

- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.
- Mukhopadhyay, S. and Gelfand, A.E. (1997). Dirichlet process mixed generalized linear models. *J. Amer. Statist. Assoc.*, 92: 633–639.
- Muller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors-in-variables. *Biometrika* 84: 535–537.
- Newton, M. A., Czado, C. and Chapell, R. (1996). Bayesian inference for semiparametric binary regression. *J. Amer. Statist. Assoc.*, 91: 142–153.
- Richardson, S. and Gilks, W. R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Amer. J. of Epidemiology*, 6: 430–442.
- Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order-Restricted Statistical Inference*. J. Wiley and Sons, New York.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4: 639–650.
- Sethuraman, J. and Tiwari, R. C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In: *Statistical Decision Theory and Related Topics III*, Eds: Gupta, S. and Berger, J. O., Springer-Verlag, New York, 2, 305–315.
- Sinha, D. (1996). Time-discrete Beta process model for interval censored survival data. Tech. Rpt., Department of Mathematics, University of New Hampshire, Durham.
- Walker, S. G. and Mallick, B. K. (1996). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. Tech. Rpt., Department of Mathematics, Imperial College, London.
- 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*. Eds: A. F. M. Smith and P. Feeman, John Wiley and Sons, New York. 363–386.

# 18

## Consistency Issues in Bayesian Nonparametrics

S. GHOSAL Vrije Universiteit, Amsterdam, The Netherlands

J. K. GHOSH Indian Statistical Institute, Calcutta, India

R. V. RAMAMOORTHI Michigan State University, East Lansing, Michigan

### 1. INTRODUCTION

The basic Bayesian model consists of a parameter  $\theta$  and a prior distribution  $\Pi$  for  $\theta$  which reflects the investigator's belief regarding  $\theta$ . This prior is updated by observing  $X_1, X_2, \dots, X_n$ , which are modelled as iid  $P_\theta$  given  $\theta$ . The updating mechanism is Bayes theorem which results in changing  $\Pi$  to the posterior  $\Pi(\cdot | X_1, X_2, \dots, X_n)$ .

One of the basic ingredients in the model is the prior distribution  $\Pi$  which should ideally correspond to the investigators's opinion about  $\theta$ . However a complete elicitation of  $\Pi$  is still not feasible, at least for high or infinite dimensional problems. In practice, a prior is adopted which accords well with the investigator's belief and is also mathematically tractable. A pragmatic choice of the prior contains a subjective component as well as a technical one.

In this chapter we are mainly concerned with consistency of the posterior. Informally, the posterior is said to be consistent at a true value  $\theta_0$  if the following holds: Suppose  $X_1, X_2, \dots, X_n$  indeed arise from  $P_{\theta_0}$ , then the posterior converges to the degenerate probability  $\delta_{\theta_0}$ . Alternatively with  $P_{\theta_0}$  probability 1, the posterior probability of any neighborhood  $U$  of  $\theta_0$  converges to 1.

Why would a Bayesian be interested in consistency? Think of an experiment in which an experimenter generates observations from a known (to the experimenter) distribution. The observations are presented to a Bayesian. It would be embarrassing if, even with large data, the Bayesian fails to come close to finding the mechanism used by the experimenter. Consistency can be thought of as a validation of the Bayesian method. It can also be interpreted as requiring that the data, at least eventually, overrides the prior opinion. Alternatively two Bayesians, with two different priors, presented with the same data eventually agree. A result of this kind relating "merging of opinions" and posterior consistency is discussed in Diaconis and Freedman (1986a). In fact, Diaconis and Freedman (1986a) (henceforth abbreviated as DF) and the ensuing discussions contain a wealth of material pertaining to posterior consistency.

An early result in posterior consistency is due to Doob (1948), who showed that posterior consistency obtains on a set of prior measure 1. This result does not settle the question of consistency for a particular  $\theta_0$  of interest. In smooth finite dimensional problems, different methods show, for example Berk (1966), that consistency obtains at all parameter points. Freedman (1963) exhibits a prior and points of inconsistency for the infinite cell multinomial. He also showed that this phenomenon is quite general and the departure from consistency can be quite dramatic. This initiated a search for priors with good consistency properties in the nonparametric case. A major impetus to Bayesian nonparametrics came from the introduction of Dirichlet Process by Ferguson (1973). Recent years have seen a surge of new priors, especially in the context of densities and semi parametric inference. Walker et al. (1997) review and discuss in detail the interpretation of parameters appearing in the priors. Hjort (1996) contains both a review and some new constructions. However questions of consistency for such priors are difficult to settle and are only beginning to receive attention. A striking negative result is given in DF. A detailed systematic study of Bayesian nonparametrics with stress on asymptotics is available in a monograph under preparation by J. K. Ghosh and R. V. Ramamoorthi (1997).

This chapter presents a brief review of some of these developments. The focus will be on elucidation of different notions of convergence and on positive consistency results.

## 2. CONSISTENCY

Let  $\mathbf{R}$  stand for real line and  $\mathcal{B}$  stand for the Borel  $\sigma$ -algebra on  $\mathbf{R}$ . The space of probability measures on  $\mathbf{R}$  will be denoted by  $\mathcal{M}$ . The space  $\mathcal{M}$  is equipped with the  $\sigma$ -algebra  $\mathcal{B}(\mathcal{M})$ —the smallest  $\sigma$ -algebra with respect to

which the functions  $\{P \mapsto P(B) : B \in \mathcal{B}\}$  are measurable. It is convenient to consider, as our parameter space, a subset  $\mathcal{P}$  of  $\mathcal{M}$ . Unless otherwise stated, the  $\sigma$ -algebra associated with  $\mathcal{P}$  will be the  $\sigma$ -algebra  $\mathcal{B}(\mathcal{M})$ , restricted to  $\mathcal{P}$ . A prior  $\Pi$  is a probability measure on  $\mathcal{P}$ . Denoting the random probability measure by  $\mathbf{P}$ , we will sometime write this as  $\mathbf{P} \sim \Pi$  and will write "given  $\mathbf{P}$ " for given  $\mathbf{P} = P$ .

Let  $X_1, X_2, \dots$  be a sequence of random variables which are given  $P$ , iid  $P$ . We will for notational simplicity use  $P$  for the distribution of each  $X_i$ , the joint distribution of  $X_1, X_2, \dots, X_n$  and the joint distribution of the entire sequence  $X_1, X_2, \dots$ . It is convenient to think of  $X_1, X_2, \dots$  as being defined on  $\Omega = \mathbf{R}^\infty$ , with  $X_i$  as the  $i$ th co-ordinate map.

Let  $\Pi$  be a prior on  $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$  and given  $P$ , let  $X_1, X_2, \dots, X_n$  be iid  $P$ . These two together determine a joint distribution for  $P, X_1, X_2, \dots, X_n$ . The resulting conditional distribution of  $P$  given  $X_1, X_2, \dots, X_n$  will be denoted by  $\Pi(\cdot | X_1, X_2, \dots, X_n)$ . Of course, the conditional distribution is not unique but in most contexts there is a natural choice and it is this version that we will work with.

The sequence  $\{\Pi(\cdot | X_1, X_2, \dots, X_n), n \geq 1\}$  is a sequence of (random) probability measures on  $\mathcal{M}$  and thus their convergence involves convergence of probabilities on the space of probabilities. Our interest is mainly in convergence to a degenerate measure  $\delta_{P_0}$  and in this case consistency can be conveniently formulated by requiring that  $\Pi(U | X_1, X_2, \dots, X_n) \rightarrow 1$  for suitable neighborhoods of  $P_0$ . Since there are a variety of topologies available on  $\mathcal{M}$ , these lead to different notions of consistency. We next briefly discuss some of these.

### 2.1 Weak Consistency

A subset  $U$  of  $\mathcal{M}$  is said to be a weak neighborhood of  $P_0$  if it contains a set of the form  $U = \{P : |\int f_i dP - \int f_i dP_0| < \varepsilon_i, i = 1, 2, \dots, k\}$ , where  $f_i, i = 1, 2, \dots, k$ , are bounded continuous functions on  $\mathbf{R}$ .

**DEFINITION 1.** The sequence  $\{\Pi(\cdot | X_1, X_2, \dots, X_n), n \geq 1\}$  is said to be weakly consistent at  $P_0$ , if there exists a  $\Omega_0 \subset \Omega$  with  $P_0(\Omega_0) = 1$  such that for  $\omega \in \Omega_0$ , for every weak neighborhood  $U$  of  $P_0$ ,  $\Pi(U | X_1, X_2, \dots, X_n) \rightarrow 1$  as  $n \rightarrow \infty$ .

When  $\mathcal{M}$  is given the weak topology,  $\mathcal{M}$  itself becomes a complete separable metric space and this in turn induces a weak topology on the space of probability measures on  $\mathcal{M}$ . Weak consistency corresponds to the convergence of the posterior in this topology. Since  $\mathcal{M}$  is a complete separable

metric space, it is meaningful to talk of the (topological) support of a prior  $\Pi$ —the smallest closed set with  $\Pi$  measure 1. It is not hard to see that  $\mathcal{B}(\mathcal{M})$  is the Borel  $\sigma$ -algebra when  $\mathcal{M}$  is viewed as a metric space.

**2.2 K-Consistency**

If  $P_1, P_2$  are probability measures on  $\mathbf{R}$ , define the Kolomogorov distance

$$d_K(P_1, P_2) = \sup_{t \in \mathbf{R}} |P_1(-\infty, t) - P_2(-\infty, t)|.$$

A K-neighborhood of  $P_0$  is a set of the form  $\{P : d_K(P_0, P) < \varepsilon\}$ .

**DEFINITION 2.** The sequence  $\{\Pi(\cdot|X_1, X_2, \dots, X_n), n \geq 1\}$  is said to be K-consistent at  $P_0$  if there exists a  $\Omega_0 \subset \Omega$  with  $P_0(\Omega_0) = 1$  such that for  $\omega \in \Omega_0$ , for every K-neighborhood  $U$  of  $P_0$ ,  $\Pi(U|X_1, X_2, \dots, X_n) \rightarrow 1$  as  $n \rightarrow \infty$ .

This consistency is of course motivated by the Glivenko-Cantelli theorem. Under the metric  $d_K$ ,  $\mathcal{M}$  is neither separable nor complete.

