BAYESIAN STATISTICS 6, pp. 27–52
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)

Information-theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems

ANDREW R. BARRON
Yale University, USA

SUMMARY

The risk of Bayes procedures (predictive densities) with Kullback-Leibler loss and the asymptotics of the posterior distribution are examined for densities with Kullback neighborhoods assigned positive prior probability. Necessary and sufficient conditions for consistency of the posterior distribution are proved. Examples reveal that the posterior distribution can behave quite peculiarly, while, in contrast, the cumulative risk of predictive densities has nice properties that can be used in advance of observing the data to help choose the prior. The underlying reason is that, through the chain rule of information theory, the cumulative risk equals the total Kullback divergence between a joint distribution in the family and the Bayes mixture, which is controlled by local properties of the prior, chiefly, how much prior mass is given to Kullback neighborhoods. In smooth parametric problems of dimension k, sample size N and Fisher information $I(\theta)$, a cumulative risk function of $(k/2)\log N + a(\theta)+$ constant is achieved asymptotically by a prior which is proportional to $|I(\theta)|^{1/2}e^{-a(\theta)}$, an exponential tilting of Jeffreys' prior using a target risk function $a(\theta)$. This prior gives mass proportional to $e^{-a(\theta)}$ to small Kullback balls around θ. A simple upper bound on cumulative risk is given by an *index of resolvability* which holds in finite samples, and it is applied to problems in model mixing, nonparametric estimation, and neural nets.

Keywords: PREDICTIVE DENSITY ESTIMATION; CUMULATIVE KULLBACK-LEIBLER RISK; INDEX OF RESOLVABILITY; CHOICE OF PRIOR; ASYMPTOTICS OF POSTERIOR.

1. INTRODUCTION

In this paper we review connections between the choice of prior and the behavior of the posterior and predictive distributions. Consistency of the posterior distribution is characterized. We focus on the cumulative Kullback risk of predictive distributions and use it to motivate choices of models and priors for parametric and nonparametric problems.

Concerning predictive distributions for a sequence of observations, the size of the cumulative Kullback risk is shown to be controlled principally by the prior probability assigned locally to small Kullback balls and is only secondarily affected by the nature of the likelihood process outside of such balls. To achieve cumulative risk approximately equal to and not greater than $a(\theta)$ with a discrete parameter set one uses prior probabilities $w(\theta) = e^{-a(\theta)}$, and to achieve cumulative risk $(k/2)\log N + a(\theta)+$constant for smooth models of parameter dimension k, sample size N, and Fisher Information $I(\theta)$, one uses a prior proportional to $|I(\theta)|^{1/2}e^{-a(\theta)}$, which gives mass proportional to $e^{-a(\theta)}$ to small Kullback balls around θ. This cumulative risk corresponds to an efficient individual risk sequence of $k/2n$ plus a summable remainder. The cumulative risk bounds are shown to hold in general for each N through an index of resolvability.

In contrast asymptotic concentration of the posterior distribution is a more delicate matter, necessitating non-local conditions. An implication for Hellinger or L_1 neighborhoods of a

density is that for the posterior distribution to asymptotically concentrate on such neighborhoods, the prior probability of the set of densities with large variation must be exponentially small.

An example is given where positive prior mass is given to Kullback neighborhoods, such that the time average Kullback risk of predictive distributions must tend to zero, however the posterior distribution does not concentrate on Kullback balls. Recognizing that the predictive density is the mean of the density function with respect to the posterior distribution, this example implies that the predictive density is obtaining its accuracy by averaging across many bad models rather than by posterior concentration.

A related example shows that the Kullback risk can increase for some sample sizes, even at parameter values given very high prior probability. It is the time average of Kullback risk that is assured good behavior, through monotonicity of the index of resolvability.

Implications of the information-theoretic analysis for gambling, prediction, data compression, reference priors and model selection are discussed. Application is given to classification and regression by neural networks.

Several main perspectives form the backdrop to our study of Bayes procedures, especially the work of Schwartz (1965) who revealed the role of uniformly consistent tests and the Kullback-Leibler support of the prior in analysis of Bayes consistency, the work of Ibragimov and Hasminskii (1973) on the information in a sample about a parameter, the work of Bernardo (1979) introducing the reference prior, the work of Dawid (1984,1992) on prequential analysis, the work of Rissanen (1984,1996) on predictive and mixture implementations of the minimum description length principle, the work of Davisson (1973) and Shtarkov (1988) on expected and worst case regret in data compression, the work of Haussler and his colleagues (1997,1998) on cumulative risk in prediction, and the work of Cover *et al.* (1990,1996) on gambling and investment interpretations. My colleague Bertrand Clarke (1989) has shaped much of the thinking about the asymptotics of the total Kullback risk of Bayes procedures, the asymptotic minimax value, and the sequences of procedures that achieve it.

2. INFORMATION THEORY PRELIMINARIES

The Kullback-Leibler divergence between two probability density functions $p(x)$ and $q(x)$ with respect to a reference measure (e.g. counting or Lebesgue) on the space of a variable X is equal to the non-negative quantity $D(p,q) = E_p \log p(X)/q(X)$, where E_p denotes expectation with respect to the distribution for X with density p. When P and Q are the corresponding distributions, we also denote the relative entropy by $D(P,Q)$. It is equal to zero only if $P = Q$.

2.1. *Average Regret*

The interpretation of the Kullback divergence is the following. Suppose in advance of observing a random variable X one is to assign a probability density function q. The aim is to produce a large value of $q(X)$, at least in terms of expected logarithm. If X is known to follow the density p, then the assignment q equal to p produces the largest expected logarithm as can be seen by noting that the expected value of the regret (difference in logarithms) $\log p(X)/q(X)$ is equal to $D(p,q)$.

If X follows some member of a family $p(x|\theta)$, $\theta \in \Theta$ of densities but the particular θ is unknown, then the expected regret using q_X is $D(p_{X|\theta}, q_X)$. [Here subscripts on the density $p_{X|\theta}$ or distribution $P_{X|\theta}$ denote that it is for the variable named X and that it is indexed by θ; the value of density at a random X is denoted $p(X|\theta)$ and the subscripts are dropped when clear from the context.] If we assign some distribution W on θ, then the choice of marginal distribution (or mixture) $p(x) = \int p(x|\theta)W(d\theta)$ minimizes the average value of $D(p_{X|\theta}, q_X)$ as is evident from the chain rule representation of $\int D(p_{X|\theta}, q_X)W(d\theta)$ as $\int D(p_{X|\theta}, p_X)W(d\theta) + D(p_X, q_X)$.

The mixed density $p(x)$ is also denoted by $m(x)$, or $p^{(W)}(x)$ when we want to make clear the dependence on the choice of the prior. The minimized average value is equal to the divergence $D(P_{\theta,X}, P_\theta \times P_X)$ between the joint distribution for θ and X and the product of the marginals when $P_\theta = W$ and this average divergence is known as the Shannon mutual information.

2.2. *Chain Rule*

Suppose a sequence of random variables $X_1, X_2, \ldots, X_N$ is assigned joint probability densities $p(X_1, \ldots, X_N)$ and $q(X_1, \ldots, X_N)$, which are written as products of conditional densities $\prod_{n=1}^N p(X_n|X^{n-1})$ and $\prod_{n=1}^N q(X_n|X^{n-1})$. The chain rule yields

$$E_{P_{X^N}} \log \frac{p(X_1, X_2, \ldots, X_N)}{q(X_1, X_2, \ldots, X_N)} = \sum_{n=1}^N E_{P_{X^n}} \log \frac{p(X_n|X^{n-1})}{q(X_n|X^{n-1})},$$

where $X^n = (X_1, \ldots, X_n)$ and for the first term in the sum there is no conditioning. Thus the total Kullback divergence between the joint distributions is a sum of expected divergences between the conditional distributions

$$D(P_{X^N}, Q_{X^N}) = \sum_{n=1}^N E_{P_{X^{n-1}}} D(P_{X_n|X^{n-1}}, Q_{X_n|X^{n-1}}).$$

In particular comparing the expected cumulative divergence between distributions for a sequence with and without use of a parameter θ we have the total divergence

$$D(P_{X^N|\theta}, P_{X^N}) = \sum_{n=0}^{N-1} E_{P_{X^n|\theta}} D(P_{X_{n+1}|X^n,\theta}, P_{X_{n+1}|X^n})$$

expressed as a sum of the Kullback risks of the predictive densities $p(X_{n+1}|X^n)$. As above, suppose a prior W is assigned to θ. Then the mixture $p(X^N) = \int p(X^N|\theta)W(d\theta)$ minimizes the average value of $D(P_{X^N|\theta}, P_{X^N})$ and this choice coincides, for each X^n with the choice of predictive density $p(X_{n+1}|X^n) = \int p(X_{n+1}|X^n, \theta)W(d\theta|X^n)$ minimizing the average of $D(P_{X_{n+1}|X^n,\theta}, P_{X_{n+1}|X^n})$ with respect to the corresponding posterior distributions $W(\theta|X^n)$ for $n = 0, 1, \ldots N-1$.

It is interesting to note that many other loss functions, such as Hellinger or L_1, lead to different estimators of the distribution that would not be the Bayesian's conditional distribution for X_{n+1} given X^n.

2.3. *Operational Interpretations*

Before continuing further with the analysis we hasten to give operational meanings to the predictive densities $p(\cdot|X^n, \theta)$ and $p(\cdot|X^n)$ for X_{n+1}. We do not require physical meaning of the parameter, in particular, it is not something to be learned. These operational meanings are perhaps most clear in gambling and data compression contexts, but I will first abstract a general predictive interpretation.

We can think of θ as labeling a family of individuals that predict in certain ways. For each time n, once we have observed $X^n = (X_1, X_2, \ldots X_n)$, the strategy θ assigns a function $p(\cdot|X_1, X_2, \ldots X_n, \theta)$ that sums (or integrates) to 1 across possible values for the next variable. For each strategy the aim is to realize large values for $p(X_{n+1}|X^n, \theta)$ and we will keep track of performance through the cumulative logarithm, summing over n less than N. As we have seen,

given a value of θ, the choice $p(X_{n+1}|X^n, \theta)$ optimizes the expected logarithm if X_{n+1} has that distribution. We think of these predictive densities merely as prediction strategies – we do not necessarily believe any one of these predictive distributions generates the data – nevertheless, we can admit that the individual predictors are behaving sensibly should they possess such beliefs as to the conditional distribution. The collection of prediction strategies for $\theta \in \Theta$ provides a target class of performance $\log p(X_1, \ldots, X_N|\theta) = \sum_{n=0}^{N-1} \log p(X_{n+1}|X^n, \theta)$ on sequences $X_1, \ldots, X_N$. We allow ourselves to do prediction by means of predictive densities $p(\cdot|X^n)$ that are outside the target class, but our aim is to have one procedure (without dependence on θ) that does well in comparison with every member of the target class.

Each strategy corresponds to a distribution on $X_1, X_2, ..., X_N$ with joint density $p(X_1, X_2, \ldots X_N)$ (summing or integrating to 1 over all X^N) which is given as a product of the predictive densities $p(X_{n+1}|X^n)$. The cumulative regret of our strategy compared to that achieved by strategy θ is thus

$$\log \frac{p(X_1, X_2, \ldots, X_N|\theta)}{p(X_1, X_2, \ldots, X_N)} = \sum_{n=0}^{N-1} \log \frac{p(X_{n+1}|X^n, \theta)}{p(X_{n+1}|X^n)}.$$

We are particularly interested in the regret achieved by Bayes predictive densities. The justification of this interest arises in part by the Bayes optimality for average expected regret (Kullback risk) discussed above. Though, as I have occasion to briefly report, there are recent justifications based on examination of the worst case value of the regret maximized over choices of $X_1, X_2, \ldots, X_N$ (see e.g., Shtarkov 1988, Barron, Rissanen, and Yu 1998, Xie and Barron 1998, Cover *et al.* 1990, 1996), here I will be content to given a more classical story based on expected regret.

