The Annals of Statistics
1998, Vol. 26, No. 6, 2157–2178

EXPONENTIAL POSTERIOR CONSISTENCY VIA GENERALIZED PÓLYA URN SCHEMES IN FINITE SEMIPARAMETRIC MIXTURES[1]

BY HEMANT ISHWARAN

Cleveland Clinic Foundation

Advances in Markov chain Monte Carlo (MCMC) methods now make it computationally feasible and relatively straightforward to apply the Dirichlet process prior in a wide range of Bayesian nonparametric problems. The feasibility of these methods rests heavily on the fact that the MCMC approach avoids direct sampling of the Dirichlet process and is instead based on sampling the finite-dimensional posterior which is obtained from marginalizing out the process.

In application, it is the integrated posterior that is used in the Bayesian nonparametric inference, so one might wonder about its theoretical properties. This paper presents some results in this direction. In particular, we will focus on a study of the posterior's asymptotic behavior, specifically for the problem when the data is obtained from a finite semiparametric mixture distribution. A complication in the analysis arises because the dimension for the posterior, although finite, increases with the sample size. The analysis will reveal general conditions that ensure exponential posterior consistency for a finite dimensional parameter and which can be slightly generalized to allow the unobserved nonparametric parameters to be sampled from a generalized Pólya urn scheme. Several interesting examples are considered.

1. Introduction. Let $f(x|\theta, y)$ denote a density taken with respect to a σ-finite measure λ on a measurable space $(\mathscr{X}, \mathscr{B})$ and let $\mathbb{P}_{\theta, y}$ denote its distribution, where (θ, y) are parameters with real valued parameters in $\Theta \otimes \mathscr{Y} \subseteq \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2}$. Notice that X, θ or y can be either univariate or multivariate. Let G_0 be some unspecified finite mixing distribution on $\mathscr{Y}$ with an unknown number of support points $k < \infty$, unknown distinct support values $y_{0,1}, \dots, y_{0,k}$ and unknown mixing probabilities $p_0 = (p_{0,1}, \dots, p_{0,k})'$, where $p_{0,j} > 0$ and $\sum_j p_{0,j} = 1$ [we will also write $y_0^k = (y_{0,1}, \dots, y_{0,k})'$ for the support vector]. Then

$$(1) \quad f_0(x) = f(x|\theta_0, G_0) = \int f(x|\theta_{0,y})\, dG_0(y) = \sum_{j=1}^{k} p_{0,j} f(x|\theta_0, y_{0,j})$$

is a finite semiparametric mixture density with respect to λ.

Received June 1997; revised August 1998.

[1] Supported in part by grant funds from the Natural Sciences and Engineering Research Council of Canada.

AMS 1991 *subject classifications*. Primary 62G20; secondary 62G05.

Key words and phrases. Dirichlet process prior, exchangeability, Pólya urn scheme, posterior consistency, relative entropy, semiparametric mixture.

Write X^n for the data $X_1, \dots, X_n$ and x^n for the observed values $x_1, \dots, x_n$ [this same superscript notation will also be used to represent other sequences of variables such as $y^n = (y_1, \dots, y_n) \in \mathscr{Y}^n$, for example]. We assume that $X_1, \dots, X_n$ are sampled independently from the distribution $\mathbb{P}_0$ having the finite semiparametric mixture density (1). A popular nonparametric Bayesian method for studying this problem is to assume that the mixing variables $Y_1, \dots, Y_n$ are conditionally independent given the distribution G, where G is distributed as a Dirichlet process (Ferguson, 1973). It is also a standard practice to assume that the data are conditionally independent given the parameters and that the parameters $Y^n = (Y_1, \dots, Y_n)$ and θ are independent r.v.'s. Therefore, writing π_θ for the prior distribution of θ and π_{Y^n} for the prior distribution of Y^n, the joint distribution for (θ, Y^n) factorizes as $\pi = \pi_\theta \times \pi_{Y^n}$. Furthermore, by integrating out G, and by the assumption of conditional independence, the Bayesian model implies the following hierarchical structure on the data:

$$
\begin{aligned}
(X_i | \theta, Y_i) &\sim {}_{\mathrm{Ind}} \mathbb{P}_{\theta, Y_i}, \qquad i = 1, \dots, n, \\
(\theta, Y^n) &\sim \pi = \pi_\theta \times \pi_{Y^n}.
\end{aligned} \tag{2}
$$

The main goal of this paper will be to study the asymptotic behavior of the posterior for θ from the Bayesian hierarchical model (2). In particular, the paper will present general conditions which ensure that the posterior for θ lies in each open neighborhood of θ_0 with exponentially high probability. The analysis will also slightly generalize the method discussed above in which $Y_1, \dots, Y_n$ are assumed to be a sample from the Dirichlet process prior by allowing $Y_1, \dots, Y_n$ to be a (not necessarily exchangeable) sample from a generalized Pólya urn scheme (an exact description of this mechanism is given in Section 2).

The reader should note carefully that the posterior analysis for θ in (2) will be based only on the prior for θ and the distribution π_{Y^n} for the Pólya urn scheme. The prior for G is left unspecified and plays no direct role in our analysis of θ. Furthermore, by focusing only on the finite-dimensional parameters (θ, Y^n), we end up (as a special case) with the posterior that Bayesians implicitly work with when they employ modern MCMC methods with the Dirichet process prior. Indeed, the very success of these methods depends upon hiding the Dirichlet process in the background while focusing instead on the much simpler task of sampling the posterior of (θ, Y^n) [see Escobar (1994) and Escobar and West (1995, 1998) for the general method as well as MacEachern (1994, 1998) for a discussion of more refined MCMC techniques].

The posterior analysis presented here is based on methodology patterned after Schwartz (1965), Barron (1988) and Clarke and Barron (1990). Related material also appears in Barron, Schervish and Wasserman (1998), and Wasserman (1998). These papers all consider the problem of exponential posterior consistency in one way or another. Schwartz (1965) discusses the problem for a general fixed parameter space, while Barron (1988), and Barron, Schervish and Wasserman (1998) study nonparametric problems with emphasis on density estimation. Clarke and Barron (1990) focus specifically on posterior consistency in parametric problems. The methods used in these papers can be adapted quite readily to our semiparametric problem, which will involve the analysis of a posterior whose dimension increases with the sample size. It is important to note, however, that although this posterior is finite-dimensional for a fixed sample size n, the problem considered here is still infinite-dimensional. This follows because the Bayesian inference is for a semiparametric mixture distribution, which is infinite-dimensional when the number of mixture points $k < \infty$ is unknown, as is studied here.

So far, very little seems to be known about the posterior consistency of the Dirichlet process in Bayesian semiparametric problems, with the exception of recent work by Diaconis and Freedman (1993), Ghoshal, Ghosh and Ramamoorthi (1997a, b) and Shen (1995). Posterior consistency is not always guaranteed in infinite dimensional problems. Indeed, although the relatively rich nature of the Dirichlet process prior would suggest that it yield a well-behaved posterior, there is a considerable amount of literature surrounding examples involving inconsistent posteriors. These counterexamples have included nonparametric problems [Freedman and Diaconis (1983)] as well as semiparametric problems [Diaconis and Freedman (1986a, b); Doss (1985)]. Nevertheless, we will see that the finite semiparametric mixture is an example of an important class of models where the use of the Dirichlet process works quite well (albeit indirectly). Indeed, we will see that the posterior for θ is exponentially consistent under fairly general conditions, and that, furthermore, this consistency holds in the more general case when the Y^n are sampled from a generalized Pólya urn scheme.

The main result describing exponential consistency is given in Theorem 3 of Section 3. Section 4 contains several examples. Sections 5–7 contain the proof of three lemmas that are needed in establishing Theorem 3. Several places in these proofs use the relative entropy measure of information (sometimes called the Kullback–Leibler divergence number). To remind the reader, if $\mathbb{P}$ and $\mathbb{Q}$ are two distributions, then the relative entropy from $\mathbb{P}$ to $\mathbb{Q}$ is defined as $K(\mathbb{P},\mathbb{Q}) = \mathbb{P}\log(d\mathbb{P}/d\mathbb{Q})$, where $d\mathbb{P}/d\mu$ and $d\mathbb{Q}/d\mu$ are the densities of $\mathbb{P}$ and $\mathbb{Q}$ taken with respect to a common dominating measure μ. Notice that the definition makes use of the linear functional notation for expectation. This same notation will be used throughout the paper when convenient, although more traditional notation will be used as well, as in $K(\mathbb{P},\mathbb{Q}) = \int \log(d\mathbb{P}/d\mathbb{Q})\, d\mathbb{P}$. Also, as convenience dictates, sets will sometimes be used as indicator functions in order to facilitate the use of the linear functional notation.

2. Generalized Pólya urn scheme.

Let $Y_1^*,\dots,Y_n^*$ be a sequence of i.i.d. r.v.'s with sample space $\mathscr{Y}$ and with distribution H whose density h is taken with respect to Lebesgue measure. We say that $Y_1,\dots,Y_n$ is a sample from a generalized Pólya urn scheme if it is generated in the following fashion:

$$\begin{aligned} &Y_1 = Y_1^*,\\ &Y_i|Y^{i-1} = \begin{cases} Y_j, & \text{with probability } (1-\alpha_i)/(i-1), \quad j=1,\dots,i-1,\\ Y_i^*, & \text{with probability } \alpha_i, \end{cases} \end{aligned} \tag{3}$$

for $i = 2, 3, \ldots, n$. In particular, this implies that the distribution for Y^n satisfies

$$d\pi_{Y^n}(y_1, \ldots, y_n) = h(y_1)\, dy_1 \prod_{i=2}^{n} \left[\alpha_i h(y_i)\, dy_i + \frac{1-\alpha_i}{i-1} \sum_{s=1}^{i-1} \delta(y_s, dy_i) \right],$$

where $\delta(y, \cdot)$ is the unit measure concentrated at y.

