Statistical Science
2016, Vol. 31, No. 4, 506–510
DOI: 10.1214/16-STS567
Main article DOI: 10.1214/16-STS592

Nonparametric Bayesian Clay for Robust Decision Bricks

Christian P. Robert and Judith Rousseau

Abstract. This note discusses Watson and Holmes [*Statist. Sci.* (2016) **31** 465–489] and their proposals towards more robust Bayesian decisions. While we acknowledge and commend the authors for setting new and all-encompassing principles of Bayesian robustness, and while we appreciate the strong anchoring of these within a decision-theoretic framework, we remain uncertain as to what extent such principles can be applied outside binary decisions. We also wonder at the ultimate relevance of Kullback–Leibler neighbourhoods into characterising robustness and we instead favour extensions along nonparametric axes.

Key words and phrases: Decision-theory, Gamma-minimaxity, misspecification, prior selection, robust methodology.

1. INTRODUCTION: THERE IS NOTHING LIKE FIRST-HAND EVIDENCE

The most sensitive aspect of both Bayesian and non-Bayesian statistics is certainly the reliance on a probabilistic model. Even Bayesian nonparametrics are based on a highly concentrated prior modelling taking place in an infinite dimension functional space. Considering the issue of misspecification from within Bayesian analysis is therefore a major undertaking that has been surprisingly overlooked in the past, to the point that it shows up in negative like a missing link in the field. Assessing the goodness of fit of a given model or exploring the consequences of working with a misspecified model have been little studied so far and there is certainly no theoretical or methodological corpus that can be acknowledged as a reference. Bayesian robustness was definitely a keyword in the 80s (Berger, Insua and Ruggeri, 1996) but the field somewhat dwindled over the years, presumably overtaken (as already pointed out by the authors) by the MCMC tsunami. This revolution in statistical computing allowed for much more ambitious modelling and the incorporation of higher variability through hierarchical structures and priors (Cappé and Robert, 2000, Robert and Casella, 2011), although this may constitute a wishful reinterpretation of the past. Another possible reason for the discontinuity in Bayesian robustness studies is that the complex mathematics involved in the formal representation of the robustness desiderata quickly get intractable. In any case, Bayesian robustness was mostly concerned in the 1990s with reducing the impact of the prior modelling, that is, it was understood in terms of robustness against the prior rather than against the model. [A different if marginal approach is adopted in Samaniego (2010), who considers a second type of prior intended to assess the performances of the original prior, albeit with limited applicability.]

In this paper, the authors aim at providing robust inference against possible misspecifications of the sampling model also, producing empirical and methodological perspectives on the issue. This attempt is most commendable and we hope it will induce others to enlarge and deepen the work in this direction. In particular, we think there is some degree of urgency in solving the conundrum of the major "Big Data" challenge, namely the impossibility of deriving a complete

Christian P. Robert is Professor, and Judith Rousseau is Professor, CEREMADE, Université Paris-Dauphine, PSL, 75775 Paris cedex 16, France (e-mail: xian@ceremade.dauphine.fr; rousseau@ceremade.dauphine.fr). Both authors are members of Laboratoire de Statistique, CREST, Paris. C. P. Robert is also affiliated as a part-time professor in the Department of Statistics of the University of Warwick, Coventry, UK.

statistical model with high dimension or complex data structure—an impossibility that differs from the "tall data" case where the sheer size of the data hinders standard algorithms; see, for example, Bardenet, Doucet and Holmes (2014). We certainly appreciate the compelling idea that robustness should be targeted to a specific decision and thus firmly set within a decision-theoretic framework. The paper starts with a refreshingly modern review of robustness in this context. The authors then advance two types of proposition towards evaluating robustness of some Bayesian procedures and possibly suggesting a more robust procedure. However, we find the empirical assessment that follows from the approach somewhat too tentative and too inconclusive to provide useful guidance to practitioners. It indeed appears that expanding the framework beyond binary decisions faces considerable conceptual and practical difficulties. We thus suggest below that more (involved) nonparametric procedures should be developed towards this very goal. In our opinion, one of the most challenging foundations of the paper is the involvement of the action a in the construction of the reference prior, which does not appear natural or acceptable to us, because the coherence of the Bayesian perspective does not seem to transfer to this approach.