The problem arises as to the basis for choice of the prior distribution on θ. Here θ indexes individuals who predict according to certain strategies and it is not clear that there is sense in this setting to the idea of a subjective choice of prior probability for θ. Indeed, one can call into question what would be meant by statements concerning the probability of sets of individuals or strategies. The point here is that, like other objects, the prior W arises operationally in constructing strategies for prediction. It is advocated that the prior be chosen to shape the relative performance of the resulting predictions as measured by the expected regret or total Kullback risk. We return to this point in Section 3.

2.4. *Data Compression*

Suppose for each X_n there is a discrete set of possible values. From information theory (see e.g. Cover and Thomas 1991), there is a uniquely decodable binary code for $X_1, \ldots, X_N$ for each choice of distribution $q(X_1, \ldots, X_N)$ that sums to not more than 1 for which the length of the codeword for $(X_1, \ldots, X_N)$ is equal to $\log 1/q(X_1, \ldots, X_N)$ rounded up to an integer. Ignoring the integer constraint (which is merely a small numerical nuisance for large N) we see that in accordance with the principles in Section 2.1 the minimal expected codelength with θ given is achieved by the code based on the distribution $p(X_1, \ldots, X_N|\theta)$ and the minimal average expected regret (code redundancy) is achieved by using the Bayes mixture $p(X_1, \ldots, X_N) = \int p(X_1, \ldots, X_N|\theta) W(d\theta)$ (Davisson 1973). So we have the same problem as discussed above. In particular the redundancy of the Bayes mixture is the same as the total Kullback divergence between $p(X_1, \ldots, X_N|\theta)$ and $p(X_1, \ldots, X_N)$. Here the chain rule has an interpretation in terms of conditional description lengths for X_{n+1} given $X_1, \ldots X_n$.

2.5. *Gambling*

Suppose that starting initially with one unit of wealth, a gambler is asked at each time n to distribute fractions $q(x|X_1,\ldots,X_n)$ of the current compounded wealth on the various possibilities x for the winning outcome X_{n+1} (e.g. horse race) at time $n+1$. Again the meaning of $q(x|X_1,\ldots,X_n)$ is operational, it is revealed by the gambler's actions, and its choice is entirely up to the gambler, subject to the condition that it is non-negative and sums to one. If the winning outcome provides $O(X_{n+1}|X_1,\ldots,X_n)$ dollars for each dollar wagered, then the total wealth at the end of N races equals $S_N = \prod_{n=0}^{N-1} q(X_{n+1}|X^n)O(X_{n+1}|X^n)$ which simplifies to $S_N = q(X_1,\ldots,X_N)O(X_1,\ldots X_N)$, where $q(X_1,\ldots,X_N) = \prod_{n=0}^{N-1} q(X_{n+1}|X^n)$ and the overall odds $O(X_1,\ldots,X_N)$ is the corresponding product of odds for individual plays. See Cover and Thomas (1991). The desire of the individual gambler is to have produced a large value of $q(X_1,\ldots,X_N)$ on the sequence that occurs.

Given a family of gambling strategies $p(X_{n+1}|X^n,\theta)$ indexed by gamblers θ, each receives wealth $S_N(\theta) = p(X_1,\ldots,X_N|\theta)O(X_1,\ldots X_N)$ where again $p(X_1,\ldots,X_N|\theta)$ is non-negative, sums to 1, and has operational meaning as the product of fractions of wealth gambled on the winning sequence of outcomes. We let these strategies form a target class for our gambling strategy. We gamble according to a sequence of predictive densities $p(X_{n+1}|X^n)$ and achieve wealth $S_N^{\text{actual}} = p(X_1,\ldots,X_N)O(X_1,\ldots,X_N)$. We want to chose a strategy that has control over the wealth ratio $S_N(\theta)/S_N^{\text{actual}} = p(X_1,\ldots,X_N|\theta)/p(X_1,\ldots,X_N)$, either for all sequences or in the sense of the expected value of the log wealth ratio (regret). Once again the Bayes strategy for the log ratio distributes wealth according to $p(X_1,\ldots,X_N) = \int p(X_1,\ldots,X_N|\theta)W(d\theta)$. This can be realized actively by gambling fractions of wealth according to the Bayes predictive distributions $p(X_{n+1}|X^n)$.

There is a passive implementation of the Bayes gambling strategy that gives the most direct operational meaning to the prior W. Before the first race we distribute our wealth among the family of gamblers, allocating $W(d\theta)$ to each gambler θ to act as an agent on our behalf. Each gambler θ compounds this initial wealth to yield $W(d\theta)p(X_1,\ldots,X_N|\theta)O(X_1,\ldots X_N)$ and summing across the agents the total wealth returned to us at the end of the N plays is $\int p(X_1,\ldots,X_N|\theta)W(d\theta)O(X_1,\ldots,X_N)$. This passively achieved wealth coincides exactly with the wealth achieved by gambling sequentially using the Bayes predictive distribution. The meaning of the prior probability $W(A)$ for each set A is the fraction of initial wealth entrusted to the gamblers with strategies θ in A. Following Cover *et al.* (1990, 1996) one may chose W so as to achieve certain wealth objectives, e.g., uniformly valid wealth ratio bounds (or regret bounds) uniformly over all outcome sequences and strategies θ.

3. INFORMATION ASYMPTOTICS AND IMPLICATIONS FOR PRIOR CHOICE

In advance of observing the data, we know the expected regret for each possible value for θ, which we have noted to be the total Kullback divergence

$$D(P_{X_1,\ldots,X_N|\theta}, P_{X_1,\ldots,X_N}) = E_{P_{X^N|\theta}} \log \frac{p(X_1,X_2,\ldots,X_N|\theta)}{p(X_1,X_2,\ldots,X_N)}.$$

The joint density assigned by the Bayes mixture with prior density $w(\theta)$ is equal to

$$p(X_1,X_2,\ldots,X_N) = \int p(X_1,X_2,\ldots,X_N|\theta)w(\theta)d\theta.$$

For smooth parametric families, such as under conditions given in Clarke and Barron (1990) in which the X_i are conditionally i.i.d. given θ, Laplace approximation of this integral reveals

that asymptotically the expected total divergence satisfies

$$D(P_{X_1,\ldots,X_N|\theta}, P_{X_1,\ldots,X_N}) = \frac{k}{2}\log\frac{N}{2\pi e} + \log\frac{|I(\theta)|^{1/2}}{w(\theta)} + o(1)$$

for parameter values internal to the parameter space in R^k. Note that such a cumulative risk corresponds to an individual risk $ED(P_{X|\theta}, P_{X|X^n})$ of $k/2n$ plus a summable remainder. Detailed second order properties of the individual Kullback risk are in Hartigan (1998). The level $k/2n$ is shown to be asymptotically efficient in Barron and Hengartner (1998).

Let's express a desired form of the Kullback risk through a function $a(\theta)$. To achieve total Kullback risk of the form $a(\theta) + C_N$ plus an asymptotically vanishing term, where C_N is a constant (depending on the sample size N and the dimension k but not on θ), we see that we are compelled to choose a prior of the form

$$w(\theta) = |I(\theta)|^{1/2} e^{-a(\theta)}/c$$

where $c = \int |I(\theta)|^{1/2} e^{-a(\theta)} d\theta$. This is the Jeffreys prior $|I(\theta)|^{1/2}$ exponentially tilted by the desired risk behavior $a(\theta)$. The interpretation, that will be illuminated further below, is that the prior assigns mass to small Kullback balls that is proportional to $e^{-a(\theta)}$.

The moral is that it is not the Fisher information, per se, that controls the total risk but rather how the mass is distributed to the Kullback balls.

With $a(\theta)$ constant, we have the locally invariant (Jeffreys) prior that gives equal mass to small Kullback balls and total risk that (except at boundary points) is in agreement with the asymptotically minimax value $(k/2)\log\frac{N}{2\pi e} + \log\int |I(\theta)|^{1/2} + o(1)$ established in Clarke and Barron (1994), in concert with Bernardo's reference prior interpretation. To build in greater accuracy for certain parameter points we assign a smaller value of $a(\theta)$ for such points and tolerate a larger cumulative risk elsewhere.

We see that the desired aim of total risk of the form $a(\theta) + C_N + o(1)$ is possible only for $a(\theta)$ for which the normalizing constant $c = \int |I(\theta)|^{1/2} e^{-a(\theta)} d\theta$ is finite. This requirement is made because, unless the prior is proper, so that $p(X^N)$ sums to not more than 1, we violate the requirements of the data compression, gambling, and prediction interpretations.

For discrete mixtures $p(X^N) = \sum_\theta p(X^N|\theta)W(\theta)$, in the case of distributions $P_{X^N|\theta}$ which for distinct pairs of $\tilde{\theta} \neq \theta$ extend to mutually singular distributions $P_{X^\infty|\tilde{\theta}}$ and $P_{X^\infty|\theta}$ on infinite sequences (e.g. through assumption of ergodicity), the asymptotics of the regret is

$$\log\frac{p(X_1,\ldots,X_N|\theta)}{p(X_1,\ldots,X_N)} = \log\frac{1}{W(\theta)} + o(1).$$

Here $o(1)$ tends to zero $P_{X^\infty|\theta}$ almost surely by a standard martingale argument (demonstrating that $\sum_{\tilde{\theta}\neq\theta} p(X^N|\tilde{\theta})W(\tilde{\theta})/p(X^N|\theta)$ tends to zero). Obtaining convergence of the expected logarithm is somewhat trickier (see e.g., Clarke and Barron 1994, Thm. 2, which assumes that a Kullback ball of small positive radius reduces to the singleton θ).

Nevertheless, for every N, uniformly over all $X_1,\ldots,X_N$, one has, by throwing away the terms in the mixture not equal to θ,

$$\log\frac{p(X_1,\ldots,X_N|\theta)}{p(X_1,\ldots,X_N)} \leq \log\frac{1}{W(\theta)}$$

and hence the total Kullback risk is bounded by

$$D(P_{X^N|\theta}, P_{X^N}) \leq \log\frac{1}{W(\theta)}.$$

Thus in the discrete parameter case we also have a direct connection between a desired bound $a(\theta)$ on the total risk and the choice of the prior $W(\theta) = e^{-a(\theta)}$, where unlike the smooth continuous parameter case there is no Fisher information term. If θ is an isolated point in the Kullback sense, then $e^{-a(\theta)}$ is the prior probability of a Kullback neighborhood.

The common answer in both the continuous and discrete frameworks is that a prior is assigned to Kullback balls that is proportional to $e^{-a(\theta)}$ to achieve total Kullback risk of shape $a(\theta)$.

4. RESOLVABILITY

Not only is the total Kullback risk more directly suited, to the applications discussed above, than is the individual Kullback risk, but also we have the good fortune that it is easier to develop suitable upper bounds for it by taking lower bounds on the mixture $m(X^N) = \int p(X^N|\tilde{\theta})W(d\tilde{\theta})$ in which we restrict the integral to convenient sets in the vicinity of hypothetical parameter values. This leads to a bound we call the index of resolvability.

