The Annals of Statistics
1994, Vol. 22, No. 4, 1763–1786

BAYESIAN NONPARAMETRIC ESTIMATION FOR INCOMPLETE DATA VIA SUCCESSIVE SUBSTITUTION SAMPLING[1]

BY HANI DOSS

Ohio State University

In the problem of estimating an unknown distribution function F in the presence of censoring, one can use a nonparametric estimator such as the Kaplan–Meier estimator, or one can consider parametric modeling. There are many situations where physical reasons indicate that a certain parametric model holds approximately. In these cases a nonparametric estimator may be very inefficient relative to a parametric estimator. On the other hand, if the parametric model is only a crude approximation to the actual model, then the parametric estimator may perform poorly relative to the nonparametric estimator, and may even be inconsistent. The Bayesian paradigm provides a reasonable framework for this problem. In a Bayesian approach, one would try to put a prior distribution on F that gives most of its mass to small neighborhoods of the entire parametric family. We show that certain priors based on the Dirichlet process prior can be used to accomplish this. For these priors the posterior distribution of F given the censored data appears to be analytically intractable. We provide a method for approximating this posterior distribution through the use of a successive substitution sampling algorithm. We also show convergence of the algorithm.

1. Introduction and summary. There are many situations arising in survival analysis and reliability theory where one is faced with the problem of dealing with incomplete data. A data set can contain some observations which are censored on the right (or equivalently on the left). A more complicated case occurs when some are censored on the right and some are censored on the left. A still more general form of censoring is interval censoring, in which associated with each observation there is a set within which the observation is known to lie. Turnbull (1974, 1976) gives a way of obtaining a nonparametric maximum likelihood estimator of a distribution function F for the general case of interval censoring. By far the most common case of censoring is right censoring, and for this case the nonparametric maximum likelihood estimator is the well-known Kaplan–Meier (1958) estimator.

There are many cases where there are physical reasons that indicate specific parametric families. Exponential distributions arise in a very large number of contexts; extreme value distributions arise frequently in reliability theory because they are the limiting distributions of the lifelengths of series or parallel systems with a large number of identically distributed components. In these

Received May 1991; revised January 1994.

[1] Research supported by Air Force Office of Scientific Research Grant F49620-94-1-0028.

AMS 1980 *subject classifications*. Primary 62G05; secondary 65U05, 60J05.

Key words and phrases. Dirichlet prior, successive substitution sampling, incomplete data, calculation of posterior distribution, Markov chain.

situations there would be a loss of efficiency if one were to use a nonparametric maximum likelihood estimator instead of using a parametric estimator.

Miller (1983) focuses attention on the loss of efficiency of the Kaplan–Meier estimator vs. a parametric estimator when the parametric model is true, and shows that this loss can be quite substantial. He also points out the obvious fact that when the parametric model is incorrectly specified, in the limit, as the sample size goes to ∞, the mean squared error of the Kaplan–Meier estimator tends to 0, but the same thing cannot be said for the parametric estimator.

This problem is complicated by the fact that a parametric model is put down only as an approximation to the true model; for example, limit theorems indicate a distribution which is only approximately normal or only approximately of an extreme value type. What one wants is an estimator that avoids the loss of efficiency due to ignoring partial information about a parametric model but that at the same time avoids the pitfalls connected with an incorrectly specified parametric model.

In this paper we consider a Bayesian framework in which we study prior distributions on F which give most of their mass to "small neighborhoods" of an *entire* parametric family. The prior distributions that we investigate are derived from the Dirichlet process priors discussed by Ferguson (1973, 1974). Before proceeding we first define these Dirichlet process priors and review their salient features. Let $\mathcal{P}$ be the space of all probability measures on the real line R. The Dirichlet process priors are probability measures on $\mathcal{P}$ parameterized by the set of all finite nonnull measures on R. Let α be a finite nonnull measure on R. The random distribution function F has a Dirichlet process prior distribution with parameter α, denoted $\mathcal{D}_\alpha$, if for every measurable partition $\{B_1, \dots, B_m\}$ of R the random vector $[F(B_1), \dots, F(B_m)]$ has the Dirichlet distribution with parameter vector $[\alpha(B_1), \dots, \alpha(B_m)]$ (here and throughout the rest of the paper, probability measures are identified with their cumulative distribution functions, and the same symbol is used to denote both a measure and its distribution function whenever convenient). When a prior distribution is put on $\mathcal{P}$, then for every $t \in R$, the quantity $F(t)$ is a random variable. Write $H = \alpha/\alpha(R)$, so that H is a probability measure on R. It turns out that if F is distributed according to $\mathcal{D}_\alpha$, then $EF(t) = H(t)$, while the quantity $\alpha(R)$ indicates the degree of concentration of $\mathcal{D}_\alpha$ around its "center" H. For example, it is well known that as $\alpha(R) \to \infty$, $\mathcal{D}_\alpha$ converges to the point mass at H in the weak topology. The Dirichlet priors have the attractive feature that the support of $\mathcal{D}_\alpha$ is the set of all probability measures whose support is contained in the support of H; in particular, if the support of H is the positive real axis, then the support of $\mathcal{D}_\alpha$ is the set of distributions of all positive random variables. Another property of the Dirichlet distributions is that they give mass 1 to the set of discrete distributions.

Susarla and van Ryzin (1976) considered the problem of right censoring, put a Dirichlet prior on the unknown distribution function F and obtained in closed form the mean of the posterior distribution of F given the data. A problem with this is that it is rare that one has a priori knowledge that the true distribution function F is close to a specified H. It is more reasonable to assume that a given parametric family holds approximately.

In the present paper we deal with the general form of censoring described earlier, and consider the situation where a parametric family $H_\theta, \theta \in \Theta \subset R^p$ is specified. We put a prior distribution on F as follows: first choose θ according to some prior measure ν; then, after having selected a number $\alpha(R) > 0$, choose F from $\mathcal{D}_{\alpha(R)H_\theta}$. The prior on F is then a "mixture of Dirichlets"; see Antoniak (1974). The posterior distribution of F given the data appears to be intractable, and interesting features of it, such as the mean of the posterior distribution of $F(t)$ for fixed t, may be very difficult to obtain in closed form. Instead, we develop a method for simulating from the posterior distribution of F given the data. Our approach is based on the method of successive substitution sampling, originally developed by Geman and Geman (1984) in the context of image reconstruction. See Gelfand and Smith (1990) for a review and description of this method and of its properties. This approach has some strong advantages. If one can simulate from the posterior distribution of F, then one can also simulate various easily interpretable quantities connected with F, such as the mean of F. One can then, by repeated simulation, obtain an estimate of the posterior distribution of such quantities. It should be mentioned that the algorithm presented in this paper provides a way of obtaining posterior distributions for various parameters of interest in the situations discussed above for the special case where there is no censoring, and the algorithm should be useful even in that case.

This paper is organized as follows. In Section 2 we describe the algorithms for simulating from the posterior distribution of F given the data. In Section 3 we illustrate our method on a data set from a clinical trial on survival after estrogen treatment for prostate cancer patients. In Section 4 we give a proof that the algorithm converges to the true posterior distribution. In Section 5 we discuss very briefly issues of consistency of the posterior distribution of F given the data.

2. The Algorithms.

2.1. *Preliminaries, successive substitution sampling and Sethuraman's construction of the Dirichlet process.* The prior on F described in Section 1 is a mixture of Dirichlets:

$$F \sim \int \mathcal{D}_{\alpha(R)H_\theta}\nu(d\theta). \tag{2.1}$$

We have $X_1, \dots, X_n$ i.i.d. $\sim F$. The X's are not directly observed; instead, we know only that $X_i \in A_i$, where the A_i's are subsets of R. In the case of standard right-censored data, A_i is a singleton if X_i is uncensored, and $A_i = (c_i, \infty)$ if X_i is censored on the right by c_i. We wish to obtain the posterior distribution of F given the (incomplete) data, and we will develop an algorithm for generating a random distribution function from this conditional distribution. A by-product of our procedure is that we will also be able to generate an observation from distributions such as $\mathcal{L}(\mu(F) \mid \text{data})$, $\mathcal{L}(\text{med}(F) \mid \text{data})$ or $\mathcal{L}(X_{n+1} \mid \text{data})$. Here,

$\mu(F) = \int x\, dF(x)$, whenever this is defined, med(F) is the median of F defined by med$(F) = \sup\{t;\ F(t) \le 1/2\}$ and X_{n+1} denotes a future observation; also, for random variables V and W, $\mathcal{L}(V \mid W)$ denotes the conditional distribution of V given W.

As mentioned in Section 1, our approach is based on the successive substitution sampling algorithm discussed by Gelfand and Smith (1990), and, before proceeding, we review this algorithm. Let $f_{Y_1,\dots,Y_p}$ be the joint distribution of the (possibly vector-valued) random variables $Y_1,\dots,Y_p$. We suppose that we do not know the form of $f_{Y_1,\dots,Y_p}$, but that we know the conditional distributions $f_{Y_i \mid Y_j, j \neq i}$, $i = 1,\dots,p$, or that at least we are able to generate observations from these conditional distributions. The objective is to generate an observation from the joint distribution of $Y_1,\dots,Y_p$, or simply an observation from the marginal distribution of Y_1. The algorithm to generate an observation from $f_{Y_1,\dots,Y_p}$ proceeds by an iterative scheme as follows. Fix an arbitrary starting point $Y_1^{(0)},\dots,Y_p^{(0)}$, generate $Y_1^{(1)}$ from $f_{Y_1 \mid Y_j, j \neq 1}(\,\cdot\,, Y_2^{(0)},\dots,Y_p^{(0)})$, generate $Y_2^{(1)}$ from $f_{Y_2 \mid Y_j, j \neq 2}(Y_1^{(1)}, \cdot, Y_3^{(0)},\dots,Y_p^{(0)})$ and so on until $Y_p^{(1)}$ is generated from $f_{Y_p \mid Y_j, j \neq p}(Y_1^{(1)},\dots,Y_{p-1}^{(1)}, \cdot)$. This completes one iteration of the scheme. After k iterations we obtain a random variable $(Y_1^{(k)},\dots,Y_p^{(k)})$. It is not difficult to see that the sequence $(Y_1^{(j)},\dots,Y_p^{(j)}), j = 1,2,\dots,$ is a Markov chain, and that $f_{Y_1,\dots,Y_p}$ is a stationary distribution of the chain. If one can establish that this chain converges to its stationary distribution, then for large $k, (Y_1^{(k)},\dots,Y_p^{(k)})$ has a distribution which is approximately equal to $f_{Y_1,\dots,Y_p}$. Thus, by repeating this algorithm independently a large number M of times, one obtains $(Y_1^{(m,k)},\dots,Y_p^{(m,k)})$, $m = 1,\dots,M$, which can be used to estimate $f_{Y_1,\dots,Y_p}$. Similarly, $Y_1^{(m,k)}$, $m = 1,\dots,M$, can be used to estimate the marginal distribution of Y_1.

