The Annals of Statistics
1996, Vol. 24, No. 1, 160–173

LOCALLY UNIFORM PRIOR DISTRIBUTIONS[1]

BY J. A. HARTIGAN

Yale University

Suppose that $X_\sigma \mid \boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \sigma^2)$ and that $\sigma \to 0$. For which prior distributions on $\boldsymbol{\theta}$ is the posterior distribution of $\boldsymbol{\theta}$ given X_σ asymptotically $N(X_\sigma, \sigma^2)$ when in fact $X_\sigma \sim N(\theta_0, \sigma^2)$? It is well known that the stated convergence occurs when $\boldsymbol{\theta}$ has a prior density that is positive and continuous at θ_0. It turns out that the necessary and sufficient conditions for convergence allow a wider class of prior distributions—the *locally uniform* and *tail-bounded* prior distributions. This class includes certain discrete prior distributions that may be used to reproduce minimum description length approaches to estimation and model selection.

1. Introduction. Suppose $X_\sigma \mid \boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \sigma^2)$ given the random variable $\boldsymbol{\theta}$. Let $\phi(x) = (1/\sqrt{2\pi})e^{-x^2/2}$. If $\boldsymbol{\theta}$ has the prior distribution P, not necessarily proper, the posterior distribution function is

$$\mathbb{P}[\boldsymbol{\theta} \le t \mid X_\sigma] = \frac{\int_{-\infty}^{t} \phi((X_\sigma - \theta)/\sigma)\, dP(\theta)}{\int_{-\infty}^{\infty} \phi((X_\sigma - \theta)/\sigma)\, dP(\theta)}.$$

(The symbol $\mathbb{P}$ is used to denote a generic probability in linear functional form. The same symbol is used for both expectation and probability, identifying probability of a set with expectation of the corresponding characteristic function.)

We say the posterior distribution is asymptotically $N(X_\sigma, \sigma^2)$ if

$$\mathbb{P}[(\boldsymbol{\theta} - X_\sigma)/\sigma \le t \mid X_\sigma] \to \int_{-\infty}^{t} \phi(u)\, du$$

in X_σ-probability as $\sigma \to 0$. There are two plausible choices for the probability distribution of X_σ: the first is the marginal distribution of X_σ for each σ; the second is the conditional distribution of $X_\sigma \mid \theta_0$ for some fixed θ_0. We will use the second choice, because it has more application to standard asymptotic problems, but the first choice is also worth exploring.

For which prior distributions P, when in fact $X_\sigma \sim N(\theta_0, \sigma^2)$, as $\sigma \to 0$, is the posterior distribution asymptotically $N(X_\sigma, \sigma^2)$? [A referee has suggested the same question, when the normality of X_σ is only asymptotic: if $(X_\sigma - \theta)/\sigma$ converges weakly to $N(0, 1)$ as $\sigma \to 0$ for each choice of θ, for what prior

Received May 1992; revised January 1995.

[1] Research partially supported by NSF Grant DMS-86-17919.

AMS 1991 *subject classification.* 62A15.

Key words and phrases. Discrete prior distributions, penalized likelihood, minimum description length.

distributions P will the posterior distribution of $(\theta - X_\sigma)/\sigma$ converge in X_σ probability to $N(0,1)$, when in fact θ_0 is true? Good question.]

The posterior distribution is exactly normal when the prior is uniform over the real line. It is well known [Walker (1969)] that the convergence occurs if the prior density has a density that is positive and continuous at θ_0 and if the posterior density is proper for some choice of σ. The prior distribution then behaves like a uniform in the neighbourhood of θ_0, and contributions outside the neighbourhood are negligible.

A prior distribution P is *locally uniform* at θ_0 if

$$\frac{P\{a\sigma < \boldsymbol{\theta} - \theta_0 \le b\sigma\}}{P\{0 < \boldsymbol{\theta} - \theta_0 < \sigma\}} \to b - a \quad \text{for each } a < b \text{ as } \sigma \to 0.$$

In n-dimensions, P is locally uniform at θ_0, if for each pair of bounded nonempty open sets S_1, S_2,

$$\frac{P\{\boldsymbol{\theta} - \theta_0 \in \sigma S_1\}}{P\{\boldsymbol{\theta} - \theta_0 \in \sigma S_2\}} \to \frac{\lambda(S_1)}{\lambda(S_2)} \quad \text{as } \sigma \to 0,$$

where λ denotes Lebesgue measure.

Say that P is *tail-bounded* at θ_0 if

$$\lim_{K\to\infty} \limsup_{\sigma\to 0} \frac{\int_{|\theta-\theta_0| \ge K\sigma} \phi((\theta - \theta_0)/\sigma)\, dP(\theta)}{P\{|\boldsymbol{\theta} - \theta_0| \le \sigma\}} = 0.$$

This condition ensures that contributions to the posterior distribution from outside σ-neighbourhoods of θ_0 are negligible. The condition would not be met, for example, if P gave zero probability or too little probability to a neighbourhood of θ_0. It will be met, for example, by a prior density bounded away from infinity and zero.

It will be shown that the posterior distribution of $\boldsymbol{\theta} \mid X_\sigma$ is asymptotically $N(X_\sigma, \sigma^2)$ if and only if P is locally uniform and tail-bounded.

