Discussion: On the Consistency of Bayes Estimates

Andrew R. Barron

The Annals of Statistics, Vol. 14, No. 1 (Mar., 1986), 26-30.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28198603%2914%3A1%3C26%3ADOTCOB%3E2.0.CO%3B2-L

The Annals of Statistics is currently published by Institute of Mathematical Statistics.

I, II. University Press, Cambridge.
LINDLEY, D. V. (1972). *Bayesian Statistics—A Review.* SIAM, Philadelphia.
LOCKETT, J. L. (1971). *Convergence in Total Variation of Predictive Distributions: Finite Horizon.* Unpublished Ph.D. dissertation, Dept. Statist., Stanford Univ.
MATTHES, T. K. and TRUAX, D. R. (1967). Tests of composite hypotheses for the multivariate exponential family. *Ann. Math. Statist.* **38**, 681–697.
NEYMAN, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36** 97–129.
NOORBALOOCHI, S. and MEEDEN, G. (1983). Unbiasedness as the dual of being Bayes. *J. Amer. Statist. Assoc.* **78** 619–623.
PRATT, J. (1965). Bayesian interpretation of standard inference statements. *J. Roy. Statist. Soc. B* **27** 169–203.
SACKS, J. (1963). Generalized Bayes solutions in estimation problems. *Ann. Math. Statist.* **34** 751–768.
SAVAGE, L. J. (1971). Elicitations of personal probability and expections. *J. Amer. Statist. Assoc.* **66** 783–801.
SAVAGE, L. J. (1972). *The Foundations of Statistics.* Dover, New York.
SCHWARTZ, L. (1965). On Bayes' procedures. *Z. Wahrsch. verw. Gebiete* **4** 10–26.
STEIN, C. S. (1955). A necessary and sufficient condition for admissibility. *Ann. Math. Statist.* **26** 518–522.
STEIN, C. S. (1981). On the coverage probability of confidence sets based on a prior distribution. Technical Report 180, Dept. Statist., Stanford Univ.
TJUR, T. (1980). *Probability Based On Randon Measures.* Wiley, New York.
VON MISES, R. (1964). *Mathematical Theory of Probability and Statistics.* (H. Geiringer, ed.). Academic, New York.
WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.
WELCH, B. L. and PEERS, H. W. (1963). On formulas for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. B.* **25** 318–329.
ZABELL, S. (1979). Continuous versions of regular conditional distributions. *Ann. Probab.* **7** 159–165.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DISCUSSION

ANDREW R. BARRON[1]

University of Illinois, Urbana

1. General remarks. Diaconis and Freedman have demonstrated some advantages and pitfalls of Bayesian inference. In summary, their results include the inconsistency of location estimates based on a Dirichlet prior; the equivalence of weak consistency and weak merging of posteriors; and an analysis of the sensitivity of the posterior to changes in the prior. In this discussion, we provide additional insight and point toward new developments. It is argued that the Dirichlet is a poor choice of prior because the Dirichlet mixture has a likelihood which is exponentially smaller than every product likelihood. We give conditions

[1]Work supported in part by NSF Grant ECS 82-11568 at Stanford University.

on the choice of the prior such that the mixture likelihood is close to virtually any product likelihood. Implications for Bayes consistency are discussed.

Consider the general problem of Bayes estimation on the line. We denote the distributions on the line by F, the product distributions on $\mathbf{R}^n$ by F^n, and the prior on distributions by μ. The mixture of product distributions is $G^{(n)} = \int F^n \mu(dF)$. (According to a Bayesian with prior μ, this $G^{(n)}$ is the distribution of the data.) Following the "what-if" principle of Diaconis and Freedman, we assume that $X_1, X_2, \ldots$ are independently drawn from a distribution F_* with a probability density function $f_*(x)$. The joint density $f_*(X_1, X_2, \ldots, X_n) = \prod_{i=1}^n f_*(X_i)$ (evaluated at the data) is called the product likelihood. Let $g(X_1, X_2, \ldots, X_n)$ be the density of the absolutely continuous component of $G^{(n)}$. Ideally, we want the mixture likelihood $g(X_1, X_2, \ldots, X_n)$ to be close to the product likelihood $f_*(X_1, X_2, \ldots, X_n)$ with high F_*^n probability, for virtually any such distribution F_*.