We now give a preliminary explanation of how in our censored data setup a certain version of the algorithm can be implemented. We take $p = 2$ and $Y_1 = F$, $Y_2 = (X_1,\dots,X_n)$. We wish to obtain $\mathcal{L}(Y_1, Y_2 \mid \text{data})$ [actually, our primary interest is in $\mathcal{L}(Y_1 \mid \text{data})$].

```
Fix starting values $F^{(0)}$ and $(X_1^{(0)},\dots,X_n^{(0)})$.
For $k = 1,\dots,K$,
    1. Generate $F^{(k)} \sim \mathcal{L}(F \mid (X_1^{(k-1)},\dots,X_n^{(k-1)}, \text{data})$.
    2. Generate $(X_1^{(k)},\dots,X_n^{(k)}) \sim \mathcal{L}((X_1,\dots,X_n) \mid F^{(k)}, \text{data})$.
```

It is not immediately clear how one would generate the random variables required in steps 1 and 2. The Dirichlet prior has already been defined in Section 1. We will now review a constructive definition of the Dirichlet prior, given in Sethuraman (1994), which is particularly convenient for simulation purposes. As in Section 1, let α be a finite nonnull measure on R and write $\alpha = \alpha(R)H$. Let $B_1, B_2, \dots$ be i.i.d. $\sim$ Beta$(1, \alpha(R))$, let $V_1, V_2, \dots$ be i.i.d. $\sim H$ and assume that the sequences $\{B_j\}$ and $\{V_j\}$ are mutually independent and are all defined on some common probability space $(\Omega, \mathcal{F}, Q)$. Let $P_j = B_j \Pi_{r=1}^{j-1}(1 - B_r)$ and form the random distribution function

$$F = \sum_{j=1}^{\infty} P_j \, \delta_{V_j}, \tag{2.2}$$

where δ_a denotes the probability measure giving unit mass to the point a.

Sethuraman (1994) showed that this F has the Dirichlet distribution with parameter measure α. [More precisely, let $\mathcal{B}_{\mathcal{P}}$ be the smallest σ-field in $\mathcal{P}$ such that the function $P \mapsto P(A)$ is measurable for each Borel set A. Then (2.2) defines a measurable map from $(\Omega, \mathcal{F}, Q)$ to $(\mathcal{P}, \mathcal{B}_{\mathcal{P}})$, and the induced distribution on F is the Dirichlet distribution with parameter measure α.]

From this construction we see that we can easily generate an F with distribution approximately equal to $\mathcal{D}_\alpha$: fix some large integer J, draw $B_1, \dots, B_J$ i.i.d. from Beta$(1, \alpha(R))$, $V_1, \dots, V_J$ i.i.d. from H and form

$$F_J = \sum_{j=1}^{J} P_j \, \delta_{V_j}. \tag{2.3}$$

However, if we let J in (2.3) be random, we can generate an F_J with sufficient accuracy to be able to generate $X_1, \dots, X_n$ which are *exactly* i.i.d. $\sim F$. This is done as follows. Recall that F is given by (2.2). Thus, to generate $X_1 \sim F$, we need to choose a random index J_1 according to the distribution

$$P\{J_1 = j\} = P_j, \qquad j = 1, 2, \dots, \tag{2.4}$$

and set $X_1 = V_{J_1}$. We choose the index J_1 by using a $U(0,1)$ random variable U_1, as follows. Let J_1 be such that $\Sigma_{j=1}^{J_1 - 1} P_j \le U_1 \le \Sigma_{j=1}^{J_1} P_j$. It is clear that J_1 satisfies (2.4). This gives a random variable X_1. To obtain $X_1, \dots, X_n$ i.i.d. $\sim F$, we repeat this n times using n independent $U(0,1)$ random variables $U_1, \dots, U_n$. The sequences $B_1, B_2, \dots$ and $V_1, V_2, \dots$ are held fixed throughout the n repetitions. Note that to do this we need only know $B_1, \dots, B_J$ and $V_1, \dots, V_J$, where $J = \max\{J_1, \dots, J_n\}$; that is, J is such that $\Sigma_{j=1}^{J} P_j \ge U_{(n)}$, where $U_{(n)} = \max\{U_1, \dots, U_n\}$. Notice that for any i, $X_i \sim H$, but the X_i's are not independent, since for $i_1 \ne i_2$, X_{i_1} and X_{i_2} are equal with positive probability.

2.2. *The posterior distribution of F when the prior on F is a simple Dirichlet.* Although our prior on F is a mixture of Dirichlets, it may be useful at this point to first describe in detail the algorithm for the simpler case where the prior on F is a single Dirichlet. So assume that $F \sim \mathcal{D}_\alpha$, where $\alpha = \alpha(R)H$ and the data are that $X_i \in A_i$, $i = 1, \dots, n$. We want to simulate an observation from $\mathcal{L}(F \mid \text{data})$. An important property of the Dirichlet process prior is that if $F \sim \mathcal{D}_\alpha$ and $X_1, \dots, X_n$ are i.i.d. $\sim F$, then

$$\mathcal{L}(F \mid X_1, \dots, X_n) = \mathcal{D}_{\alpha + \Sigma_{i=1}^{n} \delta_{X_i}}; \tag{2.5}$$

that is, the posterior distribution of F given the X's is again a Dirichlet, but with an updated parameter measure; see Ferguson (1973, 1974). This fact will play a central role in the algorithm described below.

We will need to take starting values $X_1^{(0)}, \dots, X_n^{(0)}$. It is convenient to introduce the following notation. If L is a distribution function, X is distributed according to L and A is an event, then L_A will denote the conditional distribution of X given that $X \in A$. We may take the starting values by generating $X_i^{(0)} \sim H_{A_i}$ independently for $i = 1, \dots, n$. (The issue of how to choose starting values is addressed in more detail in Section 4).

For $k = 1, \dots, K$,

1. Generate $F^{(k)} \sim \mathcal{D}_{\alpha + \Sigma_{i=1}^{n} \delta_{X_i^{(k-1)}}}$.
2. Generate $(X_1^{(k)}, \dots, X_n^{(k)}) \sim \mathcal{L}((X_1, \dots, X_n) | F^{(k)}, X_i \in A_i, i = 1, \dots, n)$.

We now describe how step 2 above is carried out. Assume without loss of generality that for $i = 1, \dots, n_u$, the ith observation is uncensored, that is, we observe $X_i = x_i$, while for $i = n_u + 1, \dots, n$, we know only that $X_i \in A_i$ for some sets A_i. Generating $X_i^{(k)}$ from $\mathcal{L}(X_i \,|\, F^{(k)}, X_j \in A_j,\ j = 1, \dots, n)$ for $i = 1, \dots, n_u$ is trivial: We just take $X_i^{(k)} = x_i$. For the censored observations we use a simple rejection method. Step 2 of the algorithm proceeds as follows.

```
2a. For $i = 1, \dots, n_u$, set $X_i^{(k)} = x_i$.
2b. For $i = n_u + 1, \dots, n$,
    (i) Generate $U_i^{(1)} \sim U(0, 1)$.
    (ii) Choose $J_i^{(1)}$ such that $\Sigma_{j=1}^{J_i^{(1)} - 1} P_j \le U_i^{(1)} \le \Sigma_{j=1}^{J_i^{(1)}} P_j$.
    (iii) If $V_{J_i^{(1)}} \in A_i$, set $X_i^{(k)} = V_{J_i^{(1)}}$; otherwise, repeat steps (i) and (ii) using independent uniforms $U_i^{(1)}, \dots, U_i^{(e_i)}$ until $V_{J_i^{(e_i)}} \in A_i$.
```

The sequences $B_1, B_2, \dots$ and $V_1, V_2, \dots$ are held fixed throughout this process. Note that to do this we need only know $B_1, \dots, B_J$ and $V_1, \dots, V_J$, where J is such that $\Sigma_{j=1}^{J} P_j \ge \max\{U_{n_u+1}^{(1)}, \dots, U_{n_u+1}^{(e_{n_u}+1)}, U_{n_u+2}^{(1)}, \dots, U_{n_u+2}^{(e_{n_u}+2)}, \dots, U_n^{(1)}, \dots, U_n^{(e_n)})\}$.

There is a slightly different way of implementing the algorithm. We may think of the data as coming in two stages, where in the first stage we observe the uncensored observations and in the second stage we observe the censored observations. We wish to find $\mathcal{L}(F \,|\, X_i, i = 1, \dots, n_u, X_i \in A_i, i = n_u + 1, \dots, n)$, where F has prior $\mathcal{D}_\alpha$, and it is clear from (2.5) that this is equivalent to finding $\mathcal{L}(F \,|\, X_i \in A_i,\ i = n_u + 1, \dots, n)$, where F has prior $\mathcal{D}_{\alpha + \Sigma_{i=1}^{n_u} \delta_{X_i}}$. This is accomplished through the algorithm described earlier, except that the effective sample size is now the number of censored observations. Although these two implementations of the algorithm are distinct, it follows from the convergence results of Section 4 that after enough iterations, a random distribution function F generated by either method will be distributed according to $\mathcal{L}(F \,|\, \text{data})$. We briefly illustrate the algorithm on the data set used by Susarla and van Ryzin (1976), which is the same as the data set used in the original paper by Kaplan and Meier (1958). The data are 0.8, 1.0+, 2.7+, 3.1, 5.4, 7.0+, 9.2, 12.1+, where a "+" denotes a censored observation. Our illustration is not intended to give any new

TABLE 1

Line 2 of the table gives the exact mean of the posterior distribution of $F(t)$. Line 3 gives the estimate of the mean of the posterior distribution of $F(t)$ obtained by successive substitution sampling. Line 4 gives the Kaplan–Meier estimate of $F(t)$

t	0.80	1.00	2.70	3.10	5.40	7.00	9.20	12.10
Exact	0.1083	0.1190	0.2071	0.3006	0.4719	0.5256	0.6823	0.7501
SSS	0.1083	0.1190	0.2084	0.3011	0.4706	0.5261	0.6802	0.7476
KME	0.1250	0.1250	0.1250	0.3000	0.4750	0.4750	0.7375	0.7375

insight into the data, but only to make a numerical comparison of our algorithm with the exact results obtained by Susarla and van Ryzin (1976). They took H to be the exponential distribution with mean 1/0.12, and for $\alpha(R)$, they considered three cases, $\alpha(R) = 4, 8$ and 16, but for the sake of brevity we make our comparison only for the case $\alpha(R) = 8$. They obtained the mean of the posterior distribution of $F(t)$ as t ranges over the eight censored and uncensored points. Our method gives an estimate of the conditional distribution of $F(t)$ for these eight points, and we took the mean of this conditional distribution to make our comparison. We took the mean number of iterations K to be 50 and the number of repetitions of the algorithm M to be 500. Table 1 compares our results with theirs, and for the sake of reference also gives the Kaplan–Meier estimate. We see that the comparison is excellent.

