BAYESIAN NONPARAMETRIC ESTIMATION VIA GIBBS SAMPLING FOR COHERENT SYSTEMS WITH REDUNDANCY[1]

BY Hani Doss, Fred W. Huffer and Kevin L. Lawson

*Ohio State University, Florida State University
and Abbott Laboratories*

We consider a coherent system S consisting of m independent components for which we do not know the distributions of the components' lifelengths. If we know the structure function of the system, then we can estimate the distribution of the system lifelength by estimating the distributions of the lifelengths of the individual components. Suppose that we can collect data under the "autopsy model," wherein a system is run until a failure occurs and then the status (functioning or dead) of each component is obtained. This test is repeated n times. The autopsy statistics consist of the age of the system at the time of breakdown and the set of parts that are dead by the time of breakdown. We develop a nonparametric Bayesian estimate of the distributions of the component lifelengths and then use this to obtain an estimate of the distribution of the lifelength of the system. The procedure is applicable to machine-test settings wherein the machines have redundant designs. A parametric procedure is also given.

1. Introduction and summary. Consider a coherent system S consisting of m independent components for which we do not know the distributions of the component lifelengths. Assume that each of the m components occupies one of two states, functioning or failed. We consider the statistical model in which each element of a sample of n replicates of S is observed until it fails. The observed data consist of the set of components that are in a failed state and the failure time of the system. The failure times of the dead components are not directly observed. The set of dead components and the system failure time comprise the "autopsy statistics" of the system. This model is usually called the autopsy model.

Two statistical problems arise in considering the autopsy model—the problems of estimating the distributions of the component lifelengths and the distribution of the entire system's lifelength. One approach to estimating the distribution of the system lifelength is to use only the observed system failure times. For example, we could use the empirical distribution function. However, such an approach ignores the (partial) information we have about the components of the system. If the structure function of the system is

known, the distribution of the lifelength of the system can, in general, be calculated from knowledge of the distributions of the component lifelengths. Hence, an alternative approach would be to estimate the distributions of the lifelengths of the m components of system S and then use the structure function of S to estimate the distribution of the system lifelength.

Clearly, the component information provided by the autopsy model is quite limited. It is reasonable to consider alternative testing procedures which provide more component information. However, the autopsy model is important when alternative testing procedures such as separate testing of components are not possible or practical. For example, it may be difficult to reproduce the conditions which exist in the functioning system when components are tested separately. For such systems, it is important to obtain every bit of information about the components when they are parts of a machine or other system. In these settings, as well as the case of certain biological systems, the autopsy model is natural. Probabilistic aspects of the autopsy model were considered by Meilijson (1981), Nowick (1990) and Antoine, Doss and Hollander (1993). Inferential aspects of the autopsy model have been considered only by Watelet (1990) and Meilijson (1994).

The U.S. Air Force's C-17 transport airplane's fuel quantity (FQ) computer is an example of a system in which the "component" is a logical subsystem whose status can be readily determined in the field. The FQ computer is a parallel system of order 2, consisting of an "A" bus and a "B" bus. When this system's data were first examined by the authors, there were approximately 2440 cumulative flying hours spread among six different prototype C-17's, each of which contained one FQ computer. We present a data analysis using our procedure in Section 3.

Consider the autopsy statistics as described above. Assume that the system we wish to study is a coherent system [see Barlow and Proschan (1981), Chapters 1 and 2]. We can use the structure function, the set of dead components and the failure time of the system to say more about the failure times of each component. Specifically, for each of the n replicates of S, we can classify the failure time of each of the m components as follows:

C1. Component failure time is greater than system failure time.

C2. Component failure time is less than system failure time.

C3. Component failure time is equal to system failure time.

C4. Component failure time is either less than or exactly equal to system failure time, but we cannot tell which.

The first two categories correspond to what are usually termed right-censored data and left-censored data, respectively. Right-censored data occur when the component is still alive when the system fails. Left-censored data occur when the component is dead and information contained in the structure function, along with information contained in the set of dead components, allows one to deduce that the component's death occurred prior to the system failure time. Similarly, the third category arises when a component is observed in a failed state and we deduce that it caused the system to fail. The

last category occurs whenever all the components in a redundant system or a redundant subsystem belong to the set of dead components. In this case, we know there is exactly one component, whose identity we do not know, with a failure time equal to that of the system, while the failure times of all the other components are strictly less than that of the system.

The last category above is problematic, particularly in frequentist settings, and is at the heart of the issue of "identifiability." (General references on identifiability are given at the end of this section.) For example, consider a parallel system of order 2. Let the distributions for components 1 and 2 be F_1 and F_2, respectively. Both components will always be dead when the system is observed in a failed state. Under the autopsy model, one can only observe the system failure time, which has distribution $F_1 F_2$. This simple redundant system is not identifiable in the frequentist sense in that it is not possible to determine F_1 and F_2 from a knowledge of the distribution of the observed data.

We use a Bayesian framework because, for the applications we have in mind, we have a small amount of data but have extensive past experience on the components in other systems. For example, in the case of the C-17 FQ computer, similar but not identical FQ computers exist in other aircraft for which there has been more extensive testing. Thus, we have strong reasons to suspect that a certain parametric family provides a good approximation to the true probability model, and we have some knowledge about the unknown parameters. A Bayes procedure makes sense here since we have only limited testing hours on just six FQ systems, each within a C-17, and we wish to estimate the distribution of the lifelength of the FQ system when it is part of the C-17.