2. The location problem. Diaconis and Freedman consider the Bayes estimation of a location parameter θ from data $X_i = \theta + \varepsilon_i$ where the ε_i are independently drawn from a distribution F. If the prior μ on distributions F is taken to be independent of θ, then the posterior distribution of θ given $X_1, X_2, \ldots, X_n$ depends on the prior μ only through the mixture distribution $G^{(n)} = \int F^n \mu(dF)$. For instance, if $G^{(n)}$ is absolutely continuous on a set A which is invariant under translations of each coordinate by θ, then the mean of the posterior is given by $\hat{\theta} = \int \theta p(\theta) g(X_1 - \theta, \ldots, X_n - \theta)\, d\theta / \int p(\theta) g(X_1 - \theta, \ldots, X_n - \theta)\, d\theta$ for $X_1, \ldots, X_n$ in A, where $p(\theta)$ is the prior density of locations.

Is the mean of the posterior consistent? As a degenerate example, suppose the prior μ is point mass at a distribution F with density f. If this prior guess is exactly right, $F = F_*$, then the posterior mean is consistent for almost every θ (and consistent for every θ if the density f_* is smooth; see Schwartz, 1965). Whereas if the prior guess is wrong, $F \neq F_*$, then Diaconis and Freedman (1986) show that the posterior mean may be inconsistent. (Surprisingly, $F \neq F_*$ does not necessarily imply inconsistent posterior means. Diaconis and Freedman state that the location estimate is consistent for essentially any F_* if F has a log-concave density.) A naive reaction to this degenerate case is, "Why consider point mass at a single distribution F, when there are priors like the Dirichlet for which all distributions are in the (weak-star) support set?"

If μ is a Dirichlet prior with absolutely continuous base measure α (having standardized density $f = \alpha'/\|\alpha\|$), then the Korwar–Hollander result (which Diaconis and Freedman (1986) uses) establishes that the distribution $G^{(n)}$ is absolutely continuous on the set of sequences with distinct X_i, and the likelihood $g(X_1, X_2, \ldots, X_n)$ is proportional to $\prod_{i=1}^n f(X_i)$. (The set with distinct X_i has probability 1 with respect to any continuous distribution F_*^n, but exponentially small probability with respect to $G^{(n)}$.) Consequently, for the location problem, the Dirichlet prior yields exactly the same estimator as the degenerate prior which places point mass at a single F! Because of this degeneracy, the Dirichlet prior is useless machinery for the location problem. We would prefer to use a

prior which (via the mixture) can simultaneously mimic a larger class of iid distributions.

3. Matching likelihoods. We need a useful definition of closeness of likelihoods. Since joint densities tend to grow (or shrink) exponentially (Barron, 1985a), the following definition is suggested as a natural property. A sequence of likelihoods $g(X_1, X_2, \ldots, X_n)$ is said to *match* likelihoods $f_*(X_1, X_2, \ldots, X_n)$ if for any $\varepsilon > 0$,

$$e^{-n\varepsilon}g(X_1, X_2, \ldots, X_n) \leq f_*(X_1, X_2, \ldots, X_n) \leq e^{n\varepsilon}g(X_1, X_2, \ldots, X_n),$$

for all n sufficiently large, with F_*^∞ probability 1. (Equivalently, $1/n$ times the log-likelihood ratio tends to zero.) Matching may be thought of as a strong merging property of mixtures. The first inequality in the definition holds without conditions (by application of Markov's inequality and the Borel–Cantelli lemma). The second inequality holds only for well designed mixtures.