2.3. *The posterior distribution of F in the general case.* We now return to the situation of main interest, where we consider an entire parametric family $H_\theta, \theta \in \Theta$, and put as prior on F the mixture of Dirichlets

$$F \sim \int \mathcal{D}_{\alpha_\theta}\, \nu(d\theta), \tag{2.6}$$

where for each $\theta \in \Theta$, $\alpha_\theta = \alpha_\theta(R)H_\theta$ and $0 < \alpha_\theta(R) < \infty$. This is slightly more general than the mixture (2.1) in that the $\alpha_\theta(R)$'s are not assumed to be all equal. This extra generality will most often not be needed, but we include it because it does not really make our formulas more complicated; see Remark 1 below. Henceforth we will assume that for each Borel set $B \subset R$, the map $\theta \mapsto \alpha_\theta(B)$ is measurable. A thorough development of the theory of mixtures of Dirichlet priors may be found in Antoniak (1974).

To carry out the algorithm in this case we will need a formula, analogous to (2.5), giving the posterior distribution of F when we observe the *complete* values $X_1, \dots, X_n$. This formula is provided by Theorem 1 below. We use the notation $\mathbf{X} = (X_1, \dots, X_n)$. Also, for a vector $\mathbf{v} \in R^n$, $\#(\mathbf{v})$ will denote the number of distinct values of $v_1, \dots, v_n$.

THEOREM 1. *Assume that for each $\theta \in \Theta$, H_θ is absolutely continuous, with a density h_θ that is continuous on R. If the prior on F is given by* (2.6), *then the* *posterior distribution of F given $X_1, \dots, X_n$ is*

$$\int \mathcal{D}_{\alpha_\theta + \Sigma_{i=1}^n \delta_{X_i}} \nu_{\mathbf{X}}(d\theta), \tag{2.7}$$

where $\nu_{\mathbf{X}}$ is the measure which is absolutely continuous with respect to ν and is defined by

$$\nu_{\mathbf{X}}(d\theta) = c(\mathbf{X})\Big(\prod\nolimits^{*} h_\theta(X_i)\Big)\frac{\big(\alpha_\theta(R)\big)^{\#(\mathbf{X})}\Gamma\big(\alpha_\theta(R)\big)}{\Gamma\big(\alpha_\theta(R)+n\big)}\nu(d\theta), \tag{2.8}$$

where the "$*$" in the product indicates that the product is taken over distinct values only, Γ is the gamma function and $c(\mathbf{X})$ is a normalizing constant.

The theorem is a special case of Lemma 1 of Antoniak (1974).

The theorem enables us to carry out the successive substitution sampling algorithm in the general case: from (2.7), we see that step 1 of the algorithm of Section 2.2 should be replaced by:

1a. Generate $\theta^{(k)}$ from $\nu_{\mathbf{X}^{(k-1)}}$.
1b. Generate $F^{(k)}$ from $\mathcal{D}_{\alpha_{\theta^{(k)}} + \Sigma_{i=1}^n \delta_{X_i^{(k-1)}}}$.

Here, we use $\mathbf{X}^{(j)}$ to denote the vector $(X_1^{(j)}, \dots, X_n^{(j)})$.

Remarks.

1. In practice, when dealing with the parametric family H_θ, $\theta \in \Theta$, it will be very convenient to take $((\alpha_\theta(R))^{\#(\mathbf{X})}\Gamma(\alpha_\theta(R))/\Gamma(\alpha_\theta(R)+n))\nu(d\theta)$ to be a conjugate prior, for then sampling from $\nu_{\mathbf{X}}$ is particularly easy. The easiest way to get $((\alpha_\theta(R))^{\#(\mathbf{X})}\Gamma(\alpha_\theta(R))/\Gamma(\alpha_\theta(R)+n))\nu(d\theta)$ to be a conjugate prior is to take $\alpha_\theta(R)$ to be constant in θ and take ν to be a conjugate prior. In this case

$$\nu_{\mathbf{X}}(d\theta) = d(\mathbf{X})\Big(\prod\nolimits^{*} h_\theta(X_i)\Big)\nu(d\theta), \tag{2.9}$$

   where $d(\mathbf{X})$ is a normalizing constant.

2. If we consider the standard and simpler Bayesian model in which $X_1, \dots, X_n$ are i.i.d. from the distribution H_θ for some $\theta \in \Theta$ and if we put the prior ν on θ, then the posterior distribution of θ given $X_1, \dots, X_n$ is

$$\nu_{\mathbf{X}}(d\theta) = e(\mathbf{X})\Big(\prod h_\theta(X_i)\Big)\nu(d\theta) \tag{2.10}$$

   for some normalizing constant $e(\mathbf{X})$, and this is the same as (2.9) except that the product is over all the X_i's. When the X_i's are generated by successive substitution sampling (step 2 of the algorithm of Section 2.2), there will be ties, especially if the $\alpha_\theta(R)$'s are small, and so there is a genuine distinction between (2.9) and (2.10).

3. It will be clear from the discussion in Section 4 that for the case where we observe only incomplete data, we have

$$\mathcal{L}(F \mid \text{data}) = \int \mathcal{D}_{\alpha_\theta + \Sigma_{i=1}^n \delta_{X_i}} \mathcal{L}(d\theta, d\mathbf{X} \mid \text{data}),$$

where $\mathcal{L}(d\theta, d\mathbf{X} \mid \text{data})$ is the conditional distribution of $(\theta, \mathbf{X})$ given the data. Thus, the posterior distribution of F is still a mixture of Dirichlets. However, this representation is not useful because there is no way to obtain the mixing measure.

3. Illustration on prostate cancer data.

We illustrate our method on a data set involving a clinical trial of 211 individuals who had Stage IV prostate cancer. These patients were treated with estrogen in a Veterans Administration Cooperative Urological Research Group study. The data are studied in Koziol and Green (1976), and the version of the data set we used is as in Hollander and Proschan (1979). Of the 211 individuals, 90 died of prostate cancer, 105 died of other diseases and 16 were still alive at the end of the study. The 105 who died of other diseases and the 16 still alive at the end of the study were treated as censored observations. One of the principal aims of the study was to determine the benefits of the estrogen therapy. It is therefore of interest to study F, the distribution of survival time if cancer of the prostate was the only cause of death. As discussed in Koziol and Green (1976), prior experience suggested that if the patients had not been treated with estrogen, the distribution of their time until death if cancer of the prostate was the only cause of death would be exponential with mean 100 months.

We analyzed this data set using our method taking our prior on F to be (2.6), where H_θ is the exponential distribution with parameter θ (θ being the reciprocal of the mean), with $\alpha_\theta(R)$ constant in θ and we considered two cases, $\alpha_\theta(R) = 1$ and $\alpha_\theta(R) = 100$. We took ν to be G(33, 3300), where G(a, b) is the Gamma distribution with shape parameter a and scale parameter b. Our reason for this is that we wished to center the prior around the family of exponential distributions. The value $\alpha_\theta(R) = 100$ corresponds roughly to the situation where we are reasonably confident about the assumption of exponentiality, whereas the value $\alpha_\theta(R) = 1$ gives a more diffuse prior. We noted that the median of an exponential distribution with parameter θ is $(\log 2)/\theta$, and if $\theta \sim$ G(33, 3300), then $P(50 \leq (\log 2)/\theta \leq 100) \doteq .95$. The reason for taking a Gamma prior on θ is that this is a conjugate family for the exponential distribution. If ν is the $G(a, b)$ distribution, then [see (2.8)] $\nu_{\mathbf{X}}$ is the G$(a + n^*, b + \Sigma^* X_i)$ distribution, where n^* is the number of distinct observations in $\mathbf{X}$, and $\Sigma^* X_i$ is the sum of the distinct X_i's. It is routine to generate observations from this distribution. We felt that these choices of prior on F were reasonable, although other choices are also possible.

A full listing of the data may be found on page 399 of Hollander and Proschan (1979). It is useful to look at Figure 2, which gives an estimate of the density of F [obtained by smoothing the Kaplan–Meier estimate of F by the method of Ramlau–Hansen (1983)]. It should be mentioned at this point that a feature of this data set is that the large observations are very heavily censored (in particular, there are only two uncensored deaths beyond 120 months and only one beyond 150 months, 158 being the last uncensored observation), so it is very difficult to nonparametrically estimate a large portion of the right tail of F. Figures 1(a) and (b) show the posterior densities of $F(60), F(120), F(150)$ and $F(180)$, the 5-year, 10-year, 12.5-year and 15-year survival rates, for $\alpha_\theta(R) = 1$ and $\alpha_\theta(R) = 100$, respectively. It is interesting to note that for $\alpha_\theta(R) = 1$ the posterior density of $F(180)$ has a spike near 1. Figures 1(a) and (b) were obtained by running the algorithm with the number of iterations K equal to 100 and the number of repetitions of the algorithm M to be 3000. Each of the plots is an average of 3000 Beta densities.

