REFERENCES

BERAN, R. (1981). Efficient robust estimation in parametric models. *Z. Wahrsch. verw. Gebiete* **55** 91–108.

DIACONIS, P. (1985). Bayesian statistics as honest work. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer.* (L. M. Le Cam and R. A. Olshen, eds.) **1** 53–64. Wadsworth, Monterey, Calif.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–26.

DOSS, H. (1985a). Bayesian nonparametric estimation of the median. I: Computation of the estimates. *Ann. Statist.* **13** 1432–1444.

DOSS, H. (1985b). Bayesian nonparametric estimation of the median. II: Asymptotic properties of the estimates. *Ann. Statist.* **13** 1445–1464.

FERGUSON, T. S. and PHADIA, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7** 163–186.

HJORT, N. L. (1984a). Nonparametric Bayes estimators of cumulative intensities in models with censoring. Research Report, Norwegian Computing Center, Oslo.

HJORT, N. L. (1984b). Bayes estimators and asymptotic efficiency in parametric counting process models. Research Report, Norwegian Computing Center, Oslo.

HJORT, N. L. (1985a). Notes on the theory of statistical symbol recognition. Research Report, Norwegian Computing Center, Oslo.

HJORT, N. L. (1985b). Semi-parametric Bayes estimators. Unpublished.

HJORT, N. L. (1985c). Contribution to the discussion of Andersen and Borgan's "Counting process models for life history data: A review." *Scand. J. Statist.* **12** 141–150.

MILLAR, P. W. (1981). Robust estimation via minimum distance methods. *Z. Wahrsch. verw. Gebiete* **55** 73–89.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequentist calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.

NORWEGIAN COMPUTING CENTER
P. B. 335 BLINDERN
N-0314 OSLO 3
NORWAY

WILLIAM S. KRASKER AND JOHN W. PRATT

Harvard University

This is one in a series of fascinating papers. They are easily read as painting a picture of modern Bayesianism in bad trouble but frequentism in fine shape. A larger historical perspective gives a different view, however. It should therefore be sketched, even if only impressionistically.

Time began in darkness and "inverse" probability. Then the Rev. Thomas Bayes let in some posthumous light. The postulate he identified and used was duly found arbitrary or ambiguous, and unfounded. Likewise Fisher's reference sets. Then Neyman and Pearson developed "objective" (frequentist) concepts even as Ramsey and de Finetti were proving that "subjective" Bayesianism was the only coherent theory possible. Soon (well before Pratt's 1961 and 1965 surveys) objective methods too were found arbitrary and theoretically and practically deficient even in the simplest situations (where uniformly most powerful tests are randomized).

But the new, true Bayesian road, though well lit, is not always smooth. Prior distributions are hard to assess, practically impossible sometimes. "Reference" priors, though convenient and objective$_2$ (subject-independent), sometimes commit theoretical and even practical improprieties, some of which—we admit it—frequentism identifies. Bayesian methodology (not theory) has progressed least in many-parameter inference. With hierarchical structures, it has done well, extending and clarifying the frequentist models and analysis of random effects. But with nonparametric structures, including ordinary sample surveys, it has done poorly.

Suppose, for simplicity, that $x_1, x_2, \dots$ are iid with continuous distribution F, possibly multivariate, and are observed infinitely accurately. A nonparametric Bayesian might expect his true prior and want an approximate prior to have these properties, among others, at least for "well-behaved" samples.

(1) $P(F$ is continuous$) = 1$.
(2) $P(\hat{F}_n$ is continuous$) = 1$, where $\hat{F}_n$ is the expectation of the posterior distribution of F, the predictive distribution of x_{n+1} given $x_1, \dots, x_n$.
(3) $\hat{F}_n - F_n$ is of order $1/n$ in a suitable sense, where F_n is the empirical distribution.
(4) A posteriori $F(x) \sim N(F_n(x), F_n(x)[1 - F_n(x)]/n)$ to order $1/n$.
(5) In intervals of order $1/n$, the information in the sample spacing is dominated by prior expectation of smoothness.