What conditions on the prior μ are sufficient for matching? Let $\|F_* - F\|$ denote the total variation distance and let $D(F_*\|F) = E \log dF_*/dF$ denote the relative entropy (Kullback–Leibler divergence). Either of the following conditions is sufficient for the mixture likelihoods $g(X_1, X_2, \ldots, X_n)$ to match the product likelihoods $f_*(X_1, X_2, \ldots, X_n)$:

(a) The prior assigns strictly positive mass to the relative entropy sets:

$$\mu\{F\colon D(F_*\|F) < \varepsilon\} > 0, \quad \text{for all } \varepsilon > 0,$$

or

(b) The prior assigns non-negligible mass to the variation distance neighborhoods in the sense that there exists ε_n with $\Sigma\varepsilon_n < \infty$ such that

$$\mu\{F\colon n\|F_* - F\| < \varepsilon_n\} \geq e^{-o(n)}.$$

Moreover, (a) and (b) each imply a local matching property. Let $g(X_1, X_2, \ldots, X_n|N)$ be the density of the absolutely continuous component of the conditional distribution $G^{(n)}(\cdot|N) = \int_N F^n(\cdot)\mu(dF)/\mu(N)$. The prior is said to locally match point mass at F_* (weakly/strongly) if for all (weak-star/variation distance) neighborhoods N the likelihoods $g(X_1, X_2, \ldots, X_n|N)$ match $f_*(X_1, X_2, \ldots, X_n)$. Note that local matching implies matching. The proofs that (a) and (b) each imply strong local matching are implicit in Schwartz (1965).

4. Bayes consistency. What are the implications of matching for consistency? Local matching implies weak consistency of the posterior, but it also implies more. Let's define weak, strong, and intermediate forms of consistency. Let $\|F_* - F\|_\pi = \Sigma_{A\in\pi}|F_*(A) - F(A)|$ be the variation distance on a partition π of the line. The total variation distance is $\|F_* - F\| = \sup_\pi \|F_* - F\|_\pi$. Sets of the form $N_\pi = \{F\colon \|F_* - F\|_\pi < \varepsilon\}$ and $N = \{F\colon \|F_* - F\| < \varepsilon\}$ are, respectively, weak-star and total variation neighborhoods of F_*. A sequence of posteriors $\mu_n = \mu(\cdot|X_1, X_2, \ldots, X_n)$ is strongly consistent for F_* if the posterior mass of

variation distance neighborhoods tends to one, $\mu_n(N) \to 1$, F_*^∞-almost surely, for any $\varepsilon > 0$, and weakly consistent if $\mu_n(N_\pi) \to 1$ a.s. for any finite partition π consisting of sets with boundary measure zero. Now for the intermediate definition: A sequence of posteriors μ_n is said to be consistent for F_* in w_n-variation if $\mu_n(N_{\pi_n}) \to 1$ a.s., where π_n is a countable partition of the line into intervals of width w_n. We require that the widths w_n tend to zero.

Why should we care about intermediate consistency? It is shown in Barron (1985b, 1986) that for any prior μ, local matching implies weak consistency and consistency in w_n-variation if $\lim nw_n > 0$ (e.g., $w_n = 1/n$). Conversely, if $\lim nw_n = 0$ then there exists a prior μ which locally matches F_* (and even satisfies property (a)), but the posterior is inconsistent in w_n-variation, $\lim \mu_n(N_{\pi_n}) = 0$. Thus w_n-consistency with w_n proportional to $1/n$ is the strongest possible consistency obtainable from the sole assumption of local matching. A consequence of this result is that Bayes estimates of the distribution need only be smoothed over intervals of width $1/n$ to obtain strongly consistent density estimates, whereas for ordinary histograms and kernel density estimates the smoothing must extend over widths w_n satisfying $nw_n \to \infty$. The proof of n^{-1}-consistency is based on finding a sequence of tests of the hypotheses $F = F_*$ versus the composite hypothesis $F \notin N_{\pi_n}$ such that the probability of error is uniformly exponentially small over all $F \notin N_{\pi_n}$. Such a test is relevant because Schwartz (1965) shows that local matching plus the existence of uniformly consistent tests implies consistency. The test statistic is essentially a weighted count of the number of empty cells, $\sum_{A \in \pi_n}(e^{nF_*(A)}I_{\{A \text{ empty}\}} - 1)$.