We discuss the resolvability bounds first in the context that the random variables $X_1, X_2, \ldots, X_N$ given θ are conditionally i.i.d. Suppose we have a family of densities $p(x|\theta), \theta \in \Theta$. If we use a prior probability distribution W, the Bayes estimators are the Bayes predictive densities $\hat{p}_n(x) = p(x|X^n) = \int p(x|\theta)W_{\theta|X^n}(d\theta)$. If hypothetically we consider the possibility that $X_1, X_2, \ldots, X_N$ are independently distributed according to some P with density $p(x)$ (which may or may not be in the family), the chain rule gives Cesaro or time average risk

$$\bar{r}_N(P) = \frac{1}{N}\sum_{n=0}^{N-1} ED(p, \hat{p}_n) = \frac{1}{N} D(P_{X^N}, M_{X^N}).$$

As we shall see from a simple bound this time average risk is made small for any P in the information support of the prior.

The information closure of the family $\{P_{X|\theta}\}$ consists of those distributions P for which the information neighborhoods $B_{\delta,P} = \{\theta : D(P, P_{X|\theta}) \leq (1/2)\delta^2\}$ are non-empty for all $\delta > 0$ and the information support of the prior consists of those P for which the information neighborhoods are assigned positive prior probability $W(B_{\delta,P}) > 0$. The Bayes estimator is said to be information consistent at P if $\bar{r}_N(P)$ tends to zero as $N \to \infty$.

The size of the risk for each N depends on how much prior probability is given to the information balls. Indeed, the following bound holds,

$$\bar{r}_N(P) \leq \min_{\delta>0}\{\frac{\delta^2}{2} + \frac{1}{N}\log 1/W(B_{\delta,P})\}.$$

The right side of this inequality is here called the *index of resolvability* of the distribution P by the mixture of distributions with prior W. An alternative expression for the index of resolvability is

$$\min_B\{\max_{\theta\in B} D(P, P_{X|\theta}) + \frac{1}{N}\log 1/W(B)\}.$$

This definition is an extension of the resolvability definition given in Barron and Cover (1991) in which W was discrete and the minimization was restricted to singleton sets $\{\theta\}$ (there $L(\theta) = \log 1/W\{\theta\}$ was interpreted as an arbitrary codelength for the parameter in a two-stage rather than mixture code for the resolution of X^N).

We note immediately from the resolvability bound that information consistency holds for any distribution in the information support of the prior. Note also that the index of resolvability

is non-increasing in N. The rate at which the index of resolvability tends to zero depends solely on how much prior mass is given to Kullback balls around P for various radii δ.

The proof of the resolvability bound on cumulative risk follows from noting that for any set B the mixture is lower bounded by $m(X^N) \geq W(B) \int_B p(X^N|\theta) W(d\theta|B)$ so that

$$E_P \log \frac{p(X^N)}{m(X^N)} \leq E_P \log \frac{p(X^N)}{\int_B p(X^N|\theta) W(d\theta|B)} + \log \frac{1}{W(B)}.$$

Then use convexity to obtain the further bound

$$\int_B D(P_{X^N}, P_{X^N|\theta}) W(d\theta|B) + \log \frac{1}{W(B)}$$

which is not greater that the resolvability using the set B, that is,

$$\max_{\theta \in B} D(P_{X^N}, P_{X^N|\theta}) + \log \frac{1}{W(B)}.$$

Dividing by N and optimizing over B produces the claimed bound. Steps in this proof are from Barron (1987) where the point was information consistency for all P in the information support of the prior. Use of the bound to express rates of convergence is in the technical report of Barron (1988). The name resolvability is more recent.

Thus the cumulative accuracy of Bayes estimators depends only on the local behavior of the prior for sets of θ with $P_{X|\theta}$ near the distribution P_X followed by the data. This simple conclusion is to be contrasted with the behavior of the posterior distribution which to asymptotically concentrate on a neighborhood of P_X requires also global conditions as in Section 8 below.

Allowing dependence in the models, the same bound holds for the Kullback rate $\bar{r}_N(P) = (1/N) D(P_{X^N}, M_{X^N})$ (the time average Kullback risk of prediction), provided the information neighborhoods are defined more generally by

$$B_\delta(P_{X^N}) = \{\theta : (1/N) D(P_{X^N}, P_{X^N|\theta}) \leq (1/2)\delta^2\}.$$

An alternative information-theoretic development of a similar bound on the total Kullback risk $D(P_{X^N}, M_{X^N})$ is obtained via chain rule expansion of the total divergence between the joint distributions $P_{X^N} \times W_\theta^{(N)}$ and $P_{X^N|\theta} W_\theta$ where the approximate posterior $W_\theta^{(N)}$ is defined to have density $e^{-D(P_{X^N}, P_{X^N|\theta})}/C_N$ with respect to the prior W_θ. The chain rule yields $D(P_{X^N}, M_{X^N}) + E_{P_{X^N}} D(W_\theta^{(N)}, W_{\theta|X^N}) = \log 1/C_N$ where $C_N = \int e^{-D(P_{X^N}, P_{X^N|\theta})} W(d\theta)$. Thus as shown in Barron (1988) and Haussler and Opper (1997) the total relative entropy risk $D(P_{X^N}, M_{X^N})$ is bounded by the quantity $\log 1/\int e^{-D(P_{X^N}, P_{X^N|\theta})} W(d\theta)$ (interpreted as a "Razor" in Balasubramanian 1997). Restricting the integral to a neighborhood, it is seen that the bound improves somewhat on the previous bound $\min_\delta \{(N/2)\delta^2 + \log 1/W(B_{\delta,P})\}$.

5. PARAMETRIC RESOLVABILITY BOUNDS

Suppose we have a finite-dimensional parametric family of densities $p(x|\theta)$, $\theta \in \Theta \subset R^k$ and that $X_1, \ldots, X_N$ are i.i.d. according to $p(x|\theta^*)$ where θ^* is in the interior of Θ and suppose the divergence $D(\theta^*, \theta) = D(P_{X|\theta^*}, P_{X|\theta})$ is twice continuously differentiable in θ

at θ^* with positive definite Hessian $J_{\theta^*,\theta}$. Let $\bar{J}_{\theta^*}$ locally dominate the Hessian for θ with $D(P_{X|\theta^*}, P_{X|\theta}) \leq (1/2)\delta^2$, so that for such θ,

$$D(P_{X|\theta^*}, P_{X|\theta}) \leq \frac{1}{2}(\theta - \theta^*)^T \bar{J}_{\theta^*}(\theta - \theta^*).$$

Then the information ball $B_{\delta,\theta^*} = \{\theta : D(P_{X|\theta^*}, P_{X|\theta}) \leq (1/2)\delta^2\}$ contains the ellipse

$$S_{\delta,\theta^*} = \{\theta : (\theta - \theta^*)^T \bar{J}_{\theta^*}(\theta - \theta^*) \leq \delta^2\}.$$

Suppose also that the prior W satisfies a near-absolute continuity property near θ^*, namely that there exists a positive $\underline{w}_{\theta^*}$ such that the prior probability of the ellipse S_{δ,θ^*} is at least $\underline{w}_{\theta^*}$ times its volume (as in the case of a prior with a density $w(\theta)$ locally bounded below by $\underline{w}_{\theta^*}$).

Now the prior probability of the information neighborhood satisfies

$$W(B_{\delta,\theta^*}) \geq W(S_{\delta,\theta^*}) \geq \underline{w}_{\theta^*} |\bar{J}_{\theta^*}|^{-1/2} v_k \delta^k$$

where v_k denotes the volume of the unit ball in R^k. Consequently, we have a bound for $(N/2)\delta^2 + \log 1/W(B_{\delta,\theta^*})$ which is optimized at $\delta^2 = k/N$, yielding

$$D(P_{X^N|\theta^*}, M_{X^N}) \leq \frac{k}{2} \log \frac{N}{k} + \frac{k}{2} + \log \left(|\bar{J}_{\theta^*}|^{1/2} / \underline{w}_{\theta^*} \right) + \log 1/v_k.$$

Thus, solely under a local quadratic behavior of the Kullback divergence, we have an upper bound that holds for all N with the desired form of dependence on θ, although with a somewhat larger constant than achieved asymptotically (under more stringent conditions). As before, the shape $a(\theta) + \text{constant}_N$ for the cumulative risk bound is arranged approximately by the choice of prior density $w(\theta)$ proportional to $|J_\theta|^{1/2} e^{-a(\theta)}$, where the Hessian of the Kullback divergence plays the role of Fisher information.

Similar bounds for dependent data models are possible as long as there is a local quadratic behavior for $(1/N)D(P_{X^N|\theta^*}, P_{X^N|\theta})$.

6. NONPARAMETRIC RESOLVABILITY BOUNDS

There are general information closure properties for some priors on infinite dimensional families of densities given in Barron, Schervish and Wasserman (1998) building on the developments in the technical report Barron (1988). The typical result involves some reference density p_0 and shows that all densities p with finite $D(p, p_0)$ are in the information support of the prior, yielding resolvability tending to zero, and hence the Bayes procedure is consistent for such target densities, where consistency is taken in the sense of time average Kullback risk tending to zero.

One may also use resolvability to examine rates of convergence.

6.1. *Minimax Rates*

The resolvability can be used to identify minimax optimal rates for density estimators as in Yang and Barron (1998b) in terms of metric entropies. Briefly, suppose a class of density functions $\mathcal{F}$ can be covered by a net of not more than $N = N_\delta$ densities, say $q_1, \ldots, q_N$, such that for every p in $\mathcal{F}$ there is a q_i with $D(p, q_i) \leq \delta^2$. The smallest such net yields the Kullback δ-entropy of the class $\mathcal{F}$ as $H_\delta = \log N_\delta$. Put a uniform prior on the net. Then using singleton sets for B we get an upper bound on the resolvability of

$$\bar{r}_N(p) \leq \min_\delta \{\delta^2 + \frac{1}{N} H_\delta\}$$

uniformly over all densities p in $\mathcal{F}$. In particular we have the bound $\bar{r}_N(p) \leq 2\delta_N^2$ when δ_N is chosen such that δ_N^2 and $(1/N)H_{\delta_N}$ are the same. As shown in Yang and Barron (1998b) this bound is tight. There resolvability is also used in the lower bound to control the mutual information that arises from application of Fano's inequality.

Barron and Hengartner (1998) use a resolvability calculation to show that the subset of densities in a given class that converge at faster than the minimax rate have a sparse cover (smaller order metric entropy).

6.2. *Model Mixing*

A practical use of the resolvability in nonparametric settings is to address the efficacy of model mixing. The idea is to take advantage of a sequence of parametric families $p(x|\theta_m, m), \theta_m \in \Theta_m$ for a sequence of model indexes $m \in \mathcal{M}$. Rather than performing model selection (for which the constants in general risk bounds can be quite horrendous and the empirical process conditions quite stringent, see Barron, Birgé, Massart (1998) or Yang and Barron (1998a)), we instead use a Bayes mixture strategy.

For each family we have prior probabilities $p(m)$, prior distributions $W_m(d\theta_m)$ conditionally for each m, and resulting mixtures $p(X^n|m) = \int p_m(X^n|\theta_m, m)W_m(d\theta_m)$. The overall mixture is $p(X^n) = \sum_m p(m)p(X^n|m)$. To relate the risk of the overall mixture to the risks of the individual mixtures, simply lower bound the sum by individual terms within it. We find that for every $P^*_{X^N}$, the Kullback rate satisfies the oracle inequality

$$(1/N)D(P^*_{X^N}, P_{X^N}) \leq \min_m \{(1/N)D(P^*_{X^N}, P_{X^N|m}) + (1/N)\log 1/p(m)\}.$$

Thus model mixing has the adaptation property of performing nearly as well as if the best resolving model m_n were known in advance.

Though this oracle inequality for model mixing is useful, I give two qualifying remarks. One is that it is for the Cesaro average of Kullback risks that we have defended here, not for the individual risk. Secondly, bounding the sum by individual terms in it does not reveal what additional advantages there may be to mixing a number of models which may have roughly equal contributions $p(m)p(X^n|m)$.