There are discrete prior distributions that are locally uniform and tail-bounded at all nonatoms of the prior in $[-1,1]$: for example, for each set of integers i, n, k, with $n \ge 1$, $0 < |k| < 2^n$, k odd, $P\{\boldsymbol{\theta} = k/2^n\} = 1/(n(n+1)2^n)$. The atoms are the binary fractions of finite length.

The discrete prior distributions may be used in penalized likelihood methods for parameter estimation and model selection. Suppose that $f(x \mid \boldsymbol{\theta})$ is the likelihood of the data x for a parameter $\boldsymbol{\theta}$ which may take values in a number of spaces of different dimensionalities corresponding to different models. Suppose that the prior density of $\boldsymbol{\theta}$ with respect to some measure over the several spaces is $p(\theta)$. The modal posterior density is achieved at the *penalized likelihood* maximum of $p(\theta)f(x \mid \theta)$, but the modal value will not necessarily occur within the true model consistently. Parameter spaces of higher dimensionality in the neighbourhood of the true model will compete for the location of the modal posterior density. For example if the prior density is 1 in the different spaces, the modal density will be the maximum of the likelihood, which will tend to be larger in the higher dimensional spaces.

However, locally uniform discrete priors may be chosen so that the mode of the posterior density occurs with asymptotic probability 1 in the true subspace.

Minimum description length methods using coding theory [Rissanen (1978, 1983, 1987, 1989), Wallace and Boulton (1968), Wallace and Freeman (1987) and Barron and Cover (1991)] take the possible estimates of a parameter in a particular subspace to be a grid of values about 1 standard error apart in the subspace. Such methods are not apparently consistent with Bayesian inference because the range of the parameter is determined by the data, and it appears as if a different prior is required for different sample sizes, since the standard errors depend on the sample sizes. For discrete random variables, we can express coding theory in statistical terms by identifying code length with negative log probability [as suggested by Dawid in the discussion of Rissanen (1987) and Wallace and Freeman (1987)]. The minimum code length description of data and parameters is at the maximum joint probability of data and parameters. For certain locally uniform discrete priors, the mode of the posterior density can occur only at a possible set of parameter values about 1 standard error apart. Thus the minimum description length prescription for estimation may be reproduced in a strict Bayesian framework with a single prior for all sample sizes.

2. Asymptotic normality equivalent to local uniformity and tail-boundedness.

We say the posterior distribution is asymptotically $N(X_\sigma, \sigma^2)$ if

$$\mathbb{P}\left[(\boldsymbol{\theta} - X_\sigma)/\sigma \le t \mid X_\sigma\right] \to \int_{-\infty}^{t} \phi(u)\, du$$

in X_σ-probability as $\sigma \to 0$.

Theorem 1. *Let $X_\sigma \mid \boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \sigma^2)$. Let P be any measure on the real line. The posterior distribution of $\boldsymbol{\theta} \mid X_\sigma$ is asymptotically $N(X_\sigma, \sigma^2)$, when $X_\sigma \sim N(\theta_0, \sigma^2)$ as $\sigma \to 0$, if and only if P is locally uniform and tail-bounded at θ_0.*

Proof. Let $\theta_0 = 0$ without loss of generality. The condition for asymptotic normality, after a change of variable, may be written

$$\frac{\int_{-\infty}^{t} \phi(u)\, dP(X_\sigma + \sigma u)}{\int_{-\infty}^{\infty} \phi(u)\, dP(X_\sigma + \sigma u)} \to \int_{-\infty}^{t} \phi(u)\, du$$

in X_σ probability as $\sigma \to 0$.

In this formulation, it is apparent that we need P to behave like a uniform near 0. If P were exactly uniform, the left-hand side would be identical to the right-hand side of the above equation. The form of convergence somewhat resembles weak convergence. The difference is that the limiting uniform distribution is improper and the approximating measures $dP(X_\sigma + \sigma u)$ are random.

Without loss of generality, we may choose a fixed unit normal Z and set $X_\sigma = \sigma Z$. We now prove a result for Z fixed that eliminates the randomness of the approximating measures. It is necessary to adapt weak convergence arguments to the improper limiting uniform distribution.

For each fixed Z,

$$f(t,Z) = \frac{\int_{-\infty}^{t} \phi(u)\, dP(\sigma(u+Z))}{\int_{-\infty}^{\infty} \phi(u)\, dP(\sigma(u+Z))} \to \int_{-\infty}^{t} \phi(u)\, du \quad \text{for all } t$$

if and only if P is locally uniform and tail-bounded.

If P is locally uniform, the distribution of $\boldsymbol{\theta}/\sigma$ given $|\boldsymbol{\theta}|/\sigma \le K$ converges weakly to a uniform on $[-K, K]$. Thus taking $|K| > |Z|$, $|Z+1|$, $|Z+b|$,

$$\frac{\int_0^b \phi(u)\, dP(\sigma(u+Z))}{P\{|\boldsymbol{\theta}/\sigma| \le K\}} \to \frac{\int_0^b \phi(u)\, du}{2K},$$

$$\frac{\int_0^b \phi(u)\, dP(\sigma(u+Z))}{\int_0^1 \phi(u)\, dP(\sigma(u+Z))} \to \frac{\int_0^b \phi(u)\, du}{\int_0^1 \phi(u)\, du}.$$

Choose $K > 2|Z|$. For $|u| > K$, $\frac{1}{2}|Z+u| < |u| < 2|Z+u|$. Setting $v = \frac{1}{2}(u + Z)$,

$$\int_{|u|>K} \phi(u)\, dP(\sigma(u+Z)) \le \int_{|v|>K/4} \phi(v)\, dP(2\sigma v).$$