We consider a Bayesian framework for estimation of the distributions of component lifelengths in which the prior distributions on each of the F_i's give most of their mass to "small neighborhoods of a parametric family." The prior distributions which we use are derived from the Dirichlet process priors discussed by Ferguson (1973, 1974). The Dirichlet process priors are probability measures on $\mathscr{P}$ parametrized by the set of all finite nonnull measures on the real line $\mathscr{R}$, where $\mathscr{P}$ is the space of all probability measures on $\mathscr{R}$. Let α be a finite nonnull measure on the Borel sets of $\mathscr{R}$. The random distribution function F is said to have a Dirichlet process prior distribution with parameter α (denoted $\mathscr{D}_\alpha$) if, for every measurable partition $\{B_1, \ldots, B_l\}$ of $\mathscr{R}$, the random vector $(F(B_1), \ldots, F(B_l))$ has the Dirichlet distribution with parameter vector $(\alpha(B_1), \ldots, \alpha(B_l))$ (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 $\mathscr{P}$, then, for every $t \in \mathscr{R}$, the quantity $F(t)$ is a random variable. Write $H = \alpha/\alpha(\mathscr{R})$, so that H is a probability measure on $\mathscr{R}$. If $F \sim \mathscr{D}_\alpha$, then $EF(t) = H(t)$, while the quantity $\alpha(\mathscr{R})$ indicates the degree of concentration of $\mathscr{D}_\alpha$ around its "center" H. For example, it is well known that as $\alpha(\mathscr{R}) \to \infty$, $\mathscr{D}_\alpha$ converges to the point mass at H in the weak topology. Ferguson (1973)

showed that the Dirichlet priors have the property that the support of $\mathscr{D}_\alpha$ is the set of all probability measures whose support is contained in the support of H. For example, if the support of H is the positive real axis, then the support of $\mathscr{D}_\alpha$ is the set of distributions of all positive random variables. Ferguson also showed that if $F \sim \mathscr{D}_\alpha$, then F is a.s. discrete.

The priors on each of the F_i's, $i = 1, \ldots, m$, that we use are mixtures of Dirichlet priors. To keep the notation less cumbersome, let $m = 2$. For component 1, we consider a parametric family H_θ, $\theta \in \Theta$, and put a mixture of Dirichlets as the prior on F_1. That is,

$$F_1 \sim \int \mathscr{D}_{\alpha_\theta} \nu(d\theta),$$

where for each $\theta \in \Theta$, $\alpha_\theta = \alpha_\theta(\mathscr{R})H_\theta$, $0 < \alpha_\theta(\mathscr{R}) < \infty$ and ν is a probability measure on Θ. Similarly, for component 2 we consider the parametric family K_ψ, $\psi \in \Psi$, and put as prior on F_2 the mixture $\int \mathscr{D}_{\beta_\psi} \mu(d\psi)$. For each $\psi \in \Psi$, $\beta_\psi = \beta_\psi(\mathscr{R})K_\psi$, $0 < \beta_\psi(\mathscr{R}) < \infty$, and μ is a probability measure on Ψ.

Our nonparametric Bayesian procedure, which uses mixtures of Dirichlets as priors on the F_i's, has two advantages. First, we protect against the problems associated with using an incorrectly specified parametric model, such as obtaining an inconsistent estimator. Second, we can avoid the loss of efficiency due to ignoring partial information we may have about a parametric model, since we use prior distributions (on each of the F_i's) that concentrate their mass around the hypothesized parametric family.

Our approach is based on the Gibbs sampling algorithm as discussed in Gelfand and Smith (1990). We now review the algorithm. Let $f_{Y_1, \ldots, Y_p}$ be the joint distribution of the (possibly vector-valued) random variables $Y_1, \ldots, Y_p$. We suppose that we do not know the form of $f_{Y_1, \ldots, Y_p}$, but that we know the conditional distributions $f_{Y_i | Y_j, j \neq i}$, $i = 1, \ldots, p$, or that at least we are able to generate observations from these conditional distributions. Suppose we want to sample observations from the joint distribution of the random variables $Y_1, \ldots, Y_p$, or simply an observation from one of the p marginals. The algorithm to generate an observation from $f_{Y_1, \ldots, Y_p}$ proceeds as follows. We fix arbitrary starting values $Y_1^{(0)}, \ldots, Y_p^{(0)}$ and then update these values. Draw $Y_1^{(1)}$ from $f_{Y_1 | Y_j, j \neq 1}(\cdot, Y_2^{(0)}, \ldots, Y_p^{(0)})$. Next, draw $Y_2^{(1)}$ from $f_{Y_2 | Y_j, j \neq 2}(Y_1^{(1)}, \cdot, Y_3^{(0)}, \ldots, Y_p^{(0)})$. Continue until we draw $Y_p^{(1)}$ from $f_{Y_p | Y_j, j \neq p}(Y_1^{(1)}, \ldots, Y_{p-1}^{(1)}, \cdot)$. We have now completed one iteration of the scheme by visiting each variable. After k iterations, we have the random variables $(Y_1^{(k)}, \ldots, Y_p^{(k)})$. The sequence $(Y_1^{(j)}, \ldots, Y_p^{(j)})$, $j = 1, 2, \ldots$, is a Markov chain and $f_{Y_1, \ldots, Y_p}$ is a stationary distribution of the chain. If one can establish that the chain converges in distribution to $f_{Y_1, \ldots, Y_p}$, then (for large k) $(Y_1^{(k)}, \ldots, Y_p^{(k)})$ has a distribution which is approximately equal to $f_{Y_1, \ldots, Y_p}$. Such observations can be used to estimate $f_{Y_1, \ldots, Y_p}$.

