Review (MR3407720) for Mathematical Reviews of "Owhadi, H. and Scovel, C. (2016) Brittleness of Bayesian inference and new Selberg formulas. Communications in Mathematical Sciences 14, no. 1, 83–145."

Michael Evans
Department of Statistics, University of Toronto

This paper is concerned with quantifying uncertainties in the estimation of quantities of interest. The problem is formulated in a Bayesian context and in essence is concerned with the robustness of certain Bayesian estimates to the choice of the prior. The "brittleness" referenced in the title refers to the fact that, depending on the class of priors used and how robustness is measured, the estimates in question can perform maximally badly. Bayesian robustness has an extensive literature, see Berger (1994) and Rios Insua and Ruggeri (2000) among many others, and while this paper is a significant contribution, various issues associated with the use of statistical inference methods in practice, suggest that the results need to be interpreted with caution.

The subject of statistical inference is about formulating rules that one is to apply to data x to determine what evidence x provides concerning questions of interest such as, what is the value of some quantity of interest (estimation) or does a quantity of interest take a particular specified value (hypothesis assessment). So a methodology is required that extracts from x an estimate of the quantity of interest, together with a measure of the accuracy of the estimate, or an indication of whether the data provides evidence either for or against a hypothesized value, together with a measure of the strength of this evidence. To this end a model $\{P_\theta : \theta \in \Theta\}$ is chosen where P_θ is a probability measure on the set $\mathcal{X}$ of possible data values and θ is a parameter indexing the possible measures. The data x was supposedly generated from one of the P_θ and its corresponding parameter value is denoted θ_{true}. The quantity of interest can then be expressed as $\phi = \Phi(\theta)$, for some function Φ defined on Θ, and inference is to be made concerning the value $\phi_{true} = \Phi(\theta_{true})$. Supposing now that Φ is real-valued, the model induces bounds $L_1 = \inf_{\theta\in\Theta} \Phi(\theta) \leq \phi_{true} \leq U_1 = \sup_{\theta\in\Theta} \Phi(\theta)$ on the unknown value ϕ_{true} which have nothing to with the data and, as such, have nothing to do with the evidence.

The data x does not arise simply by being generated from a probability distribution on $\mathcal{X}$. Rather x is produced via a process of measurement. As such, the set of possible data values is finite because all measurements are bounded and made to a finite accuracy. Furthermore, ϕ corresponds to something with a real-world interpretation and isn't just a mathematical variable. For example, if $\phi_{true} \in [0, 1]$ is the proportion of opioid users in a particular population, then it is known to be a member of a finite set of rational numbers that lie in a relatively small, and known, subinterval $[l, u]$ of $[0, 1]$, given what we know about the population and about the extent of opioid use. In essence, a statistical problem is always finite in nature.

The fact that a statistical problem is essentially finite doesn't mean that infinities cannot, or should not, be used as, for example, taking the family $\{P_\theta : \theta \in \Theta\}$ to consist of continuous distributions on an infinite $\mathcal{X}$. It does imply, however, that the use of such objects be reasonable approximations to the underlying possible distributions. There are mathematical simplicities available when employing such approximations that are too valuable to give up by being overly rigorous in the modelling. It does not make sense to argue, as a justification for using continuous distributions, that the measurements could have been made to a finer accuracy because, in the actual application, the accuracy is fixed. So discretization is not an approximation to a true continuous reality, it is the other way around in statistical problems. As such, valid choices of $\{P_\theta : \theta \in \Theta\}$ have to reflect this and cannot be too general. No real constraints are placed on the models considered in the paper.