**2.3 Consistency in Uniform Neighborhoods**

Another popular metric on  $\mathcal{M}$  is the total variation metric

$$d_T(P_1, P_2) = \sup_{B \in \mathcal{B}} |P_1(B) - P_2(B)|.$$

A total variation neighborhood of  $P_0$  is a set of the form

$$\{P : d_T(P_0, P) < \varepsilon\}.$$

**DEFINITION 3.** The sequence  $\{\Pi(\cdot|X_1, X_2, \dots, X_n), n \geq 1\}$  is said to be consistent at uniform neighborhoods of  $P_0$  if there exists a  $\Omega_0 \subset \Omega$  with  $P_0(\Omega_0) = 1$ , such that for  $\omega \in \Omega_0$ , for every total variation neighborhood  $U$  of  $P_0$ ,

$$\Pi(U|X_1, X_2, \dots, X_n) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

The space  $\mathcal{M}$  under the total variation metric is non-separable and the associated measure theoretic problems renders it uninteresting when the parameter space is all of  $\mathcal{M}$ . However, if the prior is supported by a set of densities, then the total variation metric coincides with the  $L_1$ -metric. This space is separable and in this case consistency over uniform neighborhoods is preferred over weak or K-consistency.

Let  $\Pi$  be a prior on  $\mathcal{M}$ . The expected value under  $\Pi$ ,  $E_\Pi(\mathbf{P})$  is the probability measure defined by  $\hat{P}(B) = \int P(B)\Pi(dP)$ . We will refer to the expectation of  $\mathbf{P}$  under  $\Pi(\cdot|X_1, X_2, \dots, X_n)$  as the Bayes estimate or as the predictive distribution. It is not hard to see that if the posterior is consistent at  $P_0$  in any of the above senses, then the Bayes estimate converges to  $P_0$  in the corresponding topology.

**3. GENERAL CONSISTENCY THEOREMS**

We begin with a consistency result for the multinomial.

**THEOREM 1.** Let  $\Pi$  be a prior on  $\mathcal{M}(\chi)$ , where  $\chi$  is a finite set. Then the posterior is consistent at all  $P_0$  in the support of  $\Pi$ .

This result appears as Theorem 1 in Freedman (1963). Here is a brief outline of the argument involved. Let  $\chi = \{1, 2, \dots, k\}$ , so that  $\mathcal{M}(\chi) = \{P = (P(1), P(2), \dots, P(k)) : P(i) \geq 0, \sum P(i) = 1\}$ . Let  $P_0$  be in the support of  $\Pi$ .

For any  $\varepsilon > 0$ , let

$$K_\varepsilon = \left\{ P : \sum_{j=1}^k P_0(j) \log \frac{P_0(j)}{P(j)} < \varepsilon \right\}.$$

Finite dimensionality and compactness of  $\mathcal{M}(\chi)$  ensures two things:

- (i) Every neighborhood  $V$  of  $P_0$  contains a set of the form  $K_\varepsilon$  and more crucially, every  $K_\varepsilon$  contains a neighborhood of  $P_0$ . Thus if  $P_0$  is in the support of  $\Pi$ , then  $\Pi(K_\varepsilon) > 0$  for all  $\varepsilon > 0$ .
- (ii) The convergence of

$$\frac{1}{n} \sum_{i=1}^n \log \frac{P_0(X_i)}{P(X_i)} \quad \text{to} \quad \sum_{j=1}^k P_0(j) \log \frac{P_0(j)}{P(j)}$$

is uniform in  $P$  in the sense, for any  $\varepsilon > 0$ ,

$$\sup_{P \in K_\varepsilon} \left| \frac{1}{n} \sum_{i=1}^n \log \frac{P_0(X_i)}{P(X_i)} - \sum_{j=1}^k P_0(j) \log \frac{P_0(j)}{P(j)} \right| \rightarrow 0 \quad \text{a.s. } P_0.$$

$$\inf_{P \in K_\varepsilon^c} \frac{1}{n} \sum_{i=1}^n \log \frac{P_0(X_i)}{P(X_i)} > \frac{\varepsilon}{2} \rightarrow 0 \quad \text{a.s. } P_0.$$

For any neighborhood  $V$  of  $P_0$ , let  $\varepsilon_1 > 2\varepsilon_2$  and let  $K_{\varepsilon_1}, K_{2\varepsilon_2}$  be contained in  $V$ . Then the posterior probability of  $V^c$  satisfies

$$\begin{aligned} \Pi(V^c|X_1, X_2, \dots, X_n) &= \frac{\int_{V^c} \exp \left[ \sum_{i=1}^n \log \frac{P(X_i)}{P_0(X_i)} \right] \Pi(dP)}{\int_{\mathcal{M}(\chi)} \exp \left[ \sum_{i=1}^n \log \frac{P(X_i)}{P_0(X_i)} \right] \Pi(dP)} \\ &\leq \frac{\int_{K_{\varepsilon_1}^c} \exp \left[ \sum_{i=1}^n \log \frac{P(X_i)}{P_0(X_i)} \right] \Pi(dP)}{\int_{K_{2\varepsilon_2}} \exp \left[ \sum_{i=1}^n \log \frac{P(X_i)}{P_0(X_i)} \right] \Pi(dP)} \\ &= \frac{I_{1n}(X_1, X_2, \dots, X_n)}{I_{2n}(X_1, X_2, \dots, X_n)} \quad (\text{say}) \end{aligned}$$

The uniform convergence mentioned in (ii) yields

$$\limsup_{n \rightarrow \infty} e^{n\varepsilon_2} I_{1n} = 0 \quad \text{a.s. } P_0.$$

Since  $\Pi(K_{\varepsilon_2}) > 0$ , an application of Fatou's lemma gives

$$\liminf_{n \rightarrow \infty} e^{n\varepsilon_2} I_{2n} = \infty \quad \text{a.s. } P_0.$$

These two together show that  $\Pi(V^c|X_1, X_2, \dots, X_n) \rightarrow 0$  a.s.  $P_0$ .

When  $\chi$  is infinite, even weak consistency can fail to occur in the support of  $\Pi$ . Freedman (1963) provided a dramatic example of this when  $\chi = \{1, 2, \dots\}$ .

Theorem 1 provides a passage to consistency for tail free priors. A prior  $\Pi$  on  $\mathcal{M}$  is tail free with respect to a sequence of partitions  $\mathbf{T} = \{\mathbf{T}_n, n \geq 1\}$  if,  $\mathbf{T} = \{\mathbf{T}_n, n \geq 1\} = \{\{B_\varepsilon : \varepsilon \in \{0, 1\}^n\}, n \geq 1\}$  is a nested sequence of partitions as described in Sec. 7, and if

$$\{P(B_0)\}, \{P(B_{00}|B_0), P(B_{10}|B_1)\}, \dots, \{P(B_{\varepsilon 0}|B_\varepsilon) : \varepsilon \in \{0, 1\}^k\}, \dots$$

are all independent. As we will see later, Dirichlet process and Polya tree process are tail free.

**THEOREM 2.** If  $\Pi$  is tail free with respect to  $\mathbf{T} = \{\mathbf{T}_n, n \geq 1\}$ , then (a suitable version of) the posterior is weakly consistent.

The idea behind Theorem 2 is simple. Every weak neighborhood is determined by  $\{B_\varepsilon : \varepsilon \in \{0, 1\}^k\}$  for some  $k$ . Tail free property ensures that the posterior distribution of  $\{P(B_\varepsilon) : \varepsilon \in \{0, 1\}^k\}$  given  $X_1, X_2, \dots, X_n$  is same

as the posterior given the  $2^k$ -cell multinomial with cells  $\{B_\varepsilon : \varepsilon \in \{0, 1\}^k\}$ . This puts us in the framework work of Theorem 1.

When the prior is not tail free, a useful tool to establish weak consistency is a theorem of Schwartz (1965). In this theorem too, like Theorem 1, the Kullback–Leibler numbers play an important role. A similar theorem for the infinite cell multinomial already appears in Freedman (1963).

Let  $L(\mu)$  be the set of all densities with respect to a  $\sigma$ -finite measure  $\mu$ . The Kullback–Leibler divergence of  $f$  from  $f_0$ , where both are in  $L(\mu)$  is defined as  $K(f_0, f) = \int f_0 \log(f_0/f) d\mu$ . The Kullback–Leibler neighborhood  $\{f : K(f_0, f) < \varepsilon\}$  will be denoted by  $K_\varepsilon(f_0)$ .

Let  $U$  be a set and  $f_0 \in U$ . Say that there exists a uniformly consistent sequence of tests for testing  $H_0 : f = f_0$  vs.  $H_1 : f \in U^c$ , if there exists a sequence of tests  $\phi_n(X_1, X_2, \dots, X_n)$  such that as  $n \rightarrow \infty$ ,

$$E_{f_0} \phi_n(X_1, X_2, \dots, X_n) \rightarrow 0$$

and

$$\inf_{f \in U^c} E_f \phi_n(X_1, X_2, \dots, X_n) \rightarrow 1.$$

**THEOREM 3.** (Schwartz). Let  $\Pi$  be a prior on  $L(\mu)$ . If  $f_0 \in L(\mu)$  and  $U \subset L(\mu)$  satisfy

1.  $\Pi(K_\varepsilon(f_0)) > 0$  for all  $\varepsilon > 0$ ;
  2. there exists a uniformly consistent sequence of tests for testing  $H_0 : f = f_0$  vs.  $H_1 : f \in U^c$ ;
- then  $\Pi(U|X_1, X_2, \dots, X_n) \rightarrow 1$  a.s.  $P_0$ .

The Schwartz theorem retains the flavor of Theorem 1. In Theorem 1,  $\Pi(K_\varepsilon(f_0)) > 0$  was derived as a consequence of  $f_0$  being in the support of  $\Pi$ . Condition (2) essentially plays the role of the compactness of  $\mathcal{M}(\chi)$ .

It is not hard to see that if  $U$  is a weak neighborhood of  $f_0$ , then there exists a uniformly consistent sequence of tests for testing  $H_0 : f = f_0$  vs.  $H_1 : f \in U^c$ . Since the weak neighborhoods have a countable base, Schwartz's theorem immediately yields the following corollary.

**COROLLARY 1.** If  $\Pi(K_\varepsilon(f_0)) > 0$  for all  $\varepsilon > 0$ , then the posterior is weakly consistent at  $f_0$ .

