BAYESIAN STATISTICS 6, pp. 000–000
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)
Oxford University Press, 1998

Rejoinder

=

Rejoinder

RICHARD L. SMITH
University of North Carolina, USA

I appreciate very much the comments of the four discussants, as well as those of numerous other individuals who made verbal comments about the paper. The discussants represent a variety of viewpoints in their attitude to Bayesian and frequentist statistics, and in the extent to which they accept decision theory as an appropriate framework for the assessment of statistical procedures. I believe it will help to set both the paper and the discussion in context if I present some personal opinions about how these different viewpoints interact. This will be followed by some more direct responses to the various points made by the discussants.

Why do we need a new assessment of the theoretical properties of Bayesian procedures? It is clear that Bayesian methods have been outstandingly successful as practical application of statistics — the Valencia conferences themselves present ample evidence of that. Nevertheless, it seems to me that much of the recent expansion of the subject has been driven by computational advances, especially those using Markov chain Monte Carlo algorithms, rather than by fundamental advances in Bayesian thinking. This is especially true within the field of hierarchical models. However, the old arguments have not gone away. Indeed, it seems to me that with the application of Bayesian methods to ever more complex problems, it becomes imperative to look more closely at the extent to which seemingly arbitrary choices in the specification of a prior distribution affect the results of the analysis. I have absolutely no disagreement with those Bayesians, presumably including Professor D'Agostini, who would always use genuine prior knowledge to formulate their prior distributions. It is simply my own observation that the proportion of Bayesian papers that actually do this is rather small. Hence the need for alternative, more "objective" (whatever that means) criteria for comparing Bayesian procedures.

There are, nevertheless, different approaches that one could take to this task, so it becomes necessary to defend one based primarily on frequentist reasoning, and especially on decision theory. I do not believe there is anything controversial about the notion that a sequence of probability forecasts, produced through some statistical estimation procedure, should provide a good long-run representation of the actual outcome of those forecasts. Indeed, this idea lies at the heart of Phil Dawid's theory of prequential assessment, as represented for example by Dawid (1984) as well as Dawid and Seillier-Moiseiwitsch (1993). The real difficulty lies in the specification of an appropriate probability framework for this comparison. Traditionally, frequentists have used the probability framework of independent replications of some experiment under identical conditions. Critics have pointed out that such a framework hardly ever corresponds to actual practice, and that the same observed experiment may be embedded in some replicated sequence of experiments in many different ways, possibly leading to very different assessments of the frequentist properties of a given procedure. My personal viewpoint is that this is a legitimate caveat on the use of frequentist methods to compare different procedures, but it should not stop one making the comparison. Dawid's theory represents an ideal, but hard to implement, method of comparison, whereas by adopting the viewpoint of independent experiments, one obtains comparisons which, given enough computing power, can always be made by simulation. That is my justification for this approach: it is not an ideal, and certainly not the only approach that could be adopted, but it does provide a concrete framework for comparing different procedures, and that is something which I feel to be needed.

Even within the frequentist framework there is still a choice to be made between a decision-theoretic approach and a more "inferential" approach, of which the most explored method is to assess how well the observed coverage probabilities of interval estimation procedures match the nominal coverage probabilities used to generate the intervals, as in Barndorff-Nielsen and Cox (1996) and Professor Ghosh's discussion here. It seems to me that there is something appealing about the idea that one should look within the class of Bayesian procedures for one which achieves a small coverage probability bias. Such a procedure satisfies Professor Lindley's requirement for coherence, while at the same time achieving good frequentist properties. The difficulty, as I see it, is that it is not easy to find Bayesian procedures for which the $O(n^{-1})$ coverage probability bias is 0. Ghosh's discussion provides ammunition both for and against this point of view. On the one hand, I am delighted to see that some concepts previously developed only for confidence intervals have now been extended to prediction intervals — Trevor Sweeting, in his discussion of Andrew Barron's paper, mentioned some independent results which partly overlap with those discussed by Ghosh. On the other hand, the class of models for which one can actually solve the problem, as so formulated, still seems to be rather limited. I believe it is the exception, rather than the rule, that a single prior density π will solve Ghosh's equation (2) for every α. On the other hand, an alternative might be to try to solve (2) for a single α, and then to consider the limiting case $\alpha \to 0$. This might lead to an alternative point of view somewhere between Professor Ghosh's and my own.