There are also costs for generality. For one thing the interval $[L_1, U_1]$ may be too wide in the sense that is known in the application context that ϕ_{true} lies in a much shorter subinterval $[l, u]$, as in the opioid user example. It is the case, however, that even quite constrained models for this problem will have $[L_1, U_1]$ too wide as it is not a good idea to completely disallow the possibility that $\phi_{true} \notin [l, u]$. Our knowledge about ϕ_{true} is then perhaps best expressed via a prior probability distribution Λ on Θ, as a way of putting soft constraints on the possible values for ϕ_{true}. In the opioid user example, Λ could be taken to be a beta(α, β) distribution where α and β are chosen to put a large amount of the prior mass on the interval $[l, u]$. The specification of the prior then completes the ingredients of the statistical inference problem as considered in the paper.

There is another cost to generality and this manifests as strange behavior for perfectly reasonable inference methodologies. For example, the maximum likelihood estimator has all the consistency properties you would want of an estimator in the finite context, but unless constraints are placed on $\{P_\theta : \theta \in \Theta\}$, such consistency is lost. So is this a failure for maximum likelihood methodology or are the models where this phenomenon arises deficient in some sense? Such a question is only answered through consideration of the real-world contexts where statistics is to be applied and it is reasonable to argue in favor of maximum likelihood. This leaves somewhat open the question of determining relevant conditions that the mathematical models used in inference need to satisfy to avoid such pathologies, but the answer is surely not to just accept a general mathematical formulation that does not reflect reality. A simple criterion to employ to avoid such contexts, and the ensuing doubts they raise about methodology, is to ask whether or not the paradox or counterexample in question will arise when everything is finite. If not, then it is undoubtedly the case that the negative result is due to the generality and regularity conditions are required on the ingredients in the mathematical formulation so that such problems are avoided.

To assess the robustness of Bayesian inferences, the paper introduces a class of priors Π, that includes the base prior, and considers the values $L_2 = \inf_{\Lambda \in \Pi} E_{\Lambda(\cdot\,|\,x)}(\Phi(\theta))$ and $U_2 = \sup_{\Lambda \in \Pi} E_{\Lambda(\cdot\,|\,x)}(\Phi(\theta))$ where $E_{\Lambda(\cdot\,|\,x)}$ denotes expectation with respect to the posterior distribution $\Lambda(\cdot\,|\,x)$. Clearly $L_1 \le L_2$

and $U_2 \leq U_1$. Here the paper does make a step towards making sure that the discussion is not just about mathematics but reflects a real-world context, when the posterior distribution is introduced. For it is well-known that the definition of conditional probability through the Radon-Nikodym theorem allows for nonregular conditional probabilities and these are not appropriate but just a product of the mathematical formulation. As such the paper demands that the probability of the observed data be positive, effectively restricting to the discrete case to avoid this pathology. A better approach might be to restrict the ingredients of the problem in such a way that a regular conditional probability exists and is obtained via a limiting process. In that way the benefits of using continuous distributions (as approximations) could be retained.

The size of the interval $[L_2, U_2]$ is then taken as a measure of the robustness of the Bayesian estimate $E_{\Lambda(\cdot\,|\,x)}(\Phi(\theta))$ of ϕ_{true}. It is not clear, however, why $E_{\Lambda(\cdot\,|\,x)}(\Phi(\theta))$ is to be taken as the estimate, except that is the Bayes rule when quadratic loss is employed. But why is quadratic loss the relevant loss function and why is a loss function even necessary for inference? There are other approaches to selecting an estimator based on the evidence that do not require a loss function. For scientific inference loss functions have a key weakness in that they are not checkable against the data for their suitability, in contrast to both the model and the prior. So it should be noted that the discussion of robustness here is really focused on only one estimator.