We consider two examples below. The first of these shows that the condition of Schwartz is not necessary for consistency. The second example shows that if the Schwartz condition does not hold then consistency cannot be expected in general even for finite dimensional examples. In the second example,  $P_0$  is in the weak support of the prior.

Let  $X_1, X_2, \dots, X_n$  be iid  $U(0, \theta)$  where  $\theta \in \Theta = (0, 1]$ . In this example the Kullback–Leibler divergence of every  $U(0, \theta)$  from  $U(0, 1)$  is  $\infty$ . Thus the Schwartz condition fails but if  $\Pi$  is a prior with support all of  $[0, 1]$ , then it is easy to see that the posterior is consistent at  $f_0 = U(0, 1)$ .

If the above example is modified by setting  $\Theta = (0, 1] \cup (2, 3)$ , then it can be shown, as in Ghosh and Ramamoorthi (1997), that there is a prior with  $f_0 = U(0, 1)$  in its weak support such that the posterior fails to be consistent at  $f_0 = U(0, 1)$ . In this example too, Schwartz’s condition fails to hold.

If  $U$  is a strong neighborhood, then Le Cam (1973) and Barron (1989) show that, in general, there will not exist a uniformly consistent sequence of tests for testing  $H_0 : f = f_0$  vs.  $H_1 : f \in U^c$ . The role of uniformly consistent sequence of tests is greatly clarified by the following theorem of Barron which is discussed in Barron et al. (1996) (henceforth abbreviated as BSW).

**THEOREM 4.** Let  $\Pi$  be a prior on  $L(\mu)$ ,  $f_0 \in L(\mu)$  and  $U$  is a neighborhood of  $f_0$ . Assume  $\Pi(K_\varepsilon(f_0)) > 0$  for all  $\varepsilon > 0$ . Then the following are equivalent:

1. There exists a  $\beta_0$  such that

$$P_{f_0} \{ \Pi(U^c | X_1, X_2, \dots, X_n) > e^{-n\beta_0} \text{ infinitely often} \} = 0.$$

2. There exist subsets  $V_n, W_n$  of  $L(\mu)$ , positive numbers  $c_1, c_2, \beta_1, \beta_2$  and a sequence of tests  $\{ \phi_n(X_1, X_2, \dots, X_n) \}$  such that

(a)  $U^c = V_n \cup W_n$

(b)  $\Pi(W_n) \leq c_1 e^{-n\beta_1}$

(c)  $P_{f_0} \{ \phi_n(X_1, X_2, \dots, X_n) > 0 \text{ infinitely often} \} = 0$

and

$$\inf_{f \in V_n} E_f \phi_n \geq 1 - c_2 e^{-n\beta_2}.$$

The last theorem can be used to develop sufficient conditions for posterior consistency on uniform neighborhoods. BSW provide such a condition using bracketing metric entropy. Motivated by the result of BSW and Theorem 4, we can, Ghosal et al. (1997b) prove the following:

Let  $\mathcal{F} \subset \bar{L}(\mu)$ . For  $\delta > 0$ , the  $L_1$ -metric entropy of  $\mathcal{F}$ , denoted by  $J(\delta, \mathcal{F})$ , is  $\log a(\delta)$ , where  $a(\delta)$  is the minimum over all  $k$  such that there exist  $f_1, f_2, \dots, f_k$  in  $L(\mu)$  with  $\mathcal{F} \subset \cup_{i=1}^k \{ f : \|f - f_i\|_1 < \delta \}$ .

**THEOREM 5.** Let  $\Pi$  be a prior on  $L(\mu)$  with  $\Pi(\mathcal{F}) = 1$ . Suppose  $f_0 \in L(\mu)$  and  $\Pi(K_\varepsilon(f_0)) > 0$  for all  $\varepsilon > 0$ . If for each  $\varepsilon > 0$  there is a

$\delta < \varepsilon/4$ ,  $c_1, \beta_1 > 0$ ,  $\beta < \varepsilon^2/8$  and also  $\mathcal{F}_n \subset \mathcal{F}$  such that, for all  $n$  large,

1.  $\Pi(\mathcal{F}_n^c) < c_1 e^{-n\beta_1}$ ,
2.  $J(\delta, \mathcal{F}_n) < n\beta$ ,

then the posterior is consistent at  $f_0$  on uniform neighborhoods.

#### 4. DIRICHLET PROCESSES

Dirichlet processes were introduced in the statistical context by Ferguson (1973), where he developed their basic properties and used it to provide a Bayesian interpretation of many popular nonparametric procedures. A recent review of Dirichlet process is the review article by Ferguson et al. (1996). Another good source is the text by Schervish (1995). We provide a brief summary of some of the properties for later use in this chapter.

Let  $\alpha$  be a finite measure on  $\mathbf{R}$ . A probability measure  $D_\alpha$  on  $\mathcal{M}$  is said to be a Dirichlet process with parameter  $\alpha$  if for every partition  $B_1, B_2, \dots, B_k$  of  $\mathbf{R}$  by Borel sets, the vector  $(P(B_1), P(B_2), \dots, P(B_k))$  has a  $k$ -variate Dirichlet distribution with parameter  $(\alpha(B_1), \alpha(B_2), \dots, \alpha(B_k))$ .

1.  $E_{D_\alpha}(P(A)) = \alpha(A)/\alpha(\mathbf{R})$ .

It is convenient to write  $\bar{\alpha}$  for the probability  $\alpha(\cdot)/\alpha(\mathbf{R})$ . Thus a natural choice for  $\bar{\alpha}$  is the prior belief of the distribution of  $X$ .

2. If  $P \sim D_\alpha$  and given  $P$ ,  $X_1, X_2, \dots, X_n$  are iid  $P$ , then the posterior given  $X_1, X_2, \dots, X_n$  is  $D_{\alpha + \sum \delta_{X_i}}$ .
3. Properties (1) and (2) immediately show that the Bayes estimate of  $P$  given  $X_1, X_2, \dots, X_n$  is

$$\frac{\alpha(\mathbf{R})}{\alpha(\mathbf{R}) + n} \bar{\alpha} + \frac{n}{\alpha(\mathbf{R}) + n} P_n,$$

where  $P_n$  is the empirical distribution arising from  $X_1, X_2, \dots, X_n$ . Since as  $\alpha(\mathbf{R}) \rightarrow 0$  the Bayes estimate goes to the empirical distribution,  $\alpha(\mathbf{R})$  can be thought of as a “prior sample size” or as a measure of belief in the prior. Sethuraman and Tiwari (1982) have pointed out the need for some care in this interpretation.

4.  $D_\alpha$  gives mass 1 to the set of discrete distributions. Thus  $D_\alpha$  is concentrated on a small set. On the other hand the topological support of  $D_\alpha$  is  $\{ P : \text{support}(P) \subset \text{support}(\alpha) \}$ . If  $\alpha$  has (topological) support all of  $\mathbf{R}$ , then  $D_\alpha$  will have as its support all of  $\mathcal{M}$ .
5. If  $\alpha_1$  and  $\alpha_2$  are two distinct nonatomic measures, then  $D_{\alpha_1} \perp D_{\alpha_2}$ . A slight extension shows that with a  $D_\alpha$  prior, the prior and posterior are singular with respect to each other.

6. Let  $P \sim D_\alpha$  and given  $P$ ,  $X_1, X_2, \dots, X_n$  be iid  $P$ . The marginal distribution of  $X_1, X_2, \dots, X_n$  can be interpreted as a Polya urn scheme, see Blackwell and MacQueen (1973) and Mauldin et al. (1992). The distribution of  $X_1$  is  $\bar{\alpha}$ , the distribution of  $X_2$  given  $X_1 = x_1$  is  $(\alpha + \delta_{x_1})/(\alpha(\mathbf{R}) + 1)$  and the distribution of  $X_{n+1}$  given  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  is  $(\alpha + \sum_1^n \delta_{x_i})/(\alpha(\mathbf{R}) + n)$ . One consequence of this expression is that even when  $\alpha$  is nonatomic, the probability of getting coincidences among  $X_1, X_2, \dots, X_n$  is positive. Suppose  $\alpha$  has density  $g$  with respect to Lebesgue measure. Then the joint density of  $X_1, X_2, \dots, X_n$  is given by

$$\frac{1}{[\alpha(\mathbf{R})]^{(n)}} \prod_{j: I_j=1} g(x_j) \quad (1)$$

where  $I_j = 1$  if  $x_j \notin \{x_1, x_2, \dots, x_{j-1}\}$  and 0 otherwise and  $[\alpha(\mathbf{R})]^{(n)}$  is the ascending factorial  $\alpha(\mathbf{R})(\alpha(\mathbf{R}) + 1)(\alpha(\mathbf{R}) + 2) \cdots (\alpha(\mathbf{R}) + n - 1)$ . The above expression is to be viewed as a density with respect to  $\bar{\lambda} = \sum \lambda_{B_1, B_2, \dots, B_k}$ , where  $B_1, B_2, \dots, B_k$  is a partition of  $\{1, 2, \dots, n\}$  and  $\lambda_{B_1, B_2, \dots, B_k}$  is the  $k$ -dimensional Lebesgue measure on  $\{(x_1, x_2, \dots, x_n) : x_i = x_j \text{ if } i, j \in \text{the same } B_l \text{ and } x_i \neq x_j \text{ otherwise}\}$ .

7. The Dirichlet process is tail free with respect to every partition and hence the posterior is weakly consistent. In fact, it can be shown that it is K-consistent.

## 5. MIXTURES OF DIRICHLET PROCESS

Mixtures of Dirichlet processes considered by Antoniak (1974) provide greater flexibility. In this model, a hyper parameter  $\theta$  is first chosen according to a prior  $\mu$  and given  $\theta$ ,  $P \sim D_{\alpha(\theta)}$  and given  $(P, \theta)$ ,  $X_1, X_2, \dots, X_n$  are iid  $P$ . For example, one may be ready to say that the expected value of  $P$  is normal but not be able to specify  $(\mu, \sigma)$ . In such cases it may be appropriate to take  $\theta = (\mu, \sigma)$  and  $\alpha_\theta = \alpha_\theta(\mathbf{R})N(\mu, \sigma)$ .

