Statistical Science
2016, Vol. 31, No. 4, 516–520
DOI: 10.1214/16-STS596
Main article DOI: 10.1214/16-STS592

Rejoinder: Approximate Models and Robust Decisions

James Watson and Chris Holmes

We wish to thank all of the discussants for their insightful comments. We have certainly benefitted from considering their perspective of our work. In the following rejoinder, we begin by reiterating the central tenet of our approach, followed by some general pointers to common themes arising across the discussions, and finally a point-by-point reply to some specific issues raised by individual reviewers.

The overriding objective of our work is to advocate the inclusion of decision analysis within the iterative process of scientific learning as laid out in the seminal paper of Box (1980). This iterative process proceeds firstly by a model *estimation* stage where the statistical model is updated *as if* it was true. This update, for us, takes the form of a Bayes posterior. In the approach of Box, the modeller then undertakes a second stage of model *criticism* that potentially leads to model adjustments, for example, via model elaboration, followed by re-estimation, re-criticism and so forth. In our paper, we call for the use of formal and informal (exploratory) decision analysis into the model criticism stage of Box that directly takes into account the context and rationale for the model's use and the questions under consideration.

It is important to note that we are not advocating $\pi^{\sup}_{a,C}$ as a true model for the data, nor necessarily an actual representation of beliefs, although it is interesting to see connections with the historic use of robust priors (Section 4.2). As stated by Box, model criticism and parameter estimation are distinct and demand different methodologies. It is $\psi^{\sup}_{a,C}$, the maximum expected loss occurring within a KL neighbourhood of size C at the model criticism stage that is our fundamental object of interest. For if $\psi^{\sup}_{a,C}$ is relatively large for small changes to the estimated model, then it indicates that the downstream decision analysis may be highly unstable, as for small changes in the posterior model we can observe a substantial increase in expected loss. Highlighting this at the model criticism stage allows for further diagnostics, insight and model elaboration. We believe it is incumbent on modellers to take into account the context of their models and where appropriate incorporate decision analysis into their model criticism.

Now beginning from this standpoint we consider some general themes that arose from the discussants before considering whether or not models should be discarded all together, the setting of the KL neighbourhood size and ending our rejoinder with answers to some finer points.

1. EX POST CRITICISM

1.1 Formal Model Checking: Moving Away from "Statistical Truths"

The discussion by economists Hansen and Marinacci (H&M) provides an interesting and refreshing perspective on model misspecification from outside of traditional statistics. They espouse Wald's philosophy of decision making, where the goal is no longer discovering "statistical truths" but rather to use models operationally in light of a posited objective function (loss function).[1] H&M show how concerns of model misspecification can be expressed as "aversion to ambiguity" as stated in its general form as an optimization problem where the decision maker solves

$$\max_{a\in A}\min_{\pi}\int_{\Theta} U_a(\theta)\pi(\mathrm{d}\theta) + C(\pi), \tag{1}$$

where $U_a(\theta) = -L_a(\theta)$ is the utility function, C is a penalty function, or regularization term, that encodes for *variational preferences* over probability models (Maccheroni, Marinacci and Rustichini, 2006). The form of (1) is instructive, as for KL divergence (relative entropy) and other convex C the solution of (1)

James Watson is a Postdoctoral Researcher at the Mahidol-Oxford Research Unit, Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom (e-mail: jwatowatson@gmail.com). Chris Holmes is Professor of Biostatistics, Department of Statistics, University of Oxford, United Kingdom (e-mail: cholmes@stats.ox.ac.uk).

[1] Pascal's wager is the most famous example of this.

can be calculated. Other variational preferences could be expressed, for example, L_1 neighbourhoods (Robert and Rousseau), but the decision maker may then face an intractable optimization problem over the space of probability measures on Θ.

The notion of aversion to ambiguity in econometrics is different to model estimation or model elaboration; see our rejoinder Section 1.3 below, as suggested with different flavors by Glad and Hjort, Robert and Rousseau, and Goldstein, or a more drastic rejection of modelling as suggested by Grunwald (Section 2).