Thus if P is tail-bounded,

$$\lim_{K\to\infty} \limsup_{\sigma\to 0} \frac{\int_{|u|>K} \phi(u)\, dP(\sigma(u+Z))}{\int_{|u|\le 1} dP(\sigma u)}$$
$$\le \lim_{K\to\infty} \limsup_{\sigma\to 0} \frac{\int_{|u|>K/2} \phi(u)\, dP(2\sigma u)}{\int_{|u|\le 1} dP(\sigma u)} = 0.$$

Since

$$\frac{\int_0^1 \phi(u)\, dP(\sigma(u+Z))}{\int_{|u|\le 1} dP(\sigma u)} > \frac{\phi(1)\int_0^1 dP(\sigma(u+Z))}{\int_{|u|\le 1} dP(\sigma u)} \to \frac{1}{2}\phi(1) \quad \text{as } \sigma \to 0,$$

$$\lim_{K\to\infty} \limsup_{\sigma\to 0} \frac{\int_{|u|>K} \phi(u)\, dP(\sigma(u+Z))}{\int_0^1 \phi(u)\, dP(\sigma(u+Z))} = 0.$$

Thus local uniformity and tail-boundedness together imply

$$\frac{\int_{-\infty}^{t} \phi(u)\, dP(\sigma(u+Z))}{\int_{-\infty}^{\infty} \phi(u)\, dP(\sigma(u+Z))} \to \int_{-\infty}^{t} \phi(u)\, du \quad \text{for all } t$$

for every Z, and so the convergence also takes place in probability for $Z \sim N(0,1)$:

$$\frac{\int_{-\infty}^{t} \phi(u)\, dP(X_\sigma + \sigma u)}{\int_{-\infty}^{\infty} \phi(u)\, dP(X_\sigma + \sigma u)} \to \int_{-\infty}^{t} \phi(u)\, du.$$

To show the converse, we use the fact that weakly convergent sequences have strongly convergent subsequences. Suppose that

$$f(t,Z) \to \int_{-\infty}^{t} \phi(u)\, du \quad \text{for all } t$$

in probability for $Z \sim N(0,1)$. Choose a subsequence σ_i on which, except for a set of Z-values of probability zero,

$$f(t,Z) \to \int_{-\infty}^{t} \phi(u)\, du \quad \text{for all } t.$$

For example, choose σ_i such that the convergence is accurate to 2^{-i}, with probability $1 - 2^{-i}$, for the set of t-values $-2^i, -2^i + 2^{-i}, \dots, 2^i - 2^{-i}, 2^i$.

Thus the convergence occurs for all t for at least one Z. This is sufficient to establish local uniformity and tail-boundedness.

For each $\varepsilon > 0$, divide the interval $[-K, K]$ into equal segments of length δ such that $\phi(x)/\phi(y) < 1 + \varepsilon$ whenever $|x - y| \le \delta$. Define $\phi^*(I) = \sup_I \phi(x)$. For any two segments I, J,

$$\begin{aligned}\limsup_{\sigma\to 0} \frac{\int_I dP(\sigma(u+Z))}{\int_J dP(\sigma(u+Z))} &\le \limsup_{\sigma\to 0} (1+\varepsilon) \frac{\phi^*(J) \int_I \phi(u)\, dP(\sigma(u+Z))}{\phi^*(I) \int_J \phi(u)\, dP(\sigma(u+Z))} \\ &\le (1+\varepsilon) \frac{\phi^*(J) \int_I \phi(u)\, du}{\phi^*(I) \int_J \phi(u)\, du} \\ &\le (1+\varepsilon)^2 \frac{\int_I du}{\int_J du} = (1+\varepsilon)^2.\end{aligned}$$

Combining the segments, for integer i,

$$\limsup_{\sigma\to 0} \frac{\int_0^{i\delta} dP(\sigma(u+Z))}{\int_{-K}^{K} dP(\sigma(u+Z))} \le (1+\varepsilon)^2 \frac{i\delta + K}{2K}.$$

Letting $\varepsilon, \delta \to 0$ gives

$$\limsup_{\sigma\to 0} \frac{\int_0^{b} dP(\sigma(u+Z))}{\int_{-K}^{K} dP(\sigma(u+Z))} \le \frac{b}{2K}$$

and a similar argument establishes the reverse inequality. Thus local uniformity is established.

Finally, tail-boundedness follows, once $K > |Z|$, by

$$\begin{aligned}\limsup_{\sigma\to 0}\frac{\int_{4K}^{\infty}\phi(u)\,dP\left(\frac{1}{2}\sigma u\right)}{\int_{|u|\le 1}dP(\sigma u)} &\le \limsup_{\sigma\to 0}\frac{\int_{K}^{\infty}\phi(u)\,dP(\sigma(u+Z))}{\int_{|u|\le 1}dP(\sigma u)}\\ &= \limsup_{\sigma\to 0}\frac{\int_{K}^{\infty}\phi(u)\,dP(\sigma(u+Z))}{\int_{|u|\le 1}dP(\sigma(u+Z))}\\ &\le \phi(0)\limsup_{\sigma\to 0}\frac{\int_{K}^{\infty}\phi(u)\,dP(\sigma(u+Z))}{\int_{|u|\le 1}\phi(u)\,dP(\sigma(u+Z))}\\ &= \phi(0)\frac{\int_{K}^{\infty}\phi(u)\,du}{\int_{|u|\le 1}\phi(u)\,du}\to 0 \quad \text{as } K\to\infty. \qquad \square\end{aligned}$$

3. A discrete locally uniform and tail-bounded prior. It is easy to show that P is locally uniform at θ_0 if P has a density at θ_0 that is continuous and positive, and P is not locally uniform if the density is continuous and positive in the neighborhood of θ_0, except for a jump at θ_0. We will exhibit a class of *discrete* locally uniform and tail-bounded distributions.