The reason to expect (3) and (4) is that, even in finite-parameter problems, a positive prior density affects the posterior distribution only by order $1/n$, and the likelihood dominates the prior in determining the whole posterior distribution, not merely the location at which it concentrates. That is, the sample information dominates the prior information in intervals of order $1/n^{1/2}$.

The big question is what constitutes "well-behaved" samples. Are they more than a set of probability 1 under the true prior? Less?

Diaconis–Freedman consistency demands less than (3), but demands it for almost all samples from every distribution F. This may be an important property of a Bayes rule if that rule is regarded as just another way to get a point estimate of the true parameter. However, a Bayesian wants not an estimate but the posterior distribution of the parameter given the data. Doob's theorem says that the posterior will be consistent for almost all parameter values. But consistency everywhere is neither necessary nor sufficient for a prior to be a good representation of prior beliefs.

Dirichlet priors have properties (3) and (4) (by the beta posterior of $F(x)$), but not (1), (2), or (5). Maximum possible independence is their beauty—they are manageable and consistent—but also their curse: They totally ignore smoothing, which is really the main issue, where prior information counts most. Their unsatisfactoriness is most telling in small samples (where failure to smooth matters most) but most provable in large samples (e.g., $P(x_n$ is new$) \le \|\alpha\|/(\|\alpha\| + n - 1) \to 0$ implies $P(F$ is discrete$) = 1$, which is undesirable and an easier proof than we have seen in the literature).

Now the Dirichlet happens to be consistent. Indeed, the case α constant, $\|\alpha\| = 0$ (suggested to us by Zellner), is as improper as can be yet gives the empirical distribution, while α constant, $\|\alpha\| = 1$ gives Fisher's predictive distribution (equal probability in each of the $n + 1$ intervals defined by the order statistics). But in a May 17 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. So we do not believe that bad properties of "Jeffreys-style" or symmetrized extensions of the Dirichlet should faze "practicing Bayesians" as Diaconis and Freedman imply. As far as we know, all oddities can be attributed to the priors, not to more fundamental difficulties in Bayesian philosophy. If they could not, the identification of the unidentifiable would bother us more than the inconsistency of the symmetrized Dirichlet, because guaranteed symmetry seems even more unreal to us than a priori independence of location and shape. But both seem dangerous as philosophical testing grounds.

Returning to one small part of the "big question": How much consistency should we expect or require? We might expect consistency for all absolutely continuous distributions (densities), or all lattice distributions, but we would not require it beyond what Doob guarantees for the true prior, especially now that our eyes have been opened to how much this would imply. We would no longer be surprised, let alone dismayed, by inconsistency for the kinds of samples one would get for F continuous but singular with respect to Lebesgue measure. We would be happy to restrict attention to a topologically small family of prior distributions, such as those assigning probability 1 to densities. Whether F is part of an objective probability model, or only in the mind of the beholder, the "classical"–"subjectivist" distinction of Diaconis and Freedman, seems unimportant to us, and irrelevant here, and we would look for merging of opinion to the same point as a consequence of whatever consistency is present, and merging in the full sense as a consequence of (4), not vice versa.

The results on the sensitivity of the posterior to the prior (the last part of Section 3, and Appendix B) represent an interesting approach to the problem of choosing a convenient prior $\hat{P}$ that approximates the "true" prior P, in such a way that, given the sample x, the posteriors $\hat{P}_x$ and P_x are close (compare Krasker (1984)). Under the conditions of Theorem 4, the norm of the derivative of the map T from priors to posteriors is the ratio of maximum to mean likelihood. However, T is not even continuous without the somewhat artificial assumption that $f(x|\theta)$ is bounded in θ. In addition, the results about the derivative $\dot{T}_P$ use the total-variation norm on both the priors and posteriors. This is an overly strong topology for the space of priors if the parameter indexes the set of continuous distributions on R, since the computationally feasible methods of approximating the true prior—say by a finite-dimensional parametric model, or even an extended Dirichlet process—assign probability 1 to a set that has true probability 0. As Diaconis and Freedman point out at the end of Appendix B, the results can be extended to the weak-star topology, say using the Prohorov metric d. (This requires a metric on Θ; the natural way to provide one is to identify each θ with the distribution it indexes, and use either the Prohorov or total-variation metric.) Frechet differentiability in this context requires the further assumption that $f(x|\theta)$ satisfy a Lipschitz condition in θ. (Continuity does not appear to be enough to give the necessary property that $d(f\,dP, f\,dQ) = O(d(P, Q))$.) This