The sequence of values $0 < \alpha_i < 1$ in (3) are probabilities which reflect how likely it is for a new value Y_i^* to be introduced into the sampling scheme. In particular, when

$$\alpha_i = \frac{H(1)}{H(1) + (i-1)}, \qquad i = 2, 3, \ldots, n, \tag{4}$$

then the resulting Y^n are exchangeable and describe a sample from the Dirichlet process with parameter H [Blackwell and MacQueen (1973)]. The original Blackwell and MacQueen (1973) characterization described Y_i^* as a sequence sampled from a normalized finite positive measure H. However, it has now become more common to use a proper distribution for H and to replace $H(1)$ by a constant $A > 0$, usually referred to as the precision parameter. In our treatment we will always assume that H is a proper distribution having the density h.

A sample Y^n from a Dirichlet process prior, or more generally from a generalized Pólya urn scheme, is characterized by its relatively few distinct values. This is one of the fundamental reasons for its success in the finite semiparametric mixture setting. The many ties induced by the Pólya scheme make it appear as if Y^n is an i.i.d. sample from a discrete distribution. Furthermore, the Pólya scheme also ensures that Y^n has enough distinct values necessary to encompass the k distinct values of G_0. In particular, the expected number of distinct values equals $D_n = 1 + \sum_{i=2}^{n} \alpha_i$, which guarantees a steady stream of new values as n increases [see also Antoniak (1974), page 1161]. However, as we will see, an important assumption for exponential consistency will also require that $D_n = O(\log n)$ in order to suppress Y^n from having too many distinct values. Notice that by (4) the condition holds when the sampling is from a Dirichlet process.

The choice for α_i in (4) ensures that a sample Y^n from a Dirichlet process is an exchangeable sequence, but this is not true in general for the generalized Pólya urn scheme (3). Surprisingly, the exchangeability plays a limited role in the exponential consistency for θ, although it does simplify the verification of conditions needed for Theorem 3. From a modeling perspective, nonexchangeability is certainly unappealing, but the results given here show that consistency depends more upon the Pólya mechanism generating the right number of distinct values for Y. This seems to be more important than exchangeability in modeling the finite mixture distribution G.

3. Exponential posterior consistency for $\boldsymbol{\theta}$.

The hierarchical structure (2) implies that the Bayesian marginal density for X^n is

$$m_n(x^n) = \int_{\Theta \otimes \mathscr{Y}^n} f(x^n | \theta, y^n)\, d\pi(\theta, y^n)$$

and that the posterior distribution for θ is defined by

$$\pi(U \otimes \mathscr{Y}^n | x^n) = \frac{\int_{U \otimes \mathscr{Y}^n} f(x^n | \theta, y^n)\, d\pi(\theta, y^n)}{m_n(x^n)},$$

for each U in the Borel σ-algebra for Θ.

Note. We will always assume that $f(x|\theta, y)$ is measurable in (x, θ, y). Also, here we are using a variation of the superscript notation discussed in the introduction. For example, f_0^n or $f_0(x^n)$, represents the true joint density for X^n with distribution $\mathbb{P}_0^n$, while

$$f(x^n | \theta, y^n) = \prod_{i=1}^{n} f(x_i | \theta, y_i)$$

is the joint density for $(X^n | \theta, y^n)$ with joint distribution $\mathbb{P}_{\theta, y^n} = \mathbb{P}_{\theta, y_1} \otimes \cdots \otimes \mathbb{P}_{\theta, y_n}$.

The main goal of the paper will be to show that the posterior for θ concentrates on each open neighborhood of θ_0 with high probability. That is, for each $\varepsilon > 0$, we will show that θ lies outside the open set $\Theta_\varepsilon = \{\theta: \|\theta - \theta_0\| < \varepsilon\}$ with exponentially small posterior probability (see Theorem 3 for a precise statement). In particular, establishing the exponential posterior consistency in the finite mixture problem will depend upon showing that the posterior odds satisfies the inequality

$$(5) \qquad \left(\frac{\pi(\Theta_\varepsilon \otimes \mathscr{Y}^n | X^n)}{\pi(\Theta_\varepsilon^c \otimes \mathscr{Y}^n | X^n)} = \frac{\int_{\Theta_\varepsilon \otimes \mathscr{Y}^n} f(X^n | \theta, y^n)\, d\pi(\theta, y^n)}{\int_{\Theta_\varepsilon^c \otimes \mathscr{Y}^n} f(X^n | \theta, y^n)\, d\pi(\theta, y^n)} \right) \le \exp(r\delta_n)$$

with small probability (with respect to $\mathbb{P}_0^n$). Here $r > 0$ and $\delta_n > 0$ is the exponential convergence rate satisfying

$$\delta_n = O(n) \quad \text{and} \quad \delta_n^{-1} \log n = o(1).$$

The integrals in (5) represent the Bayesian marginal densities for X^n when θ is constrained to the sets Θ_ε and Θ_ε^c (numerator and denominator, respectively). In verifying that (5) occurs with small probability, we follow the approach used in Clarke and Barron (1990), which is to compare each of the constrained Bayesian marginals to the true joint density f_0^n. When the prior density for θ is positive and continuous at θ_0, and the modified Pólya sampling scheme is "rich enough," we will see that the joint density matches the Bayesian marginal constrained to Θ_ε. If the true model $\mathbb{P}_0^n$ can be distinguished uniformly well from alternatives $\mathbb{P}_{\theta, y^n}$ for $\theta \notin \Theta_\varepsilon$, then the marginal constrained by Θ_ε^c will be exponentially smaller than the joint density. In comparing the behavior of each of these marginals to f_0^n, we will be led to the inequality (5).

The conditions needed for the comparison involving the marginal constrained by Θ_ε are as follows.

CONDITION C1.

(i) θ_0 is an interior point of Θ and $y_{0,j}$ is an interior point of $\mathscr{Y}$, for $j = 1, \dots, k$.

(ii) For a.a. $x[\mathbb{P}_0]$, the first and second partial derivatives of $\log f(x|\theta, y)$ with respect to θ and y exist and are continuous in some open neighborhood of θ_0 and some open neighborhood of $y_{0,j}$, for $j = 1, \dots, k$.

(iii) The partial derivatives in (ii) can be bounded in absolute value by a square integrable function $M \in \mathscr{L}^2(\mathbb{P}_0)$.

(iv) $\mathbb{P}_0 |\log f(x|\theta_0, y_{0,j})| < \infty$, for $j = 1, \dots, k$.

CONDITION C2.

(i) π_θ is absolutely continuous with respect to Lebesgue measure with a density that is positive and continuous at θ_0.

(ii) The density h is taken with respect to Lebesgue measure and is positive and continuous at $y_{0,j}$, for $j = 1, \dots, k$.

(iii) $\sum_{i=2}^{n} \alpha_i = O(\log n)$.

Condition C1 encourages a type of continuity for finite semiparametric mixtures in a neighborhood of $\mathbb{P}_0$. The precise form of continuity is given in Lemma 4 of Section 5 and is expressed in terms of the relative entropy. Condition C2(i) ensures that the prior density for θ is well behaved locally around θ_0, while conditions C2(ii) and (iii) ensure that the Pólya urn scheme can generate Y^n values which (approximately) mimic values sampled from G_0. In particular, Condition C2(iii) is required so that the urn scheme is hindered from generating too many distinct values of Y (remember that G_0 is a finite discrete distribution).

Conditions C1 and C2 will show that the marginal for X^n, when θ is restricted to an open set around θ_0, "locally matches" the joint density f_0^n. Lemma 1 makes this assertion more precise. Its proof is given in Section 6.

LEMMA 1. *If Conditions* C1 *and* C2 *hold*, *then for each* $r > 0$,

$$\mathbb{P}_0^n\left\{\int_{\Theta_\varepsilon \otimes \mathscr{Y}^n} f(X^n|\theta, y^n)\, d\pi(\theta, y^n) \le \exp(-r\delta_n) f_0(X^n)\right\} \tag{6}$$
$$= O\left(\delta_n^{-1} \log n\right).$$

A second ingredient for consistency requires that $\mathbb{P}_0^n$ can be distinguished exponentially well from the class

$$\left\{\mathbb{P}_{\theta, y^n} : \theta \in \Theta_\varepsilon^c \cap \Theta_n,\, y_n \in \mathscr{Y}_n^*\right\}$$

for chosen subsets $\Theta_n \subseteq \Theta$ and $\mathscr{Y}_n^* \subseteq \mathscr{Y}^n$. This is Condition C3(ii), given below, which amounts to establishing the existence of a uniformly exponentially consistent test (UEC test) between a simple hypothesis and a composite alternative hypothesis. Although this condition can sometimes be checked directly in a particular problem (see the first example in Section 4), it is usually easier to tackle the simpler problem of verifying the existence of a uniformly consistent test (UC test). This is stated as the alternative Condition C3(ii)* and requires that the Y^n are exchangeable.

