

## 8

---

*Non-Informative Priors Via Sieves and Packing Numbers*

---

S. Ghosal, J. K. Ghosh and R. V. Ramamoorthi

*Indian Statistical Institute, Calcutta, India*

*Indian Statistical Institute, Calcutta, India*

*Michigan State University, East Lansing, MI*

**Abstract:** In this paper, we propose methods for the construction of a non-informative prior through the uniform distributions on approximating sieves. In parametric families satisfying regularity conditions, it is shown that Jeffreys' prior is obtained. The case with nuisance parameters is also considered. In the infinite dimensional situation, we show that such a prior leads to consistent posterior.

**Keywords and phrases:** Non-informative prior, sieve, packing number, Jeffreys' prior, posterior consistency

---

## 8.1 Introduction

There has been some revival of interest in non-informative and automatic priors for quick, automatic Bayesian analysis as well as for providing a sort of reference point for subjective Bayesian analysis, vide Berger and Bernardo (1992) and Bernardo (1979). Some recent references are Berger and Bernardo (1989), Tibshirani (1989), Zellner (1990). Ghosh and Mukerjee (1992), Kass and Wasserman (1992), Datta and Ghosh (1995a,b), Ghosh (1994), Datta and Ghosh (1995, 1996) and Zellner (1990). These papers deal with a single parametric model. There has also been interest in such priors when one has several nested or non-nested parametric models; see, for example, Spiegelhalter and Smith (1982), Berger and Pericchi (1994, 1996) and O'Hagan (1995). Kass and Wasserman (1996) is a recent survey on non-informative and related priors.

Except for a non-informative choice of the base measure  $\alpha(\cdot)$  for a Dirichlet process prior, very little is known about non-informative priors in non-parametric or infinite dimensional problems. In the following pages, we pro-

pose and study a few methods of generating such priors using sieves and packing numbers. We also apply our ideas to parametric models.

Suppose we have a model  $\mathcal{P}$ , equipped with a metric  $\rho$ . Our preferred metric will be the Hellinger metric. To keep the discussion simple, we will initially assume that  $\mathcal{P}$  is compact. This assumption can then be relaxed in at least some  $\sigma$ -compact cases in a standard way as indicated in Section 8.2. Our starting point is a sequence  $\varepsilon_i$  diminishing to zero and sieves  $\mathcal{F}_i$ , where  $\mathcal{F}_i$  is a finite set whose elements are separated from each other by at least  $\varepsilon_i$  and has cardinality  $D(\varepsilon_i, \mathcal{P})$ , the largest  $m$  for which there are  $P_1, \dots, P_m \in \mathcal{P}$  with  $\rho(P_j, P_{j'}) > \varepsilon_i$ ,  $j \neq j'$ ,  $j, j' = 1, \dots, m$ . Clearly given any  $P \in \mathcal{P}$ , there exists  $P' \in \mathcal{F}_i$  such that  $\rho(P, P') \leq \varepsilon_i$ . Thus  $\mathcal{F}_i$  approximates  $\mathcal{P}$  within  $\varepsilon_i$  and removal of a single point from  $\mathcal{F}_i$  will destroy this property.

In the first method, we fix the sample size  $n$ , i.e., we assume we have  $n$  i.i.d.  $X_i \sim P \in \mathcal{P}$ . We choose  $\varepsilon_{i(n)}$  satisfying (1.5) of Wong and Shen (1995) or in some other suitable way. It is then convenient to think of  $\mathcal{F}_{i(n)}$  as a discrete or finite approximation to  $\mathcal{P}$ , which is as fine as is compatible with our resource measured by the sample size  $n$ . Of course, greater the value of  $n$ , the richer the approximating set we will treat as our proxy model. In the first method, our noninformative prior is simply the uniform distribution on  $\mathcal{F}_{i(n)}$ .

This seems to accord well with Basu's (1975) recommendation in the parametric case to approximate the parameter space  $\Theta$  by a finite set and then put a uniform distribution. It is also intuitively plausible that the complexity or richness of a model  $\mathcal{F}_{i(n)}$  may be allowed to depend on the sample size. Unfortunately, it follows from Heath and Sudderth (1978) that our inference based on a sample size dependent prior cannot be coherent in their sense, if one has to accept bets given any data  $X_1, \dots, X_m$ , where  $m$  is arbitrary. We therefore consider two other approaches which are coherent at least in the compact case.

In the second approach, we consider the sequence of uniform distributions  $\pi_i$  on  $\mathcal{F}_i$  and consider any weak limit point  $\pi^*$  of  $\{\pi_i\}$  as a non-informative prior. If  $\pi^*$  is unique, it is simply the uniform distribution defined and studied by Dembski (1990). It is shown in Section 8.3 that this approach leads to Jeffreys' prior in parametric problems satisfying regularity conditions. In Ghosh and Ramamoorthi (1997), where limit points  $\pi^*$  were proposed as non-informative priors, the result in Section 8.3 was stated as a conjecture. In Section 8.3, we also consider the parametric case where  $\vartheta = (\theta, \varphi)$ ,  $\theta$  alone is the parameter of interest and the conditional prior distribution of  $\varphi$  given  $\theta$ , namely  $\pi(\varphi|\theta)$ , is given. Using the natural metric in this problem, we derive a non-informative marginal prior for  $\theta$  which is similar, but not identical with the reference prior of Bernardo (1979) or Berger and Bernardo (1989) and the probability matching prior satisfying the partial differential equation mentioned in Ghosh and Mukerjee (1992).