Lipschitz condition in θ (or even continuity, for that matter) is an additional severe restriction on the set of densities $\{f(\cdot|\theta)\}$, requiring for example that they satisfy a Lipschitz condition in x, uniformly in θ. The norm of the Frechet derivative—the maximum of the ratio of change in the posterior to change in the prior, as the latter goes to zero—agrees with the formula $\|\dot{T}_P\| = f(x|\theta_{\mathrm{ML}})/\int f\,dP$ derived using the total variation norm provided $P(N_\varepsilon(\theta)) = o(\varepsilon)$, where $N_\varepsilon(\theta)$ is the ε-neighborhood of θ in Θ. (This condition should hold if Θ is more than one-dimensional.) This ostensibly shows, in situations in which the assumptions for differentiability hold, that in order to ensure $d(\hat{P}_x, P_x) \le \varepsilon$, we should select $\hat{P}$ satisfying $d(\hat{P}, P) \le \varepsilon/\|\dot{T}_P\|$. However, it is easy to show that $\|\dot{T}_P\| \to \infty$ as the sample size goes to infinity. This says in particular that in large samples the condition on $\hat{P}$ will be virtually impossible to satisfy, and says more generally that, contrary to intuition, it is in large samples that the posterior is most sensitive to the prior. We can get further insight into the local behavior of T by examining the second derivative $\ddot{T}_P$, which can be regarded as a symmetric bilinear map priors × priors → posteriors, and which will exist under the assumptions used to obtain $\dot{T}$. We find that

$$\ddot{T}_P(H,G) = -\left(\int f\,dG\Big/\left(\int f\,dP\right)^2\right) f\,dH - \left(\int f\,dH\Big/\left(\int f\,dP\right)^2\right) f\,dG$$
$$+2\left(\int f\,dH \int f\,dG\Big/\left(\int f\,dP\right)^3\right) f\,dP.$$

In particular, in the second-order expansion $T(P+H) - T(P) = \dot{T}_P(H) + \frac{1}{2}\ddot{T}_P(H,H)$, the second-order term can be important unless $\|\dot{T}_P(H)\| \ll 1$. If $\|\dot{T}_P\|$ is large, as it will be when the sample is large, the first derivative will yield a good approximation to T only too close to P to be of use.

What, in jargon natural at our institution, is the bottom line? As far as we can see, it is that satisfactory prior distributions for nonparametric problems are still unavailable and that it is naive to expect too much in certain directions. This completes our discussion and Bayesian defense against frequentist analysis. If we have referred to ourselves unseemly often, it may signify that the foundations of statistics are personal. If we have seemed unseemly to the authors, be assured that we would have much pleasure in seconding a vote of thanks to them.

REFERENCES

KRASKER, W. S. (1984). A note on selecting parametric models in Bayesian inference. *Ann. Statist.* **12** 751–757.

PRATT, J. W. (1961). Review of *Testing Statistical Hypotheses* by E. L. Lehmann, *J. Amer. Statist. Assoc.* **56** 163–167.

PRATT, J. W. (1965). Bayesian interpretation of standard inference statements. *J. Roy. Statist. Soc. Ser. B* **27** 169–203.

GRADUATE SCHOOL OF BUSINESS ADMINISTRATION
HARVARD UNIVERSITY
BOSTON, MASSACHUSETTS 02163