CONDITION C3.

(i) For each $\varepsilon > 0$ there exists a sequence of subsets $\Theta_n \otimes \mathscr{Y}_n^* \subseteq \Theta \otimes \mathscr{Y}^n$ and an $r_0 > 0$ so that eventually

$$\text{(a)}\ \pi_\theta(\Theta_n) \geq 1 - \exp(-r_0 \delta_n) \quad \text{and} \quad \text{(b)}\ \pi_{Y^n}(\mathscr{Y}_n^*) \geq 1 - \exp(-r_0 \delta_n).$$

(ii) Let $\Theta_{\varepsilon, n}^c = \Theta_\varepsilon^c \cap \Theta_n$. Then for each $\varepsilon > 0$ there exists sets A_n so that eventually, uniformly,

$$\mathbb{P}_0^n(A_n) \geq 1 - \exp(-r_0 \delta_n),$$
$$\mathbb{P}_{\theta, y^n}(A_n) \leq \exp(-r_0 \delta_n) \quad \text{where } (\theta, y^n) \in \Theta_{\varepsilon, n}^c \otimes \mathscr{Y}_n^*.$$

(ii)* The sample Y^n is exchangeable and for each $0 < \gamma < 1$ there exists a test $0 \leq \phi_n = \phi_n(X_1, \dots, X_n) \leq 1$ so that eventually, uniformly,

$$\mathbb{P}_0^n(\phi_n) \geq 1 - \gamma,$$
$$\mathbb{P}_{\theta, y^n}(\phi_n) \leq \gamma \quad \text{where } (\theta, y^n) \in \Theta_{\varepsilon, n}^c \otimes \mathscr{Y}_n^*.$$

When Condition C3 holds, Lemma 2 (which is given below) states that the marginal for X^n cannot properly match f_0^n when θ is constrained by Θ_ε^c. Condition C3(i) is included in order to be able to restrict attention to increasing subsets of $\Theta \otimes \mathscr{Y}^n$ when constructing the required UEC or UC test. In particular, the choice for $\mathscr{Y}_n^*$ will depend upon the model under study, and will typically be chosen to exclude values for y^n which can make the likelihood at $\theta \neq \theta_0$ look similar to the value at θ_0. The Pólya urn scheme should therefore sample values from these excluded y^n values with small probability, as indicated by Condition C3(i)(b). Remark 2 gives a simple method for checking Condition C3(i)(b) in certain cases. Lemma 2 also provides a rate of comparison for the case when Condition C3(ii)* is used in place of Condition C3(ii). The result follows, using a straightforward modification of ideas given in Schwartz (1965), and will rely on the exchangeability of Y^n. The proof of the lemma is deferred until Section 7.

LEMMA 2. *If Condition C*3(i), (ii) *holds, then there exists an $r' > 0$ such that*

$$(7) \quad \mathbb{P}_0^n\left\{\int_{\Theta_\varepsilon^c \otimes \mathscr{Y}^n} f(X^n|\theta, y^n)\, d\pi(\theta, y^n) \geq \exp(-r'\delta_n) f_0(X^n)\right\} = O(n^{-1}).$$

Alternatively, under Condition C3(i)–(ii)*, *the expression* (7) *holds with δ_n replaced by δ_{n*} and $O(n^{-1})$ by $O(n_*^{-1})$, where n_* is the largest integer less than or equal to $\sqrt{n}$.*

Conditions C1, C2 and C3, coupled with Lemmas 1 and 2, lead to our main result, which states that the posterior probability of Θ_ε^c is exponentially small. Observe carefully that the weaker condition of a UC test leads to a slower exponential rate of convergence.

THEOREM 3. *Suppose that Conditions* C1, C2 *and* C3(i) *and* (ii) *hold. Then for each* $\varepsilon > 0$, *there exists an* $r > 0$ *so that*

$$\mathbb{P}_0^n\{\pi(\Theta_\varepsilon^c \otimes \mathscr{Y}^n | X^n) \geq \exp(-r\delta_n)\} = O\big(\delta_n^{-1} \log n\big). \tag{8}$$

Alternatively, under Conditions C1, C2 *and* C3(i)–(ii)*, *the posterior rate* (8) *holds with* δ_n *replaced by* δ_{n_*}.

PROOF. In order to prove (8), it suffices to show that the probability of the event (5),

$$\mathbb{P}_0^n\left\{\int_{\Theta_\varepsilon \otimes \mathscr{Y}^n} f(X^n|\theta, y^n)\, d\pi(\theta, y^n) \right.$$
$$\left. \leq \exp(r\delta_n)\int_{\Theta_\varepsilon^c \otimes \mathscr{Y}^n} f(X^n|\theta, y^n)\, d\pi(\theta, y^n)\right\},$$

is of the same order as the right-hand side of (8) for some $r > 0$. Bound this probability using the upper bound

$$\mathbb{P}_0^n\left\{\exp(-r''\delta_n) f_0(X^n) \leq \exp(r\delta_n)\int_{\Theta_\varepsilon^c \otimes \mathscr{Y}^n} f(X^n|\theta, y^n)\, d\pi(\theta, y^n)\right\} + \mathbb{P}_0^n(B_n), \tag{9}$$

where $0 < r'' < r'$ (for the r' in Lemma 2) and where

$$B_n = \left\{\int_{\Theta_\varepsilon \otimes \mathscr{Y}^n} f(X^n|\theta, y^n)\, d\pi(\theta, y^n) \leq \exp(-r''\delta_n) f_0(X^n)\right\}.$$

By Lemma 1, $\mathbb{P}_0^n(B_n) = O(\delta_n^{-1} \log n)$, while the first term in (9), when reexpressed, equals

$$\mathbb{P}_0^n\left\{\exp(-(r + r'')\delta_n) f_0(X^n) \leq \int_{\Theta_\varepsilon^c \otimes \mathscr{Y}^n} f(X^n|\theta, y^n)\, d\pi(\theta, y^n)\right\}. \tag{10}$$

By Lemma 2 under Condition C3(i), (ii), this is of order $O(n^{-1})$ when $r = r' - r''$. But this is smaller than $\mathbb{P}_0^n(B_n)$, which becomes the dominating term $O(\delta_n^{-1} \log n)$ due to $\delta_n = O(n)$. This verifies (8). Under Condition C3(i)–(ii)*, substitute δ_{n_*}, for δ_n and deduce by Lemma 1 that $P_0^n(B_n) = O(\delta_{n_*}^{-1} \log n)$ and by Lemma 2 that (10) is $O(n_*^{-1})$ where $r = r' - r''$. □

REMARK 1. Inference in the semiparametric model is more difficult than in the classical parametric problem, and so it is not surprising that the rate given in Theorem 3 with $\delta_n = n$ is slower than that given in the classical problem. In both cases, the posterior probability of Θ_ε^c is exponentially small, $\exp(-rn)$, but for the semiparametric model this occurs at a $O_p(n^{-1} \log n)$ rate by Theorem 3, while for the parametric case the rate is $O_p(n^{-1})$ [Clarke and Barron (1990), Proposition 6.3].

The $\delta_n = n$, $O_p(n^{-1})$ rate is the observed rate in the classical problem, which is unlikely to be improved upon in the semiparametric mixture setting. Therefore, we might suspect that $\delta_n = n$ with an $O_p(n^{-1} \log n)$ rate is the lower bound to the rate, with the $\log n$ term representing the loss of information associated with studying the infinite-dimensional problem. However, whether this rate is achievable in each model is still unclear, and, in fact, a different rate was seen in each of the examples given in Section 4.

REMARK 2. It follows straightforwardly that H is the marginal distribution for each Y_i defined in the Pólya scheme (3). When $\mathscr{Y}_n^*$ is a product of sets $\mathscr{Y}_n^n = \mathscr{Y}_n \otimes \cdots \otimes \mathscr{Y}_n$, Condition C3(i)(b) can be easily checked by showing that $H(\mathscr{Y}_n) \geq 1 - \exp(-r\delta_n)$ for some $r > 0$. This is a simple consequence of Bonferroni's inequality,

$$\pi_{Y^n}(\mathscr{Y}_n^*) \geq 1 - n + \sum_{i=1}^{n} \mathbb{P}\{Y_i \in \mathscr{Y}_n\} = 1 - n + nH(\mathscr{Y}_n),$$

which is larger than $1 - \exp(-r_0 \delta_n)$ for some $r_0 > 0$.

REMARK 3. The assumption of exchangeability in Condition C3(ii)* can be removed when $\mathscr{Y}_n^*$ is a product of sets $\mathscr{Y}_n^n$ where $H(\mathscr{Y}_n) \geq 1 - \exp(-r\delta_n)$ for some $r > 0$. Note that by Remark 2 this will also verify Condition C3(i)(b).

4. Examples of exponential posterior consistency. Here we present several important examples of finite semiparametric mixtures for which the exponential posterior consistency in Theorem 3 holds.