A *binary fraction* distribution is a mixture of discrete uniforms in the interval $(-1, 1)$:

$$P\left\{\boldsymbol{\theta} = \frac{k}{2^n}\right\} = 2^{-n}q(n), \qquad k \text{ odd},$$

where the *denominator* probability $q(n)$ is the probability that θ is some binary fraction with minimum denominator 2^n.

THEOREM 2. *A binary fraction distribution is locally uniform and tail-bounded at* 0 *if and only if* $q(n)/\sum_{i>n}q(i)\to 0$ *as* $n\to\infty$.

PROOF. Let $Q(n) = \sum_{i>n}q(i)$. First suppose that the binary fraction distribution is locally uniform at 0. For $p > n$, there are 2^{p-n-1} fractions having denominator 2^{-p} that are between 0 and 2^{-n}. Thus

$$P\left\{0 < \boldsymbol{\theta} \le \frac{1}{2^n}\right\} = \frac{1}{2^n}q(n) + \frac{1}{2^{n+1}}Q(n).$$

Thus, as $n\to\infty$, $P\{\boldsymbol{\theta} = 2^{-n}\}$ is negligible compared to $P\{0 < \boldsymbol{\theta} < 2^{-n}\}$ if and only if $q(n)/Q(n)\to 0$. Thus if the binary fraction distribution is locally uniform, necessarily $q(n)/Q(n)\to 0$.

Now suppose that $q(n)/Q(n) \to 0$. For k odd,

$$\begin{aligned} P\left\{0 < \boldsymbol{\theta} \le \frac{k}{2^n}\right\} &= \sum_{j=1}^{k} P\left\{\frac{j-1}{2^n} < \boldsymbol{\theta} \le \frac{j}{2^n}\right\} \\ &= \sum_{j=1}^{k} P\left\{\frac{j-1}{2^n} < \boldsymbol{\theta} < \frac{j}{2^n}\right\} + \sum_{j=1}^{k} P\left\{\boldsymbol{\theta} = \frac{j}{2^n}\right\} \\ &= kP\left\{0 < \boldsymbol{\theta} < \frac{1}{2^n}\right\} + \sum_{j=1}^{k} P\left\{\boldsymbol{\theta} = \frac{j}{2^n}\right\}. \end{aligned}$$

For fixed k, $\sum_{j=1}^{k} P\{\boldsymbol{\theta} = j/2^n\} \le k^2 2^{-n} \max_{\{n-k \le m \le n\}} q(m)$ is negligible compared to $P\{0 < \boldsymbol{\theta} < 1/2^n\} = 2^{-n-1}Q(n)$, as $n \to \infty$. Thus

$$\frac{P\{0 < \boldsymbol{\theta} \le K_1/2^n\}}{P\{0 < \boldsymbol{\theta} \le K_2/2^n\}} \to \frac{K_1}{K_2}$$

uniformly in the integers $1 \le K_1, K_2 \le K$, as $n \to \infty$. For integer K, $b > 0$ fixed, define n, K_1, K_2 for each σ by

$$\frac{K}{2^{n+1}} < \sigma \le \frac{K}{2^n}, \qquad \frac{K_1}{2^n} < \sigma \le \frac{K_1 + 1}{2^n}, \qquad \frac{K_2}{2^n} < b\sigma \le \frac{K_2 + 1}{2^n}.$$

Note that $\frac{1}{2}K - 1 \le K_1 \le K$, $\frac{1}{2}Kb - 1 \le K_2 \le Kb$,

$$\frac{P\{0 < \boldsymbol{\theta} \le K_2/2^n\}}{P\{0 < \boldsymbol{\theta} \le (K_1+1)/2^n\}} \le \frac{P\{0 < \boldsymbol{\theta} \le b\sigma\}}{P\{0 < \boldsymbol{\theta} \le \sigma\}} \le \frac{P\{0 < \boldsymbol{\theta} \le (K_2+1)/2^n\}}{P\{0 < \boldsymbol{\theta} \le K_1/2^n\}}.$$

In the limit, the bounds approach $K_2/(K_1 + 1)$ and $(K_2 + 1)/K_1$, which each differ from b by less than $(K + Kb + 1)/(\frac{1}{2}K(\frac{1}{2}K - 1))$. Since K may be chosen arbitrarily large, it follows that

$$\frac{P\{0 < \boldsymbol{\theta} \le b\sigma\}}{P\{0 < \boldsymbol{\theta} \le \sigma\}} \to b$$

as required for local uniformity at 0 (negative b are treated similarly).