In the infinite dimensional case, evaluation of the limit points may prove to be impossible. However, the first approach may be used and  $\pi_{i(n)}$  may be

treated as

We r  
which pi  
 $\pi_i$ . In Se  
a certain  
which is  
Bayesian  
the third  
 $n$  at seas  
rate of c  
later.

We c  
lems. C  
other si  
continua  
will crea  
informat  
think of  
of width  
 $N(0, h)$ ,  
attentio  
paramet  
results c  
in densi

Our  
Kolmog  
also bec  
Wasser

## 8.2

Let  $K$  b  
called  $\varepsilon$   
set is ca  
be an  $\varepsilon$ -  
(or  $\varepsilon$ -ca  
bounde

Defi

treated as an approximation to a limit point  $\pi^*$ .

We now come to the third approach. Here we consider a hierarchical prior which picks up the index or hyperparameter with probability  $\lambda_i$  and then uses  $\pi_i$ . In Section 8.4, we prove in a class of infinite dimensional problems that under a certain condition on the rate of decay of  $\lambda_i$ , we obtain posterior consistency, which is a very weak form of robustness generally considered important by Bayesians. Preliminary theoretical considerations suggest that the posterior in the third approach will be close to the posterior in the first approach for large  $n$  at least when the  $\lambda_i$ 's decay suitably to make the posterior attain the optimal rate of convergence of Wong and Shen (1995). We expect to report on this later.

We offer these approaches as tentative ideas to be tried out in several problems. Computational and other considerations may require replacing  $\mathcal{F}_i$  by other sieves which need not be finite, an index  $i$  which may take value in a continuum, and distributions on  $\mathcal{F}_i$  which are not uniform. These relaxations will create a very large class of non-parametric priors, not necessarily non-informative, from which it may be convenient to elicit a prior. We would like to think of this as a fourth method. It includes such priors as a random histogram of width  $h$  introduced by Gasperini (1992), and a Dirichlet convoluted with  $N(0, h)$ , introduced by Lo (1984). These latter priors have received a lot of attention recently [see West (1992), West, Mueller and Escobar (1994)]. The parameter  $h$  here can be viewed as indexing a sieve. It is possible to obtain results on the consistency of the posterior for these and similar priors occurring in density estimation problems. These results will be presented elsewhere.

Our use of packing numbers was influenced by Wong and Shen (1995) and Kolmogorov and Tihomirov (1961). Sieves and bracketing entropy ideas have also been used in the context of posterior consistency by Barron, Schervish and Wasserman (1996).

## 8.2 Preliminaries

Let  $K$  be a compact metric space with a metric  $\rho$ . A finite subset  $S$  of  $K$  is called  $\varepsilon$ -dispersed if  $\rho(x, y) \geq \varepsilon$  for all  $x, y \in S$ ,  $x \neq y$ . A maximal  $\varepsilon$ -dispersed set is called an  $\varepsilon$ -net. An  $\varepsilon$ -net with maximum possible cardinality is said to be an  $\varepsilon$ -lattice and the cardinality of an  $\varepsilon$ -lattice is called the packing number (or  $\varepsilon$ -capacity) of  $K$  and is denoted by  $D(\varepsilon, K) = D(\varepsilon, K; \rho)$ . As  $K$  is totally bounded,  $D(\varepsilon, K)$  is finite.

Define the  $\varepsilon$ -probability  $P_\varepsilon$  on  $K$  by

$$P_\varepsilon(X) = \frac{D(\varepsilon, X)}{D(\varepsilon, K)}, \quad X \subset K. \quad (8.1)$$

It follows that  $0 \leq P_\varepsilon(\cdot) \leq 1$ ,  $P_\varepsilon(\emptyset) = 0$ ,  $P_\varepsilon(K) = 1$ ,  $P_\varepsilon(\cdot)$  is subadditive, and for  $X, Y \subset K$  which are separated at least by  $\varepsilon$ ,  $P_\varepsilon(X \cup Y) = P_\varepsilon(X) + P_\varepsilon(Y)$ .

For an  $\varepsilon$ -lattice  $S_\varepsilon$  in  $K$ , the discrete uniform distribution  $\mu_\varepsilon$  (say) on  $S_\varepsilon$  can be thought as an approximate uniform distribution on  $K$ . Because  $K$  is compact,  $\mu_\varepsilon$  will have subsequential weak limits. If all the subsequential limits are the same, then  $K$  is called uniformizable and the common limit point is called the uniform probability on  $K$ .

The following result, due to Dembski (1990), will be used in Section 8.3.