7. SOME SURPRISES

As we have seen the Cesaro time average Kullback risk of Bayes procedures is tracked by the non-increasing index of resolvability. Moreover, the Bayes prior average of the Kullback risk must also be non-increasing. Perhaps surprisingly, the individual Kullback risk $ED(P_{X_{n+1}|\theta}, P_{X_{n+1}|X^n})$ can increase for some n, even at a θ given a large prior probability.

7.1. *Increasing Kullback Risk*

We give a simple example with $n = 1$, which also serve as a precursor to an example of inconsistency given below. Let X_1 and X_2 be independent with a density of height 1 uniformly distributed on the interval $[0, 1)$ conditionally given $\theta = 0$. For $\theta = 1$ and $\theta = 2$ let the density be of height 2 on the subintervals $[0, 1/2)$ or $[1/2, 1)$, respectively. Let the prior assign equal positive weight to $\theta = 1$ and $\theta = 2$ and let it assign weight W to $\theta = 0$. Now with no data, the predictive distribution for X_1 is the mixture, which is seen to be uniform on $[0, 1)$. Thus at $\theta = 0$, the Kullback risk is $D(P_{X_1|\theta}, P_{X_1}) = 0$. But the joint mixture density for X_1 and X_2 is of course not uniform (it gives height $W + 2^2(1 - W)/2$ when X_1 and X_2 are in the same subinterval

and height W when they are in distinct subintervals) and thus $ED(P_{X_2|\theta}, P_{X_2|X_1}) > 0$. In a set of prior probability W arbitrarily close to one we attain smaller Kullback risk with no data!

7.2. *Inconsistent Posterior Distribution*

In a similar manner, mixtures of densities concentrated on $m/2$ out of m intervals can closely mimic the uniform density for sample sizes n sufficiently less than m, in which case the predictive density is accurate but the posterior distribution gives collectively large weight to densities far from the uniform. Distributing a prior on such collections of densities for a sequence of values of m leads to a proof of posterior inconsistency, even though, since positive mass will be given to Kullback neighborhoods, the predictive density is consistent in Cesaro average of Kullback risk.

Consider the densities that put height 2 on subsets of size $m/2$ of the cells $[0, 1/m)$, $[1/m, 2/m), \ldots, [(m-1)/m, 1)$. Given m let the prior put equal mass on each of these $\binom{m}{m/2}$ densities, and let prior mass $p(m)$ be put on the even integers $m = 2, 4, \ldots$ for any decreasing sequence $p(m)$ for which $p(m)$ is not exponentially small in m and $\sum_m p(m) = 1/2$. Every one of these densities has L_1 distance from the uniform equal to 1 (and hence not a small distance from the uniform in several other measures of divergence). We call this set of densities $\mathcal{F}_{bad}$. The remaining mass $1/2$ will be put on some densities close to the uniform that we specify later. Given a density our model makes $X_1, X_2, \ldots$ conditionally i.i.d.

Given m, let the random variable Y be the number of the cells of width $1/m$ that are occupied by the sample $X_1, X_2, \ldots, X_n$. For each choice U that is a union of $m/2$ of the m cells, the corresponding density for $X_1, X_2, \ldots X_n$ has height 2^n if the sample is in U and height 0 otherwise. There are $\binom{m-Y}{m/2-Y}$ choices for U which cover the sample. Consequently, for a given m, the equally weighted mixture of these densities is $2^n\binom{m-Y}{m/2-Y}/\binom{m}{m/2}$. The ratio of binomial coefficients simplifies to a product of $Y \leq n$ fractions each of which exceeds $(m/2 - Y)/m > 1/2 - n/m$, so given m the mixture density is at least

$$2^n (1/2 - n/m)^n = (1 - 2n/m)^n .$$

This bound holds uniformly for all $X_1, X_2, \ldots X_n$ for $m \geq 2n$. The mixture over m takes the sum of these weighted by $p(m)$ which will be at least as large as at a particular m_n. Now since $\lim_m (1/m) \log p(m) = 0$, there exists an $m_n \geq 2n$ with m_n/n tending to infinity sufficiently slowly that $\lim_n (1/n) \log p(m_n) = 0$, i.e., $p(m_n)$ is not exponentially small in n, and $\epsilon_n = (1/n) \log p(m_n) + \log(1 - 2n/m_n)$ tends to zero. Consequently, the mixture over all of the above densities is at least

$$p(m_n) (1 - 2n/m_n)^n = e^{-n\epsilon_n}$$

which is not exponentially small. To ease the subsequent analysis let δ_n be a positive and strictly decreasing sequence not smaller than ϵ_n arranged such that $\delta_n \to 0$ and the difference $r_n = \delta_n - \epsilon_n$ satisfies $r_n \sqrt{n} / \log\log n \to \infty$.

The remaining step to devise an example of inconsistency of the posterior is to distribute the remaining $1/2$ of the prior over some densities close to the uniform in the Kullback sense, so that the mixture over these is also not exponentially small, but arrange that this part of the mixture is eventually below $e^{-n\epsilon_n}$ with high probability. This we accomplish by considering the family of tilted densities $(\beta + 1)x^\beta$ on $(0, 1)$ with $\beta \geq 0$, which we write in the exponential family form $e^{-\theta} e^{\beta(1+\log x)}$ where the quantity in the exponent $1 + \log x$ has mean zero when x is uniformly distributed. Here $\theta = \beta - \log(1 + \beta)$ is a one-to-one correspondence for positive θ and β. The joint density for $X_1, X_2, \ldots, X_n$ for given θ takes the form $\exp\{-n\theta + \beta(\theta) S_n\}$

where $S_n = \sum_{i=1}^n (1 + \log X_i)$. This family includes the uniform density when $\theta = 0$ and densities with θ near zero are close to the uniform density in the Kullback sense.

We assign part of the prior (of total mass $1/2$) to live on the parameter $\theta \geq 0$. Here is one device to achieve our aim of positive mass in neighborhoods of $\theta = 0$ while maintaining a relatively small value of the mixture. Let δ_η for $\eta \geq 0$ be a continuous strictly decreasing function with δ_0 at least 1 such that δ_η matches the sequence δ_n given above on the integers and let $g(\theta)$ be its decreasing inverse for an interval of values of θ including $(0, 1)$. Set the prior density to be proportional to $e^{-g(\theta)}$ on $(0, 1)$ with normalizing constant $c = \int_0^1 e^{-g(\theta)}$. Now by monotonicity $g(\theta)(\tilde{\theta} - \theta)$ is not less than $g(\tilde{\theta})(\tilde{\theta} - \theta)$, so setting $\tilde{\theta} = \delta_n$ and choosing n large enough that δ_n is less than 1 we deduce that $g(\theta) + n\theta \geq n\delta_n$ for all $0 < \theta < 1$. The contribution to the Bayes mixture from these densities is $(1/c) \int_0^1 \exp\{-g(\theta) - n\theta + \beta(\theta) S_n\} d\theta$ which for all large n is less than $(1/c) \exp\{-n\delta_n + (S_n)^+ \beta_1\}$, where β_1 is the constant for which $\beta - \log(1 + \beta) = 1$ and $(\cdot)^+$ denotes the positive part.

Now the ratio of the mixture on $\mathcal{F}_{bad}$ divided by the mixture of the rest is at least

$$c \exp\{nr_n - (S_n)^+ \beta_1\},$$

which tends to infinity almost surely for $X_1, X_2, \ldots X_n$ independent uniform random variables by the law of the iterated logarithm applied to S_n. Consequently, the posterior probability of the set of densities $\mathcal{F}_{bad}$ satisfies

$$P(\mathcal{F}_{bad} | X_1, X_2, \ldots X_n) \to 1 \quad a.s.$$

This failure of the posterior to concentrate asymptotically on neighborhoods of the density does not preclude the ability of the predictive density (the posterior mean density) to be accurate. It also does not preclude the asymptotic concentration on weak neighborhoods of the distribution.

From the analysis above we can say somewhat more. Given $0 < \epsilon < 1$ let $\mathcal{F}_{\pi_m} = \{Q : \sum_{A \in \pi_m} |P(A) - Q(A)| > \epsilon\}$ be the set of distributions that have π_m-variation distance from the uniform distribution P of at least ϵ, where π_m is the partition into m equal cells. Then for $m_n/n \to \infty$ the posterior asymptotically concentrates on complements of π_{m_n} -variation neighborhoods in the sense that $P(\mathcal{F}_{\pi_{m_n}} | X_1, X_2, \ldots X_n) \to 1$ a.s. As we shall see, that is as far as we can push it, for when m_n/n is bounded the posterior will live asymptotically on π_{m_n}-variation neighborhoods.

The above analysis improves on the counterexample in Barron (1988) in that here the prior is fixed and does not change with n. It improves on the analysis in Barron, Schervish, and Wasserman (1998) in that they obtained the indicated failure of the posterior only for m_n of order at least n^2. Also here we get that the posterior probabilities of the bad sets tend to one, not just that they fail to converge to zero.

8. CONSISTENCY OF THE POSTERIOR DISTRIBUTION

In this section we illuminate general aspects of the asymptotics of posterior distributions. Suppose we assign a model $P_{X^n|\theta}$ for the distribution of data that is to be observed, indexed by a parameter θ in a parameter space to which we are to assign a prior distribution W. To assess the anticipated effect of a choice of a model and prior, we ask for each θ^* what sets A to expect the posterior distribution $P\{\Theta \in A | X^n\}$ to concentrate on asymptotically if the data should happen to follow $P_{X^n|\theta^*}$. We will assume that the model distributions have joint densities $p(X^n|\theta)$ with respect to some reference measure. We shall consider only sets A and priors W for which a good local property of the mixture is assured, namely that in probability $\int_A p(X^n|\theta) W(d\theta)$ is not exponentially smaller than $p(X^n|\theta^*)$. That is, for every $\epsilon > 0$, the

probability that $\int_A p(X^n|\theta)W(d\theta)$ is greater than $e^{-n\epsilon}p(X^n|\theta^*)$ converges to one, or what turns out to be equivalent, the sequence $(1/n)\log p(X^n|\theta^*)/\int_A p(X^n|\theta)W(d\theta)$ tends to zero in probability. For this local property to hold it is sufficient that A include Kullback balls around θ^* to which W assigns positive prior probability for each $r > 0$. We characterize for which of these sets satisfying the local property does the posterior probability $P\{\Theta \in A|X_1, X_2, \ldots X_n\}$ converges to one, exponentially fast, in probability. Let A^c denote the complement of A.

Theorem: *Suppose A satisfies the indicated local property for the mixture with prior W. Then there exists $r_0 > 0$ such that $P\{\Theta \in A^c|X^n\} \leq e^{-nr_0}$ with probability tending to one, if and only if A^c can be split into two sequences of sets, say B_n and C_n, such that there exists $r_1, r_2 > 0$ with an exponentially small $W\{\Theta \in B_n\} \leq e^{-nr_1}$ and there exists a critical set S_n with $P\{(X^n) \in S_n|\theta^*\}$ converging to zero and having uniformly exponentially small probabilities of error in a test against C_n, that is, $\sup_{\theta \in C_n} P\{(X^n) \in S_n^c|\theta\} \leq e^{-nr_2}$.*

We remark that the above result also holds if $A = A_n$ is allowed to depend on n (while retaining the local property). It also holds if we ask for asymptotics when the data follow a density $p^*(X^n)$ that is close to the family in the sense of satisfaction of the local property that $\int_A p(X^n|\theta)W(d\theta)/p^*(X^n)$ is not exponentially small in probability, in which case one writes $p^*(X^n)$ in place of $p(X^n|\theta^*)$ above.