2. GAME OF THORNS: DUALITY BETWEEN POSTERIOR AND LOSS FUNCTION

First, we want to point out that, in the specific and important setting of discrete decision spaces, and in particular for binary decisions, the propositions of the authors make a lot of sense. We are thus genuinely curious to understand how these ideas could extent to infinite and continuous decision spaces. While we are similarly strongly inclined towards a decision-theoretic approach to statistics, and hence definitely sympathetic to the perspective adopted in the paper, we also think one should keep in mind the unfortunately ignored remark by Herman Rubin (1987) that model densities and loss functions are only utilised through the product $\mathsf{prior} \times \mathsf{likelihood} \times \mathsf{loss}$ in Bayesian decision theory. Hence, in this sense, prior and loss are indistinguishable beyond this product. This reference is meant to support the point that both sampling and prior densities *and* losses should be assessed via a robustness filter, rather than solely examining the $\mathsf{prior} \times \mathsf{likelihood}$ term under the robustness magnifying glass. There exist rationality arguments and the like about the choice of a loss function (Raiffa, 1968, DeGroot, 1970, Berger, 1985). However, since most parameters are a by-product (of possibly strong relevance) of defining a model, rather than enjoying an existence of their own, it is difficult not to think of the loss function as being linked to the model (mis-)specification. In particular, this is why we do not see how "changing the likelihood changes the interpretation of the prior" (page 6) is such a "thorny issue" (page 6). The notion of loss robustness is only alluded to in the conclusion of the paper and we hope it could be considered much further in parallel with the current assessment of the model. Obviously, there is no free lunch and it is necessary to accept some of the modelling hypotheses to be able to give meaningful conclusions. However, our central comment is that the approaches seem to make a lot of sense for decisions belonging to a finite space, but may make less or at least not so obvious sense in continuous cases.

3. THE BAKER STREET IRREGULARS: ARE KULLBACK–LEIBLER NEIGHBOURHOODS A PERTINENT CHOICE TO ASSESS ROBUSTNESS?

The authors build and study semilocal measures of robustness. They are "local" in the sense that the impact of the decision is evaluated on a neighbourhood of the posterior distribution. As nicely reviewed in their Section 2, this idea is not new and Kullback–Leibler neighbourhoods have already been considered in a series of papers. The key justification found in the present paper for using Kullback–Leibler neighbourhoods is in Theorem 4.2, where some notion of coherency is invoked.

The authors build their approach from three principles. It is hard to disagree with these three principles laid out for D-open methods. However, an evaluation based on Kullback–Leibler neighbourhoods raises a few issues. The result produced in Theorem 4.1 indicates that the least favourable prior involves the exponential of the loss: this is not surprising given earlier works like Bissiri, Holmes and Walker (2013). What sounds rather confusing while central to the theme of the paper is the fact that each possible action a induces a different prior or rather a different distribution in the parameter space. To agree on this peculiar incorporation of the decision a in the least favourable model (or in posterior distribution) means that the least favourable decision is doubly *a posteriori*: it is indeed an update after both having observed the data and taken an action a. This proposal makes complete sense when evaluating the action, but it gets difficult to understand its meaning when proposing a new action. In other words, it seems to run against the grain of Bayesian principles. The persistence of the indexing in a of the notions and notation throughout the paper thus remains quite puzzling to us. (Similar puzzlement is attached to the impact of the Monte Carlo variability on the final decision, as discussed in Section 4.3.)

Besides, the least favourable posteriors only enjoy an implicit expression since the Lagrange multipliers $\lambda_a(C)$ are not explicit and it is difficult to understand how they behave with a and C.