To show that the binary fraction distribution is tail-bounded at 0, we need that

$$\lim_{K\to\infty} \limsup_{\sigma\to 0} \frac{\int_{|\theta| > K\sigma} \phi(\theta/\sigma)\, dP(\theta)}{P\{|\boldsymbol{\theta}| \le \sigma\}} = 0.$$

Let $\sigma = 2^{-p}$, let K be some positive integer. Consider contributions from each of the discrete uniform components of the binary fraction distribution:

$$S(n, \sigma) = \left\{\theta \,|\, |\theta| \ge K\sigma, \qquad \theta = \frac{k}{2^n}, k \text{ odd}\right\},$$

$$\int_{|\theta| \ge K\sigma} \phi\left(\frac{\theta}{\sigma}\right) dP(\theta) = \sum_n \sum_{\theta \in S(n,\sigma)} \phi\left(\frac{\theta}{\sigma}\right) p(\theta).$$

For $n > p$, since $\phi(x)$ is convex in the region $x \ge 1$,

$$\begin{aligned}\sum_{\theta \in S(n,\sigma)} \phi\left(\frac{\theta}{\sigma}\right) &= \sum_{|k2^{p-n}|>K} \phi(k2^{p-n})\\ &= 2\sum_{k=1+2^{n-p}K}^{\infty} \phi(k2^{p-n}) \le 2\int_K^{\infty} \phi(u)\,du\,2^{n-p},\end{aligned}$$

$$\begin{aligned}\sum_{n>p}\sum_{\theta \in S(n,\sigma)} \phi\left(\frac{\theta}{\sigma}\right)p(\theta) &\le 2[1-\Phi(K)]2^{-p}\sum_{n>p} q(n)\\ &= 2[1-\Phi(K)]2^{-p}Q(p).\end{aligned}$$

Since $P\{|\theta| \le \sigma\} \ge 2^{-p}Q(p)$,

$$\lim_{K\to\infty}\limsup_{\sigma\to 0}\frac{\sum_{n>p}\sum_{\theta\in S(n,\sigma)}\phi(\theta/\sigma)}{P\{|\boldsymbol{\theta}| \le \sigma\}} = 0.$$

For $n \le p$, using $\phi(k2^{p-n}) \le \phi(2^{p-n})^k$ gives

$$\begin{aligned}&\sum_{n\le p}\sum_{\theta \in S(n,\sigma)} \phi\left(\frac{\theta}{\sigma}\right)p(\theta)\\ &\quad\le 4\sum_{n\le p}\phi(2^{p-n})2^{-n}q(n)\\ &\quad= 4Q(p)2^{-p}\left[\sum_{p\ge n>p-J}\phi(2^{p-n})2^{p-n}q(n)/Q(p)\right.\\ &\qquad\qquad\left.+\sum_{n\le p-J}\phi(2^{p-n})2^{p-n}q(n)/Q(p)\right].\end{aligned}$$

Since $q(n)/Q(n) \to 0$, it follows that $Q(n)/Q(n+1) \to 1$ and so for some finite λ, $Q(n)/Q(n+1) \le \lambda$ all n, and then $q(n)/Q(p) \le \lambda^{p-n}$. Take $\varepsilon > 0$. The series $\sum_1^{\infty}\phi(2^j)2^j\lambda^j$ converges, so J may be chosen so that $\sum_J^{\infty}\phi(2^j)2^j\lambda^j < \varepsilon$. Next, for p large enough, $\sum_{p-J}^{p} q(n)/Q(p) < \varepsilon$. Then

$$\sum_{n\le p}\sum_{\theta \in S(n,\sigma)} \phi\left(\frac{\theta}{\sigma}\right)p(\theta) \le 8\varepsilon Q(p)2^{-p}.$$

Since this result holds for every $\varepsilon > 0$, for p sufficiently large,

$$\lim_{K\to\infty}\limsup_{\sigma\to 0}\frac{\sum_{n\le p}\sum_{\theta\in S(n,\sigma)}\phi(\theta/\sigma)p(\theta)}{P\{|\boldsymbol{\theta}| \le \sigma\}} = 0.$$

Combining the contributions from $n > p$ and $n \le p$ establishes tail-boundedness. □

One convenient *binary fraction* distribution sets $q(n) = 1/(n(n+1))$. Another possible choice of denominator distribution is the long-tailed "universal prior on the integers" [Rissanen (1983)]:

$$\log q(n) = C - \log n - \log\log n - \log\log\log n + \cdots,$$

where the summation is carried out as long as the terms remain nonnegative, which will produce locally uniform, tail-bounded distributions.

A useful improper binary fraction distribution that does not satisfy the local uniformity property sets $q(n) = 1$ all n. This prior distribution gives infinite probability to every open interval. Nevertheless the posterior probability at any binary fraction may be computed, and the posterior mode is simply computed and behaves reasonably well compared to maximum likelihood.

A distribution dense on the real line is the *binary rational* distribution which is nearly the product of two binary fraction distributions; a binary rational of type $[l, k]$ has l binary digits before the decimal point, beginning with a 1, and k binary digits after the decimal point, ending with a 1. It may be positive or negative. The probability of each binary rational of type $[l, k]$ is $1/(2^{l+k-1}l(l+1)k(k+1))$. This distribution is locally uniform and tail-bounded at nonatoms.

Another possible distribution on the rationals assigns the probability $C/(k^2l^2)$ to the rational k/l, where k, l have no common factors. I conjecture this distribution to be locally uniform in the neighbourhood of any irrational, but the calculations are abstruse.