Since  $\Pi(P|\theta, X_1, X_2, \dots, X_n)$  is  $D_{\alpha(\theta) + \sum \delta_{x_i}}$  and

$$\Pi(P|X_1, X_2, \dots, X_n) = \int \Pi(P|\theta, X_1, X_2, \dots, X_n) \Pi(d\theta|X_1, X_2, \dots, X_n),$$

the posterior distribution of  $P$  given  $X_1, X_2, \dots, X_n$  is again a mixture of Dirichlet with  $\mu$  changing to  $\mu^* = \Pi(d\theta|X_1, X_2, \dots, X_n)$  and  $\alpha(\theta)$  changing to  $\alpha(\theta) + \sum \delta_{x_i}$ . We need an expression for  $\Pi(d\theta|X_1, X_2, \dots, X_n)$  to evaluate the posterior.

If  $\mu$  has a density  $h(\theta)$  and if  $\alpha(\theta)$  has a density  $g_\theta(x)$ , then using Eq. (1), the joint density of  $(\theta, X_1, X_2, \dots, X_n)$  is

$$\frac{h(\theta)}{[\alpha(\theta)(\mathbf{R})]^{(n)}} \prod_{j: I_j=1} g_\theta(x_j)$$

and the conditional density of  $\theta$  given  $X_1, X_2, \dots, X_n$  becomes

$$c(X_1, X_2, \dots, X_n) h(\theta) \prod_{j: I_j=1} g_\theta(x_j).$$

If the true distribution  $P_0$  has a density then with  $P_0$  probability 1, the  $I_j$ 's are all equal to 1 and the conditional density of  $\theta$  given  $X_1, X_2, \dots, X_n$  becomes

$$c(X_1, X_2, \dots, X_n) h(\theta) \prod_{j=1}^n g_\theta(x_j). \quad (2)$$

Mixtures are no longer tail-free and in general one has inconsistency. Ferguson et al. (1996) has a nice example illustrating this phenomenon. Consistency issues in the case of mixtures of Dirichlet is studied in detail in Freedman and Diaconis (1983), where the following result is available.

**THEOREM 6.** If  $\alpha(\theta)(\mathbf{R}) \leq M$  for all  $\theta$ , then the mixture of  $D_{\alpha(\theta)}$  yields a weakly consistent posterior.

The mixture model, apart from being interesting in its own right, also arises in Bayesian models for semiparametric problems. Location model discussed in the next Section is one example. Another interesting type of problems is the binary regression model studied by Newton (1994).

## 6. LOCATION MODEL

Consider the location parameter problem where  $\mu$  is a prior on the location parameter  $\theta$  and  $P$  is independently chosen according to  $D_\alpha$ ; given  $(\theta, P)$ ,  $X_1, X_2, \dots, X_n$  are iid  $P_\theta$ , where  $P_\theta(A) = P(A - \theta)$ . This setup falls in the mixture model with  $\alpha_\theta(A) = \alpha(A - \theta)$ , except that our interest is in the posterior distribution of  $\theta$  given  $X_1, X_2, \dots, X_n$ , rather than in the posterior distribution of  $P$ . In other words,  $\theta$  is the parameter of interest and not just an index in the mixture distribution.

If  $\mu$  has a density  $h(\theta)$  and if  $\alpha$  has a density  $g(x)$ , when  $X_1, X_2, \dots, X_n$  are all distinct and Eq. (1) yields

$$\Pi(\theta|X_1, X_2, \dots, X_n) = \frac{h(\theta)g(x_1 - \theta)g(x_2 - \theta) \cdots g(x_n - \theta)}{\int h(t)g(x_1 - t)g(x_2 - t) \cdots g(x_n - t) dt}.$$

Barron, in his discussion of DF, notes from this expression that even though  $\theta$  is the location parameter of an unknown  $P$ , the posterior acts as though the  $X_i$ 's come from a fixed known density  $g$ . So, one would not, in general, expect consistency of the posterior.

Since the location model is not identifiable without some assumptions like symmetry, it is appropriate to consider a prior on  $\mathcal{M}^s$ —the space of distributions on  $\mathbf{R}$ , symmetric around 0. DF use a symmetrized Dirichlet process on  $\mathcal{M}^s$ . Symmetrized Dirichlet processes were first studied by Dalal (1979).

The expression for the posterior distribution for  $\theta$  is somewhat complicated (Lemma 3.1 of Diaconis and Freedman (1986b)), but DF show that asymptotic analysis of the previous case essentially holds also for symmetric  $P$ . To be more precise, in the symmetric location model, if  $\alpha$  is the Cauchy distribution then, when the true parameter is  $(0, P_0)$  where  $P_0$  has a trimodal density with compact support, for approximately half the samples the posterior concentrates near a mode  $\delta$  and another half concentrates near  $-\delta$ . The number  $\delta$  can be made as large as one wants by choosing  $P_0$  suitably. This is an example of how dramatic deviations from consistency can be. A similar phenomenon was observed by Doss (1985) under a different setup.

This is a surprising result. Addition of a single parameter seems to destroy the attractive features of the Dirichlet. That inconsistency should occur for a  $P_0$  with simple structure also comes as a surprise.

In Ghosal et al (1998), we consider the location model and prove the following:

**THEOREM 7.** Let  $\lambda$  be a prior on  $\mathcal{M}^s$ —the set of all symmetric densities and  $\mu$  be a prior on  $\mathbf{R}$ ,  $f_0$  be a symmetric density and  $\theta_0$  be a real number. Assume that  $\lambda, \mu$  satisfy the following:

1.  $\mu$  gives positive mass to every open neighbourhood of  $\theta_0$ ;
2. for every  $\varepsilon > 0$  and all sufficiently small  $|t|$ ,  $\lambda(K_\varepsilon(f_{0,t}^*)) > 0$ , where  $f_{0,t}(x) = f_0(x - t)$  and  $f_{0,t}^*(x) = (f_{0,t}(x) + f_{0,t}(-x))/2$ .

If  $f_0$  satisfies either of the following conditions, then the posterior is weakly consistent at  $(\theta_0, f_0)$  (and as a consequence, the posterior of  $\theta$  is consistent at  $\theta_0$  in the usual topology):

1.  $\lim_{t \rightarrow 0} \int f_0 \log \frac{f_0}{f_{0,t}} dx = 0$ ;
2.  $f_0$  is continuous and has compact support.

For Polya tree priors with expectation measure  $\alpha$ , a sufficient condition for  $\lambda(K_\varepsilon(f_{0,t}^*)) > 0$  will be  $\int f_{0,t} \log(f_{0,t}/\alpha) < \infty$ , or equivalently

$\int f_0 \log(f_0/\alpha_t) < \infty$  for all sufficiently small  $|t|$ , a little stronger than simply  $\int f_0 \log(f_0/\alpha) < \infty$

Thus, if the assumptions of the above theorem hold, then consistency does occur in the DF type of distributions. Since for any density  $f$ ,  $K(f, P) = \infty$  whenever  $P$  is discrete and since  $D_\alpha$  gives mass 1 to discrete densities, clearly the assumptions do not hold for the Dirichlet. In the next section we will show that Polya tree priors can be chosen to satisfy the assumptions. In Sec. 8.2, we will show that if the Dirichlet process is smoothed by convoluting with a (normal) kernel, then also the assumptions are satisfied.

Of course, the theorem does not preclude inconsistency but it does appear that in this case points of inconsistency, if they exist, will be quite complicated.

## 7. POLYA TREE PRIORS

Polya tree priors are generalization of Dirichlet process. These were discussed by Ferguson (1974). For a recent explication of their role in nonparametrics we refer to Lavine (1992, 1994). Other related references are Mauldin et al. (1992) and Schervish (1995).

Let  $E_j$  be the set of all sequences of 0's and 1's of length  $j$  and let  $E^* = \bigcup_j E_j$ .

Let  $T = \{T_n, n \geq 1\}$  be a sequence of nested partitions of  $\mathbf{R}$  into intervals such that  $\bigcup_j T_j$  generates the Borel  $\sigma$ -algebra. A bit more explicitly, let  $T_j = \{B_\varepsilon : \varepsilon \in E_j\}$ . At the  $(j+1)$ th stage, each  $B_\varepsilon$  is partitioned into  $B_{\varepsilon 0}$  and  $B_{\varepsilon 1}$ . We want each  $B_\varepsilon$  to be an interval and the  $\sigma$ -algebra generated by  $\bigcup_{\varepsilon \in E^*} B_\varepsilon$  be the Borel  $\sigma$ -algebra on  $\mathbf{R}$ .

A prior  $\Pi$  on  $\mathcal{M}$  is said to be a Polya tree prior with respect to the partition  $T$  and with parameters  $\{\alpha_\varepsilon : \varepsilon \in E^*\}$ , written as  $P \sim PT(T, \alpha)$ , if under  $\Pi$

1.  $\{P(B_{\varepsilon 0}|B_\varepsilon) : \varepsilon \in E^*\}$  are a set of independent random variables;
2. for each  $\underline{\varepsilon} \in E^*$ ,  $P(B_{\underline{\varepsilon} 0}|B_{\underline{\varepsilon}}) \sim \text{Beta}(\alpha_{\underline{\varepsilon} 0}, \alpha_{\underline{\varepsilon} 1})$ .

Here are some properties of Polya tree priors:

1. The expected value of  $P$  under  $PT(T, \alpha)$  is the probability measure  $\hat{P}$  defined by, if  $\varepsilon_k = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)$ ,  $\varepsilon_k 0 = (\varepsilon_k, 0)$  and  $\varepsilon_k 1 = (\varepsilon_k, 1)$ ,

$$\hat{P}(B_{\varepsilon_k}) = \prod_{i=1}^k \frac{\alpha_{\varepsilon_i}}{\alpha_{\varepsilon_{i-1} 0} + \alpha_{\varepsilon_{i-1} 1}}.$$

2. If  $P \sim PT(\mathbf{T}, \alpha)$  and given  $P$  if  $X_1, X_2, \dots, X_n$  are iid  $P$ , then the posterior distribution is again a Polya tree with parameters  $\alpha = \{\alpha'_\varepsilon : \varepsilon \in E^*\}$ , where  $\alpha'_\varepsilon = \alpha_\varepsilon + \sum_{i=1}^n I_{B_\varepsilon}(X_i)$ .

The above two properties together easily yield an expression for the Bayes estimate. Being tail free, weak consistency is immediate for Polya trees.

Unlike the Dirichlet, by choosing the parameters carefully, a Polya tree can be made to sit on densities. Results like the following appear in Mauldin et al. (1992) and Schervish (1995).