**Theorem 8.2.1 ((Dembski).)** *Let  $(K, \rho)$  be a compact metric space. Then the following assertions hold:*

- (a) *If  $K$  is uniformizable with uniform probability  $\mu$ , then  $\lim_{\varepsilon \rightarrow 0} P_\varepsilon(X) = \mu(X)$  for all  $X \subset K$  with  $\mu(\partial X) = 0$ .*
- (b) *If  $\lim_{\varepsilon \rightarrow 0} P_\varepsilon(X)$  exists on some convergence determining class in  $K$ , then  $K$  is uniformizable.*

**Remark 8.2.1** It is often not possible to evaluate the limit points, particularly in the case of an infinite dimensional family. Moreover, the limit point may sometimes be a degenerate measure [see, Example 2 of Dembski (1990)], which is undesirable as a non-informative prior. However, it is easy to see that if the growth of the packing number of every nonempty open set  $U$  is like that of  $K$  itself (i.e.,  $D(\varepsilon, U) \sim \mu(U)D(\varepsilon, K)$  as  $\varepsilon \rightarrow 0$  for some  $\mu(U) > 0$ ), then  $K$  cannot have any accumulation point and there is a uniform probability  $\nu$  which is positive for each open set  $U$  and is equal to  $\mu(U)$  for open  $U$  with  $\nu(\partial U) = 0$ . Moreover, if for all balls  $B(x; \varepsilon)$ ,  $x \in K$ ,  $\mu(B(x; \varepsilon)) \rightarrow 0$  as  $\varepsilon \rightarrow 0$ , then  $\nu$  is non-atomic.

To extend these ideas to non-compact  $\sigma$ -compact spaces, one can take a sequence of compact subsets  $K_n \uparrow K$  having uniform probability  $\mu_n$ ; the above specification of  $\mu$  is consistent in view of Proposition 2 of Dembski (1990). Any positive Borel measure  $\mu$  satisfying

$$\mu(\cdot \cap K_n) = \frac{\mu_n(\cdot \cap K_n)}{\mu_n(K_1)} \quad (8.2)$$

may be thought as an (improper) uniform distribution on  $K$ . Such a measure, if exists, is also uniquely determined up to a positive multiple; see Lemma 2 of Dembski (1990).

### 8.3 Jeffreys' Prior

Let  $X_i$ 's be i.i.d. with density  $f(\cdot; \theta)$  (with respect to a  $\sigma$ -finite measure  $\nu$ ) where  $\theta \in \Theta$  and  $\Theta$  is an open subset of  $\mathbf{R}^d$ . We assume that  $\{f(\cdot; \theta) : \theta \in \Theta\}$  is a regular parametric family, i.e., there exist  $\psi(\cdot; \theta) \in (L^2(\nu))^d$  ( $d$ -fold product of  $L^2(\nu)$ ) such that for any compact  $K \subset \Theta$ ,

$$\sup_{\theta \in K} \int |f^{1/2}(x; \theta + \mathbf{h}) - f^{1/2}(x; \theta) - \mathbf{h}^T \psi(x; \theta)|^2 \nu(dx) = o(\|\mathbf{h}\|^2) \quad (8.3)$$

as  $\|\mathbf{h}\| \rightarrow 0$ . Define the Fisher information by the relation

$$\mathbf{I}(\theta) = 4 \int \psi(x; \theta) (\psi(x; \theta))^T \nu(dx) \quad (8.4)$$

and assume that  $\mathbf{I}(\theta)$  is positive definite and the map  $\theta \mapsto \mathbf{I}(\theta)$  is continuous. Further, assume the following stronger form of identifiability: On every compact subset  $K \subset \Theta$ ,

$$\inf \left\{ \int (f^{1/2}(x; \theta_1) - f^{1/2}(x; \theta_2))^2 \nu(dx) : \theta_1, \theta_2 \in K, \|\theta_1 - \theta_2\| \geq \varepsilon \right\} > 0, \quad \varepsilon > 0.$$

For i.i.d. observations, it is natural to equip  $\Theta$  with the Hellinger distance defined by

$$H(\theta_1, \theta_2) = \left( \int |f^{1/2}(x; \theta_1) - f^{1/2}(x; \theta_2)|^2 \nu(dx) \right)^{1/2}. \quad (8.5)$$

The following result is the main theorem of this section.

**Theorem 8.3.1** *Fix a compact subset  $K$  of  $\Theta$ . Then for all  $Q \subset K$  with  $\text{vol}(\partial Q) = 0$ , we have*

$$\lim_{\varepsilon \rightarrow 0} \frac{D(\varepsilon, Q)}{D(\varepsilon, K)} = \frac{\int_Q \sqrt{\det \mathbf{I}(\theta)} d\theta}{\int_K \sqrt{\det \mathbf{I}(\theta)} d\theta}. \quad (8.6)$$

By using Theorem 8.2.1, we conclude from Theorem 8.3.1 that the Jeffreys measure  $\mu$  on  $\Theta$  defined by

$$\mu(Q) \propto \int_Q \sqrt{\det \mathbf{I}(\theta)} d\theta, \quad Q \subset \Theta, \quad (8.7)$$

is the (possibly improper) non-informative prior on  $\Theta$  in the sense of the second approach described in the introduction.

PROOF OF THEOREM 8.3.1. Fix  $0 < \eta < 1$ . Cover  $K$  by  $J$  cubes of length  $\eta$ . In each cube, consider the interior cube with length  $\eta - \eta^2$ .

Since by continuity, the eigenvalues of  $\mathbf{I}(\theta)$  are uniformly bounded away from zero and infinity on  $K$ , by standard arguments [see, e.g., Ibragimov and Has'minskii (1981, Theorem I.7.6)], it follows from (8.3) that there exist  $M > m > 0$  such that

$$m\|\theta_1 - \theta_2\| \leq H(\theta_1, \theta_2) \leq M\|\theta_1 - \theta_2\|, \quad \theta_1, \theta_2 \in K. \quad (8.8)$$

Given  $\eta > 0$ , choose  $\varepsilon > 0$  so that  $\varepsilon/(2m) < \eta^2$ . Any two interior cubes are thus separated at least by  $\varepsilon/m$  in terms of Euclidean distance, and so by  $\varepsilon$  in terms of the Hellinger distance.