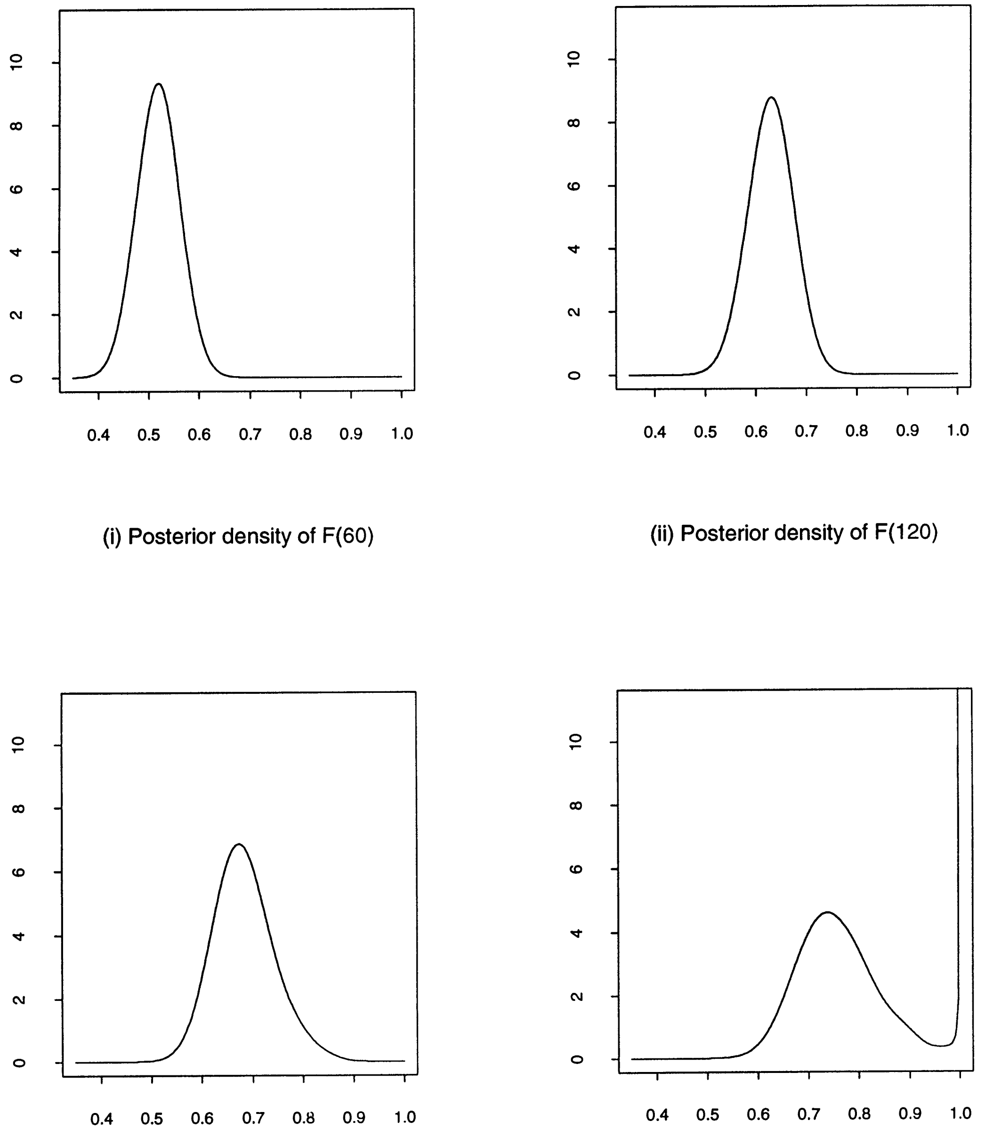
Table 2 compares our results with the standard frequentist analysis. Line 2 gives the Kaplan–Meier estimate of $F(t)$ for t equal to 60, 120 and 150. Beyond 158, the Kaplan–Meier estimate is undefined. Lines 3 and 4 of the table give the means of the posterior distribution of $F(t)$ for the four values of t, for two values of $\alpha_\theta(R)$. Line 5 gives the 95% confidence interval for $F(t)$ computed using Link's (1984) method. This method involves first finding a 95% confidence interval for the cumulative hazard function $\Lambda(t)$ and then transforming this into a confidence interval for $F(t)$ using the relationship $F(t) = 1 - \exp(-\Lambda(t))$. This method is commonly used and is available, for example, in the statistical computer package S. Lines 6 and 7 of the table give the 95% central intervals of the posterior distributions of $F(t)$, for two values of $\alpha_\theta(R)$.

For t equal to 60 and 120, as well as for other values of t in that range but not shown here, the Bayes method with $\alpha_\theta(R) = 1$ agrees very closely with the standard nonparametric frequentist method. However, differences emerge for values of t beyond which there are few uncensored observations, and it is only for those values of t that the choices made in specifying the prior had a noticeable impact. For the case $\alpha_\theta(R) = 100$ the Bayesian analysis corresponds more closely to a parametric analysis using the exponential distribution, which is to be expected. It should be mentioned that this study involves a reasonably large sample size, and so we expect that the "data swamps the prior," at least for the case $\alpha_\theta(R) = 1$.

Figures 3(a) and (b) show plots of a density estimate of the posterior distribution of a future observation, for the cases $\alpha_\theta(R) = 1$ and $\alpha_\theta(R) = 100$. These were obtained by running the algorithm for 100 iterations and first obtaining

TABLE 2
Comparison of our Bayes method with the standard frequentist procedure for estimation of $F(t)$

t	60	120	150	180
KME	0.528	0.628	0.681	
Posterior mean, $\alpha_\theta(R) = 1$	0.519	0.630	0.683	0.786
Posterior mean, $\alpha_\theta(R) = 100$	0.485	0.669	0.736	0.805
Standard 95% interval	(0.437, 0.605)	(0.528, 0.707)	(0.531, 0.783)	
Bayes 95% int., $\alpha_\theta(R) = 1$	(0.436, 0.602)	(0.540, 0.716)	(0.573, 0.815)	(0.621, 1.000)
Bayes 95% int., $\alpha_\theta(R) = 100$	(0.416, 0.555)	(0.592, 0.742)	(0.657, 0.809)	(0.724, 0.879)



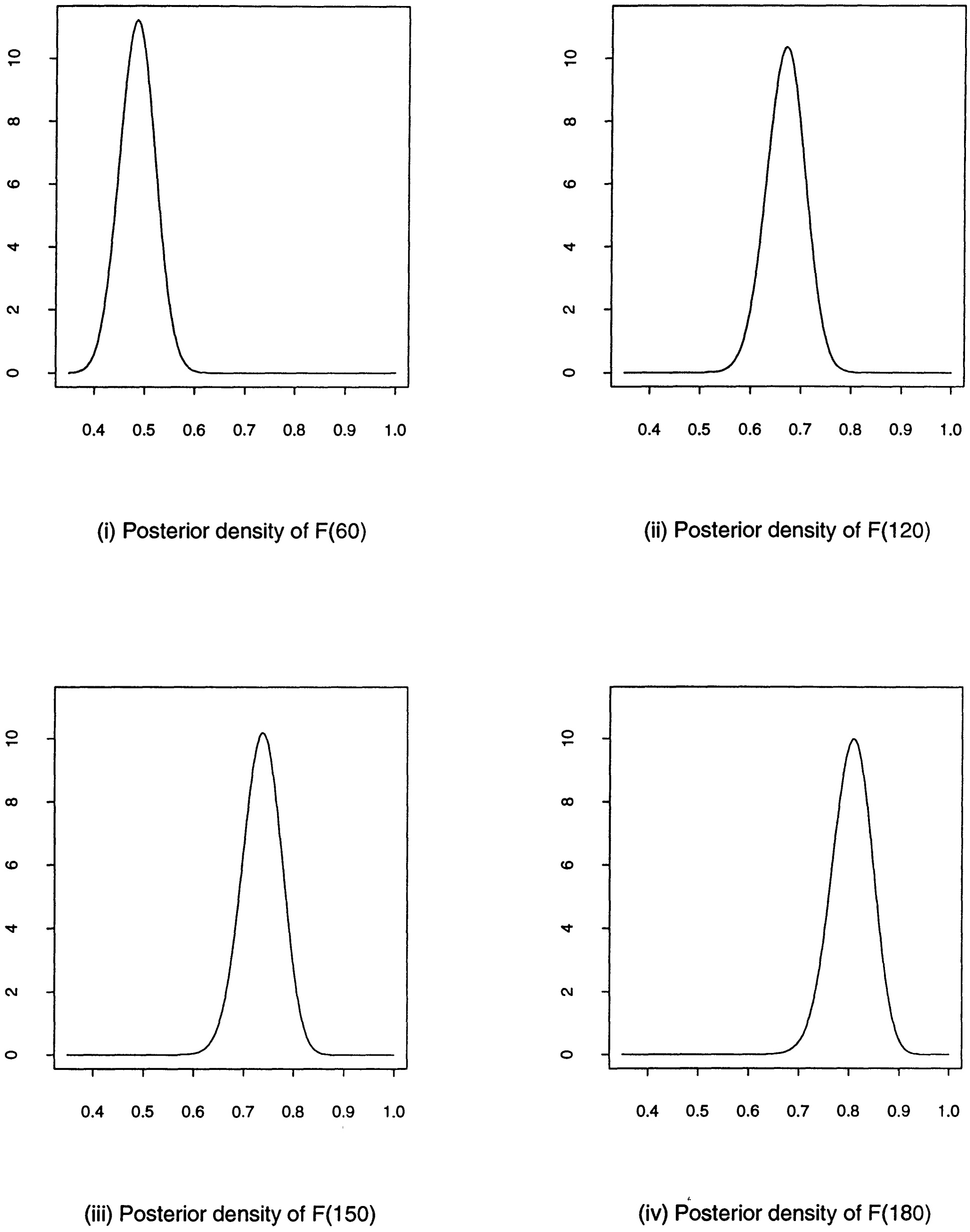
(i) Posterior density of F(60)

(ii) Posterior density of F(120)

(iii) Posterior density of F(150)

(iv) Posterior density of F(180)

FIG. 1a. *Prostate cancer data: Posterior densities of $F(t)$ given the data for four values of t, $\alpha(R) = 1$.*



(i) Posterior density of F(60)

(ii) Posterior density of F(120)

(iii) Posterior density of F(150)

(iv) Posterior density of F(180)

FIG. 1b. *Prostate cancer data: Posterior densities of $F(t)$ given the data for four values of t*, $\alpha(R) = 100$.

