DISCUSSION OF "AN OVERVIEW OF ROBUST BAYESIAN ANALYSIS" BY JAMES BERGER SELF CONTAINED: ORIGINAL DISCUSSION NOT INCLUDED*

by

Anirban DasGupta
Purdue University

Technical Report #94-3

Department of Statistics
Purdue University

February 1994

* Research supported by NSF grant DMS–9307727

◇

Discussion of
"An Overview of Robust Bayesian Analysis"
by James Berger
Self Contained: Original Discussion Not Included*

As always, Professor Berger has made a contribution that is illuminating, informative, very enjoyable, and frequently provoking. Especially useful and gratifying is the extensive bibliography, a point that deserves mention. On my part, I will do two things: I will elaborate a little on a few points made by Professor Berger in his article, but I will devote practically all of my time and space to a number of points not explicitly made in the article.

Exactly what constitutes a study of Bayesian robustness is of course impossible to define. It seems, however, that Bayesians and others alike clearly appreciate the value and importance of a study of Bayesian robustness. A few years ago, after a talk given by Persi Diaconis, Herman Rubin stood up and said that all statisticians should work only on problems of Bayesian robustness. I feel less inclined to go that far, but the comment signifies the importance of research in this area. Importance and usefulness are sometimes completely different things, however. Ultimately the value of statistical work will be judged on the basis of whether people will use methods arising out of this work. It seems natural and actually nearly inevitable on hindsight that work on Bayesian robustness started out in the form of sensitivity analysis. In fact, for a while, that is practically all one saw. These were important on a number of grounds: they certainly helped clarify questions regarding when posterior robustness will usually obtain, occasionally they helped understand the role of the dimension of the data, and to me personally they shook us by our knees and showed that apparently abstract mathematics can provide wonderful tools in obtaining answers: moment theory, Choquet capacities, operator differentials, these have been enormously useful in Bayesian robustness problems. It is not clear, at least to me personally, and on this I think I differ with Professor Berger, that beyond that sensitivity studies have any transparent and concrete use. I have not found a really satisfactory answer in my mind to the question of what should one do with the range of a posterior mean. I know there

* Research supported by NSF grant DMS–9307727

are plausible answers: continuous refinement, or use of the range as a credible interval by itself, etc. Classical robustness grew into a successful and flourishing area because they were able to answer to at least a reasonable degree the question of what is robust. All of this research would have been most likely much less influential if the results only went as far as saying the sample mean is not a good thing to use unless you have pretty much normally distributed data. They were successful in providing alternatives that were apparently acceptable: the obvious energy that went into studies of M and L estimates is a testimony to that. Having said that, it is not at all clear what would be a criterion for prescription in our area. In my paper with Mei-mei Zen, I had shown that a posterior minimax choice, coincidentally but fortunately, results again in a Bayes procedure, Bayes with respect to one of the priors one started with. But undoubtedly, we will not see this phenomenon very often in other problems. Professor Berger's due concern about whether "robust choices" are not silly from a "real Bayes" perspective therefore has to be regarded with a lot of seriousness. In spite of that, more effort should probably go into this issue than has so far.

Another point that Professor Berger (implicitly) makes and one with which I fully agree is that it is now time for us to go beyond the canonical problems. These are also the problems that are the hardest to "solve". The frequentists have the blessing that the technology of large sample theory is now so advanced that even the nastiest problem with the dirtiest model is amenable to some structured theory in the form of limit theorems. The issue of finite samples aside, this is nice within the frequentist domain. Models that people really care about: all kinds of censoring, various regressions (Cox models, many more), and the now popular semiparametric models are just a few examples that really do need to be looked at. Will the choice of a link function matter? To what extent? These are entirely different sensitivity questions we can and should ask. It may very well be that no answers are possible: a consequence I will personally find very unfortunate. But we don't know that. There is also the well understood need of a simultaneous likelihood-prior-loss(?) robustness study. But I have my doubts that much structure will ever come out in this problem: we will be only successful in seeing what we knew we will see. I will have something more to say on this later. Let me now touch on a few things that are not explicitly addressed by Professor Berger, but do seem to be natural. Questions I ask below

certainly have dash of frequentism; I therefore caution pure Bayesians that parts of what I will now show can appear to be grotesque and strange. I will give some precise theorems, mostly without proofs, because of space and also because they will all appear elsewhere.

