Thomas M. Cover B. Gopinath
Editors

Open Problems in
Communication
and Computation

3.20 ARE BAYES RULES CONSISTENT IN INFORMATION?

Andrew R. Barron

Department of Statistics
University of Illinois
Champaign, IL 61820

Bayes' rule provides a method for constructing estimators of probability density functions in both parametric and nonparametric cases. Let $X_1, X_2, \ldots, X_n$ be a random sample from an unknown probability measure P_0 with density function $p_0(x)$ with respect to a dominating measure $\lambda(dx)$. Let μ be a prior probability measure on the space of all probability measures P which have densities $p(x) = dP/d\lambda$. Then the mean of the posterior yields the following estimator of the density function

$$\hat{p}_n(x) = \hat{p}(x; X_1, X_2, \ldots, X_n) = \frac{\int p(x)(\prod_{i=1}^n p(X_i))d\mu}{\int(\prod_{i=1}^n p(X_i))d\mu}.$$

To obtain a consistency result, it is natural to require that the prior assigns positive probability to neighborhoods of the true distribution. In particular, we suppose

$$\mu\{ P : D(P_0 \| P) < \varepsilon \} > 0, \quad \text{for all } \varepsilon > 0 . \tag{1}$$

Here $D(P_0 \| P) = \int p_0(x) \log (p_0(x)/p(x)) \lambda(dx)$ is the informational divergence (also called relative entropy or Kullback-Leibler number).

1. The Problem.

Determine whether the sequence of Bayes estimators $\hat{p}_n$ converges to the true density p_0 in the sense that

$$\lim_{n\to\infty} E\, D(P_0 \| \hat{P}_n) = 0.$$

Here the expectation is with respect to P_0. It is also of interest to know

whether

$$\lim_{n\to\infty} D(P_0 \,\|\, \hat{P}_n) = 0\,, \quad P_0 \text{ almost surely.}$$

Either result would imply that the sequence of random variables $D(P_0 \,\|\, \hat{P}_n)$ converges to zero in probability.

Remark: An inequality between the information and the L^1 distance ($D(P_0 \,\|\, \hat{P}_n) \geq (1/2)(\int |\, p_0 - \hat{p}_n \,|)^2$; see [1]) shows that convergence in information implies convergence of the density estimator in the L^1 sense

$$\lim_{n\to\infty} E \int |\, p_0(x) - \hat{p}_n(x) \,|\, \lambda(dx) = 0.$$

2. Evidence for Consistency.

Does $E\, D(P_0 \,\|\, \hat{P}_n)$ tend to zero? We argue that the answer is yes along a subsequence, yes in the Cesaro sense, and yes if the posterior mean is replaced by a sample average of posterior means.

Lemma 1: *If condition (1) is satisfied then*

$$\liminf_{n\to\infty} E\, D(P_0 \,\|\, \hat{P}_n) = 0\,;$$

also

$$\liminf_{n\to\infty} D(P_0 \,\|\, \hat{P}_n) = 0\,, \quad P_0 \text{ almost surely.}$$

Moreover,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} E\, D(P_0 \,\|\, \hat{P}_k) = 0.$$

Lemma 2: *Let $\tilde{p}_n$ be an average of posterior means, that is,*

$$\tilde{p}_n(x;\, X^n) = \frac{1}{n} \sum_{k=1}^{n} \hat{p}_k(x;\, X^k)$$

where $X^k = (X_1, \ldots, X_k)$. If condition (1) is satisfied then

$$\lim_{n\to\infty} E\, D(P_0 \,\|\, \tilde{P}_n) = 0.$$

Thus the average $\tilde{p}_n = (1/n)\, \Sigma_{k=1}^{n}\, \hat{p}_k$ smooths out any humps of large D

that might lead to inconsistency. It is interesting to note that convergence still holds if the kth term in the definition of $\tilde{p}_n$ is replaced by $\hat{p}_k(\cdot\,; X^{n,k})$, where $X^{n,k}$ is any subset of the n observations of size k.