Thus consistency of the posterior distribution requires global conditions on model and prior. The prior probability of a set of "bad" models, outside which there is a uniformly consistent test, must be very small.

The above Theorem, proved in the next section, is an extension of an result of Schwartz (1965), who showed that existence of a uniformly consistent test against A^c is a sufficient condition for consistency of the posterior in the case that the local property holds. The result given here is alluded to in Barron (1986, 1989) and Barron, Schervish, and Wasserman (1998), but the proof was unpublished.

The necessary and sufficient conditions are developed to deal with the phenomenon that there does not exist a uniformly consistent test against the complement of a ball in any of the usual "metrics" for densities such as L_1, Hellinger, or Kullback-Leibler (Barron 1989). In smooth parametric cases one can be rescued by a local equivalence of Kullback-Leibler, Euclidean, and weak convergence topologies (the so-called "soundness" condition in Clarke and Barron 1990). However, to deal with nonparametric settings one needs for posterior consistency to carefully build in prior negligibility of the set of distributions that have a high degree of non-regularity. Relevant to these considerations is the result in Barron (1989) in an i.i.d. setting that for any sequence of partitions π_n there exists a uniformly consistent test (with uniformly exponentially small probabilities of error) in a test between a distribution P and the set $\{Q : \sum_{A \in \pi_n} |P(A) - Q(A)\} > \epsilon\}$ of distributions in the complement of a π_n-variation if and only if an effective cardinality of π_n is not of order larger than $O(n)$, where n is the sample size. Armed with this characterization, one can for instance take the case of densities on the unit interval and π_n the partition into n equal width intervals, and for each density q let q_{π_n} be the corresponding density that is piecewise constant on the cells in the partition. If for each ϵ there is exponentially small prior probability of the set of irregular densities q with $||q - q_{\pi_n}||_1 > \epsilon$, then the posterior distribution asymptotically concentrates on the L_1 neighborhood of any density p^* for which the prior assigns positive mass to Kullback neighborhoods.

Examples of models and priors that satisfy these consistency conditions include infinite order exponential family models with certain decay rates on the distribution of the parameters, priors that make the density function the normalization of the exponential of a Brownian motion or certain other Gaussian processes, and certain prior tree models for recursive assignment of

probabilities to cells in a refining sequence of partitions (Barron 1988, Barron, Schervish, and Wasserman 1998). In each of these cases one obtains consistency for all p^* that have finite Kullback divergence from a reference (e.g. uniform) measure.

9. PROOF OF THE CONSISTENCY THEOREM

The proof of the consistency theorem in the previous section is based on the following two lemmas that we extract from the technical report Barron (1988). In accordance with the notation there, we let $m(X^n) = \int p(X^n|\theta)W(d\theta)$ denote the mixture density and $m(X^n, A) = \int_A p(X^n|\theta)W(d\theta)$ the restriction of the mixture to sets of parameters A. The posterior distribution is $W(A|X^n) = m(X^n, A)/m(X^n)$ defined for X^n with positive $m(X^n)$. The Theorem does allow the models, the parameter set Θ, and the prior to all change with n, though such freedom is outside standard Bayes practice and not permitted in our examples, so we will not add a subscript n to W and Θ.

To prove that $W(A^c|X^n) = m(X^n, A^c)/m(X^n)$ is exponentially small with high probability when X^n follows a density $p^*(X^n)$, we follow the usual tactic as in Schwartz or Berk (1966,1970) to demonstrate that $m(X^n, A^c)/p^*(X^n)$ is exponentially small and that $m(X^n)/p^*(X^n)$ is not. The latter is a condition called merging of M^n and P^n in Barron (1986, 1988) and since $m(X^n) = m(X^n, A) + m(X^n, A^c)$ it is akin to (and in particular implied by) the local condition that $m(X^n, A)/p^*(X^n)$ not exponentially small with high probability. We use P^* to denote the hypothetical distribution on X^n at which we ask whether the posterior concentrates on sets A_n (usually taken to be a sequence of neighborhoods of P^*).

The lemmas use the following conditions for sequences of parameter sets A_n, B_n, C_n with $A_n \cup B_n \cup C_n = \Theta$ and constants a_n, b_n, c_n.

(a) Merging: $\lim_{n\to\infty} P^*\{m(X^n)/p(X^n) \geq a_n\} = 1$.

(b) Prior negligibility of B_n: $W(B_n) \leq b_n$.

(c) Existence of a uniformly consistent test against C_n: that is, for some measurable set S_n of X^n,

$$\lim_{n\to\infty} P^*\{X^n \in S_n\} = 0 \qquad \text{and} \sup_{\theta\in C_n} P\{X^n \in S_n^c|\theta\} \leq c_n.$$

Lemma 6: *Sufficiency. Suppose conditions (a), (b) and (c) are satisfied with* $\lim b_n = \lim c_n = 0$ *and let* $r_n = (b_n + c_n)/a_n$, *then for all* $\delta > 0$,

$$\limsup P^*\{W(A_n^c|X^n) > r_n/\delta\} \leq \delta.$$

Thus $W(A_n^c|X^n)$ *is of order* r_n *in probability.*

To prove the sufficiency of the conditions in the Theorem, use Lemma 6 with $b_n = e^{-nr_1}$, $c_n = e^{-nr_2}$, $a_n = e^{-n\epsilon}$ and $\delta = e^{-n\Delta}$ for positive ϵ, Δ with $\epsilon + \Delta < \min\{r_1, r_2\}$. Then r_n/δ_n tends to zero exponentially fast.

Proof of Lemma 6: The posterior probability satisfies

$$W(A_n^c|X^n) = \frac{m(X^n, A^c)}{m(X^n)} = \frac{m(X^n, A^c)/p^*(X^n)}{m(X^n)/p^*(X^n)}.$$

Consider the numerator. Let E_n be the event that $m(X^n, A_n^c)/p(X^n)$ is greater than $(b_n + c_n)/\delta$ and use the bound $P^*(E_n) \leq P^*(E_n \cap S_n^c) + P^*(S_n)$. Then successively applying Markov's inequality, the Fubini theorem for nonnegative integrands, the inclusion of A_n^c in $B_n \cup C_n$ and

(b) and (c), we have

$$\begin{aligned} P^*(E_n \cap S_n^c) &\le \frac{\delta}{b_n + c_n} \int_{S_n^c} m(X^n, A_n^c)/p(X^n) P^*(dX^n) \\ &\le \frac{\delta}{b_n + c_n} \int_{A_n^c} P(S_n^c|\theta) W(d\theta) \\ &\le \frac{\delta}{b_n + c_n} \left(\int_{B_n} W(d\theta) + \int_{C_n} P(S_n^c|\theta) W(d\theta) \right) \\ &\le \frac{\delta}{b_n + c_n} (b_n + c_n) = \delta. \end{aligned}$$

Using $P^*(S_n) \to 0$ this implies that $\limsup P^*(E_n) \le \delta$.

Finally consider the denominator. By condition (a) the event that $m(X^n)/p^*(X^n)$ is less than a_n has probability which tends to zero. The results for the numerator and denominator are combined using the union of events bound. This completes the proof of Lemma 6.

Lemma 7: *Necessity. If* $\lim P^*\{W(A_n^c|X^n) > r_n\} = 0$ *for some sequence of constants* r_n, *then for any* b_n, c_n *with product* $b_n c_n \ge r_n$, *there are sets* B_n, C_n *partitioning* A_n^c *such that conditions (b) and (c) are satisfied.*

To prove the necessity of the conditions in the Theorem, use Lemma 7 with $r_n = e^{-nr_0}$ and $b_n = c_n = e^{-nr_0/2}$. Then Lemma 7 provides the sets B_n and C_n with the desired properties. Thus together these Lemmas complete the proof of the Theorem.

Proof of Lemma 7: Set $S_n = \{X^n : W(A_n^c|X^n) > r_n\}$ which by the assumption of the Lemma satisfies $P^*(S_n) \to 0$. We note that, in S_n^c, the mixtures satisfy $m(X^n, A_n^c) \le r_n m(X^n)$.

Let $C_n = \{\theta \in A_n^c : P(S_n^c|\theta) \le c_n\}$ and $B_n = \{\theta \in A_n^c : P(S_n^c|\theta) \le c_n\}$. Then C_n clearly satisfies condition (c). Moreover, by Markov's inequality and Fubini's Theorem, the set B_n has prior probability satisfying

$$\begin{aligned} W(B_n) &\le \frac{1}{c_n} \int_{A_n^c} P(S_n^c|\theta) W(d\theta) \\ &= \frac{1}{c_n} \int_{A_n^c} m(X^n, A_n^c) \lambda(dX^n) \\ &\le \frac{r_n}{c_n} \int_{A_n^c} m(X^n) \lambda(dX^n) \\ &\le \frac{r_n}{c_n} \le b_n, \end{aligned}$$

where $\lambda(dX^n)$ is the measure dominating the family of distributions $P_{X^n|\theta}$. So condition (b) is satisfied. This completes the proof of Lemma 7.

10. NEURAL NET BOUNDS

In this section we use single hidden layer sigmoidal network models as an example setting for presentation of resolvability bounds on cumulative risk of Bayes predictive estimators.

We will consider both dichotomous response models and Gaussian error models in which the conditional distribution for the response Y given input $X = x$ has mean function $f(x)$ which we model using a neural net. In both cases there will be observations of $(X_i, Y_i)_{i=1}^N$. The inputs X_i will be i.i.d. with an arbitrary and possibly unknown distribution P_X on a given

bounded convex set B (such as the cube $[-1,1]^d$). The risk bounds we give will hold uniformly over all such P_X.

For the dichotomous response case we have $Y_i \in \{-1,1\}$, with probability of getting a 1 equal to $1/2+f(X_i)/2$. Here $f(x)$ represents the difference of the probability of getting a 1 and getting a -1, when $X=x$. For the sake of symmetry we are putting the Bernoulli distribution on $\{-1,1\}$. We will assume in this dichotomous response case that $|f(x)| \leq 1-\alpha$ is strictly less than 1. If necessary this can be arranged by mixing with a coin flip with probability α.

For the Gaussian error model we have $Y_i = f(X_i)+e_i$ where the e_i are i.i.d. Normal$(0,\sigma^2)$.

Consider the neural net model

$$f_m(x,\theta) = \sum_{j=1}^{m} c_j \psi(a_j \cdot x)$$

parameterized by $\theta = (a_j, c_j)_{j=1}^m$ with internal weight vectors a_j in R^{d+1} and external weights c_j, where $\psi(u)$ is an odd-symmetric sigmoid such as the hyperbolic tangent or $2\phi(u)-1$ where $\phi(u) = e^u/(1+e^u)$ is the logistic sigmoid. From the odd-symmetry of ψ, we restrict the c_j to be positive, without loss of generality. For simplicity an auxiliary coordinate of x is set to 1 so that the internal weights parameterize the location as well as the orientation and gain of the sigmoids. In the dichotomous response case we will clip the magnitude of $f_m(x,\theta)$ to be not greater than $1-\alpha$.

For the function f it is assumed to have a spectral norm $C_{f,B}$ which for now is assumed to be not greater than some given v. Here $C_{f,B} = \int |\omega|_B \tilde{F}(d\omega)$ is a first moment of the Fourier magnitude distribution $\tilde{F}$ and $|\omega|_B = \sup_{x\in B}|\omega \cdot x|$ is the norm of the frequency vector that is dual to the domain B for the variable X. The consequence of this assumption (established in Barron 1993) that we use is that there exists an approximation $f_m^*(x) = \sum_{j=1}^m c_j^* \psi(a_j^* \cdot x)$, with $\sum_{j=1}^m |c_j^*| \leq v$ and $|a_j^*|_B \leq \tau_m$ where τ_m is of order $\sqrt{m}\log m$, achieving

$$||f - f_m^*||^2 \leq \frac{(2v)^2}{m}$$

where the norm of the approximation error is taken in $L_2(P_X)$. Here the exterior weights may be fixed at $c_j = v/m$. This approximation bound holds more generally assuming that f/v is in the closure of the convex hull of signum functions. We make the narrower assumption of bounded spectral norm in order to have control on the magnitudes of the internal weights a_j in the model.

