The Annals of Statistics
1980, Vol. 8, No. 3, 664–672

ON APPROXIMATING PARAMETRIC BAYES MODELS BY NONPARAMETRIC BAYES MODELS

BY S. R. DALAL[1] AND GAINEFORD J. HALL, JR.

Rutgers University and The Rand Corporation

Let τ be a prior distribution over the parameter space Θ for a given parametric model P_θ, $\theta \in \Theta$. For the sample space $\mathcal{X}$ (over which P_θ's are probability measures) belonging to a general class of topological spaces, which include the usual Euclidean spaces, it is shown that this parametric Bayes model can be approximated by a nonparametric Bayes model of the form of a mixture of Dirichlet processes prior, so that (i) the nonparametric prior assigns most of its weight to neighborhoods of the parametric model, and (ii) the Bayes rule for the nonparametric model is close to the Bayes rule for the parametric model in the no-sample case. Moreover, any prior parametric or nonparametric, may be approximated arbitrarily closely by a prior which is a mixture of Dirichlet processes. These results have implications in Bayesian inference.

1. Introduction. In the usual parametric Bayes approach to point estimation problems, the Bayesian assumes that the data $\mathbf{X} = (X_1, \cdots, X_n)$ are distributed on the sample space $\mathcal{X}$ according to some measure P_θ, $\theta \in \Theta$, and that given θ, the random variables are i.i.d. Moreover, Θ is typically assumed to be a subset of some Euclidean space R^k, and often each P_θ is assumed to have a density $f(x|\theta)$ with respect to some σ-finite measure μ on $\mathcal{X}$. In the Bayes approach θ is treated as random and a prior distribution $\tau(d\theta)$ is assigned to Θ. The problem is to estimate some function $g(\theta)$ of the parameter, using the data at hand. Given a loss function $L(g(\theta), d)$, the Bayes estimate of $g(\theta)$ is the $d(x)$ which minimizes

$$\int \left[\int L(g(\theta), d(\mathbf{x})) \tau(d\theta|\mathbf{x}) \right] F_{\mathbf{X}}(d\mathbf{x}) \tag{1}$$

where $\tau(d\theta|\mathbf{x})$ is the posterior distribution of θ given the sample $\mathbf{X} = \mathbf{x}$ and $F_{\mathbf{x}}$ is the marginal distribution of X.

Ferguson ([7], [8]) has recently introduced Dirichlet processes as priors for the nonparametric Bayes estimation problems. This class of priors, besides having "large support" in the space $\mathcal{M}(\mathcal{X})$ of all probability measures over $\mathcal{X}$, leads to analytically tractable and usually easily calculable posteriors and Bayes decisions. Antoniak [1], extending Ferguson's work, showed that the use of Dirichlet processes, at times, naturally leads to the posteriors which are the mixtures of Dirichlet processes. Further the fact that a mixture of Dirichlet processes is conditionally a Dirichlet process allows one to carry many properties of Dirichlet

Received October 1976; revised July 1978.

[1]The research partially supported by U.S. Army Research Office. The revisions by this author were supported by a Grant, No. MCS-78-02160, from the National Science Foundation.

AMS 1970 *subject classifications.* Primary 62G99; secondary 60K99.

Key words and phrases. Parametric Bayes model, nonparametric Bayes model, Dirichlet process prior, mixture of Dirichlet processes, adequacy.

processes to the mixtures, and facilitates the task of finding the posterior. Several applications involving the mixtures as priors, including Bioassay, Regression, Empirical Bayes, etc., have also been discussed by Antoniak.

In this paper we show that Dirichlet processes and mixtures of Dirichlet processes can have large support in the topology of weak convergence on $\mathcal{M}(\mathcal{X})$. This is a type of "richness" property which Antoniak [1] (cf. Raiffa and Schlaifer [14], page 44) finds desirable. However this type of "richness" is not enough: we believe that "richness" of a class of priors can only be exhibited if one can show that given any prior belief one can approximate it in the class considered. We show that any prior can be approximated as closely as desired in the topology of weak convergence for distributions by a prior which is a mixture of Dirichlet processes. Thus the class MDP of mixtures of Dirichlet process priors is "rich" in this sense. We call this property *adequacy*. Using this it is shown that given any neighborhood $\mathcal{O}$ of $\{P_\theta; \theta \in \Theta\}$ in $\mathcal{M}(\mathcal{X})$, there is an MDP prior $\mathcal{P}$ approximating beliefs in τ regarding $\{P_\theta; \theta \in \Theta\}$, assigning most of its weight to $\mathcal{O}$. This can have important implications for Bayesian statistics, since it may be that the true unknown distribution P governing the data is not actually any of the P_θ, but is only in some neighborhood (relative to the weak topology on $\mathcal{M}(\mathcal{X})$) of the parametric model. Thus it is important to study priors having large support in $\mathcal{M}(\mathcal{X})$ and assigning most of their weight to neighborhoods of the parametric model.