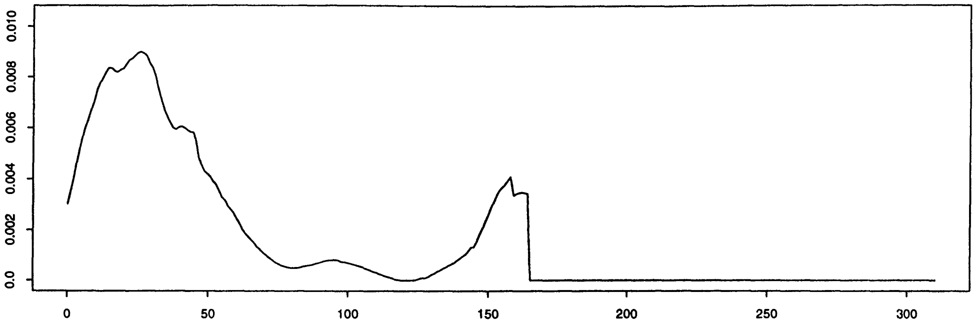


FIG. 2. *Prostate cancer data: Frequentist estimate of density of F obtained by smoothing the KME.*

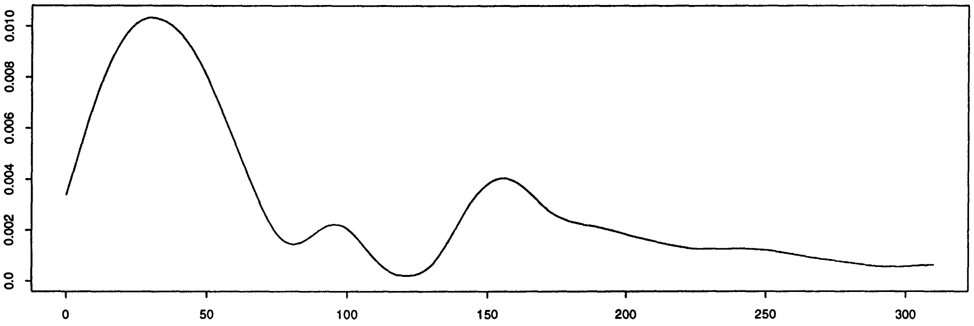


FIG. 3a. *Prostate cancer data: Density estimate of posterior distribution of a future observation given the data,* $\alpha(R) = 1$.

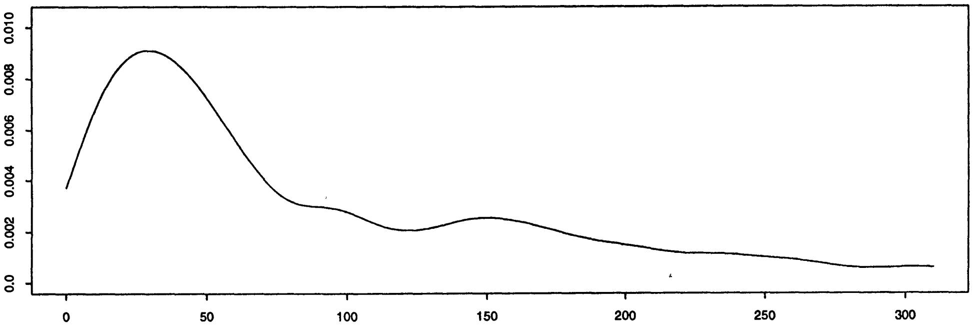


FIG. 3b. *Prostate cancer data: Density estimate of posterior distribution a future observation given the data,* $\alpha(R) = 100$.

an F from $\mathcal{L}(F \mid \text{data})$, and then generating a random variable X from this F. This was repeated independently 3000 times. A density estimate was made of the resulting 3000 points. Taking the 0.1 and the 0.9 quantiles of the distributions of $\mathcal{L}(X \mid \text{data})$ gives the 80% probability intervals [15, 250] and [13.8, 246] (the corresponding quantity in a frequentist setting is the "tolerance interval," which is difficult to interpret).

One would expect that Figure 3(b) gives a density that is closer to a power law (the distribution of a future X under the parametric Bayesian model), but that is actually true only in the part of the right tail where the data are sparse. Further analysis showed that the posterior distribution of a future observation does not look like a power law even for $\alpha_\theta(R) = 2500$. Thus, we conclude that the exponential model is not appropriate, a conclusion also reached by Hollander and Peña (1992) in their analysis of this data set.

As noted by Hollander and Proschan (1979), the survival probability for the untreated group is greater than that for the treated group in the "middle" (30–90) months. This is confirmed by our analysis, which, however, also shows the opposite for long-term survival. For example, when $\alpha_\theta(R) = 1$, the probability that a future X in the treated group is greater than 150 is 0.310, substantially larger than $\exp(-150/100) = 0.223$, the corresponding quantity for the untreated group. Even when $\alpha_\theta(R) = 100$ the Bayes estimate is 0.257, which still indicates benefit of treatment for long-term survival.

FORTRAN programs to calculate the estimates are available from the author.

4. Convergence of the successive substitution sampling algorithm. Recall that in the general case, the algorithm proceeds as follows.

Fix starting values $(\theta^{(0)}, F^{(0)}, \mathbf{X}^{(0)})$.
For $k = 1, \ldots, K$,

1a. Generate $\theta^{(k)}$ from $\nu_{\mathbf{X}^{(k-1)}}$.
1b. Generate $F^{(k)}$ from $\mathcal{D}_{\alpha_{\theta^{(k)}} + \Sigma_{i=1}^n \delta_{X_i^{(k-1)}}}$.
2. Generate $\mathbf{X}^{(k)} \sim \mathcal{L}(\mathbf{X} \mid F^{(k)}, \text{data})$.

Let $\mathcal{B}_\Theta$ be the Borel field on Θ, let $\mathcal{B}_{R^k}$ be the Borel field on R^k and recall that $\mathcal{B}_\mathcal{P}$ is the Borel field on $\mathcal{P}$ which is defined in the paragraph following (2.2). Whenever we consider a product space, the σ-field on this product space will be the product σ-field.

Note that if $(\theta^{(0)}, F^{(0)}, \mathbf{X}^{(0)})$ are starting values, we actually need to know only $\mathbf{X}^{(0)}$ to start the algorithm. Suppose that we choose $\mathbf{X}^{(0)}$ as follows. First choose $\theta \sim \nu$ and then generate $X_i^{(0)} \sim H_{\theta, A_i}$ independently for $i = 1, \ldots, n$. [Recall that if L is a distribution function and A is a set, then L_A denotes the distribution function given by $L_A(S) = L(S \cap A)/L(A)$.] Let τ denote the distribution of $\mathbf{X}^{(0)}$.

We will use the following general notation: $P_{\mathbf{X}^{(0)}}$ refers to probabilities relating to the algorithm when the starting point is $\mathbf{X}^{(0)}$; P refers to probabilities relating to the model in which $\theta \sim \nu$, the conditional distribution of F given θ is $\mathcal{D}_{\alpha_\theta}$ and, given $F, X_1, \ldots, X_n$ are i.i.d. $\sim F$.

Our objective is to show that

$$\sup_{B\in\mathcal{B}_{R^k}\times\mathcal{B}_\Theta\times\mathcal{B}_\mathcal{P}}\Big|P_{\mathbf{X}^{(0)}}\Big\{\big(\mathbf{X}^{(k-1)},\theta^{(k)},F^{(k)}\big)\in B\Big\} - P\big\{(\mathbf{X},\theta,F)\in B\,|\,\text{data}\big\}\Big| \to 0 \text{ for } [\tau]\text{-almost all } \mathbf{X}^{(0)} \tag{4.1}$$

(this convergence in total variation norm of the triples is stronger, of course, than convergence in distribution). This will imply in particular that

$$\sup_{B\in\mathcal{B}_\mathcal{P}}\big|P_{\mathbf{X}^{(0)}}\{F^{(k)}\in B\} - P\{F\in B\,|\,\text{data}\}\big| \to 0 \quad \text{for } [\tau]\text{-almost all } \mathbf{X}^{(0)}.$$

Let μ be the conditional distribution of $(\theta,\mathbf{X})$ given the data. Consider the sequence of random triples $(\mathbf{X}^{(k-1)},\theta^{(k)},F^{(k)})$, $k = 1,2,\dots$. Given a starting point, we may obtain the distribution of $(\mathbf{X}^{(k-1)},\theta^{(k)},F^{(k)})$ by conditioning on $\mathbf{X}^{(k-1)}$ and $\theta^{(k)}$. Also, we may obtain $\mathcal{L}((\mathbf{X},\theta,F)\,|\,\text{data})$ by integrating the conditional distribution of $(\mathbf{X},\theta,F)$ given $\mathbf{X}$ and θ with respect to μ. Our plan for proving (4.1) is to first establish convergence of the distribution of $(\mathbf{X}^{(k-1)},\theta^{(k)})$, which is the heart of the proof, and then show via a simple argument that this entails (4.1). For technical reasons, it is easier to first deal with the sequence $(\theta^{(k)},\mathbf{X}^{(k)})$, $k = 1,2,\dots$. Let $\mu^k_{\mathbf{X}^{(0)}}(\cdot)$ be the marginal distribution of $(\theta^{(k)},\mathbf{X}^{(k)})$ when the algorithm is started at $\mathbf{X}^{(0)}$. We will show that

$$\sup_{B\in\mathcal{B}_\Theta\times\mathcal{B}_{R^k}}\big|\mu^k_{\mathbf{X}^{(0)}}(B)-\mu(B)\big| \to 0 \quad [\tau]\text{-a.e.} \tag{4.2}$$

To do this, we will establish that $(\theta^{(k)},\mathbf{X}^{(k)})$, $k = 1,2,\dots$, satisfies the conditions of an appropriate ergodic theorem for Markov chains on general state spaces. Before proceeding, we review some definitions and concepts related to Markov chains.

Let $\{Y_k\}$ be a Markov chain on the state space $(\mathcal{Y},\mathcal{B}_\mathcal{Y})$ with transition probabilities $P_y(C)$. Let $P^k_y(C)$ be the k-step transition probabilities defined by $P^1_y(C) = P_y(C)$ and $P^k_y(C) = \int P_z(C)P^{k-1}_y(dz)$ for $k = 2,\dots$. For any set $C\in\mathcal{B}_\mathcal{Y}$ let $G_y(C) = \Sigma^\infty_{k=1}P^k_y(C)$. This is the expected number of visits to C starting from y.