Let $P_{X^N,Y^N|f}$ denote the distribution of the sample $(X_i,Y_i)_{i=1}^N$ with the (unknown) true target function f, and let $P_{X^N,Y^N|f_{m,\theta}}$ denote the corresponding distribution with $f(x)$ replaced by members of the approximating family $f_{m,\theta}(x) = f_m(x,\theta)$. Let W be a prior distribution that we assign to θ and let $P^{(W)}_{X^N,Y^N}$ denote the resulting mixture.

Our information-theoretic analysis involves examination of the total relative entropy $D(P_{X^N,Y^N|f}, P^{(W)}_{X^N,Y^N})$ which is the cumulative relative entropy risk of the Bayesian predictive distributions. The resolvability bound gives for any subset A of the parameter space

$$\frac{1}{N} D(P_{X^N,Y^N|f}, P^{(W)}_{X^N,Y^N}) \leq \max_{\theta\in A} D(P_{X,Y|f}, P_{X,Y|f_{M,\theta}}) + \frac{1}{N}\log\frac{1}{W(A)}.$$

For the Gaussian error model $D(P_{X,Y|f}, P_{X,Y|f_{m,\theta}}) = \frac{1}{2\sigma^2}||f-f_{m,\theta}||^2$, and for the dichotomous response model (using the L_1^2 and Chi-square bounds on D)

$$\frac{1}{2}||f-f_{m,\theta}||^2 \leq D(P_{X,Y|f}, P_{X,Y|f_{m,\theta}}) \leq \frac{1}{\alpha}||f-f_{m,\theta}||^2.$$

Thus our L_2 approximation bounds are ready made to bound the resolvability. The resolvability bounds for the cumulative risk of the Bayes estimators are comparable to that which was given for constrained least squares estimators in Barron (1994).

At a suitable $\theta^* = (a_j^*)_{j=1}^m$ depending on f, with norms bounded by $|a_j^*|_B \leq \tau_m$, the approximation error $||f - f_{m,\theta^*}||$ is bounded by $2v/\sqrt{m}$. Now take A to be the neighborhood of θ^* defined by $A = \{\theta : |a_j - a_j^*|_B \leq 1/\sqrt{m}, j = 1, 2, ..., m\}$, and use the triangle inequality and the fact that the sigmoid ψ is Lipshitz with $|\psi(u) - \psi(u')| \leq 2|u - u'|$ to obtain, for θ in A, that the approximation error $||f - f_{m,\theta}||$ is bounded by $||f - f_{m,\theta^*}|| + 2v/\sqrt{m}$ which is not greater than $4v/\sqrt{m}$. As a consequence of these bounds we have that

$$\frac{1}{N} D(P_{X^N,Y^N|f}, P^{(W)}_{X^N,Y^N}) \leq \frac{16v^2}{cm} + \frac{1}{N} \log \frac{1}{P\{\theta \in A\}},$$

where $c = 2\sigma^2$ in the Gaussian regression case, and $c = \alpha$ in the dichotomous regression case.

It remains to lower bound $P\{\theta \in A\}$ for a specific choice of the prior. Taking for instance a prior that makes the a_j independently uniformly distributed on $\{|a_j|_B \leq \tau_m + 1/\sqrt{m}\}$ in R^{d+1}, we have $P(A) = 1/(\sqrt{m}\tau_m + 1)^{m(d+1)}$. Consequently,

$$\begin{aligned} \frac{1}{N} D(P_{X^N,Y^N|f}, P^{(W)}_{X^N,Y^N}) &\leq \frac{16v^2}{cm} + \frac{m(d+1)}{N} \log(\sqrt{m}\tau_m + 1) \\ &= O\left(v\left(\frac{d \log N}{N}\right)^{1/2}\right) \end{aligned}$$

for $m \sim v(N/(d \log N))^{1/2}$.

Note that the second term in the bound involves the ratio of the parameter dimension $k_m = m(d+1)$ and the sample size N. Thus the bound is similar to the familiar squared approximation error plus parameter dimension divided by the sample size as in section 6.

The neural net model has a particularly nice flexibility of approximation to achieve the indicated accuracy using only order m times d parameters. In contrast, linear approximation requires exponentially many terms in d to achieve comparable accuracy for functions of bounded spectral norm (Barron 1993).

Recall that the relative entropy distance between the joint distributions is related to an average relative entropy distance between $f(x)$ and the Bayes estimates $\hat{f}_{n,Bayes}(x) = \int f_m(x,\theta)$ $p(\theta|X^n, Y^n)d\theta$, averaging over samples of size $n = 0, 1, ..., N-1$. Indeed, by the chain rule $\frac{1}{N} D(P_{X^N,Y^N|f}, P^{(W)}_{X^N,Y^N}) = \frac{1}{N} \sum_{n=0}^{N-1} ED(P_{X,Y|f}, P_{X,Y|f_{n,Bayes}})$. Let the Cesaro average of the Bayes estimates be $\hat{f}_N(x) = \frac{1}{N} \sum_{n=0}^{N-1} \hat{f}_{n,Bayes}(x)$. Then by the convexity of the relative entropy and its relationship to the squared L_2 norm, we conclude with the following bound on the mean squared error,

$$\begin{aligned} \frac{1}{\underline{c}} E||f - \hat{f}_N||^2 &\leq ED(P_{X,Y|f}, P_{X,Y|\hat{f}_N}) \\ &\leq \frac{1}{N} \sum_{n=0}^{N-1} ED(P_{X,Y|f}, P_{X,Y|f_{n,Bayes}}) \\ &= O\left(v\left(\frac{d \log N}{N}\right)^{1/2}\right), \end{aligned}$$

where $\underline{c}$ is $2\sigma^2$ in the Gaussian regression case and 2 in the dichotomous regression case.

If, as is usually the case, a bound on the spectral norm is not known in advance, one can incorporate in the prior distribution the parameter v for the sum of the external coefficients c_j. Moreover, one can mix with a prior various size models m. By such strategies, one can achieve accuracy given by the resolvability bound $C_{f,B}(d \log N/N)^{1/2}$, without prior knowledge of what size network is best. See also the discussion on model selection and mixing in Section 6.

This completes the information-theoretic proof of the accuracy of neural net estimators based on the Bayesian predictors. As a consequence of these bounds, it is sufficient to have a polynomially bounded sample size to obtain an accurate estimate of a target function with a polynomially bounded spectral norm.

The analysis of Bayes posterior mean estimates rather than optimization of penalized empirical risk is very much motivated by interest in computational issues of estimation. The idea is that while the multimodality of the empirical risk surfaces creates a major obstacle to reliable optimization, there remains the possibility to obtain Monte Carlo computations of posterior means by sampling from the posterior distribution and averaging $f_m(x, \theta)$. Various Markov chains are designed for this purpose in which the posterior distribution plays the role of the target stationary distribution, but it remains to be seen whether there is a satisfactory form of rapid convergence to stationarity suitable for accurate Monte Carlo averages for these multimodal models.

REFERENCES

Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation* **9**, 349–368.

Barron, A. R. (1986). Discussion on "The consistency of Bayes estimates" by Diaconis and Freedman. *Ann. Statist.* **14**, 26–30.

Barron, A. R. (1987). Are Bayes rules consistent in information? In *Problems in Communications and Computation* (Cover and Gopinath, eds.) New York: Springer, 85–91.

Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. *Tech. Rep.* **7**, University of Illinois.

Barron, A. R. (1989). Uniformly powerful goodness of fit tests. *Ann. Statist.* **17**, 107–124.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Information Theory* **39**, 930–945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning* **14**, 115–133.

Barron, A. R., Birgé, L. and Massart, P. (1998). Risk bounds for model selection via penalization. *Probability Theory and Related Fields* (to appear).

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Information Theory* **37**, 1034–1054 and 1738.

Barron A. R. and Hengartner, N. (1998). Information theory and superefficiency. *Ann. Statist.* (to appear).

Barron, A. R. , Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Information Theory* **44**, 2743–2760.

Barron, A. R. , Schervish, M. and Wasserman, L. (1998). The consistency of posterior distributions in nonparametric problems. *Annals of Statistics* (to appear).

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37**, 51–58.

Berk, R. H. (1970). Consistency a posteriori. *Ann. Math. Statist.* **41**, 894-906.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147.

Clarke, B. S. (1989). *Asymptotic Cumulative Risk and Bayes Risk under Entropy Loss with Applications*. Ph.D. Thesis, University of Illinois.

Clarke, B. S. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Information Theory* **36**, 453-471.

Clark, B. S. and Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Statist. Planning and Inference* **41**, 37–60.

Cover, T. M. (1990). Universal portfolios. *Mathematical Finance* **1**, 1–29.

Cover, T. M. and Ordentlich, E. (1996). Universal portfolios with side information. *IEEE Trans. Information Theory* **42**, 348-363.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Davisson, L. D. (1973). Universal noiseless coding. *IEEE Trans. Information Theory* **19**, 783–795.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *J. Roy. Statist. Soc. A* **147**, 278–292.

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian Inference. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 15–20.

Hartigan, J. (1998). The maximum likelihood prior. *Ann. Statist.* (to appear).

Haussler, D. , Kivinen, J. and Warmuth, M. (1998). Sequential prediction of individual sequences under general loss functions. *IEEE Trans. Information Theory* **44**.

Haussler, D. and Opper, M. (1997). Mutual information, metric entropy, and cumulative relative entropy risk. *Ann. Statist.* **25**, 2451–2492.

Ibragimov, I. and Hasminskii, R. (1973). On the information in a sample about a parameter. *Proc. 2nd Internat. Symp. on Information Theory.* Akademia: Kiado, Budapest, 295–309.

Rissanen J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Information Theory* **30**, 629–636.

Rissanen J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14**, 1080–1100.

Schwartz, L. (1965). On Bayes procedures. *Z. Warsch. Verw. Gebiete* **4**, 10–26.

Shtarkov, Yu. M. (1988). Universal sequential coding of single messages. *Probl. Inform. Transmission* **23**, 3–17.

Xie, Q. and Barron, A. R. (1998). Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inform. Theory* (to appear).

Yang Y. and Barron A. R. (1998a). An asymptotic property of model selection criteria. *IEEE Trans. Information Theory* **44**, 95–116.

Yang Y. and Barron A. R. (1998b). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* (to appear).

DISCUSSION

JAYANTA K. GHOSH *(Indian Statistical Institute, India, and Purdue University, USA)*

The paper provides interesting connections between Bayesian inference via entropy loss and minimum length coding and gambling. Indeed it shows asymptotically the last two problems are equivalent to the minimax entropy risk problem.

The entropy loss in Bayesian inference has been used by Lindley to define information in an experiment for a fixed prior. Bernardo has used it to define the information in a prior for a given experiment and shown how it leads to the Jeffreys and reference priors. Barron touches on these aspects and indicates how the Jeffreys prior emerges asymptotically in a minimax context.

The major new contribution of the paper is in relation to asymptotics of the entropy risk for Bayes procedures in non-parametric problems. The paper provides a lovely upper bound and shows how it can be made to converge to zero at a true density P_0 provided Schwartz's condition is satisfied, i.e, P_0 is in the Kullback-Leibler support of the prior. Of course this is only a local condition in the sense that it restricts the behavior of the prior in Kullback-Leibler neighborhoods of P_0. Using lower bounds also it can apparently be shown how a prior can be constructed to obtain optimal rates of convergence to zero for the entropy risk at P_0.