Under a different formulation due to Ferguson [7] and Doksum [6] results related to adequacy and finite additivity have been obtained by Dalal [5].

2. Preliminaries. For any topological space $\Omega, \mathcal{B}(\Omega)$ denotes the Borel σ-algebra on Ω. In the following, only nonnull regular measures on $\mathcal{B}(\Omega)$ will be considered. S_μ will denote the support of measure μ. The sample space $\mathcal{X}$ will usually be a compact Hausdorff space or a metric space. The space $\mathcal{M}(\mathcal{X})$ is the collection of all regular probability measures P on $\mathcal{X}$ endowed with the topology of weak convergence. The space $\mathcal{M}(\mathcal{M}(\mathcal{X}))$ is the collection of all probability measures $\mathcal{P}$ on $\mathcal{M}(\mathcal{X})$ (i.e., priors) together with the topology of weak convergence derived from $\mathcal{M}(\mathcal{X})$. If $\mathcal{X}$ is compact Hausdorff then $\mathcal{M}(\mathcal{X})$ is compact Hausdorff and if $\mathcal{X}$ is separable metric so is $\mathcal{M}(\mathcal{X})$ (cf. [16]).

To study nonparametric problems in a Bayesian framework, Ferguson introduced a type of random probability measure known as Dirichlet process priors. Briefly, if α is a finite measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ the random element $P \in \mathcal{M}(\mathcal{X})$ is a *Dirichlet* process (and its distribution $\mathcal{P}_\alpha \in \mathcal{M}(\mathcal{M}(\mathcal{X}))$ is a *Dirichlet process prior*) if for every $k \geq 1$ and every measurable partition $A_1, \cdots, A_k$ of $\mathcal{X}$, the distribution of $(P(A_1), \cdots, P(A_k))$ is a Dirichlet distribution with parameters $(\alpha(A_1), \cdots, \alpha(A_k))$, written $D(\alpha(A_1), \cdots, \alpha(A_k))$. The distribution of P is denoted by $P \in \mathcal{D}(\alpha)$.

Antoniak extended this definition to that of mixtures of Dirichlet processes. Let $(U, \mathcal{B}, H)$ be a probability space and assume $\alpha(\cdot, \cdot) : U \times \mathcal{B}(\mathcal{X}) \to [0, \infty)$ is a transition measure; i.e., for each $u \in U$, $\alpha(u, \cdot) = \alpha_u(\cdot)$ is a finite measure on

$(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and for each $A \in \mathcal{B}(\mathcal{X})$, $\alpha(\cdot, A)$ is measurable in $u \in U$. Then P is a *mixture of Dirichlet processes* if the conditional distribution of P given u is $\mathcal{D}(\alpha_u)$. We write $P \in \int_U \mathcal{D}(\alpha_u) H(du)$ to denote this and $\mathcal{P}_H \in \mathcal{M}(\mathcal{M}(\mathcal{X}))$ denotes the distribution of P.

Ferguson showed that if $P \in \mathcal{D}(\alpha)$ and if $\mathcal{M}(\mathcal{X})$ is given the topology of pointwise convergence (if $Q_n, Q \in \mathcal{M}(\mathcal{X})$, $Q_n \to Q$ if and only if $Q_n(A) \to Q(A)$ for all $A \in \mathcal{B}(\mathcal{X})$) then $\mathcal{P}_\alpha(\mathcal{O}) > 0$ for every open neighborhood $\mathcal{O}$ of Q if and only if $Q \ll \alpha$. Antoniak gave a sufficient condition that $\mathcal{P}_H(\mathcal{O}) > 0$ if P is an MDP (mixture of Dirichlet processes). We show that with the topology of weak convergence on $\mathcal{M}(\mathcal{X})$, $Q \in S_{\mathcal{P}_\alpha}$ if and only if $S_Q \subseteq S_\alpha$. We also give a sufficient condition that $Q \in S_{\mathcal{P}_H}$ (analogous to Antoniak's result).

In this paper we shall prefer to work with the following definition of random probability measure, used by Jagers [12].

