Statistical Science
2016, Vol. 31, No. 4, 503–505
DOI: 10.1214/16-STS563
Main article DOI: 10.1214/16-STS592

Issues in Robustness Analysis

Michael Goldstein

Abstract. How may we develop methods of analysis which address the consequences of the mismatch between the formal structural requirements of Bayesian analysis and the actual assessments that are carried out in practice? A paper by Watson and Holmes provides an overview of methods developed to address such issues and makes suggestions as to how such analyses might be carried out. This article adds commentary on the principles and practices which should guide us in such problems.

Key words and phrases: Robustness, axiomatics, coherence.

The Bayesian statistical approach is justified in two quite different ways. First, it is considered to be a principled approach, deriving procedures from a simple and natural collection of axioms. Second, it is justified, pragmatically, on the basis that it seems to give reasonable answers to otherwise difficult problems, which incorporate all of the information that we consider relevant for our solutions. However, as the problems that we analyse become increasingly complex, it becomes correspondingly more difficult to be sure either that we have represented our prior knowledge in an appropriate fashion or that our solutions are intuitively reasonable. Further, the axiomatic basis (based on the prescribed behaviour of perfectly rational individuals operating in small worlds) becomes increasingly detached from the actual conditions under which the statistical analysis is carried out.

Therefore, it is important to be clear as to the actual meaning and limitations of a Bayesian (or any other) calculation in such complex problems. Further, we must develop methods of analysis which address the consequences of the mismatch between the formal structure and the actual assessments that have been carried out. Watson and Holmes (2016) provides a splendid overview of methods developed to address these issues and makes some very interesting suggestions as to how such analyses should be carried out. As such, this provides an excellent springboard for discussion of the principles and practices which should guide us in such problems.

We first need to unpick the agenda of the paper, namely the problems that arise when "$f(x;\theta)$ may not be Nature's true sampling distribution or $\pi(\theta)$ does not reflect all aspects of prior subjective beliefs...". Does Nature have a true sampling distribution? Sometimes, perhaps, but in most cases, surely not. Our likelihoods are as much subjective judgements as are our priors. What does uncertainty in our subjective judgements correspond to? There are two fundamentally different interpretations that are habitually employed. In the first, there is a "true but unknown" collection of prior judgements that should be made, given the available prior information, and our uncertainty follows as we do not know what this true collection is. The basic problem with this interpretation is our inability to define, even in principle, this underlying truth. Second, we may be uncertain as to the value that we would assign to our prior judgements if we were able to spend more time and resource in considering the problem. Such uncertainty can be given an operational meaning and it will usually be possible to consider and quantify aspects of our beliefs about such refinements to our prior judgements. This may allow us to develop a principled approach for dealing with such uncertainties.

Consider the Ellsberg paradox, which is simple enough that it allows us to concentrate on basic principles without being distracted by technical details. In the version of the paradox described in the paper, there are two urns, each with 100 balls. In the first, half are red and half are blue. In the second, an unknown number are red and the remainder are blue. You must choose between a ticket (A) which pays \$100 if you

Michael Goldstein is Professor of Statistics, Department of Mathematical Sciences, Durham University, Science Laboratories, Stockton Road, Durham DH1 3LE, UK (e-mail: michael.goldstein@durham.ac.uk).

draw a red ball from the first urn, or a ticket (B) which pays \$100 if you draw a red ball from the second urn. In practice, people tend to prefer A to B, and the authors derive this as a robust choice, arising from the principles that they advocate. If we set aside any real world practical framing issues for the problem, and suppose, as in the paper, that you have no information about the proportion of red balls in the second urn, then this must be an example of the first interpretation of uncertainty, as there is no suggestion that further reflection and analysis would lead to an improved judgement about the proportions in the second urn.

The authors argue that their robustness assessments should be driven by real world consequences, which sounds very reasonable. However, this raises the question as to what does a preference for ticket A over ticket B actually mean? There is no particular guidance on this question that I could find in the paper, so I shall interpret the preference as meaning that if you had B, then you would pay some small amount, $w > 0$ say, to trade B for A. (If there was no such value, then it would be hard to argue that you really held such a preference.) Such preferences now have consequences. The standard Bayesian argument for such a problem is to consider, also, preferences between tickets C and D, which are the same as A and B, respectively, but with red replaced by blue. If you have the same preference between C and D as you do between A and B (and it is hard to see, from the framing of the problem, how you could not), then you would also pay w (or at least some similar positive amount) to trade D for C. Therefore, if you hold both tickets B and D, it would appear that you would pay a positive amount to trade them both for A and C. This does seem paradoxical as B and D together give exactly the same payoff (\$100), with certainty, as do A and C.