Robustness with respect to the likelihood: the role of dependence

One can make a very short case for this: real data are never really iid. Many questions suggest themselves: is the iid inference reasonably robust against moderate dependence, do noninformative or "robust" priors give some protection, etc. I will only talk about the first issue here, in the form of two results, one of which is rather surprising. For the rest of this part of the discussion, let us have the implicit understanding that we have data coming from a Gaussian process.

Theorem 1. Consider n observations coming from a weakly stationary Gaussian process with mean of each observation equal to t, and a covariance kernel given by $r(i,j) = r(|i-j|)$; let us pretend as though the covariances are arising from a continuous time function $r(x)$ (as is the case with how L estimates are defined, for instance). For estimating t using mean squared error, let $R(n)$ stand for the ratio of the Bayes risks of the iid case Bayes estimator under the true model and the iid model; assume a standard normal prior for t in this. Then $\lim R(n)$ (as n tends to infinity) exists under (frequently satisfied) conditions, and furthermore the limit equals $1 + 2.\int_0^\infty r(x)dx$.

Corollary 1. If data are coming from an Ornstein-Uhlenbeck process that we mistakenly think as iid, then we will suffer a Bayes risk 3 times as large as for the iid case even in the limit.

Proof: The covariance kernel is $r(x) = \exp(-|x|)$.

This is somewhat disappointing; even an exponentially decreasing covariance results in a loss 3 times in magnitude. With slowly varying covariance kernels (see Karamata or Feller) as is the commonly made assumption in the frequentist world, the loss will often be infinitely more! One can state a more general version of this result in terms of two general kernels, not restricting to the case when one of them is the iid case. This result, because it talks about Bayes risk, is half-frequentist. The next result is purely conditional.

Theorem 2. Consider a $100(1-\alpha)\%$ credible interval that is the correct Bayesian interval

if the data were iid. Consider the posterior probability of this interval when the process in reality is an AR(1) with parameter ϕ. Then, as n tends to infinity,

this posterior probability converges a.s. to $2.\Phi((1-\phi).z_{\alpha/2}) - 1$, it being understood that the underlying probability space uses the true marginal distribution as the measure.

Corollary 2. Mild autoregression is not much problem in this case if we have lots of data, but being close to the unit root case is disastrous.

My first version of Theorem 2 was corrected by N. D. Shyamalkumar, a graduate student at Purdue.

Group decisions: will they usually agree

Clearly this is frequentist; the apparently neat nature of the result following indicates the potental in very general problems.

Theorem 3. Consider estimating a univariate normal mean t by using a credible interval with n iid observations. Bayesian 1 has $N(0, \tau_1^2)$ and Bayesian 2 has $N(0, \tau_2^2)$ as prior for t. Denote by $C_1(X)$ the $100(1-\alpha)\%$ interval that Bayesian 1 will use if left alone. Let $P_2(X)$ denote the posterior probability of $C_1(X)$ if Bayesian 2 is forced to use $C_1(X)$ although it is not his Bayes solution. Then, for any β such that $1-\beta < 1-\alpha$,

$P_t\{P_2(X) < 1-\beta\}$ converges to zero for any t, as n tends to infinity, and in fact

$$\begin{aligned} &\lim\{n.\exp(n^2\gamma^2/2).P_t\{P_2(X) < 1-\beta\}\} \\ &= \sqrt{2/\pi}.1/\gamma, \\ \text{where } \gamma &= (\beta-\alpha)/\left\{\left(\frac{1}{\tau_1^2} - \frac{1}{\tau_2^2}\right)^2 z_{\alpha/2}\phi(z_{\alpha/2})\right\}. \end{aligned}$$

Some remarks are necessary because the statement can be baffling. The Theorem asks how often use of the other Bayesian's inference will lead to bad performance if we have lots of data. The convergence to zero is not surprising. What is surprising is the extraordinary fast convergence and even more the fact that under each t, the limit on the right side is the same. In other words, so far as pointwise limits are concerned, t vanishes altogether from the field in the long run. Is there any role of t at all? Yes indeed; the convergence is not uniform!!

Table 1 at the end gives some numbers for finite n. They clarify the finite case to some extent. I refrain from discussing.

Here is another result, which follows on use of Dini's theorem, but I am almost certain it follows from known results on the posterior CLT or even the Portmanteau theorem.

Theorem 4. In the set up of Theorem 3, let $P_1(X)$ and $P_2(X)$ denote the posterior probabilities of the common null hypothesis H_0: $t < 0$ under the two stated priors. Then the P_t joint distribution of $(P_1(X), P_2(X))$ converges weakly, as n goes to infinity, to a singular distribution supported on the main diagonal of the unit square.