Perhaps the most natural way to implement the Gibbs sampler here is to proceed as is normally done in a Bayesian analysis of missing data problems under conjugacy. That is, consider the pair (parameter θ, `missing data`): in such a setup, if we knew the missing data, we would easily be able to find the

conditional distribution of the parameter θ, and if we knew the parameter θ, we would be able to generate the missing data [see, for instance, the linkage example or the Dirichlet sampling process example in Tanner and Wong (1987)]. Indeed, this is precisely the approach taken by Doss (1994), who considers the use of Dirichlet priors for the problem of estimating an unknown distribution F in the presence of censoring. He considers random variables $X_1, \ldots, X_n \sim_{\text{iid}} F$, but only observes $X_i \in A_i$, where 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. His approach is based on a Gibbs sampling algorithm of length 2 involving $(\mathbf{X}, F)$, where $\mathbf{X} = (X_1, \ldots, X_n)$; that is, he generates $\mathbf{X}$ from $\mathscr{L}_{\text{data}}(\mathbf{X}|F)$ and F from $\mathscr{L}_{\text{data}}(F|\mathbf{X})$, where data consists of the sets $A_1, \ldots, A_n$. [Note that the subscript "data" denotes conditioning on the data; thus, $\mathscr{L}_{\text{data}}(F|\mathbf{X})$ denotes the conditional distribution of F given both $\mathbf{X}$ and the data.] The details of carrying out the steps of the algorithm rely on a constructive definition of the Dirichlet prior, given in Sethuraman (1994).

Our initial approach was simply to extend the technique of Doss (1994) to our setting; however, the approach fails since, as we show in Section 2.2, the procedure produces a Markov chain that is *reducible* and therefore does not converge to the posterior distribution. We present an entirely different algorithm that produces a Markov chain which we show to converge to the posterior.

We now give a preliminary explanation of our procedure for the case where the priors on each of the F_j's are single Dirichlets; that is, $F_j \sim \mathscr{D}_{\alpha_j}$. Let S_i, $i = 1, \ldots, n$, be the vector of lifelengths of the m components (for system i) *if we could see them all*. Let data be the set of autopsy statistics for the n systems. (It may be helpful to think of the case of parallel systems. In this case, data consists simply of the n system failure times.) The algorithm proceeds as follows. Fix arbitrary starting values $S_1^{(0)}, \ldots, S_n^{(0)}$. Generate $S_1^{(1)} \sim \mathscr{L}_{\text{data}}(S_1 \mid S_2^{(0)}, \ldots, S_n^{(0)})$. Next, generate $S_2^{(1)} \sim \mathscr{L}_{\text{data}}(S_2 \mid S_1^{(1)}, S_3^{(0)}, \ldots, S_n^{(0)})$. Continue until we generate $S_n^{(1)} \sim \mathscr{L}_{\text{data}}(S_n \mid S_1^{(1)}, \ldots, S_{n-1}^{(1)})$. We have now completed one iteration of the procedure. We repeat the procedure a large number of times and use the realizations of the chain to estimate $\mathscr{L}_{\text{data}}(S_1, \ldots, S_n)$. There are two key points that allow this procedure to produce a Markov chain which converges to the posterior distribution:

1. The conditional distribution of the lifelength of the jth component of system i depends only on the autopsy statistics for system i and on the current set of lifelengths for the jth components of S_l, $l \neq i$.
2. The joint unconditional distribution of the n lifelengths of component j can be described in full.

The estimate of $\mathscr{L}_{\text{data}}(S_1, \ldots, S_n)$ can be used to obtain an estimate of $\mathscr{L}_{\text{data}}(F_1, \ldots, F_m)$.

Note that, in contrast to Doss (1994), we deal only with the random lifelengths (of the components) and bypass entirely the problem of generating the infinite-dimensional F_j's.

In Section 2, we explore the algorithm further. In Section 2.1 we give a Gibbs sampling algorithm for our problem under parametric assumptions. In Section 2.2, we present a simple extension of the procedure of Doss (1994) to our setting and explain why this leads to a reducible Markov chain. In Section 2.3 we describe our basic Gibbs sampling algorithm, discuss a modification of it to speed up the rate of convergence and provide a rigorous proof of convergence for the case of parallel systems of order 2. In Section 2.4 we indicate the modifications necessary when dealing with an arbitrary coherent system. In Section 3 we illustrate our nonparametric procedure on data pertaining to the C-17 FQ computer.