DEFINITION. Let $(\Omega, \mathcal{F}, Q)$ be a probability space and $P : (\Omega, \mathcal{F}) \to (\mathcal{M}(\mathcal{X}), \mathcal{B}(\mathcal{M}(\mathcal{X})))$ be measurable. Then $P(\cdot)$ is a random probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

As shown in Ferguson [7], there is a version of the Dirichlet process P which satisfies the above definition. It is now necessary to show that for any random probability P and any $A \in \mathcal{B}(\mathcal{X})$, $P(A)$ is a random variable (i.e., is measurable). Let Ω be a set and $\mathcal{S}$ a class of subsets of Ω. $\mathcal{S}$ is a *π-system* if it is closed under finite intersections, and it is a *d-system* if (i) $\Omega \in \mathcal{S}$, (ii) $A, B \in \mathcal{S}$, $A \subset B$ implies $B - A \in \mathcal{S}$, (iii) the countable union of a monotone increasing sequence of members of $\mathcal{S}$ is again in $\mathcal{S}$. For any class $\mathcal{S}$ let $d(\mathcal{S})$ and $\sigma(\mathcal{S})$ denote the smallest d-system and smallest σ-algebra containing $\mathcal{S}$, respectively. We state the following two results without proof.

PROPOSITION 1. *If $\mathcal{S}$ is a π-system, then $d(\mathcal{S}) = \sigma(\mathcal{S})$.*

A proof of this is found in Jagers. Let $\mathcal{C} = \{A \in \mathcal{B}(\mathcal{X}) : P \to P(A)$ is $\mathcal{B}(\mathcal{M}(\mathcal{X}))$-measurable$\}$. We will show that $\mathcal{C} = \mathcal{B}(\mathcal{X})$ for any normal space $\mathcal{X}$. We need the following lemma (from [4]) based on Urysohn's lemma and the regularity of the measures involved.

LEMMA 1. *Let $(\mathcal{X}, \mathcal{T})$ be a normal space and $\mathcal{M}(\mathcal{X})$ be the class of all regular probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ endowed with the weak topology. Suppose A is a closed subset of $\mathcal{X}$ and for each real a let $M_a = \{P \in \mathcal{M}(\mathcal{X}) : P(A) \geqslant a\}$. Then M_a is a closed subset of $\mathcal{M}(\mathcal{X})$.*

PROPOSITION 2. *$\mathcal{C} = \mathcal{B}(\mathcal{X})$ for any normal space $\mathcal{X}$.*

PROOF. By the lemma, the class $\mathcal{K}$ of all closed sets is contained in $\mathcal{C}$. If A, B are in $\mathcal{C}$ and $A \subset B$, then $P(B - A) = P(B) - P(A)$, hence $B - A$ is in $\mathcal{C}$. Suppose $A_n \in \mathcal{C}$ and $A_n \uparrow A$; then $P(A_n) \to P(A)$, so that $A \in \mathcal{C}$. Thus $\mathcal{C}$ is a d-system containing $\mathcal{K}$ and so $d(\mathcal{K}) = \sigma(\mathcal{K}) = \mathcal{B}(\mathcal{X}) \subseteq \mathcal{C}$, completing the proof.

3. Some results on the support of MDP's. In this section we show that if $\mathcal{X}$ is either compact Hausdorff or metric, and if $F \in \mathcal{M}(\mathcal{X})$ and $P \in \mathcal{D}(\alpha)$, then $F \in S_{\mathcal{P}_\alpha}$ if and only if $S_F \subseteq S_\alpha$. Also we give a sufficient (but not necessary) condition that F belongs to the support an MDP.

Let $\mathcal{X}$ be a metric space and $F \in \mathcal{M}(\mathcal{X})$. Billingsley [2] shows that each of the following types of sets generates the topology of weak convergence for $\mathcal{M}(\mathcal{X})$ and is a basic open neighborhood of F:

(i) $\{Q : Q(C_i) < F(C_i) + \varepsilon,\ 1 \leqslant i \leqslant k\}$ where each $C_i \subseteq \mathcal{X}$ is closed, $\varepsilon > 0$, $k \geqslant 1$ arbitrary;

(ii) $\{Q : |Q(A_i) - F(A_i)| < \varepsilon,\ 1 \leqslant i \leqslant k\}$ where each A_i is an F-continuity set (i.e., $F(\partial A_i) = 0$) and $\varepsilon > 0$, $k \geqslant 1$ arbitrary.

Furthermore, if $\mathcal{X}$ is compact Hausdorff, sets of type (i) also generate the topology of weak convergence (Varadarajan [16]).

THEOREM 1. *Let $\mathcal{X}$ be metric or compact Hausdorff. Then $F \in S_{\mathcal{P}_\alpha}$ if and only if $S_F \subseteq S_\alpha$.*

PROOF. First suppose $\mathcal{X}$ is a metric space. If there is an element $x \in S_F \cap S_\alpha^c$, then there is an open neighborhood V of x such that $F(V) > 0$ and $\alpha(V) = 0$. From Ferguson [7], if $P \in \mathcal{D}(\alpha)$ then $P(V) = 0$, a.s. Thus $\mathcal{P}_\alpha\{P : P(V^c) < F(V^c) + \varepsilon\} = 0$, where $\varepsilon < 1 - F(V^c)$. Hence $F \notin S_{\mathcal{P}_\alpha}$.