THEOREM 2 (Ergodic theorem for Markov chains). *Let $\{Y_k\}$ be a Markov chain with stationary probability distribution π; that is, π is a probability measure satisfying*

$$\pi(C) = \int P_y(C)\pi(dy) \quad \textit{for all } C\in\mathcal{B}_\mathcal{Y}.$$

Suppose that there exists a set $A\in\mathcal{B}$, a probability measure ρ with $\rho(A) = 1$, a constant $\varepsilon > 0$ and an integer $n_0 \geq 1$ such that

$$G_y(A) > 0 \quad \textit{for all } y\in\mathcal{Y}, \tag{4.3}$$

$$P^{n_0}_y(C) \geq \varepsilon\rho(C) \quad \textit{for all } y\in A,\ C\in\mathcal{B}_\mathcal{Y}. \tag{4.4}$$

Assume further the aperiodicity condition

$$\begin{aligned}&\text{g.c.d.}\{m \geq 1\text{: } \textit{there is } \varepsilon_m > 0 \textit{ such that } P_y^m(C) \geq \varepsilon_m \rho(C)\\ &\qquad \textit{for each } y \in A,\ C \in \mathcal{B}_{\mathcal{Y}}\} = 1.\end{aligned} \tag{4.5}$$

Then there exists a set D_0 *with* $\pi(D_0) = 1$ *such that*

$$\sup_{C \in \mathcal{B}_{\mathcal{Y}}} \left|P_y^k(C) - \pi(C)\right| \to 0 \quad \textit{for each } y \in D_0.$$

The theorem appears in Athreya, Doss and Sethuraman (1992).

Note that (4.3) is equivalent to the condition that for any point y in the state space, the probability that the chain starting from y will eventually enter the set A is positive. Also note that the aperiodicity condition (4.5) is automatically satisfied if the n_0 appearing in (4.4) is 1.

We remark that there exist ergodic theorems for Markov chains on general state spaces for which the basic assumption is that the chain is *irreducible with respect to the stationary distribution.* This is, essentially, the condition that starting from any point in the state space, the chain has positive probability of eventually entering any set to which the stationary distribution assigns positive mass. [For a precise statement of such a theorem, see Theorem 1 of Tierney (1994).] To check this condition, one needs to be able to identify the sets which have positive mass under the stationary distribution. In our case it is very difficult to get a handle on the stationary distribution, that is, $\mathcal{L}((\mathbf{X}, \theta, F) \mid \text{data})$. Even the marginal conditional distribution $\mathcal{L}((\mathbf{X}, \theta) \mid \text{data})$ is complicated, because of the dependence among the X_i's. By contrast, the conditions required by Theorem 2 involve only the transition kernel and do not refer to the stationary distribution.

We will now apply Theorem 2 to show that under additional assumptions on the measures ν and α_θ, if the starting point of the algorithm is chosen according to τ, then the algorithm converges with probability 1.

Recall that we suppose that the observations are uncensored for $i = 1, \dots, n_u$ and are censored for $i = n_u + 1, \dots, n$. If X_i is uncensored, the set A_i is the singleton $\{X_i\}$. Also, λ will denote ordinary Lebesgue measure on R.

THEOREM 3. *Assume that there exists a set* $E_0 \subset \Theta$ *with* $\nu(E_0) > 0$, *a* $\delta > 0$ *and, for* $i = n_u + 1, \dots, n$, *disjoint sets* $E_i \subset A_i$, *with positive Lebesgue measure, such that*:

(ia) $\nu_{\mathbf{X}}(C_0) \geq \delta \nu(C_0)$ *for all* $\mathbf{X} \in A_1 \times \cdots \times A_{n_u} \times E_{n_u+1} \times \cdots \times E_n$ *and* $C_0 \subset E_0$,

(ib) $H_\theta(C_i) \geq \delta \lambda(C_i)$ *for all* $\theta \in E_0$ *and* $C_i \subset E_i$, $i = n_u + 1, \dots, n$,

(ii) $P_{\mathbf{X}^{(0)}}\{(\theta^{(k)}, \mathbf{X}^{(k)}) \in E_0 \times A_1 \times \cdots \times A_{n_u} \times E_{n_u+1} \times \cdots \times E_n$ *for some* $k \geq 1\} > 0$ *for all* $\mathbf{X}^{(0)} \in R^n$,

(iii) *there exists* $\eta > 0$ *such that*

$$\frac{\left(\alpha_\theta(R)\right)^{n-n_u}\Gamma\left(\alpha_\theta(R)+n\right)}{\Gamma\left(\alpha_\theta(R)+2n-n_u\right)} > \eta \quad \textit{for all } \theta \in E_0.$$

Then the Markov chain $\{(\theta^{(k)}, \mathbf{X}^{(k)}),\ k = 0, 1, \dots\}$ *satisfies the conditions of Theorem* 2, *with* $n_0 = 1$, $A = E_0 \times A_1 \times \cdots \times A_{n_u} \times E_{n_u+1} \times \cdots \times E_n$, *probability measure* ρ *on* A *given by* $\rho = \nu_{E_0} \times \delta_{X_1} \times \cdots \times \delta_{X_{n_u}} \times \lambda^{E_{n_u}+1} \times \cdots \times \lambda^{E_n}$ *and* $\varepsilon = \delta^{n-n_u+1}\eta\nu(E_0)\Pi_{i=n_u+1}^{n}\lambda(E_i)$. *Here*, λ^{E_i} *denotes Lebesgue measure restricted to* E_i *and normalized to be a probability measure. Sufficient conditions for* (ii) *to hold are that*

$$\nu_{\mathbf{X}}(E_0) > 0 \quad \textit{whenever } X_i \in A_i,\ i = 1, \dots, n, \tag{4.6}$$

and

$$H_\theta(E_i) > 0 \quad \textit{for all } i = n_u + 1, \dots, n,\ \textit{whenever } \theta \in E_0. \tag{4.7}$$

PROOF. We will first verify (4.4) with $n_0 = 1$. For the sake of unity of notation, we let $E_i = A_i$ for $i = 1, \dots, n_u$. Also, for an arbitrary measure K, $K_{A_i} = \delta_{X_i}$ for $i = 1, \dots, n_u$. Let $(\theta^{(0)}, \mathbf{X}^{(0)}) \in E_0 \times E_1 \times \cdots \times E_n$. Recall that to go from $(\theta^{(0)}, \mathbf{X}^{(0)})$ to $(\theta^{(1)}, \mathbf{X}^{(1)})$ we generate $\theta^{(1)} \sim \nu_{\mathbf{X}^{(0)}}$, where $\nu_{\mathbf{X}}$ is given by Theorem 1, choose F from $\mathcal{D}_{\alpha_{\theta^{(1)}} + \Sigma_{i=1}^n \delta_{X_i^{(0)}}}$ and finally generate $X_i^{(1)} \sim F_{A_i}$ independently for $i = 1, \dots, n$.

By the monotone class lemma, it suffices to check (4.4) only for rectangles. [Also, it suffices to check (4.4) only for sets C which are contained in A, since $\rho(A) = 1$.] Let $C_i \subset E_i$, $i = 0, \dots, n$. We have

$$\begin{aligned} &P_{\mathbf{X}^{(0)}}\left\{(\theta^{(1)}, \mathbf{X}^{(1)}) \in C_0 \times \cdots \times C_n\right\} \\ &\quad = \int_{C_0}\int_{\mathcal{P}}\prod_{i=1}^{n} F_{A_i}(C_i)\mathcal{D}_{\alpha_{\theta^{(1)}} + \Sigma_{i=1}^n \delta_{X_i^{(0)}}}(dF)\nu_{\mathbf{X}^{(0)}}\left(d\theta^{(1)}\right). \end{aligned} \tag{4.8}$$

Now the inner integral in (4.8) is clearly equal to

$$\begin{aligned} \int \prod_{i=n_u+1}^{n} F_{A_i}(C_i)\mathcal{D}_{\alpha_{\theta^{(1)}} + \Sigma_{i=1}^n \delta_{X_i^{(0)}}}(dF) &= \int \prod_{i=n_u+1}^{n} \frac{F(C_i)}{F(A_i)}\mathcal{D}_{\alpha_{\theta^{(1)}} + \Sigma_{i=1}^n \delta_{X_i^{(0)}}}(dF) \\ &\geq \int \prod_{i=n_u+1}^{n} F(C_i)\mathcal{D}_{\alpha_{\theta^{(1)}} + \Sigma_{i=1}^n \delta_{X_i^{(0)}}}(dF). \end{aligned} \tag{4.9}$$

Because for $i = n_u + 1, \dots, n$ the C_i's are disjoint, the last integral in (4.9) is easy to compute since it is just a joint moment for the $(n - n_u + 1)$-dimensional

Dirichlet distribution. A calculation shows that this integral is equal to

$$\begin{aligned}
&\frac{\Gamma\left(\alpha_{\theta^{(1)}}(R)+n\right)}{\Gamma\left(\alpha_{\theta^{(1)}}(R)+2n-n_u\right)}\left(\prod_{i=n_u+1}^{n}\left(\alpha_{\theta^{(1)}}+\sum_{i=1}^{n}\delta_{X_i^{(0)}}\right)(C_i)\right)\\
&\quad\geq\frac{\Gamma\left(\alpha_{\theta^{(1)}}(R)+n\right)}{\Gamma\left(\alpha_{\theta^{(1)}}(R)+2n-n_u\right)}\left(\prod_{i=n_u+1}^{n}\alpha_{\theta^{(1)}}(C_i)\right)\\
&\quad=\left(\alpha_{\theta^{(1)}}(R)\right)^{n-n_u}\frac{\Gamma\left(\alpha_{\theta^{(1)}}(R)+n\right)}{\Gamma\left(\alpha_{\theta^{(1)}}(R)+2n-n_u\right)}\left(\prod_{i=n_u+1}^{n}H_{\theta^{(1)}}(C_i)\right)\\
&\quad\geq\eta\delta^{n-n_u}\prod_{i=n_u+1}^{n}\lambda(C_i),
\end{aligned}\tag{4.10}$$

the last inequality being a consequence of conditions (iii) and (ib). Combining (4.8), (4.9) and (4.10), we see that

$$\begin{aligned}
&P_{\mathbf{X}^{(0)}}\left\{\left(\theta^{(1)},\mathbf{X}^{(1)}\right)\in C_0\times\cdots\times C_n\right\}\\
&\quad\geq\delta_{\nu_{E_0}}(C_0)\nu(E_0)\eta\delta^{n-n_u}\prod_{i=n_u+1}^{n}\left(\lambda^{E_i}(C_i)\lambda(E_i)\right)\prod_{i=1}^{n_u}\delta_{X_i}(\{X_i\}),
\end{aligned}$$

from which (4.4) follows.