More puzzling still is the feature that if the posterior distribution is extremely concentrated, as it would happen in the context of large data sets, the Kullback–Leibler neighbourhood, for a fixed radius C, will be very small, hence leading to a (probably) false impression of robustness. To illustrate this point, consider the case where the posterior distribution is close to $\mathcal{N}(\hat\theta, v/n)$ with n large (in other words, the posterior distribution satisfies a Bernstein–von Mises theorem). To understand the impact on $\lambda_a(C)$, consider the case when $L_a(\theta) = (a-\theta)^2$. Then $\pi^{\sup}_{a,C} \approx \mathcal{N}(\mu_n, v_n)$ with

$$\mu_n = (\hat\theta n/v - 2\lambda_a(C)a)/(n/v - 2\lambda_a(C))$$

and

$$v_n = (n/v - 2\lambda_a(C))^{-1}.$$

It is easy to see that when $\sqrt{n}|a-\hat\theta| \gg 1$,

$$\lambda_a(C) \approx \sqrt{\frac{nC}{2}}|a-\hat\theta|^{-1}$$

while if $a = \hat\theta$, $\lambda_a(C) \approx \sqrt{C}n/v$. In the latter case,

$$\psi^{\sup}_{(a)}(C) \approx \frac{v}{n}(1-u)^{-1}, \quad 2C = \frac{u}{1-u} + \log(1-u)$$

while in the former case,

$$\psi^{\sup}_{(a)}(C) \approx \left(\frac{v}{n} + (a-\hat\theta)^2\right)\left(1 + \frac{\sqrt{2Cv}}{\sqrt{n}|a-\hat\theta|} \cdot \left(1 - \frac{\sqrt{2Cv}}{\sqrt{n}|a-\hat\theta|}\right)^{-1}\right).$$

The difference between $\psi^{\sup}_{(a)}(C)$ and $\psi^{\inf}_{(a)}(C)$ is of the same order as the risk under π_I when $a = \hat\theta$ and is of a smaller order than the risk when $a \neq \hat\theta$. Unless C becomes quite large, the decision a that minimises $\psi^{\sup}_{(a)}(C)$ is approximately $\hat\theta$. This behaviour does suggest that the posterior leads to robust inference, while it could well be that the model is strongly mis-specified. Such a difficulty appears to be a consequence of using Kullback–Leibler neighbourhoods since L_1 neighbourhoods might have led to a different behaviour, although this is far from certain.

This naturally drives us to question the relevance of the coherence requirement of Theorem 4.2. While the coherence requirement was quite natural in Bissiri, Holmes and Walker (2013), since intended for building the likelihood, it is here unclear why the same result should be obtained directly or after having first observed a subset of observations $x_{(1)}$ and constructed a least favourable prior $\pi^{\sup}_{a,C}(x_{(1)}, a)$, prior to observing the remainder of the observations. We feel that $\pi^{\sup}_{a,C}(x, a)$ exists solely after observing x and that it makes solely sense in this very context. We cannot fathom an equivalent to this result in standard decision-theoretic point estimation. This is presumably due to the fact that, for a continuum of actions, there is not much appeal in considering a continuum of priors. It is, however, difficult to evaluate the relevance or importance of this coherence constraint. Once again, generalising to a continuous action space seems delicate without further guidance.[1]

An advantage of using the Kullback–Leibler divergence is that it is mathematically convenient, up to a certain point, but this proximity measure does not necessarily make sense from a statistical viewpoint. For instance, is there a model likelihood *and* a prior associated with *every single* distribution contained within a Kullback–Leibler neighbourhood? It is presumably the case, given that this correspondence has only to hold for a fixed value of the sample x. Furthermore, is the least favourable distribution a true posterior for *all* samples x? This is presumably not the case, unless one agrees to include "data-dependent priors" in the range of solutions. This issue may eventually prove to be the most delicate aspect of the paper, namely that the calibration of the evaluation is highly dependent on how far from the reference value we allow the posteriors to drift and that we have no clear idea about the meaning of the resulting neighbourhood, as defined at the end of Section 4.1. It seems to us that one of the reasons C is so difficult to calibrate is that Kullback–Leibler neighbourhoods are rather abstract objects. Thus, one cannot resort to intuition or subjective knowledge in the calibration process. The discussion by the authors in the conclusion of their paper touches upon this very point.

[1]Note also that the case when the decision a is the parameter θ itself or rather its estimate $\hat\theta$ as in Section 4.2 remains a puzzle to us as the notation θ in this section seems to be used with this double meaning, which makes the outcome questionable.