Conversely, assume $S_F \subseteq S_\alpha$. It suffices to show that for any neighborhood $\mathcal{O}$ of type (ii), $\mathcal{P}_\alpha(\mathcal{O}) > 0$. Form the 2^k sets obtained by intersections of the A_j and their complements, i.e., $B_{\nu_1, \ldots, \nu_k}$ for each $\nu_j = 0$ or 1 as $B_{\nu_1, \ldots, \nu_k} = \cap_{j=1}^k A_j^{\nu_j}$ where $A_j^1 = A_j$ and $A_j^0 = A_j^c$. Put $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_k)$. Now, by using the arguments similar to Proposition 3 of Ferguson [7], it suffices to show that

$$\mathcal{P}_\alpha\{|P(B_{\boldsymbol{\nu}}) - F(B_{\boldsymbol{\nu}})| < 2^{-k}\varepsilon, \forall \boldsymbol{\nu}\} > 0.$$

Note that each $B_{\boldsymbol{\nu}}$ is also an F-continuity set and let $\mathrm{Int}(A)$ denote the topological interior of A. If $\alpha(B_{\boldsymbol{\nu}}) = 0$ then $P(B_{\boldsymbol{\nu}}) = 0$, a.s. Also, $\alpha(\mathrm{Int}(B_{\boldsymbol{\nu}})) = 0$, so $F(\mathrm{Int}(B_{\boldsymbol{\nu}})) = 0$ (as $S_F \subseteq S_\alpha$). Thus $F(B_{\boldsymbol{\nu}}) = 0$ and therefore $|P(B_{\boldsymbol{\nu}}) - F(B_{\boldsymbol{\nu}})| = 0$, a.s. For those $\boldsymbol{\nu}$ with $\alpha(B_{\boldsymbol{\nu}}) > 0$, the joint distribution of the corresponding Dirichlet random variables $P(B_{\boldsymbol{\nu}})$ gives positive weight to all open sets in the set $\Sigma_{\boldsymbol{\nu} \ni \alpha(B_{\boldsymbol{\nu}}) > 0} P(B_{\boldsymbol{\nu}}) = 1$.

The case where $\mathcal{X}$ is compact Hausdorff is quite similar. The term "open set" of $\mathcal{X}$ is then replaced by "open Baire set" of $\mathcal{X}$. A net μ_λ of measures in $\mathcal{M}(\mathcal{X})$ converges to $\mu \in \mathcal{M}(\mathcal{X})$ if and only if $\mu_\lambda(A) \to \mu(A)$ for all those A for which there exist open Baire sets U_1, U_2 such that $U_1 \subset A \subset U_2^c$ and $\mu(U_2^c \cap U_1^c) = 0$ (Varadarajan [16]). A set A of this type will be called a μ continuity set. The necessary changes in the proof should now be clear.

It follows immediately from Theorem 1 that if $P \in \Sigma_{i=1}^m h_i \mathcal{D}(\alpha_i)$, $(h_i > 0)$, a finite mixture of Dirichlet processes, then $F \in S_{\mathcal{P}_H}$ if and only if $S_F \subseteq S_\alpha$, for some i. A sufficient but not necessary condition that $F \in S_{\mathcal{P}_H}$ is given in the next theorem.

THEOREM 2. *Let $\mathcal{X}$ be metric or compact Hausdorff, let $P \in \int_U \mathcal{D}(\alpha_u)H(du)$ and let $F \in \mathcal{M}(\mathcal{X})$. If there is an event $B \in \mathcal{B}$ such that $H(B) > 0$ and $S_F \subseteq S_{\alpha_u}$ for all $u \in B$, then $F \in S_{\mathcal{P}_H}$.*

PROOF. Let $\mathcal{O}$ be a neighborhood of F of type (ii). By Theorem 1, $\mathcal{P}_{\alpha_u}(\mathcal{O}) > 0$ for all $u \in B$. Thus $\mathcal{P}_H(\mathcal{O}) = \int_U \mathcal{P}_{\alpha_u}(\mathcal{O})H(du) > 0$.

To show that the above condition is not necessary, let $U = [0, 1]$, $H(du)$ be uniform on U, and $\alpha_u = \delta_u$. Then $\mathcal{D}(\alpha_u)$ is degenerate at δ_u, i.e., if $P|u \in \mathcal{D}(\alpha_u)$ then $P \equiv \delta_u$. Thus for $F \in \mathcal{M}(\mathcal{X})$ there is no set B of positive H-measure such that $S_F \subset S_{\delta_u}$.

We do not know of a necessary and sufficient condition. However, in many cases of interest, $S_{\alpha_u} = \mathcal{X}$, for all $u \in U$, so that $S_{\mathcal{P}_H} = \mathcal{M}(\mathcal{X})$.