In contrast Barron points out the problems of consistency of the posterior or optimal rate of convergence of Bayes estimate of a density require global conditions on the prior. Barron provides a theorem that gives a necessary and sufficient condition for posterior consistency for various topologies on the class of densities. He also gives a counter example where Schwartz's condition holds for a particular chosen P_0 but the posterior probability of the complement of

an L_1 or Hellinger neighborhood of P_0 does not converge to zero. Since the Schwartz condition holds the entropy risk at P_0 tends to zero.

A couple of questions and comments are in order. The entropy loss needs a careful examination since typically the only prediction risk that seems relevant is that associated with prediction of X_{n+1} given $X_1, X_2, \ldots, X_n$, not the cumulative predictive risk. In particular if the former misbehaves would or should one feel happy with the convergence of the latter to zero?

Here are two related technical questions. In Barron's counter example how does the entropy risk associated with X_{n+1} behave? Secondly, suppose one takes an arithmetic mean of the posteriors given $X_1, X_2, \ldots, X_m$ for $m = 1, 2, \ldots, n$. How does this random measure behave?

Incidently, the necessary and sufficient condition for posterior consistency has been used by Ghosal *et al.* (1997) a) to prove the L_1-norm posterior consistency at various P_0's for the most popular Bayesian method of density estimation, namely, Dirichlet mixture of normals. Also, Barron's construction of minimax priors in the non-parametric case is reminiscent of methods proposed recently in Ghosal *et al.* (1996). Also relevant is a paper on optimal rate of convergence for posterior by these authors and van der Vaart, which is under preparation.

TREVOR J. SWEETING *(University of Surrey, UK)*

This will be a relatively non-technical discussion of Barron's paper, principally because I only received it shortly before my flight to Spain! As I see it, the paper makes three main contributions: it provides an analysis of the cumulative risk of Bayes prediction; it makes a proposal for the choice of prior distribution (weight function) based on the Kullback risk function; and it presents an analysis of the consistency of posterior distributions. In the information-theoretic formulation, risk is measured by Kullback divergence, and the author carries out both asymptotic and finite sample analyses of this quantity. This is an impressive piece of work, especially the bounds obtained in Section 4. The proposal relating to the choice of prior I find somewhat less convincing however, and I will return to that in the next paragraph. Finally, the result on the consistency of posterior distributions is also impressive, especially as it succeeds in characterising posterior consistency (under the given local property). The counterexample in Section 7.2 is particularly strong, since the posterior probability of the 'bad sets' actually tends to one. A connection with the work in the previous sections of the paper is that it is possible to have reasonable predictive performance (as measured by Kullback divergence) even when the posterior distribution behaves very badly.

The author nicely motivates the information-theoretic approach, including useful review sections on data compression and gambling. For purposes of statistical analysis, I think that the information-theoretic formulation does make Bayesian sense, at least from an operational point of view. Firstly, regret is defined as $\log\{p(X)/q(X)\}$, which clearly behaves in a sensible way as an operational loss function $L(q, X)$. Secondly, although I do not believe that *estimating* the predictive density $p(\cdot \mid \theta)$ really makes much sense from a Bayesian point of view, the information-theoretic approach does at least give rise to the Bayesian predictive density, since estimation is based on a proper scoring rule.

I take it to be a major proposal in the paper that the asymptotic form of the cumulative risk given in Section 3 is a useful way of choosing the prior density w. Specifically, the cumulative risk asymptotically satisfies

$$D(P_{X^N \mid \theta}, P_{X^N}) = \frac{k}{2} \log \frac{N}{2\pi e} + a(\theta) + o(1)$$

where $a(\theta) = \log\{|I(\theta)|^{1/2}/w(\theta)\}$. Initially take an M-closed view (*cf.* Bernardo and Smith, 1994). Being accustomed to more mainstream Bayesian thinking, I find it more natural to think

about the specification of w directly and then to study the associated predictive performance, rather than vice versa. However, even if I do this, how do I really think about my 'desired form' for $a(\theta)$? This might make sense if $a(\theta)$ were related to a realistic loss function for the problem in hand, but of course it is purely operational. Furthermore, if I am comfortable with my subjective assessment of $w(\theta)$, then I cannot see why I would be unhappy with the implied form of $a(\theta)$. It clearly behaves in the right way for me, since it implies small risk where I judge $w(\theta)$ to be high, and vice versa. My conclusion is that I need to see a real worked example, which might just help my understanding! Finally, note that even if one takes an M-open view, it still makes sense to subjectively assess prior beliefs about θ as conditional probabilities on the given parametric subfamily.

It is instructive to compare the results in this paper with performance analysis via *coverage* properties. The rationale is that such results can provide additional assurance in a Bayesian analysis, at least in a long-run frequency sense. This is especially important when the prior is poorly specified or understood. Moreover, such an analysis can help identify those frequentist procedures that have some reasonable Bayesian interpretation (and those that do not!). I shall take the simplest setting of a sequence of independent and identically distributed random variables $X_1, X_2, \ldots$, a real parameter θ, and suitably regular likelihood and prior.

Consider first posterior analysis. Let $T = t_{w,\alpha}(X^n)$ be the upper α-quantile of the posterior distribution under the prior w, so that

$$P_w(\theta \leq T \,|\, X^n) = \alpha\,.$$

We ask to what extent is it true that

$$P(\theta \leq T \,|\, \theta) = \alpha \tag{1}$$

It is well-known that, to $O(n^{-1/2})$, equation (1) holds for all smooth w. That is, to a first order of approximation, the prior has no effect on the posterior distribution, and Bayesian and frequentist probability intervals are formally identical. To $O(n^{-1})$, however, (1) holds if and only if w is Jeffreys' prior (Welch and Peers, 1963), which provides a justification for using Jeffreys' prior as a sampling-based noninformative prior. (The situation is more complex when $\theta \in \mathcal{R}^k$.) A more complete analysis is given in Sweeting (1995).

Consider now the situation for predictive analysis. Let $U = u_{w,\alpha}(X^n)$ be the upper α-quantile of the predictive distribution under the prior w, so that

$$P_w(X^{n+1} \leq U \,|\, X^n) = \alpha\,.$$

Again, we would like to know the extent to which it is true that

$$P(X^{n+1} \leq U \,|\, \theta) = \alpha\,. \tag{2}$$

It turns out that, to $O(n^{-1})$, equation (2) holds for all smooth w. Note the superior coverage performance of the predictive distribution compared to the posterior distribution. This is a point of contact with the present paper, in which the analysis is in terms of predictive risk. It is also of some interest to investigate the next term in the expansion of (2) in order to see if it gives rise to a natural choice of prior based on higher-order predictive coverage. It turns out that there does indeed exist a unique prior w for which (2) holds to $O(n^{-2})$. However, in general the optimal prior depends on the level α! This fact also helps us to understand the difficulties from a frequentist perspective in constructing 'predictive distributions'. (Note added following the discussion: coincidentally, the above result was also reported at the present conference by M. Ghosh in his discussion of Smith (1999).)

Let me finish by making a practical comment on the need to consider frequency-based Bayesian performance. In my own contacts with end-users of statistical methods, especially engineers, I have found that there is often real interest in using Bayesian methods in order to formally incorporate, for example, engineering knowledge and experience into the statistical analysis. It is also clear, however, that many potential users seek assurance that the methods actually 'work'. I do not believe that this assurance can be provided entirely by a subjective Bayesian response. The research into performance analysis in the present paper and elsewhere can provide the necessary additional assurance sought, and can ultimately help us to encourage scientists and engineers to fully embrace Bayesian statistical thinking.

A. P. DAWID *(University College London, UK)*

As this paper elegantly demonstrates, there is a particularly neat fit between the Kullback-Leibler risk and the process of sequential probability forecasting. This in turn follows from the interpretation of K-L risk as the natural discrepancy function associated with the negative-log-density loss function, and the sequential factorization properties of the joint density. While Barron emphasizes overall expected loss in this paper, he does remark that much of the analysis can be performed at a more fundamental level, by developing bounds for the actual loss, over all possible outcome sequences. This "worst case analysis" is currently the subject of much attention by computer science/artificial intelligence workers, in the area of computational learning theory ("COLT"), and promises to provide a new and powerful approach to many problems of statistical interest. Links between the two communities are currently being built, and it is very much to be hoped that much traffic will cross them, in both directions.

There is an intermediate position between the overall expectation and worst case analyses, namely the "prequential approach" (Dawid, 1984, 1992) in which at any point in time we fix the data already gathered, and take a (conditional) expectation of the loss over the next observation only. The appropriate optimality property in this setting is "prequential efficiency", namely almost sure asymptotic optimality under cumulative prequential risk. It would be good to explore further the relationships between these three approaches.

The prequential approach makes it easier to handle loss functions other than logarithmic. For example, Skouras and Dawid (1998) show how we can extend the concept and properties of prequential efficiency to sequential point estimation under quadratic loss.

STEVEN N. MACEACHERN and L. MARK BERLINER *(The Ohio State University, U.S.A.)*

Barron's fine paper outlines conditions which guarantee various forms of consistency, and it implicitly suggests conditions which would result in inconsistency. One condition which rarely receives explicit attention is that on the likelihood: In order for estimative consistency to obtain, there must be enough "information" contained in the sampling distributions to enable us to distinguish between competing models, or between competing parameter values within a model. This condition is called identifiability when the data are independent and identically distributed from some sampling model. In more complex models, the assumption can appear in a more subtle form.

Berliner and MacEachern (1993) investigate estimation of the initial condition of a deterministic, dynamical system. They provide an example of a system based on the so-called baker's transformation, a 1-1 map which takes the unit square onto itself. The system consists of a sequence of points, observed with normal measurement error, through time. For this system, in spite of an explosion of Fisher information for one coordinate of the initial condition, the initial condition cannot be estimated consistently. Such results hold over a broad range of deterministic systems and measurement error distributions. MacEachern and Berliner (1995)

provide easily checkable conditions under which two deterministic systems can or cannot be distinguished.

In the context of modelling physical processes, a more sophisticated model considers the system to be a Markov chain, observed with measurement error. We can view the deterministic dynamical system as a limiting case where the amount of noise in the evolution of the system tends to 0. Inconsistency results for the deterministic system tell us that perfect reconstruction of the history of the system is often impossible not only because of measurement error, but also because of how the system evolves.

REPLY TO THE DISCUSSION

I thank the discussants for their insightful comments on the results of the paper and for bringing attention to related issues in their work. The discussants and I are in general agreement. My reply focusses on answering their questions and expands on a few of the points that they raise.

To recap the main conclusions of the paper: The local condition of positive prior mass on information neighborhoods (what Ghosh refers to as Schwartz's condition) implies that the Kullback risk of predictive densities, taking a time average across sample sizes, tends to zero according to the index of resolvability bound which nicely reveals the rate of convergence and the dependence of the risk on the choice of prior. Convergence of the posterior distribution on the parameter is more problematic, even if the local condition is assumed to be satisfied. Necessary and sufficient conditions for posterior consistency reveal that a particular global condition (related to identifiability, but in some ways stronger) is required, namely, except for a prior negligible set, a uniformly consistent test must exist against the complement of the set on which one desires the posterior to asymptotically concentrate.

Jayanta Ghosh asks whether we should be content with the favorable asymptotics of the cumulative (or time average) prediction risk if the individual prediction risk associated with prediction of X_{n+1} given $X_1, X_2, \ldots, X_n$ misbehaves. I would say, yes, we should be prepared to incur prediction loss for a few n in favor of better cumulative performance.