The only way to avoid this unpleasant consequence is to allow that your preference between A and B might change if you must also make a choice between C and D, and perhaps this is what the authors intend. However, this is problematic for several reasons. At the foundational level, in establishing the basic property of probability (that the probability of the union of disjoint events is the sum of the probabilities for the individual events), we specify probabilities for two disjoint events by considering our betting prices on these events individually. We then deduce that our betting price on the union of the events is the sum of our prices on the individual events precisely by holding the view that our betting prices on the collection of events is simply the corresponding betting prices that we have specified on the events individually. It seems problematic to carry out a principled probabilistic analysis for which the outcome of the analysis denies the basic arguments from which the axioms of probability are derived. Further, at a practical level, it would seem difficult to keep modifying our inference as we change the collection of outputs that we are concerned with. Indeed, we might even suspect that the collection of outputs selected for the analysis could have been chosen in order to achieve a preferred result for some particular outcomes of special interest.

As we are moving outside the conventional formalism, deriving robustness measures from explicit basic principles is a natural and worthwhile approach. The principles suggested in the paper raise various questions.

Principle 1b (Consequence). This specifies that we should only be concerned with sensitivity to the states, θ, which enter into the loss function, and thus restrict attention to analyses which vary the marginal prior over these states. This is not obvious to me. All the states and, therefore, the resulting inference, are linked through the prior specification. Is there some "meta theorem" to the effect that uncertainty over the prior specification for the nuisance parameters factors out of the process? If not, then it would be easy to imagine problems where the uncertainty in the "nuisance" parameter specification was the driving feature of the robustness calculation for the expected loss over θ. This would arise, for example, in any problem in which it is difficult to observe direct outputs from the likelihood of interest, but it is easy to observe outputs from likelihoods for related quantities which have an indirect, but important, relationship with the quantities of interest.

Principle 2 (Coherence). This is an important idea, based on the argument that two different analyses of the same data, and approximate joint model should lead to the same answer. However, this raises the issue as to the interpretation of the "robust" solution that the theorems derive. That the outcome satisfies a particular optimisation property for the given loss function carries no implication that this should reflect our actual posterior judgements over different options. Therefore, we cannot automatically employ that posterior distribution in any calculations beyond that of minimising the specific expected loss. However, according to my understanding, the computed posterior judgements are used in precisely this way, in the statement and derivation of the uniqueness result, Theorem 4.2, for the authors' suggested solution. This proceeds by comparing the posterior assessment of all of the data, made as a single calculation, with the two stage update in which we observe part of the data, update our judgements by the suggested formula, then update again with the remainder of the data. Do the authors have any further justifications that can be given for such an interpretation of their calculations?

I would like to add a few comments on my preferred approach to these issues. The Bayesian approach should be viewed as a model for an actual inference when dealing with a complex real world problem. For any model, we must consider the extent to which it supports its larger purpose. Typically, this will involve reducing our uncertainty as to real world system behaviour, which therefore can serve as the basis for real world decisions. For an extensive treatment of these issues, see Goldstein and Rougier (2009), and the accompanying discussion. (The paper is concerned with the use of complex computer models, but, for the purpose of the current discussion, this is simply a problem with a likelihood which is expensive to evaluate at any choice of parameter values.)

It is therefore natural and appropriate to consider explicitly the ways in which our Bayesian modelling fails to address the reasoning and analysis which would be required in order for us to have confidence in our analysis and decision making, and to incorporate such considerations explicitly into our posterior assessments. For a detailed discussion, moving from fundamental considerations to a pragmatic simulation based approach to the implementation of such principles, see Williamson and Goldstein (2015). To illustrate the basic ideas, the chosen example is based around the use of an ocean model of realistic size and complexity in order to quantify our beliefs about aspects of global mean temperature in the real ocean. In contrast with most robustness studies, this approach is directly concerned to improve the quality of the posterior uncertainty assessments, and to quantify the potential benefits from the resulting analysis. Such assessments raise important questions as to the purpose and meaning of a Bayesian inference [e.g., the analysis in Williamson and Goldstein (2015) requires us to view expectation rather than probability as the primitive for the theory and is based on explicit principles of temporal coherence]. This more general treatment allows us to move beyond sensitivity and robustness analyses, which may be interesting and revealing but have limited interpretability as a guide for decision choice.

REFERENCES

GOLDSTEIN, M. and ROUGIER, J. (2009). Reified Bayesian modelling and inference for physical systems. *J. Statist. Plann. Inference* **139** 1221–1239. MR2479863

WATSON, J. and HOLMES, C. (2016). Approximate models and robust decisions (with discussion). *Statist. Sci.* **31** 465–557.

WILLIAMSON, D. and GOLDSTEIN, M. (2015). Posterior belief assessment: Extracting meaningful subjective judgements from Bayesian analyses with complex statistical models. *Bayesian Anal.* **10** 877–908. MR3432243