4. Penalized likelihood estimation using binary fraction priors. Suppose we wish to estimate θ using the 0–1 loss function $L(d, \theta) = 1$ if $d \neq \theta$, $L(d, \theta) = 0$ if $d = \theta$. For a discrete prior distribution, the Bayes estimate is the penalized likelihood estimate $\hat{\theta}$ maximizing the posterior density or penalized likelihood $p(\theta)\phi((X_\sigma - \theta)/\sigma)$.

THEOREM 3. *For the binary fraction prior P, with nonincreasing denominator probability q, when $|X_\sigma| \leq 1$, the penalized likelihood estimate $\hat{\theta}$ is achieved by some $\theta = k/2^n$, k odd, where $2^{-n} > \sigma/2$.*

PROOF. Let $\theta_{k,n} = k/2^n$ for odd integer k. We wish to find $\hat{\theta}$ minimizing

$$L(\theta) = \tfrac{1}{2}(X_\sigma - \theta)^2/\sigma^2 - \log p(\theta).$$

For each fixed n, since the neighbouring values of $\theta_{k,n}$ are separated by an interval of length 2^{1-n}, the quantity $(X_\sigma - \theta_{k,n})^2$ has a minimum value not exceeding 4^{-n}. Also at the overall optimum $\theta_{\hat{k},\hat{n}}$, the quantity $(X_\sigma - \theta_{k,n})^2$ is nonnegative. Thus

$$\begin{aligned} -\log p(\theta_{k,\hat{n}}) \leq \min_k L(\theta_{k,\hat{n}}) &\leq \min_k L(\theta_{k,\hat{n}-1}) \\ &\leq \tfrac{1}{2}4^{1-\hat{n}}/\sigma^2 - \log p(\theta_{k,\hat{n}-1}). \end{aligned}$$

Since q is nonincreasing and $2\log 2 > 1$,

$$\begin{aligned} \hat{n}\log 2 - \log q(\hat{n}) &\leq \tfrac{1}{2}4^{1-\hat{n}}/\sigma^2 + (\hat{n}-1)\log 2 - \log q(\hat{n}-1), \\ \log 2 &\leq \tfrac{1}{2}4^{1-\hat{n}}/\sigma^2, \\ \sigma/2 &< 2^{-n} \quad \text{as required.} \qquad \square \end{aligned}$$

The effect of this theorem is that, in penalized likelihood calculations, we need consider only atoms of the prior distribution separated by at least 1 standard error. Such a separation is recommended for minimum description length estimation in Wallace and Freeman (1987). Thus the binary fraction prior gives a strict Bayesian justification of minimum description length estimation procedures.

It is of interest to know how far away the penalized likelihood estimator can be from X_σ.

THEOREM 4. *For $|X_\sigma| \le 1$, q nonincreasing, $q(n)/q(1) \ge \alpha^{n-1}$ for all n, the penalized likelihood estimate $\hat{\theta}$ satisfies*

$$|X_\sigma - \hat{\theta}| \le \sigma[1 + 2\log\sigma\,\log(2/\alpha)/\log 2]^{1/2}.$$

PROOF. We compare the optimal penalized likelihood at $\hat{n}$ with the penalized likelihood for n satisfying $2^{-n} < \sigma \le 2^{1-n}$. Assume $\hat{n} \le n$, for otherwise the inequality stated follows trivially.

$$\tfrac{1}{2}\big[X_\sigma - \hat{\theta}\big]^2/\sigma^2 + \hat{n}\log 2 - \log q(\hat{n}) \le \tfrac{1}{2}4^{-n}/\sigma^2 + n\log 2 - \log q(n),$$

$$\begin{aligned}\tfrac{1}{2}\big[X_\sigma - \hat{\theta}\big]^2/\sigma^2 &\le \tfrac{1}{2}4^{-n}/\sigma^2 + (n-1)\log(2/\alpha)\\ &\le \tfrac{1}{2} + \log\sigma\,\log(2/\alpha)/\log 2,\end{aligned}$$

from which the desired inequality follows. □

Thus the penalized likelihood estimate may be $O(\sigma\sqrt{\log\sigma})$ from X_σ.

THEOREM 5. *For $|X_\sigma| \le 1$ and if $q(n)/q(n-1) \ge 2/e$ for each n, the penalized likelihood estimate $\hat{\theta}$ satisfies*

$$|\hat{\theta}| \le 4\max[\sigma, |X_\sigma|].$$

PROOF. Suppose that $2^{-m-1} < \max[\sigma, |X_\sigma|] \le 2^{-m}$. Assume that $X_\sigma \ge 0$ without loss of generality. If $\hat{\theta} \le 2^{-m}$, the stated inequality is satisfied. If not, the optimal θ will be of form 2^{-p} for $p < m$. The penalized likelihood at $\theta = 2^{-p}$ will exceed the penalized likelihood at $\theta = 2^{-m}$, so

$$\begin{aligned}&\tfrac{1}{2}[X_\sigma - 2^{-p}]^2/\sigma^2 - \tfrac{1}{2}[X_\sigma - 2^{-m}]^2/\sigma^2\\ &\qquad \le (m-p)\log 2 + \log q(p) - \log q(m),\\ &[2^{-p} - 2^{-m}]^2 \le 2\sigma^2(m-p)[\log 2 - \log(2/e)],\\ &[2^{m-p} - 1]^2 \le 2(m-p), \qquad m - p \le 1.\end{aligned}$$

Thus $\hat{\theta} = 2^{1-m}$, which implies that $\hat{\theta} \le 4\max[\sigma, |X_\sigma|]$ as required. □

Similar bounds are obtained if $q(n)$ decreases at a different rate. The effect of this theorem is that $\hat{\theta} = O_p(\sigma)$ whenever the true value is 0. I would expect that a similar bound holds for other choices of true value.