A natural class of priors on densities is obtained by convolving the distributions drawn from the Dirichlet with a kernel of random width. These priors have been examined by Lo (1984). Here are some open questions. Are these priors strongly consistent? Do they match a large class of distributions F_*?

Another source of strongly consistent priors are those priors which assign mass to a countable set of distributions. For instance, the prior might assign mass to every histogram on dyadic intervals with rational heights. Or the prior might assign mass to every computable distribution. In either case the prior satisfies the relative entropy condition and hence it is strongly local matching for any F_* with bounded density on compact support. If a prior is root summable, $\sum_F \mu^\alpha(F) < \infty$ for some $0 < \alpha < 1$, and strongly local matching at F_*, then the posterior is strongly consistent for F_*. [See Barron (1985b, 1986)].

5. On mixtures and consistency. A useful device for incorporating a variety of possible prior beliefs is to take a countable mixture of priors. Fortunately, the local matching property and hence the consistency is preserved by countable mixtures. If at least one of the priors locally matches F_*, then so does the mixture of the priors.

In an earlier paper, Freedman and Diaconis (1983) showed that mixtures involving Dirichlet priors may be inconsistent. In particular, for probability mass functions on the positive integers, they considered the mixture of a Dirichlet prior (with uniform "stick-breaking") and a point mass at a probability mass

function $\phi(i)$ proportional to $1/i(\log i)^2$. The true probability mass function is taken to be θ^* which differs from ϕ for small i and is equal to ϕ for all large i. The posterior has the unfortunate property of concentrating at ϕ rather than in neighborhoods of θ^*. From this inconsistency, we conclude that the Dirichlet prior does not locally match θ^*. Moreover, the Dirichlet prior assigns zero mass to the relative entropy neighborhood $\{\theta: \Sigma_i \theta^*(i) \log \theta^*(i)/\theta(i) < \varepsilon\}$ for ε sufficiently small.

Freedman and Diaconis have pointed out that ϕ and θ^* have infinite entropy $H(\theta^*) = \Sigma_i \theta^*(i) \log 1/\theta^*(i)$. One might think that the inconsistency is a result of the infinite entropy; however, even if certain finite entropy mass functions are used in the construction, inconsistency will still result. It is enough that θ^* and ϕ have tails proportional to $1/i^\alpha$ where $1 < \alpha < \frac{4}{3}$. (The verification of inconsistency closely parallels Sections 2 and 3 of Freedman and Diaconis, 1983). In Freedman (1963), finite entropy appears as part of a condition for consistency. We now know that the finite entropy assumption is extraneous. It is the *relative* entropy that matters for Bayes consistency.

In summary we have discussed some inadequacies of the Dirichlet prior as revealed by the analysis of Diaconis and Freedman and we have pointed toward stronger consistency and merging results obtainable for other priors.

REFERENCES

BARRON, A. R. (1985a). The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem. *Ann. Probab.* **13** 1292–1303.

BARRON, A. R. (1985b). Logically smooth density estimation. Ph.D. Thesis, Department of Electrical Engineering, Stanford University.

BARRON, A. R. (1986). On uniformly consistent tests and Bayes consistency. Preprint.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12** 351–357.

The references to Diaconis, Freedman, and Schwartz are the same as in the paper under discussion.

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS
1409 W. GREEN STREET
URBANA, ILLINOIS 61801

JAMES BERGER

Purdue University

The very lucid paper of Diaconis and Freedman is full of stimulating ideas and discussion. The ideas fall roughly into three categories: (i) inconsistency of Bayes rule, (ii) frequentist-Bayesian interrelationships including the "what if" method, and (iii) new Bayesian devices and techniques. My comments will be grouped by these categories, and will be restricted (because of space considerations) solely to a Bayesian view of the situation.