In the context of model criticism and sensitivity analysis, we disagree with the view of Goldstein that "it would seem difficult to keep modifying our inference as we change the collection of outputs that we are concerned with". On the contrary, we believe it is incumbent on the decision maker to take into account the operational performance of the model at the model criticism stage. This underlies our Principle 1b whereby we focus on uncertainty of states included in the loss function. An interesting illustration of decision analysis within modelling comes from the Bayesian clinical trial literature where different priors may be used for different purposes:

> Note that it may be sensible to match the prior to the decision one hopes to reach; the prior should represent "an adversary who will need to be disillusioned by the data to stop further experimentation" Spiegelhalter, Freedman and Parmar (1994), as quoted in Berry et al. (2011).

1.2 Posterior Model Checking After Estimation

The general decision problem with objective function given in (1) can also be used as a basis for principled posterior model checking. Using the data twice, so called "doubly *a posteriori*" by Robert and Rousseau, disregards Bayesian principles, but posterior model checking against data is a well established and essential component of model criticism (Box, 1980). To preclude empirical model criticism after estimation would deny many established statistical procedures including the use of residual analysis, outlier detection, posterior predictive checks and calibration. If it is not feasible to check the model against out-of-sample validation data, the best resource are the data at hand. For this reason, we also disagree with Grunwald that a minimax approach with a data-dependent loss function is "unnatural" at the model criticism stage. We wholeheartedly agree with Goldstein that posterior model checking is an essential element of model construction and validation.

To reiterate, posterior model checking and criticism is fundamentally different from posterior inference or estimation. Our methods formally introduce decision theory into posterior model checking at the criticism stage, focussing on "whether our models are wrong in having missed something essential to the questions under consideration" (Hansen, 2014).

Graphical displays are also a key component to posterior model checks as discussed in Section 3 of our paper.

1.3 Model Elaboration

Model elaboration following model criticism is an important component of the Box process of iterative learning. However it would seem rather restrictive to have to pre-specify the full model elaboration a priori. Whilst we commend the ideas from Glad and Hjort, similar in spirit to those given by Carota, Parmigiani and Polson (1996), they do not provide for formal model criticism through decision analysis.

We support the use of model elaboration and model refinement in the context of model construction of π_I. In the Introduction of our paper, we state our assumption that the modeller has specified "to the best of the modeller's ability" the current model π_I. What is left, and what we believe is missing from the model criticism stage is a sensitivity analysis to those aspects central to the use and rationale for the model. Hence, we cannot agree with Glad and Hjort who seek to combine the model criticism stage within model estimation via an extension parameter to be learnt from the data "rather than by introducing extra uncertainty after the full analysis". Glad and Hjort call for "Model uncertainty first, not afterward". We would respond "Model uncertainty before *and* after estimation".

The structure proposed by Simpson et al. (2014)[2] is another approach for selecting default priors which have certain desirable "robust" properties and we support its use in the construction of π_I, but it lacks the contextual nature of a decision theoretic component to model criticism.

2. WHEN TO DISCARD MODELS?

A more fundamental question still is when to abandon models altogether. Grunwald has been a pioneer

[2]Thank you to Robert and Rousseau for pointing out this missing reference in our original paper, which was an oversight on our part.

in the development of probabilistic approaches for robust estimators, for example, Grünwald and van Ommen (2014). Recently one of us (Bissiri, Holmes and Walker, 2016) has shown how general-Bayesian updating using loss-likelihood functions of the form $e^{-\lambda \sum_{i=1}^{n} L(\theta, x_i)}$ lead to valid generalised posteriors representing subjective beliefs on the unknown value of the estimand minimising expected loss.

We believe there are situations where one approach (updating via estimators) is more suited than another (updating via likelihoods) but neither approach is better in all circumstances. In this paper we concentrate on those situations for which we shouldn't abandon models "but exercising caution in how we use them" Hansen (2014). An advantage of sticking within a model based approach is the unified framework by which one can handle various features of the data such as missing values, random effects, hierarchical structures and predictive distributions on observables. There remain open questions on how best to handle such structures within the generalised-Bayesian updates of Grunwald or Bissiri et al. (Grünwald and van Ommen, 2014, Bissiri, Holmes and Walker, 2016).

3. CHOOSING THE NEIGHBOURHOOD SIZE

As pointed out by a number of discussants and acknowledged by ourselves in the paper, there is an open question on the setting of the neighbourhood size through $C(\pi)$ in (1). Bochkina provides new insight on this, drawing on ideas from across the literature. Robert and Rousseau question the use of a fixed C for regular models with increasing sample sizes as the robust-model becomes qualitatively more and more concentrated around π_I. However, this in itself may not be such an issue; see Chapter 9 of Hansen and Sargent (2008) where they propose setting of $C(\pi)$ via detection probabilities, and an implicit implication in R&R is that the statistician must be highly confident in the accuracy of a relatively simple model if they propose to use this as their best representation of the world for increasingly large sample sizes.