For  $Q \subset K$ , let  $Q_j$  be the intersection of  $Q$  with the  $j$ -th cube and  $Q'_j$  be the intersection with the  $j$ -th interior cube,  $j = 1, \dots, J$ . Thus

$$Q_1 \cup \dots \cup Q_J = Q \supset Q'_1 \cup \dots \cup Q'_J. \quad (8.9)$$

Hence

$$\sum_{j=1}^J D(\varepsilon, Q'_j; H) \leq D(\varepsilon, Q; H) \leq \sum_{j=1}^J D(\varepsilon, Q_j; H). \quad (8.10)$$

In particular, with  $Q = K$ , we obtain

$$\sum_{j=1}^J D(\varepsilon, K'_j; H) \leq D(\varepsilon, K; H) \leq \sum_{j=1}^J D(\varepsilon, K_j; H), \quad (8.11)$$

where  $K_j, K'_j$  are analogously defined.

From the  $j$ -th cube, choose  $\theta_j \in K$ . By an argument similar to that used in the derivation of (8.8), we have for all  $\theta, \theta'$  in the  $j$ -th cube,

$$\frac{\underline{\lambda}(\eta)}{2} \sqrt{(\theta - \theta')^T \mathbf{I}(\theta_j)(\theta - \theta')} \leq H(\theta, \theta') \leq \frac{\bar{\lambda}(\eta)}{2} \sqrt{(\theta - \theta')^T \mathbf{I}(\theta_j)(\theta - \theta')}, \quad (8.12)$$

where  $\underline{\lambda}(\eta)$  and  $\bar{\lambda}(\eta)$  tend to 1 as  $\eta \rightarrow 0$ .

Let

$$\underline{H}_j(\theta, \theta') = \frac{1}{2} \underline{\lambda}(\eta) \sqrt{(\theta - \theta')^T \mathbf{I}(\theta_j)(\theta - \theta')}$$

and

$$\bar{H}_j(\theta, \theta') = \frac{1}{2} \bar{\lambda}(\eta) \sqrt{(\theta - \theta')^T \mathbf{I}(\theta_j)(\theta - \theta')}.$$

Then from (8.12), we have

$$\begin{aligned} D(\varepsilon, Q_j; H) &\leq D(\varepsilon, Q_j; \underline{H}_j), \\ D(\varepsilon, Q'_j; H) &\geq D(\varepsilon, Q'_j; \bar{H}_j). \end{aligned} \quad (8.13)$$

By (1961) on the

and

where as  $\varepsilon$  - norms that  $\tau$  that

and

Now I  $\rightarrow \int_Q$  the su result

Rem had e Kullb tion c  $K(\theta, \phi(\varepsilon)$  g can b

1.

2.

In suc limit (1995

By using the second part of Theorem IX of Kolmogorov and Tihomirov (1961), for some constants  $\tau_j, \tau'_j$  and an absolute constant  $A_d$  (depending only on the dimension  $d$ ), we have

$$D(\varepsilon, Q_j; \underline{H}_j) \sim A_d \tau_j \text{vol}(Q_j) \sqrt{\det \mathbf{I}(\theta_j)} (\underline{\lambda}(\eta))^{-d} \varepsilon^{-d} \tag{8.14}$$

and

$$D(\varepsilon, Q'_j; \overline{H}_j) \sim A_d \tau'_j \text{vol}(Q'_j) \sqrt{\det \mathbf{I}(\theta_j)} (\overline{\lambda}(\eta))^{-d} \varepsilon^{-d}, \tag{8.15}$$

where the symbol  $\sim$  signifies that the limit of the ratio of the two sides is 1 as  $\varepsilon \rightarrow 0$ . As all the metrics  $\underline{H}_j$  and  $\overline{H}_j$ ,  $j = 1, \dots, J$ , arise from elliptic norms, it can be easily concluded by making a suitable linear transformation that  $\tau_j = \tau'_j = \tau$  (say) for all  $j = 1, \dots, J$ . Thus we obtain from (8.10)–(8.15) that

$$\limsup_{\varepsilon \rightarrow 0} \frac{D(\varepsilon, Q; H)}{D(\varepsilon, K; H)} \leq \frac{\sum_{j=1}^J \sqrt{\det \mathbf{I}(\theta_j)} \text{vol}(Q_j)}{\sum_{j=1}^J \sqrt{\det \mathbf{I}(\theta_j)} \text{vol}(K'_j)} \left( \frac{\overline{\lambda}(\eta)}{\underline{\lambda}(\eta)} \right)^d \tag{8.16}$$

and

$$\liminf_{\varepsilon \rightarrow 0} \frac{D(\varepsilon, Q; H)}{D(\varepsilon, K; H)} \leq \frac{\sum_{j=1}^J \sqrt{\det \mathbf{I}(\theta_j)} \text{vol}(Q'_j)}{\sum_{j=1}^J \sqrt{\det \mathbf{I}(\theta_j)} \text{vol}(K_j)} \left( \frac{\underline{\lambda}(\eta)}{\overline{\lambda}(\eta)} \right)^d. \tag{8.17}$$