5. Model selection. Let $f(X_\sigma)$ denote the marginal density $\int (1/\sigma)\phi((X_\sigma - \theta)/\sigma)\, dP(\theta)$. The asymptotic behaviour of this density has been used in model selection to justify the Schwarz (1978) "correction factor" to adjust for differing dimensionalities of competing models. The usual asymptotics assume a prior density p that is continuous and positive near the true value θ_0. In this case, $f(X_\sigma)/p(\theta_0) \to 1$ in probability, if $X_\sigma \sim N(\theta_0, \sigma^2)$. We wish to develop the asymptotic behaviour for locally uniform and tail-bounded priors.

THEOREM 6. *If P is locally uniform and tail-bounded at θ_0, then*

$$\log f(X_\sigma)/\log \sigma \to 0$$

in probability when $X_\sigma \sim N(\theta_0, \sigma^2)$.

PROOF. Let $\theta_0 = 0$ and set $X_\sigma = \sigma Z$, where $Z \sim N(0,1)$. We first show that

$$\frac{P\{0 < \boldsymbol{\theta} < \sigma\}}{\sigma f(\sigma Z)} \to 1.$$

Following the proof of Theorem 1, for each fixed Z, t,

$$\mathbb{P}\{\boldsymbol{\theta} < \sigma(Z + t) \mid X_\sigma = \sigma Z\} \to \Phi(t).$$

Thus

$$\frac{\int_0^\varepsilon \phi(u)\, dP(\sigma(Z+u))}{\sigma f(\sigma Z)} \to \Phi(\varepsilon) - \Phi(0).$$

Since $\phi(u) < \phi(\varepsilon)$ for $0 < u < \varepsilon$,

$$\limsup_{\sigma \to 0} \phi(\varepsilon) \frac{P\{\sigma Z < \boldsymbol{\theta} < \sigma(Z + \varepsilon)\}}{\sigma f(\sigma Z)} \le \phi(0)\varepsilon.$$

By local uniformity of P, and taking ε arbitrarily small,

$$\frac{\limsup_{\sigma \to 0} P\{0 < \boldsymbol{\theta} < \sigma\}}{\sigma f(\sigma Z)} \le 1.$$

A similar argument reverses the inequality, so that

$$\frac{P\{0 < \boldsymbol{\theta} < \sigma\}}{\sigma f(\sigma Z)} \to 1.$$

To prove the theorem, it suffices to show that $\log(P\{0 < \boldsymbol{\theta} \le \sigma\}/\sigma)/\log \sigma \to 0$. Because P is locally uniform,

$$\frac{P\{0 < \boldsymbol{\theta} \le \sigma/2\}}{P\{0 < \boldsymbol{\theta} \le \sigma\}} \to \frac{1}{2} \quad \text{as } \sigma \to 0.$$

For each $\varepsilon > 0$, choose σ_0 so that

$$\frac{2P\{0 < \boldsymbol{\theta} \leq \sigma/2\}}{P\{0 < \boldsymbol{\theta} \leq \sigma\}} > 2^{-\varepsilon}$$

for $\sigma < \sigma_0$. Then, applying the inequality n times,

$$\frac{2^n P\{0 < \boldsymbol{\theta} \leq 2^{-n}\sigma_0\}}{P\{0 < \boldsymbol{\theta} \leq \sigma_0\}} > 2^{-n\varepsilon},$$

$$\frac{\log\left[P\{0 < \boldsymbol{\theta} \leq 2^{-n}\sigma_0\}/(2^{-n}\sigma_0)\right]}{\log(2^{-n}\sigma_0)} > \frac{\log\left[P\{0 < \boldsymbol{\theta} \leq \sigma_0\}/\sigma_0\right]}{\log(2^{-n}\sigma_0)} - \frac{n\varepsilon \log 2}{\log(2^{-n}\sigma_0)}.$$

The right-hand side is bounded below by -2ε for n large enough. A similar argument bounds the left-hand side above by 2ε. Thus

$$\log(P\{0 < \boldsymbol{\theta} \leq \sigma\}/\sigma)/\log\sigma \to 0 \quad \text{as } \sigma \to 0.$$

(We have shown this only for sequences of form $\sigma = 2^{-n}$, but it also follows for general sequences.) □

If the prior density has a continuous positive density at θ_0, then the stronger result $\log f(X_\sigma) \to \log p(\theta_0)$ as $\sigma \to 0$ in probability holds. However, the present theorem is strong enough to guarantee consistency of model selection for models of different dimensionality. Consider first the simplest case of a sample $X_1, \dots, X_n$ from $N(\theta, 1)$, when the prior has positive and continuous density p at θ_0. Let $f(\mathbf{X} \mid \theta)$ denote the density of the observations given θ and let $f(\mathbf{X})$ denote the corresponding marginal density of the observations. Then, with $\hat{\theta}$ the maximum likelihood estimator of θ,

$$\log\left[f(\mathbf{X})/f(\mathbf{X} \mid \hat{\theta})\right] = -\tfrac{1}{2}\log n + \log\sqrt{2\pi}\,p(\theta_0) + o_p(1).$$

Such a formula for general densities with k-dimensional parameters is given by Jeffreys (1936); it is the basis for the Schwarz (1978) correction factor, which penalizes the log likelihood in k dimensions by $-\frac{1}{2}k \log n$.