EXAMPLE (Rasch semiparametric mixture). For the interested reader, an overview of the Rasch model, as well as further references to other articles, can be found in Lindsay, Clogg and Greggo (1991). Briefly, though, the Rasch model is an exponential model used in modeling 0–1 binary outcomes. One of its many important uses is in item response studies where each individual i elicits a binary response to each of $L \geq 2$ different items or questions. If $X_{i,l}$ is the 0–1 binary response for individual i to question l, then $X_{i,l}$ has the conditional density

$$f(x_{i,l}|\theta, y_i) = \frac{\exp((\theta_l + y_i)x_{i,l})}{1 + \exp(\theta_l + y_i)}$$

for $i = 1, \dots, n$ and $l = 1, \dots, L$. Here $\theta = (\theta_1, \dots, \theta_L)$ is the vector of item response parameters measuring item difficulty, while $Y_i = y_i$ is the unique ability parameter for individual i. It is assumed that the $X_{i,l}$ are conditionally independent, given the item difficulty parameter θ and the individual parameter y_i.

A useful method for modeling heterogeneity among individuals is to assume that the Y_i are independent random variables with an unknown finite discrete distribution G_0 [see Lindsay, Clogg and Greggo (1991) for examples and motivation]. Such an assumption implies that the response values have a finite semiparametric mixture density. In particular, if $X_i = (X_{i,1}, \dots, X_{i,L})$ is the vector of binary responses for individual i, then the X_i are independent r.v.'s with the Rasch semiparametric mixture density.

$$f(x_i|\theta, G_0) = \sum_{j=1}^{k} p_{0,j} \prod_{l=1}^{L} f\big(x_{i,l}|\theta, y_{0,j}\big) = \sum_{j=1}^{k} p_{0,j} f\big(x_i|\theta, y_{0,j}\big).$$

A simple method to ensure that θ is identified is to constrain $\theta_L = 0$ to act as a baseline, although this will not guarantee identification for the mixing distribution G_0. Nevertheless, even though G_0 is unidentified, Lindsay, Clogg and Greggo (1991) show that one can still apply nonparametric maximum likelihood methods to estimate θ properly and, to a lesser extent, recover some partial information about the unknown mixing distribution. Here we will show that the Bayesian method also leads to a useful method for studying θ. In particular, we will see that the posterior for θ is exponentially consistent.

Under the baseline parameterization, $\Theta = \mathbb{R}^{L-1} \otimes \{0\}$, while $\mathscr{Y} = \mathbb{R}$ is unconstrained. Condition C1 of Theorem 3 holds straightforwardly, while Condition C2 is satisfied by the appropriate choice of prior (e.g., when π_θ and h are positive continuous densities). Therefore, in verifying the conditions for Theorem 3, the only tricky part will be in deriving the uniform test required by Condition C3. In this example, very little additional work is needed in constructing a UEC test instead of a UC test. Therefore, we will verify condition C3(ii) in the proof of exponential consistency.

The method for constructing the UEC test involves conditioning on the sufficient statistics $S_i = S(X_i) = \sum_{l=1}^{L} X_{i,l}$, for $i = 1, \dots, n$. If $\mathbb{P}_{X|S,\theta}$ denotes the conditional distribution for $(X \mid S, \theta)$, then $\mathbb{P}_{X|S,\theta}$ has the conditional density

$$f(x \mid s, \theta) = \frac{\exp(\theta' x)}{\sum_{\{x:\, s(x)=s\}} \exp(\theta' x)},$$

which is independent of y by sufficiency. Notice that the density equals one when s equals zero or L, which occurs only for those observations which are all equal to zero or all equal to one. We will avoid these observations because they provide no information for θ, and will instead base our UEC test on the discordant observations only. Indeed, it happens that the analysis can be simplified even further, by only having to consider those observations X_i where $S_i = 1$.

Let e_l denote the vector in $\mathbb{R}^L$ whose lth coordinate equals one and is zero elsewhere. Also, let $W_n = \sum_{i=1}^{n} \{S_i = 1\}$ record the number of observations with $S_i = 1$. Then our UEC test is based on the modified empirical measure $\hat{\mathbb{P}}_n(\cdot) = \hat{\mathbb{P}}_n(\cdot|X^n)$ defined by

$$\hat{\mathbb{P}}_n(B|X^n) = \begin{cases} \{e_1 \in B\}, & \text{if } W_n < n_* \\ W_n^{-1} \sum_{i=1}^{n} \{X_i \in B,\, S_i = 1\}, & \text{otherwise,} \end{cases}$$

over the measurable space $(\mathscr{E}, 2^{\mathscr{E}})$ where $\mathscr{E} = \{e_l \colon l = 1, \dots, L\}$. Thus when $W_n \geq n_*$, $\hat{\mathbb{P}}_n$ is the empirical measure based on only those observations X_i

for which $S_i = 1$, while for $W_n < n_*$, $\hat{\mathbb{P}}_n$ is the degenerate measure at e_1 (the choice for e_1 is purely arbitrary and plays no special role in the analysis).

For each $\delta > 0$, let

$$\mathscr{C}_\delta = \left\{\mathbb{P}_{X|S=1,\theta}: K\left(\mathbb{P}_{X|S=1,\theta}, \mathbb{P}_{X|S=1,\theta_0}\right) \le \delta,\ \theta \in \Theta\right\}$$

be a set of measures on $(\mathscr{E}, 2^{\mathscr{E}})$. Then $\mathscr{C}_\delta$ is a completely convex set of measures [Csiszár (1984), Definition 2.3]. Our test will be the indicator set $A_n = \{\hat{\mathbb{P}}_n \in \mathscr{C}_\delta\}$, which records whether $\hat{\mathbb{P}}_n$ is a member of $\mathscr{C}_\delta$. In verifying that this is a UEC test, we will show that A_n has exponentially large probability under $\mathbb{P}_0^n$ and exponentially small probability, uniformly, over the set of alternatives $\{\mathbb{P}_{\theta, y^n}: (\theta, y^n) \in \Theta_\varepsilon^c \otimes \mathscr{Y}^n\}$.

No measure in $\mathscr{C}_\delta$ puts mass 1 at e_1. Therefore, conditioning on $S^n = (S_1, \dots, S_n)$,

$$\begin{aligned}\mathbb{P}_{\theta, y^n}(A_n) &= \mathbb{P}_{\theta, y^n}\left\{\hat{\mathbb{P}}_n \in \mathscr{C}_\delta,\ W_n \ge n_*\right\} \\ &= \mathbb{P}_{S^n|\theta, y^n}\mathbb{P}_{X^n|S^n,\theta}\left\{\hat{\mathbb{P}}_n \in \mathscr{C}_\delta,\ W_n \ge n_*\right\}.\end{aligned} \tag{11}$$

While integrating over S^n, we need only consider those values of $S^n = s^n$ for which $W_n \ge n_*$. For a fixed one of these s^n values, the inner expectation of the last expression is evaluated over the product space formed by the W_n values of $(X^n|s^n, \theta)$ for which $s_i = 1$. These values are i.i.d. from $\mathbb{P}_{X|S=1,\theta}$, which allows us to exploit an inequality of Csiszár [(1984), Theorem 1] due to the convexity of $\mathscr{C}_\delta$ [also see Barron (1989)]. Thus, for each fixed $S^n = s^n$ value, the inner expectation in (11) is bounded by

$$\exp\left(-\sum_{\{i,\, s_i = 1\}} K\left(\mathscr{C}_\delta, \mathbb{P}_{X_i|S_i,\theta}\right)\right) \le \exp\left(-n_* K\left(\mathscr{C}_\delta, \mathbb{P}_{X|S=1,\theta}\right)\right), \tag{12}$$

where

$$K\left(\mathscr{C}_\delta, \mathbb{P}_{X|S=1,\theta}\right) = \inf_{\mathbb{Q} \in \mathscr{C}_\delta} K\left(\mathbb{Q}, \mathbb{P}_{X|S=1,\theta}\right).$$

By setting $\theta_L = 0$ we have ensured that θ is identified in $\mathbb{P}_{X|S=1,\theta}$. Hence, by the continuity in θ,

$$K\left(\mathbb{P}_{X|S=1,\theta}, \mathbb{P}_{X|S=1,\theta_0}\right) \to 0 \quad \text{if and only if } \theta \to \theta_0.$$

Therefore, for a small enough $\delta > 0$ there exists a $\delta' > 0$ so that $K(\mathscr{C}_\delta, \mathbb{P}_{X|S=1,\theta}) > \delta'$ for each $\theta \in \Theta_\varepsilon^c$. Hence, from (11) and (12),

$$\mathbb{P}_{\theta, y^n}(A_n) \le \exp(-\delta' n_*) \quad \text{for } (\theta, y^n) \in \Theta_\varepsilon^c \otimes \mathscr{Y}^n. \tag{13}$$

Similar reasoning as in (11) shows that

$$\mathbb{P}_0^n(A_n) = \mathbb{P}_{S^n|\theta_0, G_0}\mathbb{P}_{X^n|S^n,\theta_0}\left\{\hat{\mathbb{P}}_n \in \mathscr{C}_\delta,\ W_n \ge n_*\right\}. \tag{14}$$

Let $Z_{i,l} = \{X_i = e_l\}$ and $\xi_{0,l} = \mathbb{P}_{X|S=1,\theta_0}\{X_i = e_l\}$ be its expectation, for $l = 1, \dots, L$. Then for a small enough $\delta'' > 0$, the inner expectation on the right-hand side of (14) is larger than

$$\mathbb{P}_{X^n|S^n,\theta_0}\left\{\left|\sum_{\{i:\,S_i=1\}}(Z_{i,l}-\xi_{0,l})\right|\le\delta''W_n\ \text{for}\ l=1,\dots,L\right\},$$

where $W_n\ge n_*$. The $Z_{i,l}$ variables are i.i.d. Bernoulli $(\xi_{0,l})$ r.v.'s under $\mathbb{P}_{X|S=1,\theta_0}$. Therefore, by Bennett's inequality [Shorack and Wellner (1996), page 855], each of the L probabilities in the previous expression will be larger than $1-\exp(-Cn_*)$ for some $C>0$. Hence, the inner expectation in (14) occurs with exponentially high probability, uniformly over values for S^n with $W_n\ge n_*$. Conclude from this and (13) that A_n is a UEC test for $\delta_n=n_*$ with the rate $O_p\,(n_*^{-1}\log n)$.

