("Thus the (arguably correct) view that science should embrace subjective statistics"), but (b) scientific users of statistics want "objectivity," so—even though such users can't really have what they want ("Scientists hold up objectivity as the ideal of science, but often fail to achieve it in rather spectacular fashion")—(c) we Bayesian statisticians should (at least seem to) give them what they want, by branding some of our methods "objective" and using these "objective" methods as often as possible in scientific applications; because to do otherwise (i.e., to explicitly acknowledge that the desired "objectivity" is unattainable) "would lead to a considerable downgrading of the value placed on [the discipline of] statistics."

I sincerely hope that this pragmatic and somewhat market-driven philosophy is overly pessimistic. In my view, "objectivity," in the Victorian sense desired by people like Boole (1854) and Venn (1888) (writers whose views turned Fisher away from his originally Bayesian justification of maximum-likelihood methods) cannot be achieved, for the reasons given above; the goal, since subjectivity is inevitable in all human activities (scientific or otherwise), is instead *transparent* subjectivity: (a) statistical work in which the assumptions and judgments are fully in view, for everyone to consider and critique, and in which sensitivity analysis reveals stability or fragility of conclusions with respect to the assumptions and judgments, but also (b) analyses disciplined by both coherence and calibration in a way that helps us, and others working with us, to make good predictions of observables. As argued above, the kinds of priors Berger advocates sometimes lead to well-calibrated predictions and sometimes do not; the goal is not diffuseness-because-it-creates-"objectivity" but an appropriate fusion of coherence and accurate calibration. Good scientists know that Victorian scientific "objectivity" is a myth; if we continue to help them make good predictions—one of the hallmarks of successful Bayesian work—the market demand for statisticians in science and other pursuits will take care of itself.

- In his Section 4.4 Berger criticizes the practice (call this approach $(*)$) of choosing "a uniform prior over a range that includes most of the 'mass' of the likelihood function, but that does not extend too far," by expressing the concern that "the answer can still be quite sensitive to the spread of the rather arbitrarily chosen prior," and it does seem possible that there are situations—e.g., with very little data—in which this concern is justified (see, e.g., Browne and Draper (2006) and the comments on that article for discussion of topics relevant to this issue in hierarchical modeling). Berger's implicit claim in expressing this concern, however, is that the "objective" priors he favors do not suffer from similar sensitivity in the same situations. On what evidence is this implicit claim based? For instance, in the medical diagnosis example in Berger's Section 2.2, he favors Beta(α, β) priors for the relevant probabilities, with particular values of α and β close to zero (he likes $\alpha = \beta = 0.5$ because that specifies the Jeffreys prior, but gives no other justification for these precise "minimally informative" values, when presumably other choices near 0.5 (and possibly even not so near) would have proven just as "objective" by performing about as well as 0.5 in his simulation study with the

sample sizes he examined; I conjecture, by the way, that any sensible implementation of $(*)$ would have performance in Berger's Table 1 that is indistinguishable from that of the Jeffreys prior). Would he then claim that, when there is so little data that results from method $(*)$ are sensitive to details of the range of the uniform prior, his Beta(α, β) approach—with values of α and β differing from 0.5 while still proving "objective" in simulations like those in his Table 1—would not share this sensitivity?

- Goldstein's equation (1), $P_t(A) = P(A|\mathcal{B}) + R$, looks funny because its right-hand side is additive on the probability scale, appearing to leave open the possibility that R could be big or small enough to permit $P_t(A)$ to go negative or exceed 1; so his residual R must have additional constraints in addition to $E(R) = 0$?

- A question for Berger that goes beyond the narrow remit he chose for himself (his paper concerns priors on parameters of models, without any discussion of how the models were chosen): when diffuse priors are called for, given the extreme sensitivity of Bayes factors to details in how the diffuseness is specified, what role does he see for "objective" Bayes in model choice?

References

Boole, G. 1854. *An Investigation Into the Laws of Thought*. London: MacMillan. 426

Browne, W. and D. Draper. 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models (with discussion). *Bayesian Analysis* this issue. 426

GREATgroup. 1992. Feasibility, safety, and efficacy of domiciliary thrombolysis by general practitioners: Grampian Region Early Anistreplase Trial. *British Medical Journal* **305**: 548–553. 424

Leonhardt, D. 2001. Adding art to the rigor of statistical science. *The New York Times* (`www.nytimes.com`): 28 April 2001. 424

Morrison, L., P. R. Verbeek, A. McDonald, B. Sawadsky, and D. Cook. 2000. Mortality and prehospital thrombolysis for acute myocardial infarction: a meta-analysis. *Journal of the American Medical Association* **283**: 2686–2692. 425

Pocock, S. and D. Spiegelhalter. 1992. Domiciliary thrombolysis by general practitioners. *British Medical Journal* **305**: 1015. 424

Venn, J. 1888. *The Logic of Chance*. 3rd ed. London: MacMillan. 426