Remark. Of course this is expected. The result can be stated in far greater generality; consequences of such results are that with a large probability, the two Bayesian's answers hang very close together. The case of k statisticians can be done with appropriate formulation.

The valuable cases are cases where the two Bayesians differ more seriously, like normal vs. t. I believe similar results are valid there and they will appear elsewhere.

Robustness with respect to outliers: Shrinking neighborhoods

Again, so far as concrete theory goes, the frequentists within their own domain are ahead on this. I will give only one result, purported to show that it is seemingly imperative to use shrinking neighborhoods, at least in this formulation.

Theorem 5. Consider data that are $N(t,1)100(1-\varepsilon)\%$ of the times and the rest of the times we see an outlier at some x. In principle, ε can depend on the sample size n. Consider estimating t using a $N(0,1)$ prior and mean squared error as criterion. Let $r(n)$ denote the Bayes risk of the estimate that would be Bayes in the absence of outliers ($\varepsilon = 0$). Then $r(n)$ is unbounded unless $\varepsilon = O(1/n)$; it converges to zero only if $\varepsilon = o(1/n)$. If ε is $O(k/n)$ (meaning exact order), then $r(n)$ converges to k^2x^2, and hence with the usual definition, the Influence of an outlier is unbounded.

Many other questions can be raised here. I would not go into them.

To sum it up, this is another profound contribution by Professor Berger to the profession. This made me think, helped me to understand. In chapter 14 of his book, Nicholas Young (the operator theorist) writes: "a mathematical model never describes the behavior

of a system exactly...How well will an aircraft stand up to unpredictable external disturbances — gusts of wind or a stewardess wheeling a drinks trolley down the aisle? One might wonder why the idea (of robust designs) was such a late starter. Part of the answer must be that engineers were unaware of the relevant theorems and operator theorists of the engineering problem. The connection is developing rapidly ...". Perhaps there is some interest in simply exploring as a matter of scientific truth whether classical and Bayesian robustness will lead to common grounds: can one justify use of M estimates from a robust Bayesian viewpoint? From a strictly likelihood principle point of view, evidently not. But perhaps from another viewpoint. It is my feeling that that can only be good for the community as a whole. I offer my deepest gratitude to Professor Berger for again doing what he always does: open new doors.

References

Feller, W.(1973). An Introduction to Probability Theory and Applications, Vol. 2, John Wiley, New York.

Young, N.(1988). Hilbert Space, Cambridge University Press.

Zen, M. M. and DasGupta, A.(1993), Estimating a Binomial Parameter: Is Robust Bayes Real Bayes, 11, 37–60.

$$1-\beta = 0.7 \quad 1-\alpha \;=\; 0.9$$

$$\tau_1^2 = 1 \qquad \tau_2^2 = 4$$

Table 1: $P_\theta(P_2(\bar{X}) < 1-\beta)$ for given n and θ

θ	$n=2$	$n=5$	$n=10$	$n=20$	$n=30$
0.5	0.00146033	8.1124×10^{-13}	0.	0.	0.
1.0	0.0116026	1.37753×10^{-9}	0.	0.	0.
1.5	0.0590335	6.91185×10^{-7}	0.	0.	0.
2.0	0.196045	0.000103836	0.	0.	0.
2.5	0.440885	0.00477846	6.43929×10^{-15}	0.	0.
3.0	1.0	0.0703178	4.52661×10^{-10}	0.	0.
3.5	1.0	0.361147	2.75817×10^{-6}	0.	0.
4.0	1.0	0.777161	0.00152334	0.	0.
4.5	1.0	0.969992	0.0835089	0.	0.
5.0	1.0	1.0	0.578982	2.22045×10^{-16}	0.
5.5	1.0	1.0	0.962497	2.44078×10^{-9}	0.
6.0	1.0	1.0	0.999612	0.00015012	0.
6.5	1.0	1.0	1.0	0.0839436	0.
7.0	1.0	1.0	1.0	0.80429	1.99618×10^{-12}
7.5	1.0	1.0	1.0	0.99901	0.0000134143
8.0	1.0	1.0	1.0	1.0	0.0721139
8.5	1.0	1.0	1.0	1.0	0.899443
9.0	1.0	1.0	1.0	1.0	0.999971
9.5	1.0	1.0	1.0	1.0	1.0