The last portion of this section is used to summarize other relevant research in the statistical literature. First, we define the term identifiability. Suppose system S has m components and the component distributions are denoted $F_1, \ldots, F_m$. If $F_1, \ldots, F_m$ can be recovered from knowledge of the true distributions of the autopsy statistics, then we say that $F_1, \ldots, F_m$ are identifiable and that S is an identifiable system. For example, any parallel system is nonidentifiable. Meilijson (1981), Nowick (1990) and Antoine, Doss and Hollander (1993) considered probabilistic aspects of the autopsy model. In these three papers, the authors identify conditions on the structure function of the system and on the distributions of the component lifelengths that guarantee identifiability.

Relevant work on estimating the distribution of the system lifelength under the autopsy model by first estimating the distributions of the component lifelengths and then using the structure function of the system can be found in Watelet (1990) and Meilijson (1994). Watelet (1990) considered two estimators for the autopsy model. He developed nonparametric estimators for the F_j's. Meilijson (1994) estimated the parameters of these distributions from the empirical estimate of $\mathscr{L}(Z, D)$, where Z is the lifetime of the machine and D is the diagnostic set of parts that had died by time Z, by maximum likelihood from incomplete data via the EM algorithm. Meilijson assumes the F_j's are drawn from well-behaved parametric families.

Under the assumption that the failure times of the dead components are known, Doss, Freitag and Proschan (1989) also considered inferential aspects of estimating the distribution of the lifelength of the system by first estimating the distribution of the lifelengths of the system's components. In their model, they have more information than is available under the autopsy model. We also mention Arjas and Gasbarra (1994), who consider a Bayesian nonparametric approach to estimation in the random right-censorship model of survival analysis. They put a prior on the hazard rate and develop a Gibbs sampling algorithm for the numerical calculations.

2. Development and convergence of the algorithm. In order to keep the notation as light as possible Sections 2.1 and 2.3 deal only with parallel systems of order 2 (for which we have n replicates). It will be clear from the remarks in Section 2.4.1 that knowledge of how to deal with a parallel "module" makes it possible to handle an arbitrary coherent system.

2.1. *A Gibbs sampling algorithm for the autopsy model under parametric assumptions.* For the ith replicate, $i = 1, \ldots, n$, let the observed data (system failure time) be m_i, the maximum of the two component lifelengths, X_i and Y_i. Let $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} = (X_1, \ldots, X_n)$, $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Let $X_i \sim_{\text{iid}} F_\theta$ with $\theta \in \Theta$ and suppose $\theta \sim \nu$, where ν is a conjugate prior distribution on θ. Let $Y_i \sim G_\psi$ with $\psi \in \Psi$ and suppose $\psi \sim \mu$, where μ is a conjugate prior distribution on ψ. Assume that F_θ and G_ψ are absolutely continuous distribution functions. To make the discussion as easy as possible to follow, consider the case where $F_\theta = \mathscr{E}(\theta)$ (the exponential distribution with parameter θ) with $\theta \sim \nu = \mathscr{G}(a_1, b_1)$ (the gamma distribution with shape parameter a_1 and scale parameter b_1) and $G_\psi = \mathscr{E}(\psi)$ with $\psi \sim \mu = \mathscr{G}(a_2, b_2)$.

Before describing the algorithm, we introduce the following notation. If H is a distribution function, B is a set, and X is distributed according to H, then H_B will denote the conditional distribution function of X given that $X \in B$; that is,

$$(2.1) \qquad H_B(A) = H(A \cap B)/H(B)$$

when $H(B) > 0$.

The algorithm proceeds as follows.

Give arbitrary initial values to $(X, Y)_i^{(0)}$ such that $X_i^{(0)} \vee Y_i^{(0)} = m_i$. For $k = 1, \ldots, K$:

1. Generate $(\theta, \psi)^{(k)} \sim \mathscr{L}_{\text{data}}((\theta, \psi) \mid (\mathbf{X}, \mathbf{Y})^{(k-1)})$.
2. Generate $(X, Y)_i^{(k)} \sim \mathscr{L}_{\text{data}}((X, Y)_i \mid (\theta, \psi)^{(k)})$, independently for each i, $i = 1, \ldots, n$.

We now describe these two steps in more detail.

In step 1 of the algorithm, $\mathscr{L}_{\text{data}}((\theta, \psi) \mid \mathbf{Z}^{(k-1)})$ is the product of two gamma distributions, $\mathscr{G}(a_1 + n, b_1 + \sum_{i=1}^n X_i^{(k-1)})$ and $\mathscr{G}(a_2 + n, b_2 + \sum_{i=1}^n Y_i^{(k-1)})$; that is, θ and ψ are generated independently. (Note that knowledge of $\mathbf{X}$ and $\mathbf{Y}$ makes knowledge of the observed data superfluous.) To carry out step 2, first let f_θ and g_ψ be the densities of F_θ and G_ψ, respectively. Then, for each i, $i = 1, \ldots, n$, set

$$p_i^{(k)} = \frac{f_{\theta^{(k)}}(m_i)/F_{\theta^{(k)}}(m_i)}{f_{\theta^{(k)}}(m_i)/F_{\theta^{(k)}}(m_i) + g_{\psi^{(k)}}(m_i)/G_{\psi^{(k)}}(m_i)},$$

where $F_\lambda = G_\lambda = \mathscr{E}(\lambda)$. Now, generate $\mathbf{Z}^{(k)}$ by setting $(X, Y)_i^{(k)}$, $i = 1, \ldots, n$, as follows:

$$(X, Y)_i^{(k)} = \begin{cases} (m_i, V), \text{ where } V \sim G_{\psi^{(k)}, [0, m_i)}, & \text{with probability } p_i^{(k)}, \\ (V, m_i), \text{ where } V \sim F_{\theta^{(k)}, [0, m_i)}, & \text{with probability } \left(1 - p_i^{(k)}\right), \end{cases}$$

where the notation used for the conditional distribution functions is given by (2.1).

