for submission of this comment has passed. Diaconis and Freedman have done us a service in exploring the consequences of apparently innocuous assumptions so carefully.

REFERENCES

JEFFREYS, H. (1967). *Theory of Probability*. Clarendon Press, Oxford.

LINDLEY, D. V. (1972). *Bayesian Statistics, A Review*. SIAM, Philadelphia.

LINDLEY, D. V., TVERSKY, A. and BROWN, R. V. (1972). On the reconciliation of probability assessments. *J. Roy. Statist. Soc. Ser. A* **142** 146–180.

LINDSAY, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London Ser. A* **296** 639–662.

2 PERITON LANE
MINEHEAD
SOMERSET TA24 8AQ
ENGLAND

REJOINDER

P. DIACONIS AND D. FREEDMAN

Stanford University and University of California, Berkeley

Introduction. We would like to thank the discussants for their careful work. For context, we summarize our position.

(a) As a team, our motives are mixed to an unusual degree, because we differ on many issues in foundations, including the interpretation of some of our results. However, we are unanimous that the mathematics in our paper should be of interest to Bayesians, ex-Bayesians, and never-Bayesians alike.

(b) Frequentists can use the Bayesian approach, like maximum likelihood or optimality, as a powerful heuristic engine for generating statistical procedures. No such engine is foolproof, so you should always look to see how well the procedure is going to do. Even the crustiest subjectivist ought to follow this advice, when the prior is only an approximation (and possibly quite a crude one, chosen for computational convenience) to the true subjective belief. Besides its practical importance, checking operating characteristics is good, clean mathematical fun.

(c) Pitfalls in the classical approach are well known; those in the Bayesian approach perhaps less so. We have given some examples where plausible applications of Bayesian technique lead to disaster. It is particularly easy to lose your way in high dimensional parameter space.

(d) We view consistency as a useful diagnostic test. If your procedure gives the wrong answer with unlimited data, probably you will not like it so well with a finite sample either.

(e) We show how putting conditions on the underlying model and modifying the prior can sometimes rescue Bayes procedures. As a general heuristic device for developing such patches, we propose a kind of Bayesian sensitivity analysis, the "what if" method.

To our dismay, the negative reactions came from *Krasker–Pratt*, and we answer them first.

Krasker–Pratt. *Krasker and Pratt* accuse us of false cheer with respect to classical procedures. For one of us (who is seldom accused of cheer with respect to anything) the change of pace is refreshing. They cite hierarchical regression models as a Bayesian success story, despite substantial evidence to the contrary (Freedman and Navidi, 1985). And they carefully avoid committing themselves on the question of whether consistency matters. As a final point of rhetoric, they take us to task for bench-testing a Bayesian solution to the simplest textbook problem of them all: estimating a location parameter with symmetric errors. Our view is conservative: If your procedure runs into trouble on that problem, you may have worse troubles on other problems.

Their main point, however, is that "all oddities can be attributed to the priors, not to more fundamental difficulties in Bayesian philosophy." Indeed, they continue, "in a discussion session of the Seminar on Bayesian Inference in Econometrics, many found fault with the Dirichlet, no one defended it beyond consistency, and much progress was reported with more satisfactory priors."

We do not know of any substantially different priors for use in nonparametric situations, meeting the minimum tests of consistency and computability. That is why the Dirichlet and its analogs were introduced; that is why people use them; and that is why we pitched our examples in those terms. If *Krasker and Pratt* know better, they should say explicitly what priors they like. Then somebody can do the asymptotics.

The main objection to the Dirichlet seems to be that it assigns full probability to the discrete distributions: The Dirichlet priors "totally ignore smoothing, which is really the main issue, where prior information counts the most." Well, *Krasker–Pratt*, lots of luck. The Dirichlet works on a divide-and-conquer strategy. You can divide the line up into a finite number of pieces, and forget what goes on inside each piece: This reduces an infinite-dimensional problem to a finite one. Then, you can repeat the process inside each piece.

Smoothing forces data in different parts of the line to interact, and this creates a whole new level of technical complexity. Indeed, this interaction across intervals is probably what differentiates the location problem from the problem of simply estimating an unknown distribution function, where the Dirichlet performs very well indeed—despite its marked preference for discrete distributions.

On the whole, the discreteness issue seems to us to be a red herring. Consider a Dirichlet with a normal rather than Cauchy base probability: This prior still concentrates on discrete distributions but the posterior is consistent. The problem is caused by multimodal densities and a base probability which is not log convex, rather than by discreteness.

At a more speculative level, we think there are counterexamples involving priors which have full support and concentrate on smooth distributions. However, detailed calculations are difficult, and we have not done them. The first idea is to smooth the Dirichlet, as proposed by Lo. Fix a kernel density k, choose F at random from the Dirichlet, and look at the convolution $k * F$. We will use $k * D(\alpha)$ to denote the law of $k * F$ when F has the law $D(\alpha)$. Such priors are necessarily inconsistent, because they do not have full support: Indeed, $\max k * F \leq \max k$.