Example (Weibull semiparametric mixture). There has been some recent interest in studying the relationship between identification constraints for Y and inference for θ in the Weibull semiparametric mixture. This research has mostly focused on how tail bounds on Y translate into rates of estimation for θ. Heckman and Singer (1984) used a moment constraint on Y to verify consistency using nonparametric maximum likelihood estimation, while Honoré (1990) and Ishwaran (1996b) constructed estimators having polynomial rates under more stringent moment conditions. An exact relationship between the rate of estimation for θ and tail bounds on Y is given in Ishwaran (1996a). As we will see, the behavior of Y will also play a role in the Bayesian approach, with the exponential posterior consistency for θ depending upon the tail behavior of the distribution of H used in the Pólya urn scheme.

For convenience, we will work with a log-transform of the Weibull semiparametric mixture. Therefore, we will consider the i.i.d. r.v.'s

$$X_i=\log\left(\frac{W_i}{Y_i}\right)^{1/\theta}=\frac{1}{\theta}(\log W_i-\log Y_i),$$

where $W_i\sim\exp(1)$ are independent of $Y_i\sim G_0$. Hence, $(X\mid\theta,y)$ has the conditional density

$$f(x\mid\theta,y)=y\theta\exp(\theta x-y\exp(\theta x))\quad\text{where } x\in\mathbb{R}\text{ and }(\theta,y)\in\mathbb{R}^+\otimes\mathbb{R}^+.$$

In applying Theorem 3, we will verify Condition C3(ii)* by constructing a UC test between $\mathbb{P}_0^n$ and

$$\{\mathbb{P}_{\theta,y^n}\colon\theta\in\Theta_\varepsilon^c,\ y_n\in\mathscr{Y}_n^*\},$$

where $\mathscr{Y}_n^*$ will be the product of sets $\mathscr{Y}_n^n$, with $\mathscr{Y}_n=\{n^{-\delta}\le Y\le n^{\delta}\}$ chosen to control the tail behavior of Y. Later we will see that an appropriate choice for 2δ is $\varepsilon/\theta_0<1$. The UC test will be based on the indicator set

$$A_n=\left\{K_1^{-1}\le\sum_{i=1}^n Z_i\le K_1\right\},$$

where $Z_i = \{X_i \le -\log(K_2 n)/\theta_0\}$ are Bernoulli r.v.'s and K_1, K_2 are positive constants. By choosing K_1 large enough and K_2 small enough, we can make $\mathbb{P}_0^n(A_n)$ arbitrarily close to one and $\mathbb{P}_{\theta, y^n}(A_n)$ arbitrarily small, uniformly over $\Theta_\varepsilon^c \otimes \mathscr{Y}_n^*$.

Checking Conditions C1 and C2 is straightforward. Therefore, we can proceed to Condition C3. By Remark 2, Condition C3(i)(b) holds for $\delta_n = n^\delta$ when $H(\mathscr{Y}_n) \ge 1 - \exp(-rn^\delta)$ for some $r > 0$. This condition is easily satisfied by many continuous densities, and so we will assume it is true here. To determine the value of $\mathbb{P}_{\theta, y^n}(A_n)$ over $(\theta, y^n) \in \Theta_\varepsilon^c \otimes \mathscr{Y}_n^*$, we will consider the two cases when $\theta \ge \theta_0 + \varepsilon$ and $\theta \le \theta_0 - \varepsilon$.

When $\theta \ge \theta_0 + \varepsilon$, Markov's inequality gives the bound

$$\mathbb{P}_{\theta, y^n}(A_n) \le \mathbb{P}_{\theta, y^n}\left\{\sum_{i=1}^n Z_i \ge K_1^{-1}\right\} \le K_1 \sum_{i=1}^n \mathbb{P}_{\theta, y_i} Z_i. \tag{15}$$

However,

$$\mathbb{P}_{\theta, y_i} Z_i = 1 - \exp\Bigl(-y_i (K_2 n)^{-\theta/\theta_0}\Bigr). \tag{16}$$

By $2\delta = \varepsilon/\theta_0$, this is less than $1 - \exp(-Cn^{-(1+\delta)})$ for some $C > 0$ when $y_i \in \mathscr{Y}_n$ and $\theta \ge \theta_0 + \varepsilon$. Therefore, the right-hand side of (15) is uniformly $o(1)$.

Now suppose that $\theta \le \theta_0 - \varepsilon$. If $\mu_n = \sum_{i=1}^n \mathbb{P}_{\theta, y_i} Z_i$, then $\mu_n \ge Cn^\delta$ by (16) and our choice for δ. Hence, for $\theta \le \theta_0 - \varepsilon$ there exists some $C > 0$ (a generic constant) such that

$$\mathbb{P}_{\theta, y^n}(A_n) \le \mathbb{P}_{\theta, y^n}\left\{\sum_{i=1}^n Z_i - \mu_n \le K_1 - \mu_n\right\} \le \mathbb{P}_{\theta, y^n}\left\{\left|\sum_{i=1}^n Z_i^*\right| \ge C\mu_n\right\}, \tag{17}$$

where $Z_i^* = Z_i - \mathbb{P}_{\theta, y_i} Z_i$. But by Bennett's inequality [Shorack and Wellner (1996), page 855], the right-hand side of (17) is no larger than

$$2\exp\left(-\frac{(C\mu_n)^2}{2\mu_n(1 + C/3)}\right) \le 2\exp(-Cn^\delta) \quad \text{uniformly.}$$

Therefore, $\mathbb{P}_{\theta, y^n}(A_n)$ is arbitrarily small, uniformly over $(\theta, y^n) \in \Theta_\varepsilon^c \otimes \mathscr{Y}_n^*$.

To choose K_1 and K_2, note that

$$\mathbb{P}_0 Z_i = G_0\left[1 - \exp\left(-\frac{Y}{K_2 n}\right)\right] = \frac{\gamma_0}{n}(1 + o(1)),$$

where $\gamma_0 = G_0(Y)/K_2 > 0$. Therefore, $\sum_{i=1}^n Z_i \rightsquigarrow \text{Poisson}(\gamma_0)$ under $\mathbb{P}_0$. Hence

$$\mathbb{P}_0^n(A_n) \to \mathbb{P}\bigl\{K_1^{-1} \le \text{Poisson}(\gamma_0) \le K_1\bigr\},$$

for each noninteger $K_1 > 1$. By choosing K_1 large enough and K_2 small enough, we can make the right-hand side arbitrarily close to one. This verifies that A_n is a UC test and establishes Condition C3(i)–(ii)*. Consequently, Theorem 3 can be applied with $\delta_n = n_*^\delta$ for the rate $O_p\ (n_*^{-\delta} \log n)$.

Note. Here δ depends upon ε and therefore the size of the neighborhood around θ_0.

Example (Paired exponentials). Another interesting example is the semiparametric paired exponential model studied by Lindsay (1985). This model arises through data $X_i = (X_{i,1}, X_{i,2}) = (W_{i,1}/Y_i, W_{i,2}/(\theta Y_i))$, where $W_{i,1}, W_{i,2}$ are independent exp(1) r.v.'s, independent of $Y_i \sim G_0$. The parameter θ is interpreted as the ratio of the hazard rates of a sample of paired exponential variables.

The conditional density of $(X_i \mid \theta, y_i)$ equals

$$f(x_i \mid \theta, y_i) = y_i \exp(-y_i x_{i,1}) \theta y_i \exp(-\theta y_i x_{i,2}) \quad \text{for } x_i \in \mathbb{R}^+ \otimes \mathbb{R}^+,$$

where $(\theta, y_i) \in \mathbb{R}^+ \otimes \mathbb{R}^+$. Verifying Conditions C1 and C2 is straightforward. In verifying Condition C3(ii), one can simplify the problem by constructing a UEC test based on the transformed data

$$Z_i = \frac{X_{i,1}}{X_{i,2}} = \frac{\theta W_{i,1}}{W_{i,2}}, \qquad i = 1, \dots, n.$$

By transforming the data, the construction of the UEC test is made simpler because it will only require distinguishing $\mathbb{P}_{Z|\theta_0}$ exponentially well from $\mathbb{P}_{Z|\theta}$, over $\theta \in \Theta_\varepsilon^c$. The methodology for finding such a test is fairly straightforward. For example, apply the methods in Clarke and Barron [(1990), Proposition 6.2]. Doing this will show that the posterior is exponentially consistent with $\delta_n = n$ for the rate $O_p\,(n^{-1} \log n)$.