To estimate $\mathscr{L}_{\text{data}}((\theta, \psi))$, we use the sequence of generated $\mathbf{Z}^{(k)}$'s to approximate the mixture

$$\int \mathscr{L}_{\text{data}}((\theta, \psi) \mid \mathbf{Z}) \, d\mathscr{L}_{\text{data}}(\mathbf{Z})$$

by, say, $(1/K)\sum_{j=1}^{K} \mathscr{L}_{\text{data}}((\theta, \psi) \mid \mathbf{Z}^{(j)})$.

If ν and μ are not conjugate priors for F_θ and G_ψ, respectively, but they are continuous distribution functions with univariate log-concave density functions, then we can still apply the above algorithm by using an efficient rejection scheme of Gilks and Wild (1992).

2.2. *A naïve extension of the algorithm of Doss* (1994). Doss (1994) uses a Gibbs sampling algorithm of length 2 involving $(\mathbf{X}, F)$. The prior he puts on F is actually a mixture of Dirichlets, but we shall consider a single Dirichlet prior for our naïve extension, since if the procedure fails for a single Dirichlet, it will fail a fortiori for a mixture.

Suppose we have n replicates of a parallel system of order 2. Let $\mathbf{Z}, \mathbf{X}$ and $\mathbf{Y}$ be defined as in Section 2.1. Suppose $X_i \sim_{\text{iid}} F$ and $Y_i \sim_{\text{iid}} G$, $i = 1, \ldots, n$. Suppose $F \sim \mathscr{D}_\alpha$ and $G \sim \mathscr{D}_\beta$, where $\mathscr{D}_\alpha$ and $\mathscr{D}_\beta$ are Dirichlet priors with parameter measures α and β. We observe $\texttt{data} = (m_1, \ldots, m_n)$, where $m_i = X_i \vee Y_i$, $i = 1, \ldots, n$. Our goal is to estimate $\mathscr{L}_{\text{data}}(F, G)$.

We know that the conditional distribution of (F, G) given $(\mathbf{X}, \mathbf{Y})$ is

$$(2.2) \qquad \mathscr{L}_{\text{data}}((F, G) \mid \mathbf{Z}) = \mathscr{D}_{\alpha + \sum_{i=1}^{n} \delta_{X_i}} \otimes \mathscr{D}_{\beta + \sum_{i=1}^{n} \delta_{Y_i}},$$

where $\otimes$ denotes product measure; that is, F and G are independent. Also, given an updated (F, G), we can generate a random $\mathbf{Z}$ conditional on the data. To run the Gibbs sampling algorithm, we first set initial values $\mathbf{Z}^{(0)}$ and $(F, G)^{(0)}$. Then, for some large K, execute the following loop for $k = 1, \ldots, K$:

1. Draw $(F, G)^{(k)}$ from $\mathscr{L}_{\text{data}}((F, G) \mid \mathbf{Z}^{(k-1)})$.
2. Draw $\mathbf{Z}^{(k)}$ from $\mathscr{L}_{\text{data}}(\mathbf{Z} \mid (F, G)^{(k)})$.

To carry out the second step in the above loop, we need to compute the probability that each component's lifelength takes on the observed maximum, since one of the values must. Consider the case where $n = 1$. Let the observed maximum be denoted by m. We assume that $m > 0$ and that α and β are continuous measures. Suppose the initial value of $\mathbf{Z}^{(0)}$ is $(X_1, Y_1)^{(0)} = (m, V)$, for some $V < m$. These initial values give rise to $(F, G)^{(1)}$ via the first step of the algorithm. By (2.2), $F^{(1)} \sim \mathscr{D}_{\alpha + \delta_m}$, but by definition of the Dirichlet prior, this implies that $F^{(1)}(\{m\})$ is distributed as a beta distribution with parameters $(\alpha + \delta_m)(\{m\})$ and $(\alpha + \delta_m)(\{m\}^c)$, and therefore $F^{(1)}(\{m\}) > 0$ with probability 1. Next, recall from the previous section that if $F \sim \mathscr{D}_\gamma$, then $EF(A) = \gamma(A)/\gamma(\mathscr{R})$ for any Borel set A. Since $G^{(1)} \sim \mathscr{D}_{\beta + \delta_V}$, we have

$$EG^{(1)}(\{m\}) = \frac{\beta + \delta_V}{\beta(\mathscr{R}) + 1}(\{m\}) = 0.$$

Thus, $G^{(1)}(\{m\}) = 0$ with probability 1. In other words, with probability 1, $F^{(1)}$ has an atom at m and $G^{(1)}$ does not, from which it is clear that

$$P\{X_1^{(1)} = m \mid X_1^{(1)} \vee Y_1^{(1)} = m\} = 1.$$

As we continue to run the algorithm, $X_1^{(2)}, X_1^{(3)}, \ldots$ will each take on the value m. The case where $(X_1, Y_1)^{(0)} = (V, m)$ is handled by symmetry. Thus, we see that the starting point does not get "washed out" and therefore the algorithm produces a Markov chain that cannot converge to the posterior distribution.