**THEOREM 8.** Suppose that  $\lambda$  is a continuous probability measure such that  $\lambda(B_{\varepsilon_k}) = 2^{-k}$  for all  $k$  and further  $\alpha_{\varepsilon_k} = a_k$ . If  $\sum a_k^{-1} < \infty$ , then the resulting polya tree gives mass 1 to the set of all distributions absolutely continuous with respect to  $\lambda$ .

In the discussion of the location model and strong consistency we required priors on densities which give positive mass to Kullback–Leibler neighborhoods. The following theorem proved in Ghosal et al. (1998) shows that this can be achieved for Polya tree priors. A weaker result also appears in Lavine (1994).

**THEOREM 9** Suppose that  $\lambda$  is a continuous probability measure such that  $\lambda(B_{\varepsilon_k}) = 2^{-k}$  for all  $k$  and further  $\alpha_{\varepsilon_k} = a_k$ . If  $\sum a_k^{-1/2} < \infty$  then the then for any density  $f_0$  (with respect to  $\lambda$ ) with  $\int f_0 \log f_0 < \infty$ , we have, for all  $\varepsilon > 0$ ,  $PT(\mathbf{T}, \alpha)(K_\varepsilon(f_0)) > 0$ .

Thus Polya tree priors (for suitable  $\alpha_\varepsilon$ ) provides an example of priors mentioned in Theorem 7. As for consistency in uniform neighborhoods, BSW construct an appropriate sieve and show that if  $\alpha_{\varepsilon_k} = 8^k$ , then the posterior is strongly consistent. The result of BSW suggests using such a prior for the location model. We expect that if the tails of the prior  $\mu$  for the location parameter decay rapidly then this approach would yield consistency (of the pair  $(\theta, f)$ ) in uniform neighborhoods for a wide class of “true” distributions.

## 8. DIRICHLET MIXTURES

While Polya tree priors can be made to sit on densities, it is not possible to ensure the densities in the support to have enough smoothness properties. Priors on smooth families of densities can be constructed via Dirichlet mixtures, a method suggested by Lo (1994).

Let  $\Theta$  be a parameter set, typically  $\mathbf{R}$  or  $\mathbf{R}^2$ . Let  $K(x, \theta)$  be a kernel, i.e., for each  $\theta$ ,  $K(x, \theta)$  is a probability density in  $x$  and jointly measurable in  $(x, \theta)$ . For any probability  $P$  on  $\Theta$ , let  $K(x, P) = \int K(x, \theta)P(d\theta)$ . Lo's method consists of choosing a mixture  $K(x, P)$  at random according to a Dirichlet prior. Note that when  $K(x, \theta)$  is a location family  $K(x, P)$  is just the convolution  $K * P$ . Typically, such location families will also have a scale parameter  $\sigma$ , which plays a role similar to that of window length in kernel density estimation. We look at the structure of the posterior before looking at specific kernels.

Formally, we have a hierarchical Bayes model which consists of  $P \sim D_\alpha$  and given  $P$ ,  $X_1, X_2, \dots, X_n$  are iid  $K(\cdot, P)$ ; our interest is in the posterior distribution  $P$  given  $X_1, X_2, \dots, X_n$ . This is the approach taken in West (1992) and West et al. (1994).

It is convenient to view the observations  $X_1, X_2, \dots, X_n$  as arising from  $P$  and  $(\theta_1, X_1), (\theta_2, X_2), \dots, (\theta_n, X_n)$ , where  $P \sim D_\alpha$  and given  $P$ , the pairs  $(\theta_1, X_1), (\theta_2, X_2), \dots, (\theta_n, X_n)$  are independent with  $\theta_i$  having the distribution  $P$  and given  $\theta_i$ ,  $X_i$  is an observation from  $K(\cdot, \theta_i)$ . The latent variables  $\theta_i$ , while unobserved, serve as a useful tool in describing and simulating the posterior. Indeed, West and others have set up efficient computing algorithms based on Gibbs sampling.

Letting  $\Pi(P|X_1, X_2, \dots, X_n)$  stand for the posterior distribution of  $P$  given  $X_1, X_2, \dots, X_n$  and by  $H(\theta|X_1, X_2, \dots, X_n)$ , the distribution of  $(\theta_1, \theta_2, \dots, \theta_n)$  given  $X_1, X_2, \dots, X_n$ ,

$$\begin{aligned} \Pi(P|X_1, X_2, \dots, X_n) &= \int \Pi(P|(\theta_1, X_1), (\theta_2, X_2), \dots, (\theta_n, X_n)) \\ &\quad \times H(d\theta|X_1, X_2, \dots, X_n). \end{aligned}$$

Since  $P$  and the  $X_i$ 's are conditionally independent given the  $\theta_i$ 's,

$$\Pi(P|(\theta_1, X_1), (\theta_2, X_2), \dots, (\theta_n, X_n)) = D_{\alpha + \Sigma \delta_{\theta_i}},$$

and hence

$$\Pi(P|X_1, X_2, \dots, X_n) = \int D_{\alpha + \Sigma \delta_{\theta_i}}(P) H(d\theta|X_1, X_2, \dots, X_n).$$

Let  $\alpha(\mathbf{R}) = M$  and  $\bar{\alpha} = \alpha/M$ . Denote by  $\hat{G}_n(\cdot, \theta)$  the empirical distribution based on  $(\theta_1, \theta_2, \dots, \theta_n)$ . The Bayes estimate of  $P(\cdot)$  then becomes

$$\frac{M}{M+n} \bar{\alpha}(\cdot) + \frac{n}{M+n} \int \hat{G}_n(\cdot, \theta) H(d\theta|X_1, X_2, \dots, X_n)$$

and the Bayes estimate of the density at  $x$ ,  $E(K(x, \theta)|X_1, X_2, \dots, X_n)$ , becomes

$$\frac{M}{M+n} K_0(x) + \frac{n}{M+n} \frac{1}{n} \sum_{i=1}^n \int K(x, \theta_i) H(d\theta|X_1, X_2, \dots, X_n),$$

where  $K_0(x)$  is the prior expectation  $\int K(x, \theta) \bar{\alpha}(d\theta)$ . The Bayes estimate is thus composed of a part attributable to the prior and a part due to observations. Ferguson (1983) remarks that the second term in the above convex combination, namely  $n^{-1} \sum_{i=1}^n \int K(x, \theta_i) H(d\theta | X_1, X_2, \dots, X_n)$  can be viewed as a partially Bayesian estimate with the influence of the prior guess removed. The evaluation of the above quantities depend on  $H(\theta | X_1, X_2, \dots, X_n)$ . The joint prior for  $(\theta_1, \theta_2, \dots, \theta_n)$  is, from Eq. (1),

$$\frac{\alpha(d\theta_1)}{\alpha(\mathbf{R})} \times \frac{(\alpha(d\theta_2) + \delta_{\theta_1})}{\alpha(\mathbf{R}) + 1} \times \dots \times \frac{(\alpha(d\theta_n) + \sum_{i=1}^{n-1} \delta_{\theta_i})}{\alpha(\mathbf{R}) + n}$$

Further, the likelihood given  $(\theta_1, \theta_2, \dots, \theta_n)$  is  $\prod_{i=1}^n K(X_i, \theta_i)$ . Hence  $H$  can be written down using standard Bayes formula. Some algebra yields the following expression for the Bayes estimate of the density at  $x$ :

$$\frac{M}{M+n} \int K(x, \theta) \bar{\alpha}(d\theta) + \frac{n}{M+n} \sum_p W(p) \times \sum_{i=1}^{N(p)} \frac{e_i}{n} \frac{\int K(x, \theta) \prod_{l \in C_i} K(X_l, \theta) \alpha(d\theta)}{\int \prod_{l \in C_i} K(X_l, \theta) \alpha(d\theta)}$$

where  $p = \{C_1, C_2, \dots, C_n\}$  is a partition of  $\{1, 2, \dots, n\}$ ,  $N(p)$  being the cardinality of  $p$ ,  $e_i$  is the number of elements in  $C_i$ ,

$$W(p) = \frac{\phi(p)}{\sum_{p'} \phi(p')}, \quad \text{and} \quad \phi(p) = \prod_{i=1}^{N(p)} \left\{ (e_i - 1)! \int \prod_{l \in C_i} K(X_l, \theta) \alpha(d\theta) \right\}.$$

Consistency of the Bayes estimates in these models is studied in Ghorai and Rubin (1982).

A different kind of application of Dirichlet mixtures is made by Brunner and Lo (1989), who estimate a decreasing density on the positive half-line by using a Dirichlet mixture of uniform densities. By a well-known theorem of Khintchine and Shepp, any decreasing density on the positive half-line is given by a mixture  $\int \theta^{-1} I\{0 \leq x \leq \theta\} P(d\theta)$  and conversely. Brunner and Lo (1989) induce a prior on the space of these densities by putting a Dirichlet prior on  $P$ . A symmetric (about zero) strongly unimodal density, which is precisely a symmetrization of a decreasing density on the positive half-line, is often a reasonable model for error distribution. Brunner and Lo (1989) also consider the estimation of a symmetric (about an unknown location) strongly unimodal density on the real line, by using the above Dirichlet mixture of uniform and some prior on the location of the symmetry. Brunner (1992) used a similar idea when the densities are not necessarily symmetric.

Brunner (1995) applied the idea of Brunner and Lo (1989) to the linear regression problem.

### 8.1 Random Histograms

The simplest kernel is  $K(x, \theta) = h^{-1}$  if both  $x$  and  $\theta$  are between  $(ih, (i+1)h)$ . This corresponds to choosing a histogram with bins  $(ih, (i+1)h)$ . A useful method suggested by Gasparini (1992) is to start with a prior  $\mu$  with density  $m(h)$  for  $h$  and given  $h$ , pick a histogram with bin width  $h$  along the following lines: Given  $h$ , let  $\alpha_h = M_h \bar{\alpha}_h$  be a finite measure on integers and given  $h$ , let  $P \sim D_{\alpha_h}$ . Define the random histogram by

$$f_{h,P} = \sum_{i=-\infty}^{\infty} \frac{P(\{i\})}{h} I_{[ih, (i+1)h]}.$$

If  $n_i(h)$  is the number of  $X_1, X_2, \dots, X_n$  in the bin  $[ih, (i+1)h]$ , it is not hard to see that the posterior is of the same form as the prior with  $\alpha_h$  updated to  $\alpha_h + \sum \delta_{n_i(h)}$  and  $m(h)$  changing to

$$m^*(h) = \frac{m(h) \prod_{i=1}^{\infty} (\alpha_h(\{i\}))^{(n_i(h)-1)}}{M_h + n}.$$

The predictive density with no observation is given by  $\hat{f}(x) = \int f_h(x) m(h) dh$ , where  $f_h(x) = h^{-1} \sum_{i=-\infty}^{\infty} \bar{\alpha}_h(\{i\}) I_{[ih, (i+1)h]}(x)$ .