The alternative, decision-theoretic, point of view has the advantage that it provides a concrete framework to perform explicit comparisons among competing procedures in reasonably complicated problems. I hope that those who disagree with this approach will nevertheless be stimulated to pursue other forms of comparison.

I applaud Professor Lindley's demonstration that the plug-in approach to the exponential prediction problem is incoherent from a Bayesian point of view. This confirms what I had suspected all along, that the true-blooded Bayesian would not even think about using the plug-in approach. The question this raises, for me, is to what extent the same result holds for other classes of prediction problems. To clarify my own viewpoint on this, I am not trying to advocate that in practice one should adopt the plug-in approach — this seems to be bad whether viewed through Bayesian or frequentist eyes. However the fact that there are some points of view by which the plug-in approach appears to be better than some natural classes of Bayesian procedures ought to act as a caution against the unthinking adoption of Bayesian procedures.

I turn finally to Professor Dey's discussion, which is decision-theoretic and which has led me to develop some extensions of the theory, which I will outline here and present more fully in the revision of Smith (1997). I consider in particular the LINEX loss, Dey's equation (3), which when $a^* < 0$ has the property that it penalizes underprediction of ϕ more heavily than overprediction by the same proportion.

For the exponential problem of Section 2, with gamma prior leading to the Bayes estimator $\hat\phi_B^L$, Dey's equation (7), we have

$$\mathrm{E}\{L(\phi,\hat\phi_B^L)\} = \frac{(a^*z\theta)^2}{2n} + \frac{1}{n^2}\left[\left\{\frac{5}{2}+2a+\frac{a^2}{2}-(3+a)b\theta+\frac{b^2\theta^2}{2}\right\}(a^*z\theta)^2+\frac{1}{3}(a^*z\theta)^3\right]. \tag{1}$$

The corresponding result for the MLE is

$$\mathrm{E}\{L(\phi,\hat\phi_{MLE})\} = \frac{(a^*z\theta)^2}{2n} + \frac{1}{n^2}\left\{\frac{5}{2}(a^*z\theta)^2-\frac{7}{6}(a^*z\theta)^3+\frac{1}{8}(a^*z\theta)^4\right\}. \tag{2}$$

The expression in (2) contains a term in $(a^*z\theta)^4$ which is absent from (1) and therefore confirms that in this case, in contrast to those discussed in the paper, $\hat\phi_B^L$ really does dominate $\hat\phi_{MLE}$ in the tail as $z\to\infty$! If one introduces a tail weighting ϕ^{-r} in this case (note that this is equivalent to replacing b by $b-rz$), then the optimal r is 0, and moreover, the same result holds in the general setting of (27), provided $G=1$. For $G\neq 1$, the optimal r is $3(G-1)$.