Condition (4.3) follows immediately from condition (ii) of the theorem. We will show that (4.6) and (4.7) ensure that

$$P_{\mathbf{X}^{(0)}}\left\{\left(\theta^{(1)},\mathbf{X}^{(1)}\right)\in E_0\times\cdots\times E_n\right\}>0\quad\text{for all }\mathbf{X}^{(0)}\in R^n,\tag{4.11}$$

from which it will follow that (4.6) and (4.7) are sufficient but far from necessary to ensure condition (ii) of the theorem. We rewrite the probability in (4.11) as

$$\begin{aligned}
&\int_{E_0}\int_{\mathcal{P}}\prod_{i=1}^{n}F_{A_i}(E_i)\mathcal{D}_{\alpha_{\theta^{(1)}}+\Sigma_{i=1}^{n}\delta_{X_i^{(0)}}}(dF)\nu_{\mathbf{X}^{(0)}}\left(d\theta^{(1)}\right)\\
&\quad\geq\int_{E_0}\left(\alpha_{\theta^{(1)}}(R)\right)^{n-n_u}\frac{\Gamma\left(\alpha_{\theta^{(1)}}(R)+n\right)}{\Gamma\left(\alpha_{\theta_{(1)}}(R)+2n-n_u\right)}\left(\prod_{i=n_u+1}^{n}H_{\theta^{(1)}}(E_i)\right)\nu_{\mathbf{X}^{(0)}}\left(d\theta^{(1)}\right)
\end{aligned}$$

and we see that by (4.6), (4.7) and the fact that $0<\alpha_{\theta^{(0)}}(R)<\infty$ for all $\theta\in\Theta$, this is the integral of a positive function over a set of positive measure, and so is positive.

The aperiodicity condition (4.5) follows from the fact $n_0 = 1$. As mentioned earlier, the conditional distribution of $(\mathbf{X}, \theta)$ given the data is a stationary distribution. This concludes the proof of the theorem. □

REMARKS.

1. In practice, we will have $\alpha_\theta(R)$ constant in θ, and so condition (iii) will automatically be satisfied, and condition (ia) will involve the simpler measure given by (2.9).
2. Condition (ib) involves Lebesgue measure on R and therefore precludes the possibility that the sets E_i have infinite Lebesgue measure. We could have stated condition (ib) in terms of different measures, but we did not do so because in practice the sets E_i will be chosen so that they are compact sets and Lebesgue measure will then be the most natural measure to use.
3. The conditions of the theorem are actually *extremely* easy to verify for the standard parametric families, if we take $\alpha_\theta(R)$ to be constant and take ν to be a conjugate prior. In particular, condition (ia) then involves a posterior that is available in closed form.
4. In the statement of the theorem the sets $E_{n_u+1}, \dots, E_n$ are assumed disjoint. We do this so that we can obtain more easily a lower bound for $P_{\theta^{(0)}}\{\mathbf{X}^{(1)} \in C_1 \times \cdots \times C_n\}$ in the proof of the theorem. In practice, if we can find nondisjoint sets $E_{n_u+1}, \dots, E_n$ which satisfy the conditions of the theorem, then we should also be able to find disjoint sets that work also. It is possible to prove a version of the theorem without the assumption that the sets $E_{n_u+1}, \dots, E_n$ are disjoint, but the gains in doing so are outweighted by the complications that arise in the proof.

Recall that μ is the conditional distribution of $(\theta, \mathbf{X})$ given the data, and that τ is the distribution on $\mathbf{X}^{(0)}$ described just prior to (4.1). Given a starting point $(\theta^{(0)}, \mathbf{X}^{(0)})$, we need to know only $\mathbf{X}^{(0)}$. Let $\mu^{\mathbf{X}}(\cdot \mid \text{data})$ and $\mu^{\theta}(\cdot \mid \text{data})$ denote the marginal distributions of $\mathbf{X}$ and θ, respectively, when $(\theta, \mathbf{X}) \sim \mu$; that is, $\mu^{\mathbf{X}}(\cdot \mid \text{data})$ and $\mu^{\theta}(\cdot \mid \text{data})$ are the conditional distributions of $\mathbf{X}$ and θ, respectively, given the data. From Theorems 2 and 3 we conclude that there exists a set $D_0 \subset R^n$ with the property that $\mu^{\mathbf{X}}(D_0 \mid \text{data}) = 1$ and such that

$$\sup_{B \in \mathfrak{B}_\Theta \times \mathfrak{B}_{R^n}} \left|\mu^k_{\mathbf{X}^{(0)}}(B) - \mu(B)\right| \to 0 \quad \text{for all } \mathbf{X}^{(0)} \in D_0.$$

So we need to generate the starting point from $\mu^{\mathbf{X}}(\cdot \mid \text{data})$. Unfortunately, $\mu^{\mathbf{X}}(\cdot \mid \text{data})$ is unknown. If, however, we can establish that τ is absolutely continuous with respect to $\mu^{\mathbf{X}}(\cdot \mid \text{data})$, then it will suffice to generate the starting point from the prior τ. The proposition below gives conditions that ensure this absolute continuity.

PROPOSITION 1. *If for every* $\mathbf{X}$ *such that* $X_i \in A_i$, $i = 1, \dots, n$, *we have*

$$\prod_{i=1}^{n} h_\theta(X_i) > 0 \quad \textit{for } [\nu]\textit{-a.e. } \theta, \tag{4.12}$$

then

$$\tau \ll \mu^{\mathbf{X}}(\,\cdot \mid \textit{data}). \tag{4.13}$$

PROOF. Here is a brief outline of the proof. We first show that condition (4.12) implies that the prior ν is absolutely continuous with respect to $\mu^{\theta}(\cdot \mid \text{data})$. Then we establish that for fixed θ, the distribution $H_{\theta, A_1} \times \cdots \times H_{\theta, A_n}$ [this is the distribution of $(X_1^{(0)}, \dots, X_n^{(0)})$ when $X_i^{(0)} \sim H_{\theta, A_i}$ and the $X_i^{(0)}$'s are independent] is absolutely continuous with respect to the conditional distribution of $\mathbf{X}$ given θ and the data, and we argue that this proves (4.13).

More formally, the proof proceeds as follows. Suppose that $S \subset R^n$ is a set such that $P\{\mathbf{X} \in S \mid \text{data}\} = 0$. Then

$$\int P\{\mathbf{X} \in S \mid \text{data}, \theta\} \mu^{\theta}(d\theta \mid \text{data}) = 0,$$

and thus

$$P\{\mathbf{X} \in S \mid \text{data}, \theta\} = 0 \quad \text{for } \left[\mu^{\theta}(\,\cdot \mid \text{data})\right]\text{-a.e. } \theta. \tag{4.14}$$

We will now show that $\nu \ll \mu^{\theta}(\cdot \mid \text{data})$. Suppose $S_0 \subset \Theta$ is such that $\mu^{\theta}(S_0 \mid \text{data}) = 0$. By conditioning on $\mathbf{X}$, we obtain

$$\int_{A_1 \times \cdots \times A_n} P\{\theta \in S_0 \mid \text{data}, \mathbf{X}\} d\mu^{\mathbf{X}}(d\mathbf{X} \mid \text{data}) = 0,$$

and in particular there exists a point $\mathbf{X}_0 \in A_1 \times \cdots \times A_n$ such that $P\{\theta \in S_0 \mid \text{data}, \mathbf{X}_0\} = 0$; that is, $\nu_{\mathbf{X}_{(0)}}(S_0) = 0$, where $\nu_{\mathbf{X}}$ is given by Theorem 1. From (4.12) this implies $\nu(S_0) = 0$, and we conclude that $\nu \ll \mu^{\theta}(\cdot \mid \text{data})$. Therefore, (4.14) gives

$$P\{\mathbf{X} \in S \mid \text{data}, \theta\} = 0 \quad [\nu]\text{-a.e. } \theta.$$

We rewrite this as

$$\sum_{r=1}^{n} P\{\mathbf{X} \in S \mid \text{data}, \theta \, \#(\mathbf{X}) = r\} P\{\#(\mathbf{X}) = r \mid \text{data}, \theta\} = 0 \quad [\nu]\text{-a.e. } \theta.$$

It is easy to see that $P\{\#(\mathbf{X}) = n \mid \text{data}, \theta\} > 0$ $[\nu]$-a.e. θ, and this implies

$$P\{\mathbf{X} \in S \mid \text{data}, \theta, \#(\mathbf{X}) = n\} = 0 \quad [\nu]\text{-a.e. } \theta. \tag{4.15}$$

Now it is not hard to see that, conditional on the data, θ and $\#(\mathbf{X}) = n$, the distribution of the X_i's is that of n independent random variables with distributions H_{θ, A_i}. We argue as follows. This distribution is obtained by first choosing $F \sim \mathcal{D}_{\alpha_\theta}$ and generating $X_1, \dots, X_n$ independently from F. We then condition on the event $\{\#(\mathbf{X}) = n\}$. At this stage, these X's are i.i.d. from H_θ [if $F \sim \mathcal{D}_\alpha$ where α is nonatomic and $X_1, \dots, X_n$ are i.i.d. $\sim F$, then, conditional on $\#(\mathbf{X})$, the distinct X's are i.i.d. from α normalized to be a probability measure; this is a well-known fact, and can be seen from Sethuraman's construction, for example]. We now further condition on the event $X_i \in A_i$, $i = 1, \dots, n$, and it is clear that the distribution we obtain is that of n independent random variables with distributions H_{θ, A_i}. Thus, (4.15) is rewritten as $(H_{\theta, A_i} \times \cdots \times H_{\theta, A_n})(S) = 0$ for $[\nu]$-a.e. θ, and therefore

$$\int (H_{\theta, A_i} \times \cdots \times H_{\theta, A_n})(S)\nu(d\theta) = 0.$$

This is the assertion that $\tau(S) = 0$, and so the proposition is proved. □

We will now see that convergence of the distributions of the triples follows easily from the convergence of the distributions of the pairs.