For locally uniform and tail-bounded priors, this is weakened to

$$\log\left[f(\mathbf{X})/f(\mathbf{X} \mid \hat{\theta})\right] = -\tfrac{1}{2}\log n + o_p(\log n).$$

Related formulae appear in Dawid (1984) and Rissanen [(1987), Theorem 4.1]. The leading order term in $\log n$ is the same in both approximations, and it is this term that determines consistency in model selection. For example, for X and θ two dimensional,

$$\log\left[f(\mathbf{X})/f(\mathbf{X} \mid \hat{\theta})\right] = 2\left(-\tfrac{1}{2}\log n\right) + o_p(\log n).$$

Now suppose that $\theta^* = (\theta_1, 0)$ defines a one-dimensional subspace of the original two-dimensional parameter space. Then

$$\log\left[f^*(\mathbf{X})/f(\mathbf{X} \mid \hat{\theta}^*)\right] = -\tfrac{1}{2}\log n + o_p(\log n).$$

The terms $f(\mathbf{X} \mid \hat{\theta})$ and $f(\mathbf{X} \mid \hat{\theta}^*)$ differ by $O_p(1)$ when θ_0 lies in the smaller subspace, by the standard theory of likelihood ratio tests. Thus

$$\log[f^*(\mathbf{X})/f(\mathbf{X})] = \tfrac{1}{2}\log n + o_p(\log n)$$

for locally uniform and tail-bounded priors at θ_0 in each of the parameter spaces, when θ_0 lies in the smaller subspace. Thus the posterior probability of the smaller subspace given $\mathbf{X}$ approaches 1 as $n \to \infty$.

If estimation is done by maximizing the posterior density, it is of interest to know when the value of θ maximizing $p(\theta)f(\mathbf{X} \mid \theta)$ lies in the same parameter subspace as the true value θ_0. The maximizing θ lies in the correct subspace with probability approaching 1 if an appropriate discrete locally uniform and tail-bounded prior is chosen for the various subspaces. For example, suppose that $X_1, \dots, X_n$ are two-dimensional random variables sampled from the circular normal $N(\theta, I)$, where the set Ω of θ values is the plane and the subspace $\Omega^* = \{\theta_1, \theta_2) \mid \theta_2 = 0\}$ is to be allowed for. Choose binary fraction priors for θ_1 and θ_2 holding independently. For example,

$$P\{\theta_1 = k_1/2^{n_1}, \theta_2 = k_2/2^{n_2}\} = 1/[2^{n_1+n_2}n_1(n_1+1)n_2(n_2+1)]$$

for $n_1, n_2 \geq 1$ and k_1, k_2 odd. The subspaces Ω and Ω^*, with $\Omega \supset \Omega^*$ are assumed to have prior probabilities 0.5, and the specified binary fraction priors hold within each of the subspaces.

Suppose now that the true value is $\theta = 0$. We have shown in Section 4 that both $\hat{\theta}$ and $\hat{\theta}^*$ are $O_p(1/\sqrt{n})$. The log likelihoods $\log f(\mathbf{X} \mid \hat{\theta})$ and $\log f(\mathbf{X} \mid \hat{\theta}^*)$ for the posterior modes $\hat{\theta}, \hat{\theta}^*$ therefore differ by $O_p(1)$.

Also, for Ω, the log prior density for $\hat{\theta}$ is

$$\begin{aligned}\log|\hat{\theta}_1| + \log|\hat{\theta}_2| - 2\log\big[-\log|\hat{\theta}_1|\big] - 2\log\big[-\log|\hat{\theta}_2|\big] + O_p(1)\\ = -\log n - 4\log\log n + O_p(1).\end{aligned}$$

For Ω^*, the log prior density for $\hat{\theta}^*$ is

$$\log|\hat{\theta}_1^*| - 2\log\big[-\log|\hat{\theta}_1^*|\big] + O_p(1) = -\tfrac{1}{2}\log n - 2\log\log n + O_p(1).$$

Thus the lower dimensional log posterior differs from the higher dimensional log posterior by $\frac{1}{2}\log n + 2\log\log n + O_p(1)$, which has the same first order term as the Schwarz correction factor. This means that we favour the estimate in the lower dimensional subspaces about the same using penalized likelihoods as we do using the marginal probabilities of the observations under the different models, and so we can rely on simple optimization of the penalized likelihood to identify the correct model.

REFERENCES

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

Dawid, A. P. (1984). Present position and potential developments: Some personal views, statistical theory, the prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292.

JEFFREYS, H. (1936). Further significance tests. *Proceedings of the Cambridge Philosophical Society* **32** 416–445.

RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

RISSANEN, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** 416–431.

RISSANEN, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc. Ser. B* **49** 223–239.

RISSANEN, J. (1989). *Stochastic Complexity in Statistical Enquiry*. World Scientific Publishers, NJ.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

WALKER, A. M. (1969). Asymptotic behaviour of posterior distributions. *J. Roy. Statist. Soc. Ser. B* **31** 80–88.

WALLACE, C. S. and BOULTON, D. M. (1968). An information measure for classification. *Comput. J.* **11** 185–194.

WALLACE, C. S. and FREEMAN, P. R. (1987). Estimation and inference by compact coding (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 240–265.

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520