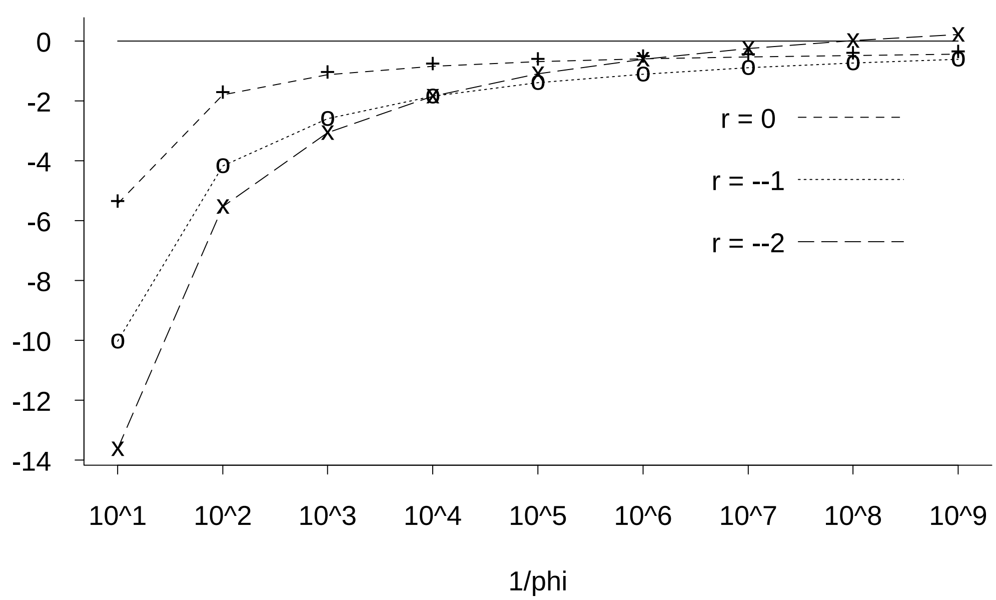

Figure 1. *Normalized values of $\mathcal{A}_L$ for the Poisson-GPD model (same parameters as in Figure 2 of the main paper) for LINEX loss with $a^*=-.4498$. Computed for the Bayes estimator with $r=0$ and the tail-weighted estimators with $r=-1$ and $r=-2$.*

As an example of the general case, we calculate the corresponding $\mathcal{A}_L$ function for the Poisson-GPD model, exactly as in Figure 2 of the paper. The Bayes estimator here solves the Bayes decision problem for LINEX loss under the prior density proportional to λ^{-1}, as in the paper. In this case we took $a^*=-.4498$, which corresponds to the property that an underprediction of the estimated probability by a factor of 10 is penalized exactly twice as much as an overprediction by a factor of 10. However, a very similar curve is obtained for other negative values of a^*. The precise function plotted here is $\mathcal{A}_L\phi^4/e_0^2$ where ϕ is the probability being estimated and e_0 is the asymptotic variance as $n\to\infty$ of $n\hat\phi_{\mathrm{MLE}}^{-a^*}$ — this form of normalization turns out to yield stable asymptotic results as

$\phi \downarrow 0$. Finally, we repeat the computations for tail-weighted estimators with $r = -1$ and $r = -2$: for this problem, it appears that tail weighted estimators with negative r are more competitive than those with positive r, and indeed it turns out that they beat the estimator with $r = 0$ for much of the range of ϕ, though ultimately (as $\phi \downarrow 0$) $r = 0$ is best. All three estimators convincingly beat the MLE over the range plotted except for $r = -2$ and $\phi < 10^{-8}$ – this is the only portion of Figure 1 for which $\mathcal{A}_L > 0$ for any of the estimators. To date, no attempt has been made to confirm these results by simulation. Nevertheless, as they stand they provide convincing asymptotic evidence that for this loss function, Bayesian estimators easily beat the plug-in approach when assessed from the point of view of frequentist risk.

One can also look more closely at (1). Within the term in curly brackets, the dependence on θ can be eliminated by setting $b = 0$, and then, the optimal a is -2, in other words, the Jeffreys prior $\pi(\theta) \propto \theta^{-1}$. As pointed out by Jim Berger, the same prior is optimal for squared logarithmic loss (15), since in this case it leads to the optimal invariant procedure, the equivalence of the invariance-prior Bayes and frequentist best invariant procedures in the case of invariant loss functions being a standard result of decision theory (Berger 1985, Chapter 6).

In conclusion, I thank all the discussants for their contributions, and the conference organisers for the opportunity to present this paper in Valencia.

REFERENCES

Barndorff-Nielsen, O.E. and Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli* **2**, 319–340.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis* (second edition). Springer, New York.

Dawid, A.P. (1984). Statistical theory: the prequential approach (with discussion). *J.R. Statist. Soc. A* **147**, 278–292.

Seillier-Moiseiwitsch, F. and Dawid, A.P. (1993). On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc.* **88**, 355–359.

Smith, R.L. (1997). Predictive inference, rare events and hierarchical models. Unpublished manuscript, University of North Carolina, Chapel Hill.