4. The adequacy of mixtures of Dirichlet processes. In this section we give several theorems which will enable us to approximate parametric Bayes models by nonparametric models using Dirichlet processes and MDP's. We give the proof for $\mathcal{X}$ compact Hausdorff and sketch the proof for $\mathcal{X}$ Polish (i.e., a Borel subset of a complete separable metric space).

Let $\mathcal{F}(\mathcal{X}) = \{\Sigma_1^n a_i\delta_{x_i} : a_i \geqslant 0,\ \Sigma_1^n a_i = 1,\ x_i \in \mathcal{X} \forall i,\ 1 \leqslant i \leqslant n,\ n \geqslant 1\}$, i.e., $\mathcal{F}(\mathcal{X})$ is the class of all (atomic) probability measures with finite support on $\mathcal{X}$. $\overline{\mathcal{F}}(\mathcal{X})$ denotes the weak closure of $\mathcal{F}(\mathcal{X})$ in $\mathcal{M}(\mathcal{X})$.

LEMMA. *If $\mathcal{X}$ is compact Hausdorff then $\overline{\mathcal{F}(\mathcal{M}(\mathcal{X}))} = \mathcal{M}(\mathcal{M}(\mathcal{X}))$.*

PROOF. By repeated applications of the Riesz representation theorem and weak compactness of the closed unit sphere it follows that $\mathcal{M}(\mathcal{M}(\mathcal{X}))$ is compact Hausdorff. Now apply the Krein-Milman theorem (Robertson and Robertson [15]) to $\mathcal{M}(\mathcal{X})$ to conclude that $\overline{\mathcal{F}(\mathcal{M}(\mathcal{X}))} = \mathcal{M}(\mathcal{M}(\mathcal{X}))$.

THEOREM 1. *Let $\mathcal{X}$ be a compact Hausdorff space and let $\{\alpha_\lambda\}$ be a net of measures on $\mathcal{X}$ such that $\alpha_\lambda(\mathcal{X}) \uparrow \infty$ and $\alpha_\lambda(\cdot)/\alpha_\lambda(\mathcal{X}) \equiv G_\lambda(\cdot) \to_w F$. Then $\mathcal{P}_{\alpha_\lambda} \to_w \delta_F$.*

PROOF. We only need show that for any open set $\mathcal{O}$ containing F in $\mathcal{M}(\mathcal{X})$, $\mathcal{P}_{\alpha_\lambda}(\mathcal{O}) \to 1$. Assume $\mathcal{O}$ is a basic open set of type (ii) of Section 3, containing F. Again form the 2^k sets B_ν obtained from intersection of the A_i and A_i^c and note

$$\mathcal{P}_{\alpha_\lambda}(\mathcal{O}) \geqslant \mathcal{P}_{\alpha_\lambda}\{|P(B_\nu) - F(B_\nu)| < 2^{-k}\varepsilon, \forall \nu\}. \tag{1}$$

Since $\{P(B_\nu)\}$ have a joint Dirichlet distribution and since $\mathcal{E}_{\alpha_\lambda}P(A) = \alpha_\lambda(A)/\alpha_\lambda(\mathcal{X})$ and $\operatorname{Var}_{\alpha_\lambda}P(A) = \alpha_\lambda(A)\alpha_\lambda(A^c)/[\{\alpha_\lambda(\mathcal{X})\}^2\{\alpha_\lambda(\mathcal{X}) + 1\}]$ for any event A and $\alpha_\lambda(A)/\alpha_\lambda(\mathcal{X}) \to F(A)$ for any F-continuity set A, it is clear that the probability on the right-hand side of (1) converges to 1 as $n \to \infty$, completing the proof.

Theorem 1 combined with the lemma immediately yield the following result.

THEOREM 2. *If $\mathcal{X}$ is compact Hausdorff then $\overline{\text{MDP}} = \mathcal{M}(\mathcal{M}(\mathcal{X}))$.*

The method of proof of Theorem 1 together with the Portmanteau theorem and the fact that $\mathcal{F}(\mathcal{X})$ is weak dense in $\mathcal{M}(\mathcal{X})$ is $\mathcal{X}$ is Polish (Billingsley [2]) produce the analogous result for $\mathcal{X}$ Polish:

THEOREM 3. *If $\mathcal{X}$ is a Polish space then* $\overline{\text{MDP}} = \mathcal{M}(\mathcal{M}(\mathcal{X}))$.

The last theorem of this section is a result for MDP's analogous to Theorem 1. It will be useful in the next section on applications.