In some specific settings, it should be possible to create neighbourhoods by looking at some quantity of interest and setting bounds or limits on the posterior values or variability for this quantity. An alternative would be to resort to an approximate Bayesian computation (Marin et al., 2011) type of robustness where only posteriors that have the ability to predict the actual data (or the original optimal decision) are allowed within the corresponding neighbourhood. While this approach is only vaguely defined, and may be delicate to implement in practice, it carries a most natural kind of proximity.

As a side remark, we appreciate the notion of evaluating the impact of single observations on the inference, as it creates a decision-based outlier and leverage assessment perspective, but we cannot quantify how much one can conclude about the appropriateness of the model from the divergence measure, the more because it looks terribly similar to an harmonic mean estimator (Chopin and Robert, 2007).

In connection with this remark, and even before we read the short Section 4.2.3, the introduction of C-admissible actions reminded us of Γ-minimax procedures, mostly studied in the 1980s and early 1990s. This notion did not attract a large flock of followers at the time, because it is quite delicate to figure out whether or not a procedure is Γ-minimax. However, once we went through Section 4.2.3, we got mildly confused as we could not see how the action enters the choice of the minimax prior. Nor why it should.[2] Clearly, C-admissibility as defined here is a much more convoluted notion, which not only involves the range of acceptable priors but also a new least favourable prior and the optimal decision under the original prior. In a possibly vague sense, this original prior constitutes the centre of the Kullback–Leibler ball.[3] The various proposals of Section 4.2 highlight the links between the proposed approach and other proposals for robust inference, but they also suggest that the calibration of C can only be achieved on a case by case basis.

[2]We note that the current paper extends the original Γ-minimax problem by considering a collection of posteriors rather than priors. While this makes complete sense from a conditioning perspective, it does not necessarily lead to coherent answers since the least favourable priors are then data dependent.

[3]In computational terms, using the centre of the ball for importance sampling (page 17) may be fraught with danger as the Kullback proximity does not guarantee tail behaviour, and hence finite variance. Hence, using the importance weights to calibrate the Kullback neighbourhood cannot be recommended without further assessment.

4. TURNING TO NONPARAMETRIC BRICKS

A nonparametric approach via (e.g.) Dirichlet priors is an expected if welcomed thread. For one thing, it sounds more genuine for inferential purposes, when compared with the call to Kullback–Leibler neighbourhoods. This is particularly compelling when considering a candidate posterior as the functional parameter of the Dirichlet (hyper) prior—although the use of the term "prior" clashes with the fact that the Dirichlet process is centred at the posterior π_I. Looking at the variability or at the distribution of the loss under a Dirichlet process centred at the posterior distribution constitutes quite an interesting and elegant proposal, but we wonder about the type of *neighbourhoods* of the posterior distributions induced by this approach. (We also appreciate the Monte Carlo convenience offered by the Dirichlet process representation described in Section 4.3.1.)

How do these "neighbourhoods" compare to the Kullback–Leibler neighbourhoods of the first part? Bayesian nonparametrics suffer from the well-known drawback that functional priors are almost irremediably concentrated on very small regions of the functional space, and thus do not necessarily reflect much of a range of possible posteriors. Obviously, in a nonparametric framework, the notion of neighbourhood is much weaker than in the parametric case. One can envision these neighbourhoods as soft versus hard neighbourhoods, by analogy to soft versus hard thresholding. When the authors state in Section 4.3.2 that an action optimal under the functional parameter of the Dirichlet prior will remain optimal (in expectation) under the random measure distributed from this Dirichlet, this may constitute a second-tier property in that it only stands in expectation.

We find the proposition of studying the probability of changing the optimal decision particularly relevant, in particular in the context of discrete and finite decision spaces. However, we are unsure of what happens or would happen in the context of continuous decisions. Indeed, in this latter case, $L_a(\theta)$ also follows a Dirichlet process centred at the posterior distribution of $L_a(\theta)$. Hence, if q_τ is the τth quantile of $L_a(\theta)$ under the posterior π_I, then $Q(L_a(\theta) \le q_\tau) \sim \text{Beta}(\alpha(1-\tau), \alpha\tau)$, which depends only on τ and on the mass of the Dirichlet process α. There is no dependence whatsoever on π_I, which seems to imply it presents little interest. Thus, we wonder which relevant quantities could be produced in continuous settings towards a better understanding of the variability of the loss function under deviations from the posterior.