Now let  $\eta \rightarrow 0$ . By the convergence of Riemann sums,  $\sum_{j=1}^J \sqrt{\det \mathbf{I}(\theta_j)} \text{vol}(Q_j) \rightarrow \int_Q \sqrt{\mathbf{I}(\theta)} d\theta$  and  $\sum_{j=1}^J \sqrt{\det \mathbf{I}(\theta_j)} \text{vol}(Q'_j) \rightarrow \int_Q \sqrt{\mathbf{I}(\theta)} d\theta$  and similarly for the sums involving  $K_j$ 's and  $K'_j$ 's. Also,  $\underline{\lambda}(\eta) \rightarrow 1$  and  $\overline{\lambda}(\eta) \rightarrow 1$ , so the desired result follows. ■

**Remark 8.3.1** During a discussion Prof. Hartigan remarked that Jeffereys had envisaged constructing non-informative priors by approximating  $\Theta$  with Kullback-Liebler neighborhoods, and asked us if the construction in the last section can be carried out using the K-L neighborhoods. Since the K-L divergence  $K(\theta, \theta') = \int f_\theta \log \frac{f_\theta}{f_{\theta'}}$  is not a metric there would be obvious difficulties in formalizing the notion of an  $\varepsilon$ -net. However, if the family of densities  $\{f_\theta : \theta \in \Theta\}$  have well-behaved tails such that, for any  $\theta, \theta', K(\theta, \theta') \leq \phi(H(\theta, \theta'))$ , where  $\phi(\varepsilon)$  goes to 0 as  $\varepsilon$  goes to 0, then any  $\varepsilon$ -net  $\{\theta_1, \theta_2 \dots \theta_k\}$  in the Hellinger metric can be thought of as a K-L net in the sense that

1.  $K(\theta_i, \theta_j) > \varepsilon$  for  $i, j = 1, 2, \dots, k$ ,
2. for any  $\theta$  there exists an  $i$  such that  $K(\theta_i, \theta) < \phi(\varepsilon)$ .

In such situations, the above theorem allows us to view the Jeffereys' prior as a limit of uniform distributions arising out of K-L neighborhoods. Wong and Shen (1995) show that suitable tail behavior for all  $\theta, \theta', \int_{(f_\theta/f_{\theta'} \geq \exp \frac{1}{\delta})} f_\theta (\frac{f_\theta}{f_{\theta'}})^\delta < M$ .

We now consider the case when there is a nuisance parameter. Let  $\theta$  be the parameter of interest and  $\varphi$  be the nuisance parameter, and we assume for simplicity that both are real-valued. We can write the information matrix as

$$\begin{pmatrix} I_{11}(\theta, \varphi) & I_{12}(\theta, \varphi) \\ I_{21}(\theta, \varphi) & I_{22}(\theta, \varphi) \end{pmatrix}, \tag{8.18}$$

where  $I_{12} = I_{21}$ . In view of Theorem 8.3.1, it is natural to put the prior  $\pi(\varphi|\theta) = \sqrt{I_{22}(\theta, \varphi)}$  for  $\varphi$  given  $\theta$ . So we need to construct a non-informative marginal prior for  $\theta$ . First let us assume that the parameter space is compact. With  $n$  i.i.d. observations, the joint density of the observations given  $\theta$  only is given by

$$g(\mathbf{x}^n; \theta) = (c(\theta))^{-1} \int \prod_{i=1}^n f(x_i; \theta, \varphi) \sqrt{I_{22}(\theta, \varphi)} d\varphi, \tag{8.19}$$

where  $c(\theta) = \int \sqrt{I_{22}(\theta, \varphi)} d\varphi$  is the constant of normalization. Let  $I_n(\theta; g)$  denote the information for the family  $\{g(\mathbf{x}^n; \theta) : \theta \in \Theta\}$ . Under adequate regularity conditions, it can be shown that the information per observation  $I_n(\theta; g)/n$  satisfies

$$\lim_{n \rightarrow \infty} I_n(\theta; g)/n = (c(\theta))^{-1} \int I_{11.2}(\theta, \varphi) \sqrt{I_{22}(\theta, \varphi)} d\varphi = J(\theta) \text{ (say)}, \tag{8.20}$$

where  $I_{11.2} = I_{11} - I_{12}^2/I_{22}$ . Let  $H_n(\theta, \theta + h)$  be the Hellinger distance between  $g(\mathbf{x}^n; \theta)$  and  $g(\mathbf{x}^n; \theta + h)$ . Locally as  $h \rightarrow 0$ ,  $H_n^2(\theta, \theta + h)$  behaves like  $I_n(\theta; g)h^2$ . Hence by Theorem 8.3.1, the non-informative (marginal) prior for  $\theta$  would be proportional to  $\sqrt{I_n(\theta; g)}$ . In view of (8.20), passing to the limit as  $n \rightarrow \infty$ , the (sample size independent) marginal non-informative prior for  $\theta$  should be taken to be proportional to  $(J(\theta))^{1/2}$ , and so the prior for  $(\theta, \varphi)$  is proportional to  $J(\theta)\pi(\varphi|\theta)$ . Generally, for a noncompact parameter space, we proceed like Berger and Bernardo (1989). Fix a sequence of compact sets  $\Lambda_l$  increasing to the whole parameter space. Put  $\Phi_l(\theta) = \{\varphi : (\theta, \varphi) \in \Lambda_l\}$  and normalize  $\pi(\varphi|\theta)$  on  $\Phi_l(\theta)$  as

$$p_l(\varphi|\theta) = (c_l(\theta))^{-1} \pi(\varphi|\theta) I\{\varphi \in \Phi_l(\theta)\}, \tag{8.21}$$

where  $c_l(\theta) = \int_{\Phi_l(\theta)} \sqrt{I_{22}(\theta, \varphi)} d\varphi$  is the constant of normalization, as before. The marginal non-informative prior for  $\theta$  at stage  $l$  is then defined as

$$\pi_l(\theta) = \sqrt{\int I_{11.2}(\theta, \varphi) p_l(\varphi|\theta) d\varphi}. \tag{8.22}$$