THEOREM 4. *Let $\mathcal{X}$ be a Polish space or a compact Hausdorff space. Let $\{\alpha_n(u, \cdot);\ u \in U,\ n \geq 1\}$ be a sequence of transition measures on $\mathcal{X}$ such that for each $u \in U$, $\alpha_n(u, \mathcal{X}) \uparrow \infty$ as $n \uparrow \infty$ and $\alpha_n(u, \cdot)/\alpha_n(u, \mathcal{X}) \equiv G_n(u, \cdot) \to_w F(u, \cdot) \in \mathcal{M}(\mathcal{X})$, where $F(u, \cdot)$ is a transition probability. Let H_n, $n \geq 0$, be probability measures on $(U, \mathcal{B})$ and σ be a σ-finite measure on $(U, \mathcal{B})$ such that if $h_n = dH_n/d\sigma$, $n \geq 0$, then $h_n \to h_0$ a.s. $-[\sigma]$. If $\mathcal{P}_{H_n} = \int_U \mathcal{D}(\alpha_n(u, \cdot))H_n(du)$ and $\mathcal{Q}_0 = \int_U \delta_{F(u,\cdot)} H_0(du)$ then $\mathcal{P}_{H_n} \to_w \mathcal{Q}_0$.*

PROOF. Let $\mathcal{O}$ be open in $\mathcal{M}(\mathcal{X})$. It suffices to show $\liminf_n \mathcal{P}_{H_n}(\mathcal{O}) \geq \mathcal{Q}_0(\mathcal{O})$. If $E = \{u \in U : F(u, \cdot) \in \mathcal{O}\}$ then $\mathcal{Q}_0(\mathcal{O}) = H_0(E)$. Thus by Fatou's lemma,

$$
\begin{aligned}
\liminf_n \mathcal{P}_{H_n}(\mathcal{O}) &= \liminf_n \int_U \mathcal{P}_{\alpha_n(u,\cdot)}(\mathcal{O}) h_n(u)\sigma(du) \\
&\geq \int_U \liminf_n \{\mathcal{P}_{\alpha_n(u,\cdot)}(\mathcal{O}) h_n(u)\}\sigma(du) \\
&\geq \int_E \lim_n \{\mathcal{P}_{\alpha_n}(\mathcal{O}) h_n(u)\}\sigma(du) \\
&= \int_E 1 \cdot h_0(u)\sigma(du) = H_0(E)
\end{aligned} \tag{2}
$$

since for each $u \in E$, $\lim_n \mathcal{P}_{\alpha_n(u,\cdot)}(\mathcal{O}) = 1$.

In particular if $H_n \equiv H_0$, for all n, then $\int_U \mathcal{D}(\alpha_n(u, \cdot))H_0(du)$ converges weakly to $\int_U \delta_{F(u,\cdot)} H_0(du)$, where $\sigma = H_0$ and $h_0 \equiv 1$.

5. Interpretation and applications of the main results. We turn now to the question of approximating parametric Bayes models. Let $\{\mathcal{P}_\theta;\ \theta \in \Theta\}$, $\tau(d\theta)$ be the Bayes model described in the introduction. Assume that $\mathcal{X}$ is compact Hausdorff or Polish (e.g., Euclidean space). Take $U = \Theta$ and $H(d\theta) = \tau(d\theta)$ for the MDP. Let $M : \Theta \to (0, \infty)$ be measurable and put $\alpha(\theta, \cdot) = M(\theta)P_\theta(\cdot)$. Then for any $\varepsilon > 0$ and any open neighborhood $\mathcal{O}$ of the set $\{P_\theta;\ \theta \in \Theta\}$ in $\mathcal{M}(\mathcal{X})$, it is possible to find $M(\theta)$ so that

$$
\mathcal{P}_\tau(\mathcal{O}) \geq 1 - \varepsilon. \tag{1}
$$

To show this let $\mathcal{O}_\theta$ be an open neighborhood of P_θ of type (ii) of Section 3 so that $\mathcal{O}_\theta \subseteq \mathcal{O}$, and choose $M(\theta)$ so large that $\mathcal{P}_{\alpha(\theta,\cdot)}(\mathcal{O}_\theta) \geq 1 - \varepsilon$ (Theorem 1, 2, Section 4). Then $\mathcal{P}_\tau(\mathcal{O}) \geq \int \mathcal{P}_{\alpha(\theta,\cdot)}(\mathcal{O}_\theta)\tau(d\theta) \geq 1 - \varepsilon$.