Nevertheless, it remains an interesting open question to resolve whether the individual prediction risk must converge to zero under conditions favorable for the cumulative risk. Namely, if the data follow a density p in the information support of the prior, does it follow that $ED(p, \hat{p}_n)$ where $\hat{p}_n$ is the Bayes (predictive) density estimator?

In the example where the prior is based on a large number of erratic densities (of height 2 placed on each choice of half of an even number of equal-spaced cells in $[0, 1]$), we indeed found that when the data follow the uniform distribution the posterior probability of L_1 or Hellinger neighborhoods of the uniform converges to zero rather than one. Nevertheless, the predictive density (which is an average of these erratic densities with respect to the posterior distribution) does converge in information to the uniform, in the time averge sense, since the uniform is in the information support. Moreover, in response to Jayanta's question concerning this example, calculations show that the predictive density $\hat{p}_n(x)$ converges to $p(x)$ in probability for each x, and this without taking the time average. Briefly, the analysis is as follows. For each even number of intervals m one evaluates the associated predictive density. This predictive density equals 0 when the number of occupied cells Y is greater than $m/2$. Otherwise, when $Y \leq m/2$, the predictive density is 2 if x is in an occupied cell and it is $(m - 2Y)/(m - Y)$ if x is in an unoccupied cell. This value is near 1 if x is in an unoccupied cell and $m/2$ is large compared to n. Thus one finds that with high probability, uniformly over all m greater than a large multiple of n, the predictive density is close to 1. For the models with m less than a multiple of n one uses the existence of a uniformly consistent test to ensure that the contribution to the posterior mixture is negligible.

Finally, Jayanta suggests the estimator $\bar{p}_N(x) = (1/N)\sum_{n=0}^{N-1}\hat{p}_n(x)$, which is the time average of the Bayes predictive densities for a given prior W, and asks about its behavior. I am delighted to reply that it has an individual risk sequence $ED(p,\bar{p}_N)$ which is bounded by the index of resolvability. Indeed, by convexity of the Kullback-Leibler divergence, $ED(p,\bar{p}_N)$ is not greater than $(1/N)\sum_{n=0}^{N-1} ED(p,\hat{p}_n)$ which by the chain rule equals the total Kullback rate $(1/N)D(P_{X^N}, P_{X^N}^{(W)})$, so the suggested estimator has individual risk that mirrors the favorable time average risk of the Bayes predictive density. This was noted in Barron (1987) and it is used in Yang and Barron (1998) to achieve the minimax rates. In conversation, John Hartigan has suggested the geometric mean of the Bayes predictive densities, $(\prod_{n=0}^{N-1}\hat{p}_n(x))^{1/N}$, which exactly achievies risk $(1/N)\sum_{n=0}^{N-1} ED(p,\hat{p}_n)$ (by Jensen's inequality the geometric mean integrates to less than one, so normalization to produce a probability density estimate yields some reduction in this risk). Though they have delightful frequentist properties I am not sure what to make of the arithmetic and geometric means of the predictive densities from a Bayes standpoint. Decision-theoretically, there must exist a prior W_N for which the individual Kullback risk of the resulting predictive density is not larger than the bound attained by these estimators, though at present I do not see how to exhibit it. Despite these intriguing risk properties of averaged estimators, use of the predictive density associated with a fixed prior with suitable average properties seems preferable.

Trevor Sweeting questions the proposal to use the asymptotic shape $a(\theta)$ of the total Kullback risk (which in the parametric case equals $\log\{|I(\theta)|^{1/2}/w(\theta)\}$ plus a constant depending on N), to help assess the choice of prior. My point is that one is not necessarily automatically equipped with a prior as a subjective distribution on θ. Priors as well as the parameters on which they live sometimes arise only to provide procedures for certain actions concerning possible data, such as compression, gambling or prediction. In such a setting, there is nothing to base the choice of prior other than what we understand about the behavior of these procedures. Here I add to this understanding by providing the relationship between the choice of prior and the total Kullback risk. Trevor writes, "this might make sense if $a(\theta)$ were related to a realistic loss function for the problem in hand." Conveniently, the loss function is clearly realistic for compression, it is suitable for gambling when the aim is to achieve growth rate optimality, and Trevor admits it is realistic for prediction. Trevor notes that $a(\theta)$ behaves in a sensible way, it implies risk is small where $w(\theta)$ is set to be high, and vice versa.

If θ exists as an object about which we can and do have probabilistic beliefs, then, yes, it is natural to specify w directly. Then, if we are comfortable with our subjective assessment, I agree that we should be happy with the consequent form $a(\theta)$ of the total risk of the predictive density. Moreover, in this case we may be led to go beyond estimating the predictive density to estimate θ and to inquire (as we have done) about the behaviour of the posterior distribution for various possibilities for the true θ. On the other hand, if the parameter is merely an index we concoct for distributions we may use on data, the θ-centric perspective is misplaced and it is then better to center attention on distributions assigned to x (for which we have advocated the use of Bayes predictive densities) and to choose w to give the behaviour we desire.

The form $w(\theta) = |I(\theta)|^{1/2}e^{-a(\theta)}/c$ suggests consideration of priors that are in between the extremes of providing "default" asymptotic minimax procedures (when $|I(\theta)|^{1/2}$ is integrable) and "subjective" priors that ignore the behavior of the model. For example, for k-dimensional multivariate normal means (in which $|I(\theta)|^{1/2}$ is constant and not integrable), we might choose $w(\theta)$ to be any of several choices (e.g. multivariate t or Cauchy) in which $a(\theta)$ is approximately $k(1+\epsilon)\log||\theta||$ for large $||\theta||$, for some $\epsilon > 0$. Here a choice of $k\log||\theta||$ or smaller (for large $||\theta||$) would not provide integrability. This magnitude is best possible in the sense that for any

integrable prior density and for each large radius r, the risk term $a(\theta)$ must be at least $k \log ||\theta||$ for most θ of radius not exceeding r. If a subjective prior is available one might still desire a $(1/2, 1/2)$ mix of the subjective prior with such a just barely integrable prior to avoid larger than necessary regret for large magnitude θ while incurring not more than one additional bit of regret than would be achieved with exclusive use of the subjective prior.

In the oral discussion *José Bernardo* pointed out that priors of the same form $|I(\theta)|^{1/2}e^{-a(\theta)}$ arise in his reference prior formulation subject to constraints.

Philip Dawid draws further attention to the worst case analysis of regret that parallels the expected log story surprisingly closely. An advantage of this framework is that one can take a completely operational view. No distribution need exist governing data. Here I summarize the setting and results that most closely correspond to results of the paper. One simply has a family of strategies $p(X^N|\theta) = \prod_{n=0}^{N-1} p(X_{n+1}|X^n, \theta)$ for gambling, compression or prediction. With hindsight the ideal strategy that makes the most money, yields the shortest codelength, and yields the least cumulative "log-loss" of prediction corresponds to the parameter value $\hat{\theta} = \hat{\theta}(X^N)$ maximizing the total likelihood. Now at each step $n < N$ this choice depends on future X_{n+1}, so it is not realizable. One may ask for the strategy (possibly outside of the family) $q(X^N) = \prod q(X_{n+1}|X^n)$ that minimizes the worst case regret $\max_{X^N} p(X^N|\hat{\theta}(X^N))/q(X^N)$ (here it does not affect the optimization whether we put in a logarithm or not). The minimum over all choices of $q(X^N)$ that sum to 1 is the normalized maximum likelihood $p^{NML}(X^N) = p(X^N|\hat{\theta}(X^N))/C_N$ where $C_N = \sum_{X^N} p(X^N|\hat{\theta}(X^N))$ (Starkov 1988). The minimized worst case regret is this normalization constant C_N and for smooth families the asymptotics of the minimax log regret is $\log C_N = (k/2)\log N/2\pi + \log \int |I(\theta)|^{1/2} + o(1)$ which is closely related to the minimax expected log regret discussed in Section 3. Indeed, as shown in Barron, Rissanen and Yu (1998), the normalized maximum likelihood $p^{NML}(X^N)$ and the mixture with respect to Jeffreys prior $p^{Jeffreys}(X^N)$ are asymptotically indistinguishable in information, in the sense that the total divergence $D(P_{X^N}^{Jeffreys}, P_{X^N}^{NML})$ tends to zero as $N \to \infty$. Thus Bayes procedures retain a key role in the worst case regret asymptotics. To some extent one may think of the choice of the prior in the same manner as before. Laplace approximation leads to a pointwise log regret of $(k/2)\log N/2\pi + \log\{|\hat{I}(\hat{\theta})|^{1/2}/w(\hat{\theta})\} + o(1)$ where $\hat{I}(\theta)$ is the empirical Fisher information and modifications of the mixture (to deal with sequences with $\hat{I}(\hat{\theta})$ much different from $I(\hat{\theta})$) lead to log regret not larger than $(k/2)\log N/2\pi + \log |I(\hat{\theta})|^{1/2}/w(\hat{\theta}) + o(1)$ (Takeuchi and Barron 1998, where the remainder term is negligible uniformly over possible data sequences). Thus to achieve a log regret function of a constant plus $a(\hat{\theta})$, a function of $\hat{\theta}$ achieving the best value with hindsight, we use the (modified) Bayes mixture with prior $w(\theta) = |I(\theta)|^{1/2}e^{-a(\theta)}/c$ as before.

As Phil Dawid points out computational learning theorists have made substantial contributions to worst case cumulative regret analysis, deriving suitable order bounds (but without identification of constants) and taking specific advantage of Bayes-like procedures for general loss functions. See the work by Vladimir Vovk and refinements developed in Haussler, Kivinen, and Warmuth (1998). Phil also nicely points to developments involving almost sure analysis of cumulative conditional expected loss that permit quantification of efficient procedures in his prequential setting.

Steven MacEachern and Mark Berliner contribute discussion of their work on the nonidentifiability of initial conditions in certain dynamical systems. Identifiability conditions expressed through conditions on consistency of tests are essential for consistency of posterior distributions as we showed. However, identifiability is not critical for time average consistency of Bayes predictive distributions in information (because it is the predictive distribution of the data we

are after, not the values of parameters which may index nearly indistinguishable distributions). As we have seen an information support condition is enough. Provided one uses the appropriate conditional Kullback divergences in the chain rule expansion, this information consistency can be established for certain types of dependent processes following the pattern in section 4. I suspect that it is possible to use this pattern of analysis for the dynamical systems provided some non-zero noise is included in the model.

I conclude by thanking the discussants for a stimulating exchange.

ADDITIONAL REFERENCES IN THE DISCUSSION

Berliner, L.M. and MacEachern, S.N. (1993). Examples of inconsistent Bayes procedures based on observations on dynamical systems. *Statistics & Probability Letters* **17**, 355–360.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1996). Non-Informative priors via sieves and packing numbers. *Advances in Decision Theory and Applications*. (S. Panchpakesan and N. Balakrishnan, eds.) Boston: Birkhouser, 119–132.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997). Posterior consistency of dirichlet mixtures in density estimation. *Tech. Rep.*, Purdue University.

MacEachern, S. N. and Berliner, L. M. (1995). Asymptotic inference for dynamical systems observed with error. *J. Statist. Planning and Inference* **46**, 277–292.

Skouras, K. and Dawid, A. P. (1998). On efficient point prediction systems. *J. Roy. Statist. Soc. B* **60**, 765–780.

Smith, R. L. (1999). Bayesian and frequentist approaches to parametric predictive inference. *In this volume*.

Sweeting, T. J. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82**, 1–23.

Takeuchi, J.-I. and Barron, A. R. (1998). Mixture models achieving asymptotically optimal coding regret. *Tech. Rep.*, Yale University.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. B* **35**, 318–329.