2.3. *The algorithms we propose*: *detailed description and convergence for parallel systems of order* 2. Before introducing the setup and notation needed for the algorithm, we briefly describe the "extended Pólya urn scheme" as given in Blackwell and MacQueen (1973). Define a sequence $\{T_1, T_2, \ldots\}$ of random variables as a *Pólya sequence with parameter* α if, for every $B \subset \mathscr{R}$, we have $P(T_1 \in B) = \alpha(B)/\alpha(\mathscr{R})$, and for every n,

$$(2.3) \qquad P(T_{n+1} \in B \mid T_1, \ldots, T_n) = \frac{\alpha(B) + \sum_{i=1}^n \delta_{T_i}(B)}{\alpha(\mathscr{R}) + n}.$$

Blackwell and MacQueen proved that if $\{T_1, T_2, \ldots\}$ is a Pólya sequence with parameter α, then the empirical distribution of $\{T_1, \ldots, T_n\}$ converges a.s. to a limiting discrete measure H. Furthermore, $H \sim \mathscr{D}_\alpha$. Also, given H, the random variables $T_1, T_2, \ldots$ are iid $\sim H$. In addition, for every n,

$$(2.4) \qquad\qquad T_1, \ldots, T_n \text{ are exchangeable.}$$

Recall that X_i and Y_i are the lifelengths of components 1 and 2 in system i. For the X_i's, we consider a parametric family H_θ, $\theta \in \Theta \subset \mathscr{R}^{d_1}$, and put a mixture of Dirichlets as the prior on F. That is,

$$(2.5) \qquad\qquad F \sim \int \mathscr{D}_{\alpha_\theta} \nu(d\theta),$$

where, for each $\theta \in \Theta$, $\alpha_\theta = \alpha_\theta(\mathscr{R})H_\theta$, $0 < \alpha_\theta(\mathscr{R}) < \infty$ and ν is a probability measure on Θ. Similarly, for the Y_i's, we consider the parametric family K_ψ, $\psi \in \Psi \subset \mathscr{R}^{d_2}$, and put the mixture $\int \mathscr{D}_{\beta_\psi} \mu(d\psi)$ as the prior on G, where for each $\psi \in \Psi$, $\beta_\psi = \beta_\psi(\mathscr{R})K_\psi$, $0 < \beta_\psi(\mathscr{R}) < \infty$ and μ is a probability measure on Ψ. We will assume that, for each θ and ψ, H_θ and K_ψ are absolutely continuous, with continuous densities. Given F and G, $X_1, X_2, \ldots$, are iid $\sim F$, and $Y_1, Y_2, \ldots$, are iid $\sim G$. Also, for every n, $X_1, \ldots, X_n$ are exchangeable, as are $Y_1, \ldots, Y_n$, by (2.4). We assume that F is independent of G. It follows that $(X_1, \ldots, X_n)$ is independent of $(Y_1, \ldots, Y_n)$. We observe

$$\texttt{data} = (m_1, \ldots, m_n), \quad \text{where } X_i \vee Y_i = m_i, i = 1, \ldots, n.$$

Our goal is to estimate $\mathscr{L}_{\text{data}}(F, G)$. As in Section 2.2, knowledge of $(\mathbf{X}, \mathbf{Y})$ makes knowledge of $\texttt{data}$ superfluous.

2.3.1. *The basic Gibbs sampling algorithm.* Recall that S_i, $i = 1, \ldots, n$, is the vector of lifelengths for system i. For the case of a parallel system of order 2, $S_i = (X_i, Y_i)$. To unify the notation, let $S_0 = (\theta, \psi)$ and let $\mathbf{S} = (S_0, \ldots, S_n)$.

The algorithm proceeds as follows. Fix arbitrary starting values $S_0^{(0)}, \ldots, S_n^{(0)}$. Then cycle through the $N = n + 1$ elements of $\mathbf{S}$ in order; that is, at time t, we update element $K = K(t) = (t - 1) \bmod N$ by generating

$$(2.6) \qquad S_K^{(t)} \sim \mathscr{L}_{\text{data}}\left(S_K \mid S_j^{(t-1)} \text{ for } j \neq K\right).$$

At time jN, each component of $\mathbf{S}$ has been updated j times. Our algorithm generates vectors $\mathbf{S}^{(t)}$, $t = 1, 2, \ldots$, where $\mathbf{S}^{(t)}$ is the same as $\mathbf{S}^{(t-1)}$ except for the one element $S_K^{(t)}$ which has been updated at time t according to (2.6). Note that this notation differs slightly from that of Section 1, where $\mathbf{S}^{(t)}$ was formed by updating *all* the elements of $\mathbf{S}^{(t-1)}$.