Thus we can find a mixture of Dirichlet processes prior which will assign most of its mass to neighborhoods of $\{P_\theta;\ \theta \in \Theta\}$. Moreover, if there is a set $B \subseteq \Theta$ such that $\tau(B) > 0$ and $S_{\alpha(\theta,\cdot)} = \mathcal{X}$ for all $\theta \in B$, then $S_{\mathcal{P}_\tau} = \mathcal{M}(\mathcal{X})$, so the true unknown distribution F governing the data (which may only be known to be

"near" the model $\{P_\theta; \theta \in \Theta\}$ in the topology of weak convergence) will lie in $S_{\mathcal{P}_\tau}$ (Theorem 2, Section 3). In many cases of interest, it will happen that $S_{P_\theta} = \mathcal{X}$ for all $\theta \in \Theta$. Further note that the MDP thus obtained approximates (parametric beliefs in) τ (in the weak convergence sense) when τ is viewed as a prior over $\{P_\theta; \theta \in \Theta\}$, rather than merely on Θ. Thus in this broad sense a nonparametric Bayes model can be found approximating the given parametric Bayes model.

As mentioned earlier, it is advantageous to have an approximating sequence of MDP's for any specified prior $\mathcal{Q}$, since MDP's usually facilitate computations. This is not to say that the sequences in the earlier construction are suitable for numerical computations. In fact the problems of finding efficient, computationally convenient approximating sequences, the rates of convergence, etc., need further investigation before an attempt at numerical approximation can be made. However, a practical advice to the Bayesian seeking a suitable prior is that he should restrict his attention to a small class, the class of mixtures of Dirichlet processes.

To understand the nature of this result assume that $\mathcal{X}$ is a Polish space with metric ρ. Then $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}(\mathcal{M}(\mathcal{X}))$ are Polish spaces with Prohorov distances ρ_1, ρ_2 respectively. We may restrict attention to sequences of measures. To elaborate further we state the following theorem given in Pyke [13]. An elementary treatment can also be found in Billingsley [3].

THEOREM. (Skorokhod-Dudley). *If (S, m) is a separable metric space, and $\{P_n\}$, P are probability measures thereon, $P_n \to_w P$ implies the existence of a probability space $(\Omega^*, \mathcal{A}^*, P^*)$ and measurable functions $X_n^* : \Omega^* \to S$ and $X^* : \Omega^* \to S$ such that $P_n = P^* X_n^{-1}$ and $P = P^* X^{-1}$ and $m(X_n^*, X^*) \to 0$ a.s.*

Thus if $\mathcal{Q}$ is a given prior probability measure and if $\mathcal{P}_n \to_w \mathcal{Q}$ is an approximating sequence of MDP's then by this theorem there exists a probability space $(\Omega^*, \mathcal{A}^*, P^*)$ and measurable mappings $\zeta_n : \Omega^* \to \mathcal{M}(\mathcal{X})$, $\zeta : \Omega^* \to \mathcal{M}(\mathcal{X})$ such that $P^* \zeta_n^{-1} = \mathcal{P}_n$, $P^* \zeta^{-1} = \mathcal{Q}$ and $\rho_1(\zeta_n(\omega^*), \zeta(\omega^*)) \to 0$ a.s. $-[P^*]$. In addition for each ω^*, $\zeta_n(\omega^*)$, $\zeta(\omega^*)$ are probability measures on $(\mathcal{X}, \mathcal{M}(\mathcal{X}))$ and since convergence in ρ_1 norm is equivalent to weak convergence on $\mathcal{M}(\mathcal{X})$, for almost all $\omega^* - [P^*]$ there exists a probability space $(\Omega_{\omega^*}, \mathcal{A}_{\omega^*}, Q_{\omega^*})$ and measurable mappings $\{X_{n,\omega^*}\}_{n=1}^{\infty}$ and X_{ω^*} from Ω_{ω^*} into $\mathcal{X}$ such that $\zeta_n(\omega^*) = Q_{\omega^*} X_{n,\omega^*}^{-1}$, $\zeta(\omega^*) = Q_{\omega^*} X_{\omega^*}^{-1}$ and $\rho(X_{n,\omega^*}(\omega), X_{\omega^*}(\omega)) \to 0$ a.s. $-[Q_{\omega^*}]$.

Thus we have the interpretation that given a prior probability measure $\mathcal{Q}$ on a Polish space $\mathcal{X}$, there is a sequence $\mathcal{P}_n$ of MDP's such that samples (on $\mathcal{X}$) obtained from realizations (which are probability measures on $\mathcal{X}$) of $\mathcal{Q}$ and $\mathcal{P}_n$ are arbitrarily close. The interpretation is analogous to that in the usual Bayesian parametric problems.

However, this result has some limitations. One would like to know whether $\mathcal{P}_n \to_w \mathcal{Q}$ implies that the joint distribution of random variables $(P_n(B_1), \cdots, P_n(B_k))$ converges in law to the distribution of $(Q(B_1), \cdots, Q(B_k))$. The answer in general is negative. Another difficulty with these results is that

without imposing further conditions one cannot approximate the posterior as a limit of corresponding posteriors of the approximating priors.