5. CONCLUSION: THE GAME IS AFOOT!

While the authors have already uncovered several interesting new avenues for exploring Bayesian robustness, we want to point out yet another avenue associated with the notion of penalised complexity (PC) priors, as proposed by Simpson et al. (2014), In fact, we think penalised complexity is eminently relevant to the robustness issue as this perspective tackles the prior specification side covered in Section 2.2. The starting point in Simpson et al. (2014) is a base model, out of which possibly robust extensions can be constructed. While there is no automated derivation of this base model, it corresponds to the operational model mentioned in the current paper. Simpson et al. (2014) further rely on a functional distance from the base and make the quite natural proposal of setting a prior on this distance. Even though there is no decision-theoretic aspect to be found in this proposal, an extension in this direction is certainly feasible.

In conclusion, we commend the authors for this foray into Bayesian robustness and for producing such a novel perspective. Formalising this aspect of Bayesian analysis is absolutely essential for methodological and practical purposes, even when not required by foundational arguments. We also recognise that the proposals made in the paper are mostly exploratory, rather than directive, primarily aiming at representing the variability of the Bayesian output when some of its components are uncertain or misspecified. Again, we stress that this represents an important step in the rational and objective evaluation of Bayesian procedures and we congratulate the authors for initiating such a path. Some of the proposals made therein may require further investigation in terms of Bayesian coherence, but they open a different perspective on how to envision Bayesian decision making from a broader viewpoint.

ACKNOWLEDGEMENTS

Research partly supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2012–2015 grant ANR-11-BS01-0010 "Calibration".

The main bulk of those comments were written during the Bayesian week workshop held at CIRM, Luminy, France, March 1–4, 2016, in a most supportive and serene working environment. We thank the organisers of this meeting for this great opportunity. We are also grateful to Peter Green for his remarkable patience and to the editorial team for helpful suggestions towards the final version of this discussion.

REFERENCES

BARDENET, R., DOUCET, A. and HOLMES, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proc*. 31*st Intern*. *Conf*. *Machine Learning* (*ICML*) 405–413.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York. MR0804611

BERGER, J., INSUA, D. and RUGGERI, F. (1996). Bayesian robustness. In *Proceedings of the* 2*nd International Workshop Held in Rimini*, *May* 22–25, 1995. IMS, Hayward, CA. MR1478676

BISSIRI, P., HOLMES, C. and WALKER, S. (2013). A general framework for updating belief distributions. Technical report, available at arXiv:1306.6430v1.

CAPPÉ, O. and ROBERT, C. P. (2000). Markov chain Monte Carlo: 10 years and still running. *J. Amer. Statist*. *Assoc*. **95** 1282–1286. MR1825276

CHOPIN, N. and ROBERT, C. (2007). Comments on "Estimating the integrated likelihood via posterior simulation using the harmonic mean identity". In *Bayesian Statistics* (O. U. P. Bernardo et al., eds.). *Oxford Sci*. *Publ*. **8** 371–416. Oxford Univ. Press, Oxford. MR2433201

DEGROOT, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York. MR0356303

MARIN, J., PUDLO, P., ROBERT, C. and RYDER, R. (2011). Approximate Bayesian computational methods. *Stat*. *Comput*. **21** 279–291.

RAIFFA, H. (1968). *Decision Analysis*: *Introductory Lectures on Choices Under Uncertainty*. Addison-Wesley, Reading, MA.

ROBERT, C. and CASELLA, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statist*. *Sci*. **26** 102–115. MR2849912

RUBIN, H. (1987). A weak system of axioms for "rational" behavior and the nonseparability of utility from prior. *Statist*. *Decisions* **5** 47–58. MR0886877

SAMANIEGO, F. J. (2010). *A Comparison of the Bayesian and Frequentist Approaches to Estimation*. Springer, New York. MR2664350

SIMPSON, D. P., MARTINS, T. G., RIEBLER, A., FUGLSTAD, G.-A., RUE, H. and SØRBYE, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. Available at arXiv:1403.4630.

WATSON, J. and HOLMES, C. (2016). Approximate models and robust decisions. *Statist*. *Sci*. **31** 465–489.