We note that the posterior mean density is the best possible estimator from the point of view of the Bayes risk (with loss function given by the informational divergence). Thus if any estimator exists which is Bayes risk consistent, then the posterior mean is Bayes risk consistent.

Lemma 3: *Among all probability density estimators based on the data X^n, the posterior mean density estimator $\hat{p}_n(x; X^n)$ minimizes the Bayes risk*

$$R_n = \int E_P\, D(P \,\|\, \hat{P}_n)\, d\mu.$$

Moreover, the Bayes risk R_n is a decreasing sequence. Thus

$$\lim_{n\to\infty} R_n \text{ exists.}$$

It is not known if this limit is zero. Although the average risk is decreasing, the risk $E_P\, D(P \,\|\, \hat{P}_n)$ might *increase* for some P and some n. If we could ensure that $E_{P_0}\, D(P_0 \,\|\, \hat{P}_n)$ were decreasing, then by Lemma 1 we would have $\lim E_{P_0}\, D(P_0 \,\|\, \hat{P}_n) = 0$.

Doob [2] used martingale arguments to establish Bayes consistency results. The drawback is that the results only show convergence for distributions in a set of prior measure one, and there is no known method for determining whether a given distribution is in this set. Nevertheless, the following result is readily obtained.

Lemma 4: *Except for a set of distributions P which has μ measure zero, if condition (1) is satisfied for P then*

$$\lim_{n\to\infty} D(P \,\|\, \hat{P}_n) = 0\,, \quad P \text{ almost surely.}$$

The following result is proved in Barron [3] using the technique of Schwartz [4]. It was first obtained by Freedman [5] in the discrete case (under the extra condition of finite entropy $H(P_0)$).

Lemma 5: *If condition (1) is satisfied then the posterior distribution* $\mu_n(\cdot | X^n)$ *asymptotically concentrates on open neighborhoods of the true distribution* P_0, *that is,*

$$\lim_n \mu_n(\{P \in N\} | X^n) = 1, \ P_0 \text{ almost surely.}$$

This result assumes that the neighborhoods N are open with respect to the topology of setwise convergence of probability measures. (For instance, N could be $\{P: \sum_A |P_0(A)-P(A)| < \varepsilon\}$, where the sum is for A in a countable partition of the sample space.)

Finally, we mention that for parametric problems, Strasser [6] has shown under condition (1) and other mild assumptions that if the maximum likelihood estimator is consistent, then Bayes rules are also consistent. Although consistency in the information sense is not usually addressed in the parametric setting, the usual conditions for the consistency of the MLE are sufficiently restrictive that convergence of the parameter estimators $\hat{\theta} \to \theta$ implies $D(P_\theta \| P_{\hat{\theta}}) \to 0$.

3. Evidence Against Consistency.

In Barron [7] it will be shown that there exist priors which satisfy (1),

$$\mu\{P: D(P_0 \| P) < \varepsilon\} > 0 \text{ for all } \varepsilon > 0,$$

yet the posterior distribution given X^n asymptotically concentrates outside D neighborhoods of the true P_0, that is, for some $\varepsilon > 0$,

$$\lim_n \mu_n(\{P: D(P_0 \| P) < \varepsilon\} | X^n) = 0, \ P_0 \text{ almost surely.}$$

Proof of Lemma 1 and Lemma 2. Let P^n denote the product measure with joint probability density function $p(x^n) = \prod_{i=1}^n p(x_i)$ and let M^n denote the mixture of these distributions obtained using the prior μ. This mixture has joint density function

$$m(x^n) = \int p(x^n) \, d\mu.$$

We first show that condition (1) implies that the informational divergence between P_0^n and M^n has a rate tending to zero; that is,

$$\lim_{n\to\infty} \frac{1}{n} D(P_0^n \parallel M^n) = 0.$$

Given $\varepsilon > 0$, let $N = \{ P : D(P_0 \parallel P) < \varepsilon \}$. Now the divergence rate is

$$\frac{1}{n} D(P_0^n \parallel M^n) = \frac{1}{n} E \log \frac{p_0(X^n)}{\int p(X^n) d\mu} \leq \frac{1}{n} E \log \frac{p_0(X^n)}{\int_N p(X^n) d\mu}$$

$$= \frac{1}{n} E \log \frac{p_0(X^n)}{\int_N p(X^n) d\mu/\mu(N)} + \frac{1}{n} \log \frac{1}{\mu(N)}.$$