5. Continuity at $\mathbb{P}_0$.

Lemma 1 asserts a type of local matching between the Bayesian marginal and the true joint density. The lemma is essential in the proof of the exponential consistency stated in Theorem 3 and relies on a form of continuity for finite semiparametric mixtures at the true mixture $\mathbb{P}_0$. This type of continuity can be expressed in terms of the relative entropy and depends upon regularity conditions for the density $f(x \mid \theta, y)$, such as those expressed in Condition C1. Here we give a statement and a proof of this local continuity. The proof of Lemma 2 will be given in Section 6.

Lemma 4. *Suppose that Condition* C1 *holds and that* $\Theta_n \otimes \mathscr{Y}_n^k$ *is a shrinking open neighborhood of* (θ_0, y_0^k) *and* $\mathscr{P}_n$ *a set of probability vectors shrinking to* p_0. *Let* $\mathbb{Q}_{\theta, y^k, p}$ *be the distribution for the mixture density*

$$q(x|\theta, y^k, p) = \sum_{j=1}^k p_j f\big(x|\theta, y_j\big) \tag{18}$$

$$\text{where } \big(\theta, y^k\big) \in \Theta \otimes \mathscr{Y}^k,\ p_j > 0,\ \sum_{j=1}^k p_j = 1.$$

Then, uniformly over $\Theta_n \otimes \mathscr{Y}_n^k \otimes \mathscr{P}_n$,

$$K\big(\mathbb{P}_0, \mathbb{Q}_{\theta, y^k, p}\big) = O\Big(\big\|\big(\theta - \theta_0, y^k - y_0^k, p - p_0\big)\big\|^2\Big). \tag{19}$$

Furthermore,

$$K\big(\mathbb{P}_0, \mathbb{P}_{\theta, y_j}\big) < \infty \tag{20}$$

uniformly over (θ, y_j) *values in the neighborhood of Condition* C1, *for* $j = 1, \dots, k$.

Proof. First, let us prove (19). By definition, the relative entropy equals

$$K\big(\mathbb{P}_0, \mathbb{Q}_{\theta, y^k, p}\big) = \mathbb{P}_0 \log\left(\frac{f_0(X)}{q(X|\theta, y^k, p)}\right).$$

We can assume that (θ, y^k) is close enough to (θ_0, y_0^k) so that Condition C1 can be applied. Therefore, by Condition C1(ii), the first- and second-order partial derivatives for $f(x|\theta, y)$ exist for the θ and y values of interest for a.a. $x[\mathbb{P}_0]$. Hereafter, ignore the set with measure zero where the derivatives may not exist. With $\delta = (\theta - \theta_0, y^k - y_0^k, p - p_0)'$, a two-term Taylor series expansion around (θ_0, y_0^k, p_0) gives

$$\log\left(\frac{f_0(x)}{q(x|\theta, y^k, p)}\right) = \delta'\nabla_0(x) + \delta' V_0(x)\delta + \|\delta\|^2 r(x, \theta, y^k, p), \tag{21}$$

where ∇_0 is the vector of first-order partial derivatives and V_0 is the matrix of second-order partial derivatives, both evaluated at (θ_0, y_0^k, p_0). The remainder term is defined by

$$\|\delta\|^2 r(x, \theta, y^k, p) = \delta'\big[V(x, \theta, y^k, p) - V_0(x)\big]\delta,$$

where $V(x, \theta, y^k, p)$ is the matrix of second-order partial derivatives evaluated at some point between (θ, y^k, p) and (θ_0, y_0^k, p_0).

By Condition C1(iii), each element in V can be bounded in absolute value by an $\mathscr{L}^2(\mathbb{P}_0)$ function. For example, consider an element such as

$$\begin{aligned}\frac{\partial^2}{\partial\theta_s\theta_l}&\log\left(\frac{f_0(x)}{q(x \mid \theta, y^k, p)}\right)\\ &= -q(x \mid \theta, y^k, p)^{-1} \sum_{j=1}^k p_j \frac{\partial^2}{\partial\theta_s\theta_l} f(x \mid \theta, y_j)\\ &\quad + q(x \mid \theta, y^k, p)^{-2} \sum_{j=1}^k p_j \frac{\partial}{\partial\theta_s} f(x \mid \theta, y_j) \sum_{j'=1}^k p_j' \frac{\partial}{\partial\theta_l} f(x \mid \theta, y_{j'}).\end{aligned}$$

Because $q(x \mid \theta, y^k, p) \ge p_j f(x \mid \theta, y_j)$ for each j, the right-hand side can be bounded in absolute value by

$$\begin{aligned}\sum_{j=1}^k &\left|\frac{\partial^2}{\partial\theta_s\theta_l}\log f(x \mid \theta, y_j)\right| + \sum_{j=1}^k \left|\frac{\partial}{\partial\theta_s}\log f(x \mid \theta, y_j)\frac{\partial}{\partial\theta_l}\log f(x \mid \theta, y_j)\right|\\ &+ \sum_{j=1}^k \left|\frac{\partial}{\partial\theta_s}\log f(x \mid \theta, y_j)\right| \sum_{j'=1}^k \left|\frac{\partial}{\partial\theta_l}\log f(x \mid \theta, y_{j'})\right|.\end{aligned}$$

By Condition C1(iii), this can be bounded by the $\mathbb{P}_0$-integrable function $kM + (k + k^2)M^2$. The other terms in V can be handled in a similar way (keeping in mind that the p_j are close enough to $p_{0,j}$ to ensure that they are bounded away from zero). Therefore, by the dominated convergence theorem and the continuity of $V(x, \cdot, \cdot, \cdot)$ [Condition C1(ii)], $\mathbb{P}_0 r(X, \theta, y^k, p) \to 0$ as $\delta \to 0$.

Taking the $\mathbb{P}_0$-expectation of the right- and left-hand sides of (21) gives

$$K(\mathbb{P}_0, \mathbb{Q}_{\theta, y^k, p}) = \mathbb{P}_0(\delta' \nabla_0) + \mathbb{P}_0(\delta' V_0 \delta) + o(\|\delta\|^2).$$

Both $\delta' \nabla_0$ and $\delta' V_0 \delta$ must be integrable by Condition C1(iii), which means that $\mathbb{P}_0(\delta' \nabla_0)$ equals zero or it is the dominating term as $\delta \to 0$. However, because (θ_0, y_0^k) is an interior point, it must be the case that $\mathbb{P}_0(\delta' \nabla_0) = 0$, otherwise the left-hand side (a positive distance) would be negative for certain values of δ.

Note. The components of $\delta' \nabla_0$ related to p must have zero expectation, because

$$\begin{aligned}\mathbb{P}_0(p - p_0)' & \left[\frac{\partial}{\partial p_j} \log q(X \mid \theta_0, y_0, p_0) \right]_{j=1,\dots,k} \\ &= \sum_{j=1}^{k} (p_j - p_{0,j}) \int f(x \mid \theta_0, y_{0,j})\, d\lambda(x) = 0.\end{aligned}$$

Thus,

$$K(\mathbb{P}_0, \mathbb{Q}_{\theta, y^k, p}) = \mathbb{P}_0(\delta' V_0 \delta) + o(\|\delta\|^2) = O(\|\delta\|^2) \quad \text{uniformly}.$$

To prove (20), first use the bound

$$K(\mathbb{P}_0, \mathbb{P}_{\theta, y_j}) \le \sum_{j'=1}^{k} \mathbb{P}_0 \left| \log f(X \mid \theta_0, y_{j',0}) \right| + \left| \mathbb{P}_0 \log f(X \mid \theta, y_j) \right|. \tag{22}$$

The sum on the right-hand side is finite by Condition C1(iv). To work out the remaining term, apply a one-term Taylor series expansion to the integrand [this is justified by Condition C1(ii)]:

$$\log f(x \mid \theta, y_j) = \log f(x \mid \theta_0, y_{j,0}) + (\theta - \theta_0, y_j - y_{j,0})' \nabla(x, \theta, y_j),$$

where $\nabla(x, \theta, y_j)$ is the vector of first-order partial derivatives of $\log f(x \mid \theta, y_j)$ evaluated at some point between (θ, y_j) and $(\theta_0, y_{0,j})$. This term can be bounded by a $\mathbb{P}_0$-integrable function using Condition C1(iii). From this and Condition C1(iv), deduce that the Taylor series expansion is uniformly integrable over the neighborhood of (θ, y_j) values. Therefore, the right-hand side of (22) is uniformly finite. □

6. Locally matched marginal. This section provides a proof of Lemma 1 and makes use of the local continuity expressed in Lemma 4 of Section 5.