In our algorithm, the updating of $S_i^{(t)}$, $i = 1, \ldots, n$, is accomplished by the use of the following lemma, the proof of which is a calculation.

LEMMA 1. *Suppose A and B are distribution functions on $[0, \infty)$ satisfying $A = A_c + A_d$ and $B = B_c + B_d$, where A_c and B_c are absolutely continuous with continuous derivatives denoted A_c' and B_c', and A_d and B_d are discrete. Let $X^\star$ and $Y^\star$ be independent with distributions A and B, respectively. For $m > 0$, we can generate a pair (X, Y) with*

$$(2.7) \qquad (X, Y) \sim \mathscr{L}\left((X^\star, Y^\star) \mid X^\star \vee Y^\star = m\right)$$

by the mixing procedure we now describe.

Define probability distributions on $[0, m)$ by

$$\bar{A}_c(v) = \frac{A_c(v)}{A_c(m)}, \qquad \bar{B}_c(v) = \frac{B_c(v)}{B_c(m)},$$

$$\bar{A}_d(v) = \frac{A_d(v)}{A_d(m-)}, \qquad \bar{B}_d(v) = \frac{B_d(v)}{B_d(m-)}.$$

If the denominator in any of the equations above is 0, then we set the corresponding probability distribution to δ_0. Let V_1, V_2, V_3, V_4 be random variables with distributions $\bar{A}_c, \bar{B}_c, \bar{A}_d, \bar{B}_d$, respectively. We now take

$$(2.8) \qquad (X, Y) = \begin{cases} (V_1, m), & \text{with probability } p_1, \\ (m, V_2), & \text{with probability } p_2, \\ (V_3, m), & \text{with probability } p_3, \\ (m, V_4), & \text{with probability } p_4, \\ (m, m), & \text{with probability } p_5, \end{cases}$$

where the mixing probabilities $p_1, \ldots, p_5$ are given by the following formulas. If $A_d(\{m\}) + B_d(\{m\}) > 0$, then

$$p_1 = \frac{A_c(m) B_d(\{m\})}{D}, \qquad p_2 = \frac{A_d(\{m\}) B_c(m)}{D},$$

$$p_3 = \frac{A_d(m-) B_d(\{m\})}{D}, \qquad p_4 = \frac{A_d(\{m\}) B_d(m-)}{D}$$

and

$$p_5 = \frac{A_d(\{m\}) B_d(\{m\})}{D},$$

with D such that $\Sigma_{i=1}^5 p_i = 1$. *If* $A_d(\{m\}) = B_d(\{m\}) = 0$, *then*

$$p_1 = \frac{A_c(m) B_c'(m)}{D}, \qquad p_2 = \frac{A_c'(m) B_c(m)}{D}, \qquad p_3 = \frac{A_d(m) B_c'(m)}{D},$$

$$p_4 = \frac{A_c'(m) B_d(m)}{D}, \qquad p_5 = 0,$$

with D such that $\Sigma_{i=1}^5 p_i = 1$.

By (2.3) and (2.4), for the tth step of the algorithm, the conditional distribution of X_i, given the other $(n-1)$ X_l's and θ, is given by

$$(2.9) \qquad A_i^{(t)} = \frac{\alpha_{\theta^{(t)}} + \Sigma_{l \neq i} \delta_{X_l^{(t)}}}{\alpha_{\theta^{(t)}}(\mathscr{R}) + n - 1}.$$

Similarly, for the tth step of the algorithm, the conditional distribution of Y_i, given the other $(n-1)$ Y_l's and ψ, is

$$(2.10) \qquad B_i^{(t)} = \frac{\beta_{\psi^{(t)}} + \Sigma_{l \neq i} \delta_{Y_l^{(t)}}}{\beta_{\psi^{(t)}}(\mathscr{R}) + n - 1}.$$

Thus we can generate $S_i^{(t)} = (X_i^{(t)}, Y_i^{(t)})$ by using Lemma 1 with $A = A_i^{(t)}$, $B = B_i^{(t)}$ and $m = m_i$.

To update $S_0^{(t)}$ in the above algorithm, we will need formulas for updating the "mixing measures" ν and μ. The formula for the conditional distribution of θ given $\mathbf{X}$ is well known. Proposition 1 below is a special case of Lemma 1 of Antoniak (1974). The formula for the conditional distribution of ψ given $\mathbf{Y}$ will be evident by symmetry. The notation $\#(\mathbf{v})$ is used to denote the number of distinct values in the vector $\mathbf{v}$.

PROPOSITION 1.   *Assume that, for each* $\theta \in \Theta$, H_θ *is absolutely continuous, with a density* h_θ *that is continuous on* $\mathscr{R}$. *If the prior of F is given by* (2.5), *then the* (*marginal*) *posterior distribution of* θ *given* $\mathbf{X} = \mathbf{x}$ *is*

$$(2.11) \qquad \nu_{\mathbf{x}}(d\theta) = c(\mathbf{x}) \left(\prod^{\mathrm{dist}} h_\theta(x_i) \right) \frac{(\alpha_\theta(\mathscr{R}))^{\#(\mathbf{x})} \Gamma(\alpha_\theta(\mathscr{R}))}{\Gamma(\alpha_\theta(\mathscr{R}) + n)} \nu(d\theta),$$

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

From (2.11) and the independence of F and G, we can update $S_0^{(t)}$ by independently generating $\theta^{(t)}$ and $\psi^{(t)}$ from $\nu_{\mathbf{x}^{(t)}}$ and $\mu_{\mathbf{y}^{(t)}}$, respectively.