In general, we feel more work is needed in this important problem.

4. POINT BY POINT RESPONSE

We now turn to some specific points raised by individual reviewers.

4.1 Grunwald: Data Dependent Losses, Maximin or Minimax

Grunwald questions the use of loss functions for "*already observed* data", and later on highlights a potential conflicting recommendation to either minimise or maximise the expected loss in the neighbourhood (Sections 4.2.1, 4.2.2). We acknowledge that this was at best unclear in the paper. To us, it seems perfectly sensible to use already observed data within the model criticism stage (minimax) when exploring decision robustness. Whereas in the absence of a likelihood during the estimation stage, a loss-function can be used within a general update as replacement for the log-likelihood (maximin) (Bissiri, Holmes and Walker, 2016).

4.2 Goldstein: Ellsberg Paradox, Subaddivity and Coherence

The paradox described by Goldstein for the Ellsberg double urn scenario is an example of the notion of "coherent risk measures" often found in finance and actuarial science (Artzner et al., 1999). Risk measures are clearly not probability measures but functionals defined on loss functions. This notion of coherence requires a *subaddivity* property whereby the "risk" evaluation of a grouping of actions is less or equal to the sum of risks of the individual actions. The Value at Risk is a famous example of a noncoherent risk measure [see Section 3.1 and Artzner et al. (1999)]. We argue that although this coherence property is natural in finance (e.g., portfolio selection) it has no place in standard Bayesian decision theory where the decision maker is looking to choose a specific state in the action space (not a subset for example). Goldstein arrives at the paradox by changing the action space midway (initially $\mathcal{A} = \{A, B, C, D\}$ and subsequently $\mathcal{A} = \{A\&C, B\&D\}$). Our treatment of the Ellsberg paradox (Section 4.1.3) showed how actions with equal posterior expected loss could be differentiated using their variance, a consequence of local minimax decision making. If the action space is redefined, then the evaluation (e.g., decision maker's criticism via local minimax) should be recalculated. This resolves the paradox as now both the actions $A\&C$ and $B\&D$ have the same expectation and variance in loss, which is zero. It therefore seems wrong to state that our procedure "denies the basic arguments from which the axioms of probability are derived". This example nicely illustrates how local minimax can be used to score actions and thereby regularise ill-posed problems such as Ellsberg (1961).

Goldstein observes that $\pi^{\sup}$ "carries no implication that this should reflect our actual posterior judgements... beyond that of minimising the expected loss". We would mainly agree with this statement, as outlined in the opening of this rejoinder. However, the coherence property is important in ensuring that we arrive at the same value of the expected loss, $\psi^{\sup}$, regardless of the order by which the data is presented.

4.3 Glad and Hjort: Model Elaboration, Concept Drift and Dirichlet Processes

Glad and Hjort provide an illustrative example in the optimal situation where the a priori model elaboration happens to match the truth, $y = \beta_0 + \beta_1 x + \varepsilon$, and is compared to a simpler model $y = \beta_0 + \varepsilon$. Their example takes $n = 100$ pairs (y_i, x_i) with $x_i = i/n, i = 1, \ldots, n$. The task is to then predict y_{n+1} at $x_{n+1} = 1 + 1/n$. Using flat priors their MAP estimates of β_0, β_1 are unbiased and on average the dotted red curve in Figure 1 should be centred on the true value of 2.349. Their model elaboration adjusts for the underprediction by the base model at x_{n+1}. However, suppose the true model was in fact $y = \beta_0 + \theta(x - \bar{x})^2$, so that $\mathrm{Cov}(x - \bar{x}, (x - \bar{x})^2) = 0$, in this case the elaboration parameter is centred at $\delta = 0$ and their green dotted curve would lay over the solid black curve suggesting no model miss-fit. This is precisely where careful posterior model checking and diagnostics can reveal systematic deficiencies in the posited model and its elaboration.

In our situation, if we suspected model misspecification and a loss on prediction then following Section 4.2.1 in our paper we might explore the predictive distribution in y_{n+1} using a weighted regression with weights $\Delta(u_i) = -|x_{n+1} - x_i|$ giving

$$\pi^{\sup} \propto e^{-\sum_i \Delta(u_i) L(y_{n+1}, \theta)} \pi_I(\theta),$$

where the loss could be L_2 loss to data for example. This would improve prediction and identify potential under or over prediction by the baseline model, $y = \beta_0 + \varepsilon$, regardless of knowing the truth or not.