Proof of Lemma 1. For each open set $U \subseteq \Theta$, let

$$m_n(x^n \mid U) = \frac{1}{\pi_\theta(U)} \int_{U \otimes \mathscr{Y}^n} f(x^n \mid \theta, y^n)\, d\pi(\theta, y^n) \tag{23}$$

denote the density for the marginal when π_θ is conditioned on $\theta \in U$. Write $M_n(\cdot \mid U)$ for its distribution. To verify (6), first divide by f_0^n inside the probability, invert both sides and then increase the bound by restricting θ to the smaller set $\Theta_n = \{\theta\colon \|\theta - \theta_0\| \le n^{-1/2}\}$. Now multiply both sides by $\pi_0(\Theta_n)$, take logs and then increase the bound even further by taking the absolute value of the log ratio term involving the densities. Therefore, the left-hand side of (6) is less than

$$\mathbb{P}_0^n \left\{ \left| \log \frac{f_0(X^n)}{m_n(X^n \mid \Theta_n)} \right| \ge r\delta_n + \log \pi_\theta(\Theta_n) \right\}. \tag{24}$$

The density for π_θ is continuous and positive around θ_0 by Condition C2(i). Therefore, $\pi_\theta(\Theta_n) = O(n^{-d_1/2})$ (recall that Θ is d_1-dimensional) and

$$r\delta_n + \log \pi_\theta(\Theta_n) \ge C\delta_n \quad \text{for some } C > 0,$$

because $\delta_n^{-1} \log n = o(1)$ (hereafter C will be used for a generic positive constant). Apply Markov's inequality to (24) in order to obtain the upper bound

$$C\delta_n^{-1} \mathbb{P}_0^n \left| \log \frac{f_0(X^n)}{m_n(X^n \mid \Theta_n)} \right| \le C\delta_n^{-1} \big[K(\mathbb{P}_0^n, M_n(\cdot \mid \Theta_n)) + 2\exp(-1) \big], \tag{25}$$

where the last inequality follows by bounding the negative part of the log-ratio using the inequality $u \log(u) \ge -\exp(-1)$ [this same trick is used in Clarke and Barron (1990) in their equation (6.3)].

To bound the relative entropy in (25), we use a lower bound for m_n by restricting the range of integration for y^n. We will restrict attention to the set of y^n values whose first k values are all different and constrained by

$$y^k \in \mathscr{Y}_n^k = \big\{ y^k \in \mathscr{Y}^k \colon \|y_j - y_{0,j}\| \le n^{-1/2} \quad \text{for } j = 1, \dots, k \big\}$$

and whose last $n - k$ values are chosen from the set of unique values $\{y_1, \dots, y_k\}$. That is, the last $n - k$ values $y_{k+1}, \dots, y_n$ are repetitions from y^k, for each $y^k \in \mathscr{Y}_n^k$. The key idea underlying this constraint to m_n is to select a subset of $\mathscr{Y}^n$ which occurs with nonnegligible probability, so that the $Y_{k+1}, \dots, Y_n$ values sampled from this set approximate an i.i.d. sample from G_0. In allowing only Y^k to be unique, we ensure that $(Y_{k+1}, \dots, Y_n \mid Y^k = y^k)$ are exchangeable r.v.'s with the discrete sample space $\{y_1, \dots, y_k\}$. Notice that these values lie in a shrinking neighborhood of the support points $y_{0,j}$ for G_0.

Call $W_1, \dots, W_n$ a finite Pólya sequence with a Pólya($\mathscr{S}$) distribution if W_1 is sampled uniformly from the set of values $\mathscr{S}$, followed by W_2 sampled uniformly from $\mathscr{S} \cup \{W_1\}$, and so forth, ending with W_n sampled uniformly from $\mathscr{S} \cup \{W_1\} \cup \cdots \cup \{W_{n-1}\}$. In particular, when $Y_{k+1}, \dots, Y_n$ are conditioned to take only sampled values from a fixed set of y^k values, then the conditioned sequence has a Pólya(y^k) distribution. This is helpful because it allows us to exploit a well-known connection between finite Pólya sequences and the Dirichlet distribution. In particular, Theorem 2.3 in Mauldin, Sudderth and Williams (1992) implies that for the conditioned Y_j, there exists a random probability vector P such that

$$\begin{aligned}\left(Y_j \mid y^k, P = p\right) &\sim_{\text{ind}} \text{Bernoulli}_k(y^k, p), \qquad j = k+1, \dots, n, \\ P &\sim \text{Dirichlet}_k(1, \dots, 1),\end{aligned}$$

where $\text{Bernoulli}_k(y^k, p)$ is the discrete distribution with support y^k and probability vector p (k-dimensional).

Therefore, by conditioning on y^k and $P = p$ and by restricting the values for y^n as described above, $Y_{k+1}, \dots, Y_n$ must have the same distribution as an i.i.d. sample from a discrete distribution with support y^k and probability vector p. In particular, by restricting attention to our constrained set of Y^n values, we obtain the lower bound

$$\begin{aligned}m_n(x^n \mid \Theta_n) \geq C_n^* \int d\pi_{\theta, n}(\theta) \int_{\mathscr{Y}_n^k} dH^k(y^k) f(x^k \mid \theta, y^k) \\ \times \int dD_k(p) \prod_{i=k+1}^{n} q\left(x_i \mid \theta, y^k, p\right),\end{aligned}$$

where $\pi_{\theta, n} = \pi_\theta / \pi_\theta(\Theta_n)$ is the prior π_θ conditioned on $\theta \in \Theta_n$, the mixture density q is defined by (18), D_k is the distribution for a $\text{Dirichlet}_k(1, \dots, 1)$ and

$$C_n^* = \alpha_2 \cdots \alpha_k (1 - \alpha_{k+1}) \cdots (1 - \alpha_n)$$

is the probability that $Y_{k+1}, \dots, Y_n$ are repetitions from the set of unique values Y^k.

Restrict p to values in

$$\mathscr{P}_n = \left\{ p = (p_1, \dots, p_k) \colon |p_j - p_{0,j}| \leq n^{-1/2}, \, p_j > 0, \, \sum_{j=1}^{k} p_j = 1 \right\}$$

in order to bound $m_n(x^n \mid \Theta_n)$ even further. Let $H_n^k = H^k / H^k(\mathscr{Y}_n^k)$ be the distribution for H^k conditioned on $y^k \in \mathscr{Y}_n^k$ and $D_{k,n} = D_k / D_k(\mathscr{P}_n)$ the distribution for D_k conditioned on $p \in \mathscr{P}_n$. Then,

$$\begin{aligned}m_n(x^n \mid \Theta_n) \geq C_n \iiint d\pi_{\theta, n}(\theta) \, dH_n^k(y^k) \, dD_{k,n}(p) f(x^k \mid \theta, y^k) \\ \times \prod_{i=k+1}^{n} q\left(x_i \mid \theta, y^k, p\right),\end{aligned}$$

where $C_n = C_n^* H^k(\mathscr{Y}_n^k) D_k(\mathscr{P}_n)$.

Using the previous bound, Jensen's inequality, and then Fubini's theorem to interchange the order of integration, we can bound the relative entropy in (25) by

$$\begin{aligned} K(\mathbb{P}_0^n, M_n(\cdot \mid \Theta_n)) &\le \mathbb{P}_0^n \iiint d\pi_{\theta,n}(\theta)\, dH_n^k(y^k)\, dD_{k,n}(p) \\ &\qquad \times \left[\frac{f_0(X^n)}{f(x^k \mid \theta, y^k)\prod_{i=k+1}^n q\big(x_i \mid \theta, y^k, p\big)} \right] - \log C_n \log \\ &= \iiint d\pi_{\theta,n}(\theta)\, dH_n^k(y^k)\, dD_{k,n}(p) \\ &\qquad \times \left[\sum_{i=1}^k K\big(\mathbb{P}_0, \mathbb{P}_{\theta, y_i}\big) + (n-k) K\big(\mathbb{P}_0, \mathbb{Q}_{\theta, y^k, p}\big) \right] - \log C_n \end{aligned}$$

By Lemma 4, the first k relative entropy terms are uniformly finite, while the entropy term containing $\mathbb{Q}_{\theta, y^k, p}$ is uniformly $O(n^{-1})$. Therefore,

$$\begin{aligned} K(\mathbb{P}_0^n, M_n(\cdot \mid \Theta_n)) &\le C \iiint d\pi_{\theta,n}(\theta)\, dH_n^k(y^k)\, dD_{k,n}(p) - \log C_n \\ &= C - \log C_n. \end{aligned}$$

By Condition C2(iii) and the continuity and positivity of H^k and D_k around y_0^k and p_0, $-\log C_n = O(\log n)$. Hence, by (24) and (25), we can bound the left-hand side of (6) by

$$O\big(\delta_n^{-1} \log n\big) + O\big(\delta_n^{-1}\big) = O\big(\delta_n^{-1} \log n\big). \qquad \square$$

7. UC and UEC tests. This section provides a proof of Lemma 2. Being able to substitute the simpler assumption of a UC test (Condition C3(ii)*) in place of the assumption of a UEC test [Condition C3(ii)] is made possible by a simple modification to a result given in Schwartz [(1965), Lemma 6.1]. Some of these details are sketched below.