THEOREM 4. *If the conditions of Theorem* 3 *and Proposition* 1 *are satisfied, then* (4.1) *holds.*

Our main tool is the following lemma.

LEMMA 1. *Let μ_k be a sequence of measures and let μ be a measure on the measurable space $(\Omega, \mathcal{F})$ such that*

$$\sup_{C \in \mathcal{F}} |\mu_k(C) - \mu(C)| \to 0. \tag{4.16}$$

Let Φ be the collection of all measurable real-valued nonnegative functions which are bounded above by 1. *Then*

$$\sup_{\phi \in \Phi} \left| \int \phi \, d\mu_k - \int \phi \, d\mu \right| \to 0. \tag{4.17}$$

This lemma is well known and, in fact, many authors prefer (4.17) to (4.16) as the definition of convergence in total variation norm. We omit the straightforward proof.

PROOF OF THEOREM 4. When the conditions of Theorem 3 and Proposition 1 are satisfied, then, if the starting point $\mathbf{X}^{(0)}$ is chosen according to τ, the distribution of the Markov chain $(\theta^{(1)}, \mathbf{X}^{(1)}), (\theta^{(2)}, \mathbf{X}^{(2)}), (\theta^{(3)}, \mathbf{X}^{(3)}), \dots$ converges in total variation norm to $\mathcal{L}((\theta, \mathbf{X}) \mid \text{data})$. For the purpose of proving (4.1), it is more convenient to work with the chain $(\mathbf{X}^{(0)}, \theta^{(1)}), (\mathbf{X}^{(1)}, \theta^{(2)}), (\mathbf{X}^{(2)}, \theta^{(3)})$, and we show below the fairly intuitive fact that this chain converges in total variation norm as well. More precisely, let $\tilde{\mu}^k_{\mathbf{X}^{(0)}}(\cdot)$ be the marginal distribution of $(\mathbf{X}^{(k-1)}, \theta^{(k)})$ when the algorithm is started at $\mathbf{X}^{(0)}$ and let $\tilde{\mu}$ be the conditional distribution of $(\mathbf{X}, \theta)$ given the data. Then

$$\sup_{B \in \mathcal{B}_{R^k} \times \mathcal{B}_\Theta} \left| \tilde{\mu}^k_{\mathbf{X}^{(0)}}(B) - \tilde{\mu}(B) \right| \to 0 \quad [\tau]\text{-a.e.}, \tag{4.18}$$

a statement analogous to (4.2). To see (4.18) let $B \in \mathcal{B}_{R^k} \times \mathcal{B}_\Theta$. Then

$$P_{\mathbf{X}^{(0)}}\left\{ (\mathbf{X}^{(k-1)}, \theta^{(k)}) \in B \right\} = \int P_{\mathbf{X}^{(0)}}\left\{ (\mathbf{X}^{(k-1)}, \theta^{(k)}) \in B \mid \mathbf{X}^{(k-1)} \right\} \times P_{\mathbf{X}^{(0)}}\left(d\mathbf{X}^{(k-1)}\right). \tag{4.19}$$

Theorem 3 implies in particular that for $[\tau]$-a.e. $\mathbf{X}^{(0)}$, the conditional distribution of $\mathbf{X}^{(k-1)}$ given $\mathbf{X}^{(0)}$ converges in total variation norm to $\mathcal{L}(\mathbf{X}\,|\,\text{data})$. Let $\phi_B(\mathbf{x})$ denote $P_{\mathbf{X}^{(0)}}\{(\mathbf{X}^{(k-1)}, \theta^{(k)}) \in B\,|\,\mathbf{X}^{(k-1)} = \mathbf{x}\}$. This is clearly nonnegative and bounded above by 1, and so by Lemma 1 the right side of (4.19) converges to $\int \phi_B(\mathbf{x})\mu^{\mathbf{X}}(d\mathbf{x}|\text{data})$, which we recognize as $\tilde{\mu}(B)$.

To prove (4.1), we argue in a similar way. For $B \in \mathcal{B}_{R^k} \times \mathcal{B}_\Theta \times \mathcal{B}_\mathcal{P}$ we have

$$\begin{aligned} &\left|P_{\mathbf{X}^{(0)}}\left\{\left(\mathbf{X}^{(k-1)}, \theta^{(k)}, F^{(k)}\right) \in B\right\} - P\{(\mathbf{X}, \theta, F) \in B\,|\,\text{data}\}\right| \\ (4.20) \quad &= \left|\int P_{\mathbf{X}^{(0)}}\left\{\left(\mathbf{X}^{(k-1)}, \theta^{(k)}, F^{(k)}\right) \in B\,\middle|\,\mathbf{X}^{(k-1)}, \theta^{(k)}\right\}\tilde{\mu}^k_{\mathbf{X}^{(0)}}\left(d\mathbf{X}^{(k-1)}, d\theta^{(k)}\right)\right. \\ &\qquad \left. - \int P\{(\mathbf{X}, \theta, F) \in B\,|\,\mathbf{X}, \theta\}\mathcal{L}(d\mathbf{X}, d\theta)\right|. \end{aligned}$$

Letting $\phi_B(\mathbf{x}, \theta)$ denote $P_{\mathbf{X}^{(0)}}\{(\mathbf{X}^{k-1)}, \theta^{(k)}, F^{(k)}) \in B|\mathbf{X}^{(k-1)} = \mathbf{x}, \theta^{(k)} = \theta\}$, we see that the expression to the right of the equals sign in (4.20) is

$$(4.21) \qquad \left|\int \phi_B(\mathbf{x}, \theta)\tilde{\mu}^k_{\mathbf{X}^{(0)}}(d\mathbf{x}, d\theta) - \int \phi_B(\mathbf{x}, \theta)\tilde{\mu}(d\mathbf{x}, d\theta)\right|.$$

Note that the same ϕ_B appears inside the two integrals in (4.21). Since $0 \le \phi_B(\mathbf{x}, \theta) \le 1$, (4.18) and Lemma 1 together imply (4.1). □

5. Remarks on the consistency of the posterior distribution of F given the data. We begin by defining the notion of consistency of the posterior distribution of F in the present context. We assume that $\{A_i\}$ is an infinite sequence of sets that is fixed in advance of the experiment. Let data(n) denote the data at time n. We say that the posterior distribution of F is *consistent* at $F^{(0)}$ if for $X_1, X_2, \ldots$ i.i.d. $\sim F^{(0)}$, we have, with probability 1, $\mathcal{L}(F\,|\,\text{data}(n))$ converges in distribution to the point mass at $F^{(0)}$. The convergence is in the weak topology on $\mathcal{P}$ [this is the topology generated by the σ-field $\mathcal{B}_\mathcal{P}$ defined in the paragraph following (2.2)]. We shall say that the posterior is *consistent* if it is consistent for every $F^{(0)} \in \mathcal{P}$. This notion of consistency refers to the posterior and not to estimators.

Diaconis and Freedman (1986a, b) and Doss (1985a, b) studied the question of consistency in the following setting. We observe data $X_1, \ldots, X_n$ i.i.d. $\sim K_\theta$, where $K_\theta(x) = K(x - \theta)$, and both θ and K are unknown. A prior is put on the pair (K, θ) by taking $\theta \sim \nu$, $K \sim \mathcal{D}_\alpha$, and taking θ and K be independent. Here, ν is an arbitrary probability measure on R. In this situation the marginal posterior distribution of θ, $\nu(\cdot\,|\,X_1, \ldots, X_n)$, can be obtained and Diaconis and Freedman (1986b) and Doss (1985b) show that this posterior can be inconsistent in the following sense: there exists a pair $(K^{(0)}, \theta^{(0)}) \in \mathcal{P} \times R$ such that if $X_1, X_2, \ldots$ are i.i.d. $\sim K^{(0)}_{\theta^{(0)}}$, then $\nu(\cdot\,|X_1, \ldots, X_n)$ can fail to converge to the point mass at $\theta^{(0)}$ with probability 1.

On the surface it appears that these results show that in our case also the posterior is inconsistent, since our situation generalizes theirs in two ways: the parametric family K_θ that we consider is not restricted to be a location family, and we do not necessarily observe all the X_i's. However, there is an important difference between our setup and theirs. We are interested in consistency of $\mathcal{L}(F \mid \text{data}(n))$ [i.e., $\mathcal{L}(K_\theta \mid \text{data}(n))$] and this may happen even if the marginal posterior distribution of θ is inconsistent. Diaconis and Freedman [(1983), Section 5], show that in the setup where we observe complete data $X_1, \dots, X_n \sim F^{(0)}$, the posterior is consistent if $\alpha_\theta(R)$ is bounded. This leads us to believe that it should be possible to establish consistency of the posterior for at least some models that involve censoring.

Acknowledgments. I am very grateful to B. Narasimhan and F. Huffer for their help in the preparation of this paper, and to G. Li for help with the computer programs in an early draft of the paper.

REFERENCES

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

ATHREYA, K. B., DOSS, H. and SETHURAMAN, J. (1992). On the convergence of the Markov chain simulation method. Preprint.

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

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

DIACONIS, P. and FREEDMAN, D. A. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68–87.

DOSS, H. (1985a). Bayesian nonparametric estimation of the median. I. Computation of the estimates. *Ann. Statist.* **13** 1432–1444.

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

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

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

GELFAND, A. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.* **85** 398–409.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

HOLLANDER, M. and PEÑA, E. (1992). A chi-square goodness-of-fit test for randomly censored data. *J. Amer. Statist. Assoc.* **87** 458–463.

HOLLANDER, M. and PROSCHAN, F. (1979). Testing to determine the underlying distribution using randomly censored data. *Biometrics* **35** 393–401.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

KOZIOL, J. A. and GREEN, S. B. (1976). A Cramér–von Mises statistic for randomly censored data. *Biometrika* **63** 465–474.

LINK, C. (1984). Confidence intervals for the survival function using Cox's proportional-hazard model with covariates. *Biometrics* **40** 601–609.

MILLER, R. G., JR. (1983). What price Kaplan–Meier? *Biometrics* **39** 1077–1081.

RAMLAU-HANSEN, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* 11 453–466.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650.

SUSARLA, V. and VAN RYZIN, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71** 897–902.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762.

TURNBULL, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* **69** 169–173.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored, and truncated data. *J. Roy. Statist. Soc. Ser. B* **38** 290–295.

DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
1958 NEIL AVENUE
COLUMBUS, OHIO 43210