As an example of applications, let the loss function $L(\cdot, \cdot)$ for the parametric Bayes model have the form $L(\theta, d) = \rho(g(\theta) - d)$ where $g : \Theta \to R$ and $\rho : R \to R$ is differentiable with $\Psi = \rho'$. For $Q \in \mathfrak{M}(R)$, define $T_\rho(Q)$ as the solution to

$$\int \Psi(x - t)Q(dx) = 0 \tag{2}$$

whenever the solution exists and is unique (Huber [11]). For the parametric model, the Bayes rule for the no sample problem is $d^* = T_\rho(\pi_g)$ where π_g is the distribution of $g(\theta)$ under $\tau(d\theta)$.

Define $\hat{g}(P_\theta) = g(\theta)$, where we assume the map $\theta \to P_\theta$ is (weak) continuous. If $\hat{g}$ can be extended to all of $\mathfrak{M}(\mathfrak{X})$ in such a way that $\hat{g}$ is continuous a.s.-$[\mathfrak{Q}_0]$, where $\mathfrak{Q}_0 = \int \delta_{P_\theta} \tau(d\theta)$ then $\hat{\pi}_g \to_w \pi_g$, $\hat{\pi}_g$ denoting the distribution of $\hat{g}(P)$ under $\mathfrak{P}_\tau$. Thus we may approximate the parametric model as closely as desired by nonparametric model so that the distribution of $\hat{g}(P)$ is as close as desired to the distribution of $g(\theta)$. If T_ρ is continuous at π_g, the Bayes rule $T_\rho(\hat{\pi}_g)$ will be close to $T_\rho(\pi_g)$. Suppose, for instance, that P_θ is $\mathfrak{N}(\theta, 1)$, $\theta \in R$, and $g(\theta) = \hat{g}(P_\theta) = \theta$, the mean of P_θ. Take $\tau(d\theta) = \mathfrak{N}(0, \sigma^2)$, σ^2 known. We may extend $\hat{g}$ to all of $\mathfrak{M}(\mathfrak{X})$ by setting $\hat{g}(P) = S_\beta(P)$, the β-trimmed mean, $0 < \beta < \frac{1}{2}$. Then $S_\beta(P_\theta) \equiv \theta$ and $\hat{g}$ is continuous on $\mathfrak{M}(\mathfrak{X})$. Thus $\hat{\pi}_g$ converges weakly to π_g as $\mathfrak{P}_H$ converges weakly to $\mathfrak{Q}_0$. Moreover, if Ψ in (2) is bounded, continuous and strictly monotone increasing, then $T_\rho(\hat{\pi}_g) \to T_\rho(\pi_g)$ (Hampel [10]).

Acknowledgments. This research was carried out independently by each of the authors, and preliminary versions were circulated. However, due to the similarity of results, the research is published as a joint paper. The results of Dalal were extensions of those obtained in his thesis, and he is grateful to his advisor Professor W. J. Hall, and to Professor J. H. B. Kemperman for some helpful discussions. Both authors would like to thank Professor W. J. Hall for constructive comments on the first draft.

REFERENCES

[1] ANTONIAK, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174.

[2] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[3] BILLINGSLEY, P. (1971). *Weak Convergence of Measures: Applications in Probability*. *SIAM*, Philadelphia.

[4] DALAL, S. R. (1975). Some contributions to Bayes nonparametric decision theory. Ph.D. Dissertation. University of Rochester, New York.

[5] DALAL, S. R. (1978). On the adequacy of mixtures of Dirichlet processes. *Sankhyā, Ser.* A **40** 185–191.

[6] DOKSUM, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probability* **2** 183–201.

[7] FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

[8] FERGUSON, T. S. (1974). Prior distributions on space of probability measures. *Ann. Statist.* **2** 615–629.

[9] HALL, G. J. (1975). On approximating parametric Bayes Models by nonparametric Bayes models. Technical report, University of Texas, Austin.

[10] HAMPEL, F. (1971). A qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896.

[11] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

[12] JAGERS, P. (1974). *Aspects of Random Measures and Point Processes. Advances in Probability and Related Topics.* (P. Ney and S. Port, eds.) Vol. 3 179–240.

[13] PYKE, R. (1968). Applications of almost surely convergent constructions of weakly convergent processes. Boeing Scientific Research Laboratories. Math. Note No. 570.

[14] RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory.* Massachusetts Institute of Technology Press, Cambridge.

[15] ROBERTSON, A. P. and ROBERTSON, W. J. (1964). *Topological Vector Spaces.* Cambridge University Press.

[16] VARADARAJAN, V. S. (1965). Measures on topological spaces. *Amer. Math. Soc. Transl.* **48** 161–228.

DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
NEW BRUNSWICK, NEW JERSEY 08903

RAND CORPORATION
1700 MAIN STREET
SANTA MONICA, CALIFORNIA 90406