Here all the expectations are with respect to P_0^n. By the convexity of the informational divergence this is

$$\leq \int_N \frac{1}{n} D(P_0^n \parallel P^n) \, d\mu/\mu(N) + \frac{1}{n} \log \frac{1}{\mu(N)}$$

$$= \int_N D(P_0 \parallel P) \, d\mu/\mu(N) + \frac{1}{n} \log \frac{1}{\mu(N)}.$$

By the definition of N this is

$$\leq \varepsilon + \frac{1}{n} \log \frac{1}{\mu(N)}.$$

Letting $n \to \infty$ then $\varepsilon \to 0$ shows that indeed

$$\lim_{n\to\infty} \frac{1}{n} D(P_0^n \parallel M^n) = 0.$$

Now we need to relate this to the convergence of density estimators.

Let $\hat{p}_n(x_{n+1})$ be our density estimate at the point x_{n+1} based on the data $X^n = x^n$. We can write this as

$$\hat{p}_n(x_{n+1}) = \frac{\int p(x_{n+1}, x^n) d\mu}{\int p(x^n) d\mu} = \frac{m(x_{n+1}, x^n)}{m(x^n)} = m(x_{n+1} \mid x^n).$$

The last expression is sometimes called the predictive density. It is the

conditional density function for X_{n+1} given X^n. Note that with respect to M^n the data $X_1, X_2, \ldots, X_n$ are no longer independent (but they are exchangeable).

Now by the chain rule

$$\frac{1}{n} D(P_0^n \| M^n) = \frac{1}{n} \sum_{k=1}^{n} E \log \frac{p_0(X_k)}{m(X_k \mid X^{k-1})}.$$

The terms in the sum are just $E\, D(P_0 \| \hat{P}_k)$. Thus

$$\frac{1}{n} D(P_0^n \| M^n) = \frac{1}{n} \sum_{k=1}^{n} E\, D(P_0 \| \hat{P}_k).$$

But we have shown that condition (1) implies that this tends to zero. Thus the $E\, D(P_0 \| \hat{P}_n)$ tends to zero in the Cesaro sense. Since the terms are positive this implies that we have convergence to zero along a subsequence. This implies convergence in probability along a subsequence and hence almost sure convergence along a further subsequence. This completes the proof of Lemma 1.

For Lemma 2, use the convexity of divergence once more to obtain

$$E\, D(P_0 \| \tilde{P}_n) = E\, D(P_0 \| \frac{1}{n} \sum \hat{P}_k) \leq \frac{1}{n} \sum_{k=1}^{n} E\, D(P_0 \| \hat{P}_k)$$

which tends to zero. This completes the proof.

REFERENCES

[1] I. Csiszár, "Information-type Measures of Difference of Probability Distributions and Indirect Observations," *Studia Sci. Math. Hungar.* 2, pp. 299-318 (1967).

[2] J.L. Doob, "Application of the Theory of Martingales," *Colloq. Int. CNRS*, pp. 22-28 (1949).

[3] A.R. Barron, "Discussion on the Consistency of Bayes Estimates," *Ann. Statistics*, 14, pp. 26-30 (1986).

[4] L. Schwartz, "On Bayes' Procedures," *Z. Wahrsch. verw. Gebiete*, 4, pp. 10-26 (1965).

[5] D. Freedman, "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case," *Ann. Math. Statistics*, 34, pp. 1386-1403 (1963).

[6] H. Strasser, "Consistency of Maximum Likelihood and Bayes Estimates," *Ann. Statistics*, 9, pp. 1107-1113 (1981).

[7] A.R. Barron, "On Uniformly Powerful Tests and Bayes Consistency," in preparation.