Let  $\theta_0$  be a fixed value of  $\theta$ . The (joint) non-informative prior is finally defined as

$$\lim_{l \rightarrow \infty} \frac{(c_l(\theta))^{-1} \pi_l(\theta) \pi(\varphi|\theta)}{(c_l(\theta_0))^{-1} \pi_l(\theta_0) \pi(\varphi|\theta_0)}, \tag{8.23}$$

Non

assu  
obta  
then  
the p  
and  
com  
belie

8.4

In th  
the t

The  
ger d  
 $\infty$ . I  
prob  
addir

then  
 $X_1, X$

PROC  
 $P_0$  st  
that

Since  
Helli  
and si  
from t

To est

assuming that the above limit exists for all  $\theta$ . Informally, the prior for  $\theta$  is obtained by taking the average of  $I_{11,2}(\theta, \varphi)$  (with respect to  $\sqrt{I_{22}(\theta, \varphi)}$ ) and then taking the square-root. The reference prior of Berger and Bernardo or the probability matching prior takes averages of other functions of  $\sqrt{I_{11,2}(\theta, \varphi)}$  and then transforms back. So they all have common structure and are worth comparing through examples. In the examples of Datta and Ghosh (1995b), we believe that they reduce to the same prior.

### 8.4 An Infinite Dimensional Example

In this section, we show that in a certain class of infinite dimensional families, the third approach mentioned in the introduction leads to consistent posterior.

**Theorem 8.4.1** *Let  $\mathcal{P}$  be a family of densities where  $\mathcal{P}$ , metrized by the Hellinger distance, is compact. Let  $\varepsilon_n$  be a positive sequence satisfying  $\sum_{n=1}^{\infty} n^{1/2} \varepsilon_n < \infty$ . Let  $\mathcal{F}_n$  be an  $\varepsilon_n$ -net in  $\mathcal{P}$ ,  $\mu_n$  be the uniform distribution on  $\mathcal{F}_n$  and  $\mu$  be the probability on  $\mathcal{P}$  defined by  $\mu = \sum_{n=1}^{\infty} \lambda_n \mu_n$ , where  $\lambda_n$ 's are positive numbers adding upto unity. If for any  $\beta > 0$*

$$\lim_{n \rightarrow \infty} e^{\beta n} \frac{\lambda_n}{D(\varepsilon_n, \mathcal{F}_n)} = \infty, \tag{8.24}$$

then the posterior distribution based on the prior  $\mu$  and i.i.d. observations  $X_1, X_2, \dots$  is consistent at every  $p_0 \in \mathcal{P}$ .

PROOF OF THEOREM 8.4.1. Fix a  $p_0 \in \mathcal{P}$  and a neighborhood  $U$  of  $p_0$ . Let  $P_0$  stand for the probability corresponding to the density  $p_0$ . We need to show that

$$\frac{\int_{U^c} \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} \mu(dp)}{\int_{\mathcal{P}} \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} \mu(dp)} \rightarrow 0 \text{ a.s. } [P_0]. \tag{8.25}$$

Since  $\mathcal{P}$  is compact under the Hellinger metric, the weak topology and the Hellinger topology coincide on  $\mathcal{P}$ . Thus  $U$  is also a weak neighborhood of  $p_0$  and since there is a uniformly consistent test for  $p = p_0$  against  $p \notin U$ , it follows from the arguments of Schwartz (1965) and Barron (1986), that for some  $\beta > 0$ ,

$$e^{n\beta} \int_{U^c} \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} \mu(dp) \rightarrow 0 \text{ a.s. } [P_0]. \tag{8.26}$$

To establish (8.25), it suffices to show that for every  $\beta > 0$ ,

$$e^{n\beta} \int_{\mathcal{P}} \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} \mu(dp) \rightarrow \infty \text{ a.s. } [P_0]. \tag{8.27}$$

Let  $p_n$  and  $p_{0n}$ , respectively, stand for the joint densities  $\prod_{i=1}^n p(X_i)$  and  $\prod_{i=1}^n p_0(X_i)$ . Let  $\|\cdot\|_r$  and  $H(\cdot, \cdot)$  denote the  $L^r$ -norms,  $r = 1, 2$ , and Hellinger distance respectively. Then

$$\begin{aligned} E_{P_0} \left| \frac{p_n}{p_{0n}} - 1 \right| &= \|p_n - p_{0n}\|_1 \\ &= \|(p_n^{1/2} + p_{0n}^{1/2})(p_n^{1/2} - p_{0n}^{1/2})\|_1 \\ &\leq \|p_n^{1/2} + p_{0n}^{1/2}\|_2 H(p_n, p_{0n}) \\ &\leq 2\sqrt{1 - \int p_n p_{0n}} \\ &= 2\sqrt{1 - \left(\int p p_0\right)^n} \\ &= 2\sqrt{1 - \left(1 - \frac{1}{2}H^2(p, p_0)\right)^n}. \end{aligned}$$