In view of the conjugate property, the predictive density given observations  $X_1, X_2, \dots, X_n$  can be easily written down. For any  $f_0$ , let  $f_{0,h}$  be the approximation  $f_0$  by histograms of bin width  $h$ , i.e., if  $x$  is in  $(ih, (i+1)h]$  then  $f_{0,h}(x) = (1/h) \int_{ih}^{(i+1)h} f(y) dy$ . The calculations in Gasparini (1992) can be used to show:

**THEOREM 10** Suppose the prior satisfies

1.  $\alpha_h$  is a probability measure for all  $h$ ;
2. For each  $h$ , there exists a constant  $K_h$  such that

$$\frac{\alpha_h(\{j-1\})}{\alpha_h(\{j\})} < K_h \quad \text{for} \quad j = \dots, -1, 0, 1, 2, \dots.$$

Then the posterior is weakly consistent at any  $f_0$  satisfying  $\int x^2 f_0(x) dx < \infty$  and  $\lim_{h \rightarrow 0} \int f_0(x) \log(f_{0,h}/f_0) dx = 0$ .

Under additional conditions, Gasparini (1992) shows that the Bayes estimate is strongly consistent. We expect that the techniques of BSW and the

Sec. 3 would enable in identifying densities for which the posterior is consistent on uniform neighborhoods.

Choice of  $m$  which is positive in a neighborhood of  $\theta$  will allow for a wide variability in the choice of histograms and will ensure that the prior has all of  $L_1$  as support. If  $f_0$  is one's prior belief about the density, then an appropriate choice of  $\bar{\alpha}_h$  would be

$$\bar{\alpha}_h(\{i\}) = \int_{ih}^{(i+1)h} f_0(x) dx.$$

Of course, this choice will lead to a prior expected density which may not equal  $f_0$  but can be viewed as an approximation to  $f_0$ . A different Bayesian histogram is proposed by Hartigan (1996). Another method is to consider for each  $k$ , a mixture of  $k$  many beta random variables. A suitable choice of the parameters of the beta leads to choosing the distribution function at random through Bernstein polynomials. This method was proposed by Diaconis and has been investigated by Petrone (1997).

## 8.2 Dirichlet Mixtures of Normal Densities

A natural choice for the kernel is the normal density  $\phi_\sigma(x - \theta) = (1/\sqrt{2\pi}\sigma) \exp[-(x - \theta)^2/2\sigma^2]$ . The  $\sigma$  in the kernel is analogous to the window length in density estimation. It may be chosen empirically or in a fully Bayesian way by selecting a prior on  $\sigma$ . We start looking at the case when  $\sigma$  is fixed. If  $P$  is chosen according to  $D_\alpha$  then, as before, this set up yields the pair  $(\theta_1, X_1), (\theta_2, X_2), \dots, (\theta_n, X_n)$  of latent variables  $\theta_1, \theta_2, \dots, \theta_n$  which are iid  $P$ , and  $X_i \sim N(\theta_i, \sigma^2)$ .

The posterior calculations can be carried out as before. A convenient choice of  $\bar{\alpha}$  is the conjugate prior  $N(\mu, \tau^2)$ . If  $\alpha = M\bar{\alpha}$ , Ferguson (1983) argues that as  $M \rightarrow \infty$ , the Bayes estimate with the influence of the prior removed, converges to  $1/n \sum_{i=1}^n f(x|X_i)$ , where  $f(x|X_i)$  is the Bayes estimate of  $K(x, \theta)$  based on a single observation  $X_i$  and when  $\theta$  has the prior  $\bar{\alpha}$ . In particular when

$$\bar{\alpha} = N(\mu, \tau^2), f(x|X_i) = N\left(x \mid \frac{\mu\sigma^2 + \tau^2 X_i}{\sigma^2 + \tau^2}, \frac{\sigma^2 + 2\tau^2}{\sigma^2 + \tau^2}\right).$$

Ferguson (1983) notes that "this yields a variable kernel estimate with constant window size, but centered at a point between  $X_i$  and  $\mu$  as is typical of shrinkage estimates".

The sample  $X_1, X_2, \dots, X_n$  may also be viewed as conditionally independent (given  $\theta_1, \theta_2, \dots, \theta_n$ )  $N(\theta_i, \sigma^2)$  variables, where the means  $\theta_1, \theta_2, \dots, \theta_n$  are drawn from an uncertain  $P$  which is itself distributed as  $D_\alpha$ . Given

$\theta_1, \theta_2, \dots, \theta_n$ , the next value  $\theta_{n+1}$  is a new value with probability  $M/(M+n)$  and is one of the previous ones with probability  $n/(M+n)$ . Thus if  $M$  is small, typically the  $n$  observations arise from few, much fewer than  $n$ , normal populations. This view is adopted by West, Escobar and others (West (1992), West et al. (1994)), where they effectively demonstrate the use of the Dirichlet mixture model in many applications.

In general, we expect consistency with the mixture only when the bandwidth is allowed to take arbitrarily small values. Suppose that the bandwidth is also given a prior having  $\theta$  in its support. Below we present some results of Ghosal et al. (1997b) regarding consistency of these mixtures.

Since the prior is on densities, consistency on uniform neighborhoods is the appropriate notion. The tool we use is Theorem 5 which involves two parts—the positive prior mass for Kullback–Leibler neighborhoods and sieves with suitable metric entropy. The first two results are concerned about the first issue, among which the Theorem 11 is about consistency at compactly supported densities and has a neat form. For this result, actually we neither need that the mixture is Dirichlet nor the kernel is normal. The only facts used are that the compactly supported density is in the weak support of the distribution of the mixtures and the kernel is positive and continuous.

**THEOREM 11.** Let  $f_0$  be a density having compact support contained in the support of  $\alpha$ . Suppose that  $\lim_{\sigma \rightarrow 0} \int f_0 \log(f_0/f_{0,\sigma}) = 0$ , where  $f_{0,\sigma} = \phi_\sigma * f_0$ . Then  $\Pi(K_\varepsilon(f_0)) > 0$  for all  $\varepsilon > 0$ .

The analogue of Theorem 12 when  $\alpha$  has support all of  $\mathbf{R}$  is somewhat involved and as in Shyamalkumar (1996), requires that the tail behavior of  $f_0$  and  $\alpha$  be related. Loosely speaking, a result of Doss and Sellke (1982) shows that there are functions  $l_1, u_1$  and  $l_2, u_2$  such that with probability 1 under  $D_\alpha$ , the upper tail  $P(x, \infty)$  is greater than  $l_1(x)$  and  $P(x + k \log x, \infty)$  is less than  $u_1(x)$ . The functions  $l_2, u_2$  deal similarly with the lower tails. For any  $\sigma > 0$ , set

$$L_\sigma(x) = \begin{cases} \phi_\sigma(k \log x)(l_1(x) - u_1(x)), & \text{if } x > 0, \\ \phi_\sigma(k \log x)(l_2(x) - u_2(x)), & \text{if } x < 0. \end{cases}$$

**THEOREM 12.** If

1.  $\lim_{\sigma \rightarrow 0} \int f_0 \log(f_0/f_{0,\sigma}) = 0$ ;
2.  $\lim_{a \rightarrow \infty} \int_{-\infty}^{\infty} f_0(x) \log\left(\frac{f_{0,\sigma}(x)}{\int_{-a}^a \phi_\sigma(x - \theta) f_0(\theta) d\theta}\right) dx = 0$ ,

3. for all  $\sigma$ ,

$$\lim_{M \rightarrow \infty} \int_{|x| > M} f_0(x) \log \left( \frac{f_0(x)}{L_\sigma(x)} \right) = 0,$$

then  $\Pi(K_\varepsilon(f_0)) > 0$  for all  $\varepsilon > 0$ .

For example, when  $\alpha$  is double exponential, we may choose any  $k > 2$  and the requirements of the theorem are satisfied if  $f_0$  has finite moment generating function in an open interval containing  $[-1, 1]$ . If  $\alpha$  is normal, the condition needed is the integrability of  $x(\log x)e^{x^2/2}$  with respect to  $f_0$ .

For strong consistency, we estimate the  $L_1$ -metric entropies of sets of the form  $\{\phi_\sigma * P : P[-a, a] = 1\}$  and  $\{\phi_\sigma * P : P[-a, a] \geq 1 - \delta\}$ . Then using Theorem 5, we can establish the following two consistency theorems.

**THEOREM 13.** Let  $\alpha$  have compact support. Suppose that the prior  $\mu$  for  $\sigma$  has bounded support and satisfies  $\mu\{\sigma < t\} \leq c_1 \exp[-c_2/t]$  for some  $c_1, c_2$ . Then  $\Pi(K_\varepsilon(f_0)) > 0$  for all  $\varepsilon > 0$  implies that the posterior is strongly consistent at  $f_0$ .

**THEOREM 14.** Suppose the prior  $\mu$  for  $\sigma$  has bounded support. For any  $\delta > 0$ , let  $a_n$  be such that for some  $\beta_1$ , for all large  $n$ ,

$$D_\alpha\{P : P[-a_n, a_n] < 1 - \delta\} < e^{-n\beta_1}.$$

For an  $\eta > 0$ , suppose that there is a sequence  $\sigma_n \downarrow 0$  be such that  $a_n/\sigma_n < n\eta$  and  $\mu\{\sigma < \sigma_n\} \leq e^{-n\beta_0}$  for some  $\beta_0 > 0$ . Then  $\Pi(K_\varepsilon(f_0)) > 0$  for all  $\varepsilon > 0$  implies that the posterior is strongly consistent at  $f_0$ .

If for example,  $\alpha$  is chosen as a normal distribution and  $\sigma^2$  is given an inverse gamma prior, then  $a_n$  in Theorem 14 is of the order  $\sqrt{n}$  and  $\sigma_n = C/\sqrt{n}$  for a suitable (large)  $C$  (depending on  $\delta$ ) satisfies the conditions of Theorem 14.