Setting concerns with the choice of estimator aside, consider the possible values for $[L_2, U_2]$. The brittleness of Bayesian inference is interpreted to mean that $L_2 \approx L_1$ and $U_2 \approx U_1$. If $\inf_{\theta\in\Theta} \Phi(\theta)$ and $\sup_{\theta\in\Theta} \Phi(\theta)$ are finite and assumed by Φ, then the prior Λ concentrated on $\{\theta : \Phi(\theta) = \inf_{\theta\in\Theta} \Phi(\theta)\}$ gives $L_2 = L_1$ and Λ concentrated on $\{\theta : \Phi(\theta) = \sup_{\theta\in\Theta} \Phi(\theta)\}$ gives $U_2 = U_1$. So, if Π is too large, brittleness isn't informative. Accordingly, the family Π of priors used is key to determining the relevance of the interval $[L_2, U_2]$ for measuring robustness and various families have been considered in the literature.

The specific problem of estimating the mean of an arbitrary distribution concentrated on $[0, 1]$ is discussed in the paper. The family Π is then taken to be the set of all probability distributions such that the sequence given by the first n moments $(\mu_1(\theta), \ldots, \mu_n(\theta))$ of a distribution P_θ on $[0, 1]$, is uniformly distributed on the space of all possible such truncated moment sequences. The brittleness result is established in that context. It is not at all clear, however, why this Π is an appropriate family of priors to employ for the purpose of assessing robustness. Without a clear justification the result has little impact as it may simply be too large a family of priors.

At this point it is well to reflect back on the purpose of a theory of statistical inference. Certainly data and the questions of scientific interest come first. The role of the ingredients specified by the statistician, here $\{P_\theta : \theta \in \Theta\}$ and Λ, is to formalize the reasoning process from the data to answering the questions of interest. The formalization arises through the specification of a theory that leads to, for example, an estimate of a quantity of interest and an assessment of the error in this estimate. The value of such a formal theory resides in part, in how convincing it is in expressing the evidence appropriately. While there is debate about what that theory is, it is worth remarking that that there is a theory that leads to an estimate of ϕ_{true} that is completely robust to the marginal prior Λ_Φ, although not to the conditional prior on nuisance parameters. Furthermore, the assessment of the uncertainty in the estimate also possesses optimal robustness properties, A discussion of such an approach to inference can be found in Evans (2015). The overall point here is that the rules for inference play a significant role in determining robustness, with some approaches leading to much more robust inferences than others. There is no reason to always think of basing estimation on minimizing quadratic loss, as the developments in the paper implicitly do.

Suppose that a theory of inference has been settled on and so an estimate of ϕ_{true} has been determined. That is not the end of the story. For there are still the ingredients that go into the theory, beyond the data, to produce the inferences, namely, $\{P_\theta : \theta \in \Theta\}$ and Λ. If these are grossly in error, then surely the inferences are of suspect validity even if the inferences possess optimal robustness properties. The model $\{P_\theta : \theta \in \Theta\}$ is considered suspect if the data x is surprising for every P_θ and the prior is suspect if there is prior-data conflict, namely, the truth lies in the tails of the prior. There are consistent methods for checking models and for checking priors. Logically one checks a model first and, if the model passes, then checks the prior. If a model or prior fails, then modifications are required. Interestingly, it is typically much clearer how one is to select a new prior to avoid the conflict, based on the idea of one prior being weakly informative with respect to another. Also, it is interesting to note that it is precisely when a prior is in conflict with the data that it can be expected that inferences will be highly nonrobust with respect to the prior. A relevant reference that discusses these issues is Evans (2015).

The paper is interesting and contains results of value relevant to Bayesian robustness. In a sense the formulation of the problem seems to be too mathematically general to regard the results as entirely convincing. It is acknowledged, however, that finding the precise regularity conditions needed to make the use of very general models compatible with a satisfactory theory of inference is not an easy task, but it is a necessary one.

Berger, J. O. (1994) An overview of robust Bayesian analysis (with discussion). Test, 3, 5–124.

Evans, M. (2015) Measuring Statistical Evidence Using Relative Belief. Monographs on Statistics and Applied Probability 144, CRC Press, Taylor & Francis Group.

Rios Insua, D. and Ruggeri, F. (2000) Robust Bayesian Analysis. Springer-Verlag.