Using the elementary inequality  $1 - x^n \leq n(1 - x)$  for  $0 \leq x \leq 1$ , we observe that if  $p \in B(p_0, \varepsilon_n) = \{p : H(p_0, p) < \varepsilon_n\}$ , then

$$E_{P_0} \left| \frac{p_n}{p_{0n}} - 1 \right| \leq \sqrt{2n\varepsilon_n},$$

and so

$$\begin{aligned} E_{P_0} \left| \int_{B(p_0, \varepsilon_n)} \left( \frac{p_n}{p_{0n}} - 1 \right) \mu(dp) \right| &\leq \int_{B(p_0, \varepsilon_n)} E_{P_0} \left| \frac{p_n}{p_{0n}} - 1 \right| \mu(dp) \\ &\leq \sqrt{2n\varepsilon_n} \mu(B(p_0, \varepsilon_n)). \end{aligned}$$

Hence

$$\begin{aligned} P_0 \left\{ \int_{\mathcal{P}} \left( \frac{p_n}{p_{0n}} \right) \mu(dp) \leq \frac{1}{2} \mu(B(p_0, \varepsilon_n)) \right\} \\ \leq P_0 \left\{ \int_{B(p_0, \varepsilon_n)} \left( \frac{p_n}{p_{0n}} \right) \mu(dp) \leq \frac{1}{2} \mu(B(p_0, \varepsilon_n^*)) \right\} \\ \leq P_0 \left\{ \left| \int_{B(p_0, \varepsilon_n)} \left( \frac{p_n}{p_{0n}} - 1 \right) \mu(dp) \right| > \frac{1}{2} \mu(B(p_0, \varepsilon_n)) \right\} \\ \leq \frac{2E_{P_0} \left| \int_{B(p_0, \varepsilon_n)} \left( \frac{p_n}{p_{0n}} - 1 \right) \mu(dp) \right|}{\mu(B(p_0, \varepsilon_n))} \\ \leq \sqrt{8n\varepsilon_n}. \end{aligned}$$

By the construction,  $\exp[n\beta] \mu(B(p_0, \varepsilon_n)) \geq \exp[n\beta] \lambda_n / D(\varepsilon_n, \mathcal{P})$ , which goes to infinity in view of Assumption (8.24). An application of the Borel-Cantelli Lemma yields (8.27). ■

Non-Infor

**Remark**  
for sieves  
proof tha  
(8.24) for  
Precis  
 $\mu = \sum_{n=1}^{\infty}$   
with

the poste

A useful

where 0  
 $\gamma > 1$ . If  
becomes

If  $\varepsilon_n = \varepsilon$   
then (8.  
posterior  
An  $\varepsilon$   
density

$\mathcal{P} = \{$

where  $r$   
Theorem  
Hence t  
posterior

**Acknow**  
tional E  
India.  
9307727  
1R01 G

**Remark 8.4.1** Consistency is obtained in the last theorem by requiring (8.24) for sieves whose width  $\varepsilon_n$  was chosen carefully. However it is clear from the proof that consistency would follow for sieves with width  $\varepsilon_n \downarrow 0$  by imposing (8.24) for a carefully chosen subsequence.

Precisely, if  $\varepsilon_n \downarrow 0$ ,  $\mathcal{F}_n$  is an  $\varepsilon_n$ -net,  $\mu$  is the probability on  $\mathcal{P}$  defined by  $\mu = \sum_{n=1}^{\infty} \lambda_n \mu_n$  and  $\delta_n$  is a positive summable sequence, then by choosing  $j(n)$  with

$$\varepsilon_{j(n)} \leq \sqrt{\frac{2}{n}} \delta_n, \tag{8.28}$$

the posterior is consistent, if

$$\exp[n\beta] \frac{\lambda_{j(n)}}{D(\varepsilon_{j(n)}, \mathcal{P})} \rightarrow \infty. \tag{8.29}$$

A useful case corresponds to

$$D(\varepsilon, \mathcal{P}) \leq A \exp[c\varepsilon^{-\alpha}], \tag{8.30}$$

where  $0 < \alpha < 2/3$  and  $A$  and  $c$  are positive constants,  $\delta_n = n^{-\gamma}$  for some  $\gamma > 1$ . If in this case,  $j(n)$  is the smallest integer satisfying (8.28), then (8.29) becomes

$$\exp[n\beta - c\varepsilon_{j(n)}^{-\alpha}] \lambda_{j(n)} \rightarrow \infty. \tag{8.31}$$

If  $\varepsilon_n = \varepsilon/2^n$  for some  $\varepsilon > 0$  and  $\lambda_n$  decays no faster than  $n^{-s}$  for some  $s > 0$ , then (8.31) holds. Moreover, the condition  $0 < \alpha < 2$  in (8.30) is enough for posterior consistency in probability.