PROOF OF LEMMA 2. First we will prove (7) under Condition C3(i) and (ii). To bound the left-hand side of (7), divide both sides inside the set by $\pi_\theta(\Theta_\varepsilon^c) f_0(X^n)$, take square roots and then apply Markov's inequality. Therefore, remembering the definition (23) for $m_n(\cdot \mid \Theta_\varepsilon^c)$, we obtain the upper bound

$$\begin{aligned} &\mathbb{P}_0^n \left\{ \sqrt{\frac{m_n(X^n \mid \Theta_\varepsilon^c)}{f_0(X^n)}} \ge \frac{\exp(-r'\delta_n/2)}{\sqrt{\pi_\theta(\Theta_\varepsilon^c)}} \right\} \\ (26) \qquad &\le \exp\left(\frac{r'\delta_n}{2}\right) \mathbb{P}_0^n \sqrt{\frac{m_n(X^n \mid \Theta_\varepsilon^c)}{f_0(X^n)}} \\ &\le \exp\left(\frac{r'\delta_n}{2}\right) \Big[1 - v\big(\mathbb{P}_0^n, M_n(\cdot \mid \Theta_\varepsilon^c)\big)^2\Big]^{1/2}, \end{aligned}$$

where $v(\mathbb{P}, \mathbb{Q})$ is the total variation distance between distributions $\mathbb{P}$ and $\mathbb{Q}$,

$$v(\mathbb{P}, \mathbb{Q}) = \tfrac{1}{2}\|\mathbb{P} - \mathbb{Q}\|_1 = \sup_{B \in \mathscr{B}} | \mathbb{P}B - \mathbb{Q}B |$$

[a similar inequality is used by Schwartz (1965) in the proof of her Theorem (6.1)]. Remembering that $\delta_n^{-1} \log n = o(1)$, we can show that (7) is $O(n^{-1})$ by verifying that

$$(27) \qquad v\big(\mathbb{P}_0^n, M_n(\cdot \mid \Theta_\varepsilon^c)\big) \ge 1 - C\exp(-r_0 \delta_n) \quad \text{for } 0 < r' < r_0,$$

where $C > 0$ represents a generic constant. This will prove (7) for any r' constrained as above.

However, by Condition C3(ii), we know there exist sets A_n such that eventually, uniformly,

$$\mathbb{P}_0^n(A_n) \ge 1 - \exp(-r_0 \delta_n),$$
$$\mathbb{P}_{\theta, y^n}(A_n) \le \exp(-r_0 \delta_n) \quad \text{where } (\theta, y^n) \in \Theta_{\varepsilon, n}^c \otimes \mathscr{Y}_n^*.$$

Therefore,

$$\begin{aligned} v\big(\mathbb{P}_0^n, M_n(\cdot \mid \Theta_\varepsilon^c)\big) &\ge \mathbb{P}_0^n(A_n) - M_n(A_n \mid \Theta_\varepsilon^c) \\ &\ge 1 - \exp(-r_0 \delta_n) \\ (28) \qquad &\quad - \frac{1}{\pi_\theta(\Theta_\varepsilon^c)}\left[\int_{\Theta_{\varepsilon,n}^c \otimes \mathscr{Y}_n^*} d\pi(\theta, y^n)\mathbb{P}_{\theta, y^n}(A_n) + \pi(B_n)\right] \\ &\ge 1 - 2\exp(-r_0 \delta_n) - \frac{\pi(B_n)}{\pi_\theta(\Theta_\varepsilon^c)}, \end{aligned}$$

where $B_n = (\Theta_\varepsilon^c \otimes \mathscr{Y}^n) \cap (\Theta_{\varepsilon,n}^c \otimes \mathscr{Y}_n^*)^c$. However, by condition C3(i),

$$\pi(B_n) \le \pi_{Y^n}(\mathscr{Y}_n^{*c}) + \pi_\theta(\Theta_n^c) \le 2\exp(-r_0 \delta_n).$$

Therefore, by (26), (27) and (28), the left-hand side of (7) is $O(n^{-1})$.

To prove a variation of (7) under Condition C3(i)–(ii)*, replace δ_n in (26) and (27) by δ_{n_*}. The proof in Lemma 6.1 of Schwartz (1965) can be easily extended to allow for independent, but not identically distributed r.v.'s. Follow the same construction in the proof, but for convenience only use the first $m = n_*^2$ observations $X_1, \dots, X_m$. Now block these values into n_* groups, each of size n_*. Using Schwartz's method and applying Condition C3(ii)* to each of these blocks, it easily follows that there exist sets A_n (based only upon $X_1, \dots, X_m$) and an $r'' > 0$, such that eventually, uniformly,

$$\mathbb{P}_0^n(A_n) = \mathbb{P}_0^m(A_n) \ge 1 - \exp(-r'' n_*),$$
$$\mathbb{P}_{\theta, y^n}(A_n) = \mathbb{P}_{\theta, y^m}(A_n) \le \exp(-r'' n_*) \quad \text{where } (\theta, y^m) \in \Theta_{\varepsilon, n_*}^c \otimes \big(\mathscr{Y}_{n_*}^*\big)^{n_*}.$$

Hence, following the same type of argument leading to (28), we find that

$$v\big(\mathbb{P}_0^n, M_n(\cdot \mid \Theta_\varepsilon^c)\big) \ge 1 - 2\exp(-r'' n_*) - \pi(B_n^*)/\pi_\theta(\Theta_\varepsilon^c),$$

where $B_n^* = (\Theta_\varepsilon^c \otimes \mathscr{Y}^m) \cap (\Theta_{\varepsilon, n_*}^c \otimes (\mathscr{Y}_{n_*}^*)^{n_*})^c$. A similar argument as before shows that

$$\pi(B_n^*) \le \Big[1 - \pi_{Y^m}\Big(\big(\mathscr{Y}_{n_*}^*\big)^{n_*}\Big)\Big] + \Big[1 - \pi_\theta(\Theta_{n_*})\Big].$$

However, by Bonferroni's inequality and the assumed exchangeability of Y^n

$$\begin{aligned}\pi_{Y^m}\Big(\big(\mathscr{Y}^*_{n_*}\big)^{n_*}\Big) &= \pi_{Y^m}\big[(Y_1,\dots,Y_{n_*})\in\mathscr{Y}^*_{n_*},\dots,(Y_{m-n_*+1},\dots,Y_m)\in\mathscr{Y}^*_{n_*}\big]\\ &\geq 1-n_*+n_*\pi_{Y^{n_*}}\big[(Y_1,\dots,Y_{n_*})\in\mathscr{Y}^*_{n_*}\big].\end{aligned}$$

Therefore, by Condition C3(i), $\pi(B^*_n) = O(n_*\exp(-r_0\delta_{n_*}))$ and

$$v\big(\mathbb{P}^n_0, M_n(\cdot\mid\Theta^c_\varepsilon)\big) \geq 1-2\exp(-r''n_*)-Cn_*\exp(-r_0\delta_{n_*}) \quad \text{for some } C>0.$$

By choosing $r' > 0$ to be small enough, deduce that the left-hand side of (7) is $O(n_*^{-1})$. □

Acknowledgments. The author is grateful for the helpful comments from the Associate Editor and the two referees.

REFERENCES

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

BARRON, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Dept. Statistics, Univ. Illinois, Champaign.

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

BARRON, A. R., SCHERVISH, M. and WASSERMAN, L. (1998). The consistency of posterior distributions in nonparametric problems. Unpublished manuscript.

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355.

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

CSISZÁR, I. (1984). Sanov property, generalized *I*-projection and a conditional limit theorem. *Ann. Probab.* **12** 768–793.

DIACONIS, P. and FREEDMAN, D. (1986a). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–26.

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.

DOSS, H. (1985). Bayesian nonparametric estimation of the median (II): asymptotic properties of the estimates. *Ann. Statist.* **13** 1445–1464.

ESCOBAR, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268–277.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588.

ESCOBAR, M. D. and WEST, M. (1998). Computing nonparametric hierarchical models. *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statist.* **133** 1–22. Springer, Berlin.

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

FREEDMAN, D. and DIACONIS, P. (1983). On inconsistent Bayes estimates in the discrete case. *Ann. Statist.* **11** 1109–1118.

GHOSHAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1997a). Consistency issues in Bayesian nonparametrics. Technical report, Michigan State Univ.

GHOSHAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1997b). Posterior consistency of Dirichlet mixtures in density estimation. Technical report, Michigan State Univ.

HECKMAN, J. and SINGER, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52** 271–320.

HONORÉ, B. E. (1990). Simple estimation of a duration model with unobserved heterogeneity. *Econometrica* **58** 453–473.

ISHWARAN, H. (1996a). Identifiability and rates of estimation for scale parameters in location mixture models. *Ann. Statist.* **24** 1560–1571.

ISHWARAN, H. (1996b). Uniform rates of estimation in the semiparametric Weibull mixture model. *Ann. Statist.* **24** 1572–1585.

LINDSAY, B. (1985). Using empirical partially Bayes inference for increased efficiency. *Ann. Statist.* **13** 914–931.

LINDSAY, B., CLOGG, C. C. and GREGO, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc.* **86** 96–107.

MACEACHERN, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23** 727–741.

MACEACHERN, S. N. (1998). Computational methods for mixture of Dirichlet process models. *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statist.* **133** 23–44. Springer, Berlin.

MAULDIN, R. D., SUDDERTH, W. D. and WILLIAMS, S. C. (1992). Pólya trees and random distributions. *Ann. Statist.* **20** 1203–1221.

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

SHEN, X. (1995). On the properties of Bayes procedures in general parameter spaces. Unpublished manuscript, Ohio State Univ.

SHORACK, G. R. and WELLNER, J. A. (1996). *Empirical Processes with Applications to Statistics*. Wiley, New York.

WASSERMAN, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statist.* **133** 293–304. Springer, Berlin.

CLEVELAND CLINIC FOUNDATION
DEPARTMENT OF BIOSTATISTICS AND EPIDEMIOLOGY W64
9500 EUCLID AVENUE
CLEVELAND, OHIO 44195
E-MAIL: ishwaran@bio.ri.ccf.org