It seems more natural to consider the location and scale of the base measure of Dirichlet prior for the mixing distribution as unknown hyperparameter with some specified prior. Under some conditions, consistency continues to hold for such priors.

Another way of handling  $\sigma$  is to treat the pair  $(\theta, \sigma)$  as the a hyper parameter and consider a Dirichlet prior for the distribution of  $(\theta, \sigma)$ . This would yield Dirichlet mixtures of normal over both the location and scale parameters. Ferguson (1983) carries out an analysis when the  $\bar{\alpha}$  is the normal-gamma conjugate prior.

Interestingly, Theorems 11 and 12 have an important implication in the location problem discussed in Sec. 6. If we smooth the (symmetrized) Dirichlet process used by DF by a normal kernel with variance  $h^2$  and  $h$  is chosen from a prior distribution on  $(0, \infty)$  having 0 in its support, then the posterior distribution of the location parameter is consistent if the true error density is symmetric, satisfies conditions of Theorem 7 and  $K(f_0, f_0 * \phi_h) \rightarrow 0$  as  $h \rightarrow 0$ , where  $\phi_h(\cdot)$  stands for the normal density with mean 0 and variance  $h^2$ . This follows from Theorem 7 by observing the following facts:

- (1) Since the normal kernel is symmetric, smoothing a symmetrized Dirichlet process is same as symmetrizing a smoothed Dirichlet process;
- (2) If  $\Pi$  is a prior and  $f_0$  is a symmetric density with  $\Pi(K_\varepsilon(f_0)) > 0$ , then the symmetrization  $\tilde{\Pi}$  of  $\Pi$  also satisfies  $\tilde{\Pi}(K_\varepsilon(f_0)) > 0$  (see Lemma 4.1 of Ghosal et al. (1998)). The posterior density for the location parameter is a smooth function in this case, as opposed to the posterior based on a Polya tree prior. On the other hand, computation is much more involved in this case.

## 9. GAUSSIAN PROCESS PRIORS

These priors introduced by Lenk (1988) are generalizations of a construction of Leonard (1978). The idea is to start with a Gaussian process  $\{Z(x, \omega), x \in I\}$  on an interval  $I$  and define a random density on  $I$  through

$$f(x, \omega) = \frac{\exp[Z(x, \omega)]}{\int_I \exp[Z(t, \omega)] dt}.$$

The Gaussian process has as parameters the mean  $\mu(x)$  and the covariance kernel  $\sigma(x, y)$ . Lenk introduces an additional parameter  $\xi$  to obtain a conjugate family.

Let  $\mu(x)$  be a continuous mean function and  $\sigma(x, y)$  be continuous and positive definite and let  $\{Z(x, \omega) : x \in I\}$  be a Gaussian process with mean  $\mu(x)$  and covariance kernel  $\sigma(x, y)$ . It is convenient to introduce the intermediate process  $W(x, \omega) = \exp[Z(x, \omega)]$ . Denote the distribution of  $W$  by  $LN(\mu, \sigma, 0)$  (LN stands for lognormal). For each  $\xi$  define a process or equivalently a probability measure  $LN(\mu, \sigma, \xi)$  on  $(\mathbf{R}^+)^I$  by

$$\frac{dLN(\mu, \sigma, \xi)}{dLN(\mu, \sigma, 0)}(\omega) = \frac{[\int_I W(x, \omega) dx]^\xi}{C(\mu, \sigma, \xi)}.$$

The function on  $(\mathbf{R}^+)^I$  defined by  $f(x, \omega) = W(x) / \int W(t) dt$  gives a random density and the distribution of this density under  $LN(\mu, \sigma, \xi)$  is denoted by  $LNS(\mu, \sigma, \xi)$ .

Suppose  $f \sim LNS(\mu, \sigma, \xi)$  and given  $f, X_1, X_2, \dots, X_n$  are iid  $f$ . Then the posterior is  $LNS(\mu^*, \sigma, \xi^*)$ , where

$$\mu^*(x) = \mu(x) + \sum_{i=1}^n \sigma(x_i, x), \quad \xi^* = \xi - n.$$

This expression, though not entirely convincingly, allows  $\xi$  to be thought of as a "pseudo sample size". Similarly if  $\sigma(x, y)$  is of the form  $\rho(|x - y|)$ , then  $\rho$  might be considered as a strength of belief about the prior.

In Ghosh and Ramamoorthi (1997), consistency is shown when the Gaussian process is a standard Brownian motion. Consistency issues in general case is under investigation.

### 10. CENSORED DATA

Dirichlet process and Polya trees provide an elegant framework for the Bayesian analysis of right censored data. The model under consideration consists of  $X_1, X_2, \dots, X_n$  and  $Y_1, Y_2, \dots, Y_n$  positive iid random variables with distributions  $F$  and  $G$  respectively. The  $X$ 's correspond to life times and the  $Y$ 's to censoring times; the observations are  $(Z_i, \delta_i)$ , where  $Z_i = (X_i \wedge Y_i)$ ,  $\delta_i = I_{[X_i < Y_i]}$ ,  $i = 1, 2, \dots, n$ . The goal is to make inference on  $F$ , the distribution of the life time.

Susarla and van Ryzin (1976) investigate the case when  $F \sim D_\alpha$  and later Blum and Susarla (1977) show that the posterior distribution of  $F$  can be obtained as a mixture of Dirichlet process. The mixture representation is cumbersome and an alternative representation can be obtained as a Polya tree with the partition depending on  $(Z_1, \delta_1), (Z_2, \delta_2), \dots, (Z_n, \delta_n)$ . We describe this representation below.

Let  $(Z_1, \delta_1), (Z_2, \delta_2), \dots, (Z_n, \delta_n)$  be the observations and the distinct values of the censored observations, arranged in increasing order, be denoted by  $Z^* = (Z_{(1)}, Z_{(2)}, \dots, Z_{(k)})$ . Construct a sequence  $T(Z^*)$  of nested partitions as follows:

$$B_0 = (0, Z_{(1)}], \quad B_1 = (Z_{(1)}, \infty),$$

$$B_{10} = (Z_{(1)}, Z_{(2)}], \quad B_{11} = (Z_{(2)}, \infty),$$

and in general if  $\mathbf{1}_m$  is a string of  $m$  1's, then for  $m \leq k - 1$ ,

$$B_{\mathbf{1}_m 0} = (Z_{(m)}, Z_{(m+1)}]$$

and

$$B_{\mathbf{1}_{m+1}} = (Z_{(m+1)}, \infty).$$

The other  $B_\epsilon$  are partitioned arbitrarily into two intervals so that the partition  $T(Z^*)$  satisfies the conditions in Sec. 7.

**THEOREM 15.** Let  $F \sim D_\alpha$ . Then the posterior is Polya tree with respect to the partition  $T(Z^*)$  and with parameters

$$\alpha_{\epsilon_1, \epsilon_2, \dots, \epsilon_k}^* = \alpha(B_{\epsilon_1, \epsilon_2, \dots, \epsilon_k}) + \sum_{i=1: \delta_i=1}^n I_{B_{\epsilon_k}}(Z_i) + C_{\epsilon_k},$$

where  $C_{\epsilon_k}$  is the number of pairs  $(Z_i, 0)$ 's such that the interval  $(Z_i, \infty)$  is contained in  $B_{\epsilon_k}$ .

Note that if  $B_{\epsilon_k} = (Z_k, \infty)$ , then  $C_{\epsilon_k}$  is just the number of individuals on test at time  $Z_k$ .

The Polya tree representation of the posterior gives consistency.

**THEOREM 16.** If  $F \sim D_\alpha$ , then the posterior distribution of  $F$  is weakly consistent.

This and other related consistency results appear in Rajagopalan (1997).

Important classes of priors such as the simple homogeneous process (Ferguson and Phadia (1979)), and the extended gamma process (Dykstra and Laud (1981)) have been used for analyzing right censored data. These priors are neutral to the right in the sense of Doksum (1974). Another interesting class of priors for right censored data is the class of beta process introduced by Hjort (1990). These have been further generalized and studied by Muliere and Walker (Muliere and Walker (1997), Walker and Muliere (1997)). We expect the posterior consistency to hold in these cases.

### 11. NON-INFORMATIVE PRIORS

Except for the choice  $\alpha(\mathbf{R}) \rightarrow 0$  for a Dirichlet process, there is very little development towards a notion (or notions) of non-informative priors for infinite dimensional models. This section summarizes Ghosal et al. (1997a), where some tentative proposals were made.

Let  $\mathcal{F}$  be a family of densities equipped with, say, the Hellinger metric  $d_H(f, g)$  defined by  $d_H^2(f, g) = \int (\sqrt{f} - \sqrt{g})^2 dx$ . Assume further that  $\mathcal{F}$  is compact. For each  $\epsilon > 0$ , let  $\mathcal{F}_\epsilon$  be an  $\epsilon$ -sieve, i.e., a maximal set with the property  $\mathcal{F}_\epsilon \subset \mathcal{F}$  and  $d_H(f, g) > \epsilon$  for every  $f, g$  in  $\mathcal{F}_\epsilon$ . Since  $\mathcal{F}$  is compact,  $\mathcal{F}_\epsilon$  is a finite set. Denote by  $D(\epsilon, \mathcal{F})$  the cardinality of  $\mathcal{F}_\epsilon$ .

Since  $\mathcal{F}_\varepsilon$  is, in a sense, an  $\varepsilon$ -approximation of  $\mathcal{F}$ , a natural approach would be to take the uniform distribution on  $\mathcal{F}_\varepsilon$ , say  $\Pi_\varepsilon$ , as an approximation to whatever might be considered as the "uniform" distribution on  $\mathcal{F}$ . In practice,  $\varepsilon$  should depend on the sample size to reflect the disposition to entertain more complex models when a large sample is available. This approach, while attractive, would, in view of the dependence of the prior on sample size, run into problems of incoherence in the sense of Heath and Sudderth (1978).

Another approach would be to take any limit point  $\Pi^*$  of  $\{\Pi_\varepsilon : \varepsilon > 0\}$  as a non informative prior. If  $\Pi^*$  is unique, it is precisely the uniform distribution defined and studied by Dembski (1990). In Ghosal et al. (1997a), it is shown that in finite dimensional regular models this approach leads to Jeffreys' prior.