Glad and Hjort point to the issue of using the Dirichlet process to draw distributions from a KL neighbourhood centered at π_I. As we mention at the start of Section 4.3 and reference an accompanying paper where this is studied in depth (Watson, Nieto-Barajas and Holmes, 2016), the KL divergence of a random draw will be infinite. Hence, we don't use the KL divergence as a metric to define the neighbourhood but the L_1 distance, centred at the loss distribution induced by π_I.

4.4 Robert and Rousseau: Continuous Actions, Leverage and Dirichlet Draws

Robert and Rousseau wonder how our methods translate for continuous action spaces and argue that they have a discrete flavour to them. Indeed, the examples we provide are all with discrete action spaces. However, the use of (1) for continuous parameter spaces is widespread, for example, in the robust control and econometrics literature (Whittle, 1990, Hansen and Sargent, 2008); moreover, Γ-minimax priors have historically be used for continuous problems.

Regarding the uniqueness of models for all neighbourhood sizes, see Ahmadi-Javid (2011). Regarding identification of high-leverage points for decision sensitivity, the performance of the proposed estimator used in Section 3.4 should not impact on whether the estimand is useful, in much the same way that one would be wrong to criticise the use of Bayes factors and marginal likelihoods because some people estimate them using harmonic means. Regarding the comment of centering the Dirichlet process on $L_a(\theta)$, and hence "There is no dependence on $\pi_I \ldots$". This is incorrect. For as $\{\pi_I, a\}$ change so does the distribution of $L_a(\theta)$, hence it is more correct to write $q_\tau^{(a,\pi_I)}$ to define the value of the τ'th quantile of the loss distribution $L_a(\theta)$ under π_I which will depend on the action a.

5. CONCLUSION

We believe that decision analysis has an important role to play in Box's model criticism in situations where models are used operationally to assist the choice of actions. To deny this masks a key component of uncertainty that directly affects sensitivity of decisions to model misspecification. We would like to thank all of the discussants for their comments and we hope that collectively this work will motivate others to explore the integration of decision analysis within posterior model criticism and robustness.

REFERENCES

AHMADI-JAVID, A. (2011). An information-theoretic approach to constructing coherent risk measures. In *IEEE International Symposium on Information Theory Proceedings* (*ISIT*) 2125–2127. IEEE, New York.

ARTZNER, P., DELBAEN, F., EBER, J.-M. and HEATH, D. (1999). Coherent measures of risk. *Math. Finance* **9** 203–228. MR1850791

BERRY, S. M., CARLIN, B. P., LEE, J. J. and MÜLLER, P. (2011). *Bayesian Adaptive Methods for Clinical Trials*. CRC Press, Boca Raton, FL. MR2723582

BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745

CAROTA, C., PARMIGIANI, G. and POLSON, N. G. (1996). Diagnostic measures for model criticism. *J. Amer. Statist. Assoc.* **91** 753–762. MR1395742

ELLSBERG, D. (1961). Risk, ambiguity, and the Savage axioms. *Q. J. Econ.* 643–669.

GRÜNWALD, P. and VAN OMMEN, T. (2014). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. Preprint. Available at arXiv:1412.3730.

HANSEN, L. P. (2014). Nobel lecture: Uncertainty outside and inside economic models. *J. Polit. Econ.* **122** 945–987.

HANSEN, L. P. and SARGENT, T. J. (2008). *Robustness*. Princeton University Press, Princeton, NJ.

MACCHERONI, F., MARINACCI, M. and RUSTICHINI, A. (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica* **74** 1447–1498. MR2268407

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

SPIEGELHALTER, D. J., FREEDMAN, L. S. and PARMAR, M. K. B. (1994). Bayesian approaches to randomized trials. *J. Roy. Statist. Soc. Ser. A* **157** 357–416. MR1321308

WATSON, J., NIETO-BARAJAS, L. and HOLMES, C. (2016). Characterising variation of nonparametric random probability models using the Kullback–Leibler divergence. *Statistics*. To appear. Available at arXiv:1411.6578.

WHITTLE, P. (1990). *Risk-Sensitive Optimal Control*. Wiley, Chichester. MR1093001