An example of this kind is the following class of densities considered in density estimation [see, e.g., Wong and Shen (1995)]:

$$\mathcal{P} = \{f = g^2 : g \in C^r[0, 1], \int g^2(x) dx = 1, \|g^{(j)}\|_{\text{sup}} \leq L_j, j = 1, \dots, r, \\ |g^{(r)}(x_1) - g^{(r)}(x_2)| \leq L_{r+1}|x_1 - x_2|^m\},$$

where  $r$  is a positive integer.  $0 \leq m \leq 1$  and  $L_j$ 's are fixed constants. By Theorem XV of Kolmogorov and Tihomirov (1961),  $D(\varepsilon, \mathcal{P}) \leq \exp[c\varepsilon^{-1/(r+m)}]$ . Hence the hierarchical prior constructed in Theorem 8.4.1 leads to consistent posterior.

**Acknowledgements.** Research of the first author was supported by the National Board of Higher Mathematics, Department of Atomic Energy, Bombay, India. Research of the second author was supported by NSF grant number 9307727. Research of the third author was supported by NIH grant number 1R01 GM49374.

---

## References

1. Barron, A. R. (1986). Discussion of "On the consistency of Bayes estimates" by P. Diaconis and D. Freedman, *Annals of Statistics*, **14**, 26–30.
2. Barron, A. R., Schervish, M. and Wasserman, L. (1996). The consistency of posterior distributions in non parametric problems, *Preprint*.
3. Basu, D. (1975). Statistical information and likelihood, *Sankhyā, Series A*, **37** 1–71.
4. Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors, *Journal of the American Statistical Association*, **84**, 200–207.
5. Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors (with discussions), In *Bayesian Statistics V* (Eds., J. M. Bernardo *et al.*), pp. 35–60.
6. Berger, J. O. and Pericchi, L. (1994). Intrinsic Bayes factor for model selection and prediction in general linear model, *Preprint*.
7. Berger, J. O. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association*, **91**, 109–121.
8. Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussions), *Journal of the Royal Statistical Society, Series B*, **41**, 113–147.
9. Datta, G. S. and Ghosh, J. K. (1995a). On priors providing frequentist validity for Bayesian inference, *Biometrika*, **82**, 37–46.
10. Datta, G. S. and Ghosh, J. K. (1995b). Noninformative priors for maximal invariant in group models, *Test*, **4**, 95–114.
11. Datta, G. S. and Ghosh, M. (1995). Some remarks on noninformative priors, *Journal of the American Statistical Association*, **90**, 1357–1363.
12. Datta, G. S. and Ghosh, M. (1996). On the invariance of noninformative priors, *Annals of Statistics*, **24** (to appear).
13. Dembski, W. A. (1990). Uniform probability, *Journal of Theoretical Probability*, **3**, 611–626.
14. Gasparini, M. (1992). Bayes nonparametrics for biased sampling and density estimation, *Ph.D thesis*, University of Michigan.

15. C  
(

16. C

C  
1

17. C

C

18. F  
a

19. I

/

20. F

r

N

21. F

b

1

22. F

c

S

3

23. L

o

24. C

J

25. S

4

26. S

a

S

27. T

E

28. V

r

3

15. Ghosh, J. K. (1994). *Higher Order Asymptotics*, NSF-CBMS Regional Conference Series in Probability and Statistics 4, IMS, Hayward, CA.
16. Ghosh, J. K. and Mukerjee, R. (1992). Non-informative priors (with discussions), In *Bayesian Statistics, 4* (Eds., J. M. Bernardo *et al.*), pp. 195–210.
17. Ghosh, J. K. and Ramamoorthi R. V. (1997). *Lecture notes on Bayesian asymptotics*, Under preparation.
18. Heath, D. and Sudderth, W. (1978). On finitely additive priors, coherence and extended admissibility, *Annals of Statistics*, **6**, 333–345.
19. Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*, New York: Springer-Verlag.
20. Kass, R. E. and Wasserman, L. (1992). A reference Bayesian test for nested hypotheses with large samples, *Technical Report, #567*, Carnegie Mellon University.
21. Kass, R. E. and Wasserman, L. (1996). The selection of prior distribution by formal rules, *Journal of the American Statistical Association*, **96**, 1343–1370.
22. Kolmogorov, A. N. and Tihomirov, V. M. (1961).  $\epsilon$ -entropy and  $\epsilon$ -capacity of sets in function spaces, *American Mathematics Society Transl. Ser. 2*, **17**, 277–364. [Translated from Russian: *Uspekhi Mat. Nauk*, **14**, 3–86, (1959).]
23. Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates, *Annals of Statistics*, **12**, 351–357.
24. O'Hagan, A. (1995). Fractional Bayes factors for model comparisons, *Journal of the Royal Statistical Society, Series B*, **57**, 99–138.
25. Schwartz, L. (1965). On Bayes procedures, *Z. Wahrsch. Verw. Gebiete*, **4**, 10–26.
26. Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information, *Journal of the Royal Statistical Society, Series B*, **44**, 377–387.
27. Tibshirani, R. (1989). Noninformative priors for one parameter of many, *Biometrika*, **76**, 604–605.
28. Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs, *Annals of Statistics*, **23**, 339–362.

29. West, M. (1992). Modelling with mixtures, In *Bayesian Statistics, 4*, (Eds., J. M. Bernardo *et al.*), pp. 503-524.
30. West, M., Muller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation, In *Aspects of uncertainty: A Tribute to D. V. Lindley*, pp. 363-386.
31. Zellner, A. (1990). Bayesian methods and entropy in economics and econometrics, In *10 International MaxEnt Workshop*, University of Wyoming.