In keeping with the spirit of the mixture models studied earlier, yet another alternative is to view  $\varepsilon$  as a hyper parameter and consider a hierarchical prior  $\lambda$  for  $\varepsilon$ . In Ghosal et al. (1997a), the following consistency result is proved for such priors.

**THEOREM 17.** Let  $\mathcal{F}$  be a family of densities where  $\mathcal{F}$ , metrized by the Hellinger distance, is compact. Let  $\varepsilon_n$  be a positive sequence satisfying the condition  $\sum_{n=1}^{\infty} n^{1/2} \varepsilon_n < \infty$ . Let  $\mathcal{F}_n$  be an  $\varepsilon_n$ -sieve in  $\mathcal{F}$ ,  $\mu_n$  be the uniform distribution on  $\mathcal{F}_n$  and  $\mu$  be the probability on  $\mathcal{F}$  defined by  $\mu = \sum_{n=1}^{\infty} \lambda_n \mu_n$ , where  $\lambda_n$ 's are positive numbers adding upto unity. If for any  $\beta > 0$

$$\lim_{n \rightarrow \infty} e^{\beta n} \frac{\lambda_n}{D(\varepsilon_n, \mathcal{F}_n)} = \infty, \quad (3)$$

then the posterior distribution based on the prior  $\mu$  and iid observations  $X_1, X_2, \dots, X_n$  is consistent at every  $f_0 \in \mathcal{F}$ .

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

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

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 17 leads to consistent posterior. With a little modification of the construction of the prior, in this

case it is possible to achieve a rate of convergence of the order  $n^{-(r+m)/(2(r+m)+1)}$  of the posterior distribution, which is optimal.

Diaconis and Freedman (1993) consider a binary regression problem and show that a hierarchical prior based on uniform distributions (on certain finite dimensional sets) leads to a consistent posterior under a condition on the rate of decay of the hierarchical weights. This prior is somewhat similar in spirit to what we considered above.

While we have demonstrated the feasibility of obtaining consistency for a variety of popular priors, consistency by itself is not adequate. Given a consistency result, one would want results on rates of convergence or simulations to get an idea of how large an  $n$  is required to get convergence to  $\delta_{P_0}$  to a desired level of accuracy. More precisely, given  $P_0, \varepsilon, \eta$  and  $U$ , one would want an  $n_0$  such that for  $n > n_0$ ,  $\Pi(U|X_1, X_2, \dots, X_n) \geq 1 - \varepsilon$  with  $P_0$ -probability greater than  $1 - \eta$ . Such results or simulations would be relatively easy to get for tail free priors and weak neighborhoods  $U$ . Similar results in the context of Theorem 3 or 5 will require more work. We will return to these topics elsewhere.

## REFERENCES

- Antoniak, C. (1974). Mixtures of Dirichlet processes with application to Bayesian non-parametric problems. *Ann. Statist.* 2: 1152–1174.
- Barron, A. R. (1989). Uniformly powerful goodness of fit tests. *Ann. Statist.* 17: 107–124.
- Barron, A. R., Schervish, M. and Wasserman, L. (1996). The consistency of posterior distributions in non parametric problems. Technical report, Carnegie Mellon University. *Ann. Statist.* To appear.
- Berk, R. (1966). Limiting behavior of the posterior distribution when the model is incorrect. *Ann. Math. Statist.* 37: 51–58.
- Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *Ann. Statist.* 1: 353–355.
- Blum, J. and Susarla, V. (1977). On the posterior distribution of a Dirichlet process given randomly right censored observations. *Stoch. Processes Appl.* 5: 207–211.
- Brunner, L. J. (1992). Bayesian nonparametric methods for data from a unimodal density. *Statist. Probab. Lett.* 14: 195–199.

- Brunner, L. J. (1995). Bayesian linear regression with error terms that have symmetric unimodal densities. *J. Nonparameteric Statist.* 4: 335–348.
- Brunner, L. J. and Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.* 17: 1550–1566.
- Dalal, S. R. (1979). Dirichlet invariant processes and application to nonparametric estimation of symmetric distributions. *Stoch. Process Appl.* 9: 99–107.
- Dembski, W. A. (1990). Uniform probability. *J. Theoret. Probab.* 3: 611–626.
- ✓ Diaconis, P. and Freedman, D. (1986a). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* 14: 1–67.
- Diaconis, P. and Freedman, D. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.* 14: 68–87.
- Diaconis, P. and Freedman, D. (1993). Nonparametric binary regression: a Bayesian approach. *Ann. Statist.* 21: 2108–2137.
- Doksum, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* 2: 183–201.
- Doob, J. L. (1948). Application of the theory of martingales. *Coll. Int. du CNRS, Paris*, 22–28.
- Doss, H. (1985). Bayesian nonparametric estimation of the median. II. Asymptotic properties of the estimates. *Ann. Statist.* 13: 1445–1464.
- Doss, H. and Sellke, T. (1982). The tails of probabilities chosen from a Dirichlet prior. *Ann. Statist.* 10: 1302–1305.
- Dykstra, R. L. and Laud, P. W. (1981). A Bayesian nonparameteric approach to reliability. *Ann. Statist.* 9: 356–367.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* 1: 209–230.
- Ferguson, T. S. (1974). Prior distribution on the spaces of probability measures. *Ann. Statist.* 2: 615–629.
- Ferguson, T. S. (1983). Bayesian density estimation by mixtures of Normal distributions. In *Recent advances in Statistics* (Rizvi M., Rustagi, J. and Siegmund, D., Eds.) 287–302.
- Ferguson, T. S. and Phadia, E. G. (1979). Bayesian nonparameteric estimation based on censored data. *Ann. Statist.* 7: 163–186.

- Ferguson, T. S., Phadia, E. G. and Tiwari, R. (1996). Bayesian nonparametric inference. In *Current issues in Statistical inference. Essays in honor of D. Basu* (Ghosh, M. and Pathak, P. K., Eds.) 127–150.
- Freedman, D. (1963). On the asymptotic distribution of Bayes estimates in the discrete case I. *Ann. Math. Statist.* 34: 1386–1403.
- Freedman, D. and Diaconis, P. (1983). On inconsistent Bayes estimates in the discrete case. *Ann. Statist.* 11: 1109–1118.
- Gasparini, M. (1992). *Bayes Nonparametrics for biased sampling and density estimation*. Ph. D. thesis, University of Michigan.
- Ghorai, J. K. and Rubin, H. (1982). Bayes risk consistency of nonparametric Bayes density estimates. *Austral. J. Statist.* 24: 51–66.
- Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997a). Noninformative priors via sieves and consistency. In *Advances in Statistical Decision Theory and Applications* (S. Panchapakesan and N. Balakrishnan, Eds.) Birkhauser, Boston, 1997, 119–132.
- Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997b). Posterior consistency of Dirichlet mixtures in density estimation. Technical report # WS-490, Vrije Universiteit, Amsterdam. *Ann. Statist.* To appear.
- Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1998). Consistent semiparametric estimation about a location parameter. *J. Statist. Plan. Inf.* To appear.
- Ghosh, J. K. and Ramamoorthi R. V. (1997). *Lecture notes on Bayesian asymptotics*. Under preparation.
- Hartigan, J. A. (1996). Bayesian histograms. In *Bayesian statistics 5* (Bernardo J. et al. Eds.) 211–222.
- Heath, D. and Sudderth, W. (1978). On finitely additive priors, coherence and extended admissibility. *Ann. Statist.* 6: 333–345.
- Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* 18: 1259–1294.
- Hjort, N. L. (1996). Bayesian approaches to non- and semiparametric density estimation. In *Bayesian statistics 5* (Bernardo J. et al., Eds.) 223–253.
- Kolmogorov, A. N. and Tihomirov, V. M. (1961).  $\varepsilon$ -entropy and  $\varepsilon$ -capacity of sets in function spaces. *Amer. Math. Soc. Transl. Ser. 2*: 17 277–364. (Translated from Russian: *Uspekhi Mat. Nauk* 14: 3–86, (1959).)

- Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* 20: 1222–1235.
- Lavine, M. (1994). More aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* 22: 1161–1176.
- Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* 1: 38–53.
- Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.* 83: 509–516.
- Leonard, T. (1978). Density estimation, stochastic processes, and prior information. *J. Roy. Statist. Soc., Ser. B* 40: 113–146.
- Lo, A. Y. (1994). On a class of Bayesian nonparametric estimates I: Density estimates. *Ann. Statist.* 12: 351–357.
- Mauldin, R. D., Sudderth, W. D. and Williams, S. C. (1992). Polya trees and random distributions. *Ann. Statist.* 20: 1203–1221.
- Muliere, P. and Walker, S. G. (1997). A Bayesian nonparametric approach to survival analysis using Polya trees. *Scand. J. Statist.* 24: 231–240.
- Newton, M. A. (1994). A diffuse prior limit in semiparametric binary regression. Technical Report # 936, Department of Statistics, University of Wisconsin, Madison.
- Rajagopalan, K. Srikanth (1997). *Posterior consistency in some Bayesian nonparametric problems*. Ph. D. Thesis, Michigan State University.
- Schwartz, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* 4: 10–26.
- Sethuraman, J. and Tiwari, R. (1982). Convergence of Dirichlet measures and interpretation of their parameters. In *Statistical Decision Theory and Related Topics. III 2* (Gupta, S. S. and Berger, J. O., Eds.), Academic Press, New York, 305–315.
- Schervish, M. J. (1995). *Theory of Statistics*. Springer Series in Statistics. Springer-Verlag, New York.
- Shyamalkumar, N. D. (1996). *Contributions to Bayesian Nonparametrics and Bayesian Robustness*. Unpublished Ph. D. Thesis, Purdue University.
- Petrone, S. (1997). Bayesian density estimation using Bernstein polynomials. Preprint.

- Susarla V. and van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* 71: 897–902
- Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M. (1997). Bayesian nonparametric inference for random distributions and related functions. Preprint.
- Walker, S. G. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Polya-urn scheme. *Ann. Statist.* 25: 1762–1780.
- West, M. (1992). Modeling with Mixtures. In *Bayesian Statistics 4*: (J. M. Bernardo et al., Eds.) 503–524.
- 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*. Eds: A. F. M. Smith and P. Feeman, John Wiley and Sons, New York. 363–386.
- Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* 23: 339–362.