The next idea is to consider mixtures. Fix a sequence k_n of kernel densities tending to point mass at 0, for example, normal with mean 0 and variance $1/n$. Consider nonnegative weights w_n adding to 1. Look at

$$\Sigma w_n k_n * D(\alpha_n).$$

This prior has full support, but concentrates on smooth densities. We think it is consistent for some choices of w_n and α_n, and inconsistent for others (if w_n tends to 0 sufficiently rapidly, and the mass of α_n tends to infinity). Furthermore, the prior may be consistent for estimating an unknown distribution function, but inconsistent in the location problem, where the unknown distribution function is just a nuisance parameter.

Here is another construction, which starts from the Dirichlet but forces the random distribution function to be absolutely continuous; compare Kraft's (1964) modification of Fabius (1964).

(i) Use a Dirichlet on the integers to distribute mass to the intervals $[n, n+1)$, but not within.
(ii) Within $[n, n+1)$, use a beta to randomly split the mass between the left half and right half; the beta may depend on n.
(iii) Keep on going, with different betas at different stages. Make the variances of the betas decay to zero very rapidly, so the random distribution function is nearly equal to its expectation; this can be, for example, either normal or Cauchy.

If the decay of the variances is rapid enough, almost all the random distributions will be smooth (absolutely continuous, maybe differentiable except at binary rationals). This prior will be consistent for estimating an unknown distribution function on the line, by the divide-and-conquer argument. Now consider the location problem; we guess this prior is consistent when its expectation is the normal, and inconsistent with the Cauchy.

The real mathematical issue, it seems to us, is to find computable Bayes procedures and figure out when they are consistent and when they are inconsistent. We wish *Krasker and Pratt* would use their considerable talents to help solve the problem, instead of burying it deeper in a pile of rhetoric.

Hartigan. Is *Hartigan* part of the problem or part of the solution? He seems to reject the idea that dimensionality of the parameter space matters. He is being uncharacteristically disingenuous, when he reproduces Freedman's (1963) original counterexample using a countable set of parameters. The only sensible way to think about those parameters is as a sequence in an infinite-dimensional space—the way it was set up originally. The question about entropy neighborhoods is answered (positively) in that same paper. For our part, we confess to not giving any careful, formal treatment of dimensionality. This leaves a clear field for *Hartigan*.

Berger. *Berger* asks, "How likely is it for one to encounter a consistency problem in practice?" In nonparametric problems, we think that inconsistency will be the rule not the exception, unless great care is taken in specifying the prior. Even in high-dimensional problems, details of the prior can have substantial and unanticipated effects on the behavior of the procedure.

Berger and others sometimes suggest that subjective judgements can and should be quantified as probability distributions. This is a cornerstone article of faith for some Bayesians. In certain problems, this kind of quantification is surely possible and helpful. However, in other problems, it may not be. Indeed, the attempt to develop a full-blown subjective probability distribution may be counterproductive, while informal use of intuition could help.

Berger (and *Krasker–Pratt* too) wonder whether inconsistency on a null set matters. We think it can. Man may be the measure of all things, but then one man's null set can be another's support. Freedman (1965) used category because it is neutral between measures, and because sets large in the sense of category are large for "most" measures.

Finally, *Berger* asks for more details on our claim that estimates of the form proposed by Box–Tiao, Fraser, and Johns can be inconsistent. The proposed estimators are all of the form

$$\hat{\theta} = \frac{\int \theta \Pi f\left[(X_i - \theta)/\sigma; \lambda\right] \nu(d\theta, d\sigma, d\lambda)}{\int \Pi f\left[(X_i - \theta)/\sigma; \lambda\right] \nu(d\theta, d\sigma, d\lambda)},$$

where $f(x; \lambda)$ is a family of densities and ν is formal prior. For example, Fraser takes the family of all t-densities: const$(1 + x^2)^\lambda$.

These estimates are Bayes rules based on non-log-convex densities. We have checked that for some choices of prior ν, the rules are inconsistent. We have not shown that the rules actually suggested are inconsistent, but we believe them to be.

Derivatives. *Le Cam*, *Krasker–Pratt*, and *Clayton* all make useful comments about derivatives. As they indicate, it is possible to calculate higher-order and even mixed partial derivatives with respect to the prior, the model, and the loss function. *Clayton* asks for a derivative of the predictive distribution with respect to the prior. We find this easiest to think about in the context of exchangeable processes. There, the predictive distribution is a linear function of the posterior, so the calculus is straightforward: If P_θ^∞ is the product measure at parameter θ, μ is the prior on θ, and $\dot{\mu}_x(d\theta)$ is the derivative of the posterior given x, the derivative of the predictive distribution with respect to the prior is the signed measure

$$\int P_\theta^\infty(\cdot)\dot{\mu}_x(d\theta).$$

Envoi. We will not comment further on the many other interesting points raised. We hope discussants not singled out for reply will be relieved rather than insulted, on the theory that no news is good news. Finally, we warmly thank the Editor, Associate Editor, and discussants for their encouragement and support.

REFERENCES

FREEDMAN, D. A. and NAVIDI, W. C. (1986). Regression models for adjusting the 1980 census. *Statist. Sci.* **1** 3–11.

KRAFT, C. H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probab.* **1** 385–388.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD CALIFORNIA 94305

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720