2.3.2. *Modification of the algorithm to speed up the rate of convergence.* Atoms in $A_i^{(t)}$ (or $B_i^{(t)}$) that are not at one of the values m_j are, strictly speaking, transient, but under some conditions it may take many iterations before these disappear completely [for instance, when $\alpha(R)$ is small]. To the extent that they are persistent, this corresponds to slow mixing of the Markov chain produced by our algorithm.

This problem has been encountered by other authors. In a different context, Bush and MacEachern (1996) [see also West, Müller and Escobar (1994)] have shown that it is possible to add to each cycle of the basic Gibbs sampler an extra step that moves the locations of the clusters of observations, in order to improve the rate of convergence. Their ideas can be adapted to our framework.

We now explain how in our model we can append an extra step which moves the clusters in such a way that the posterior distribution of **S** is still a stationary distribution for the new Markov chain.

Let us describe the "group structure" among the X's by the vector $U = (u_1, \ldots, u_n)$ defined by $u_i = \min\{j: X_j = X_i\}$. Let $\mathscr{U}$ be the set of distinct values among $u_1, \ldots, u_n$. (Thus, each element in $\mathscr{U}$ corresponds to a group.) Similarly, define V and $\mathscr{V}$ to be the "groups" and their distinct values for the Y's.

The new chain is described as follows. Given the current value of $\mathbf{S}^{(jN)}$, we run the basic Gibbs sampling chain described earlier, and obtain a new value $\mathbf{S}^{((j+1)N)}$. We now retain from $\mathbf{S}^{((j+1)N)}$ only the values of θ and ψ, the group structures U and V, and the indicators $I(X_i = m_i)$ and $I(Y_i = m_i)$. We do not retain any other information about the X's and the Y's.

Next we generate a vector **X** from the conditional distribution of **X** given these values of θ, U, the indicators $I(X_i = m_i)$ and the data, and similarly independently generate a vector **Y**. From standard properties of the Dirichlet process, we see that this is done as follows. Define (random) sets A_i and B_i by

$$A_i = \begin{cases} \{m_i\}, & \text{if } X_i = m_i, \\ (0, m_i), & \text{if } X_i < m_i, \end{cases} \quad \text{and} \quad B_i = \begin{cases} \{m_i\}, & \text{if } Y_i = m_i, \\ (0, m_i), & \text{if } Y_i < m_i, \end{cases}$$

and define the intersections

$$A_k^* = \bigcap_{i: u_i = k} A_i \quad \text{for } k \in \mathscr{U} \quad \text{and} \quad B_k^* = \bigcap_{i: v_i = k} B_i \quad \text{for } k \in \mathscr{V}.$$

We generate X_k from $\alpha_\theta(\cdot \mid A_k^*)$, independently for $k \in \mathscr{U}$ [here, the $\alpha_\theta(\cdot \mid A_k^*)$'s are the obvious conditional distributions]. Similarly, independently generate the values Y_k, $k \in \mathscr{V}$. The full vectors **X** and **Y** are now obtained by respecting the group structures U and V.

Convergence of the algorithm that includes the extra step follows directly from convergence of the basic Gibbs sampling algorithm; see the remark following the proof of Theorem 1. We note a substantial increase in the rate of convergence for the modified algorithm in the example of Section 3.

This modification of the algorithm presents additional advantages when estimating certain quantities. Consider for instance the problem of estimating the distribution of a future X observation. If we use the basic Gibbs sampling algorithm for L cycles, we can, at the end of cycle j, generate an observation from the probability distribution

$$(2.12) \qquad D_j = \frac{\sum_{i=1}^n \delta_{X_i^{(jN)}} + \alpha_{\theta^{(jN)}}}{n + \alpha_{\theta^{(jN)}}(\mathscr{R})}.$$

The empirical distribution of these L points can be used to estimate the distribution of a future observation given the data. However, if we wish to estimate the density of the absolutely continuous component of this distribution, we are forced to use a smoothing operation, such as a kernel density estimate. Rao–Blackwellization here consists of taking the average of the L distributions in (2.12). Again, estimation of the absolutely continuous component of the distribution of a future observation requires a smoothing operation.

On the other hand, if we use the algorithm with the added step, we can in each cycle form the probability distribution C_j defined by

$$(2.13) \qquad \begin{aligned} C_j(t) &= E\Big\{ D_j(t) \mid U, V, \{A_i\}_{i=1}^n, \{B_i\}_{i=1}^n, \theta, \psi \Big\} \\ &= \frac{\big(\sum_{k \in \mathscr{U}} a_k \alpha_\theta(t \mid A_k^*)\big) + \alpha_\theta(t)}{n + \alpha_\theta(\mathscr{R})}, \end{aligned}$$

where $a_k = \#\{j \colon u_j = k\}$, and compute the average $\overline{C}(t) = (1/L)\sum_{j=1}^L C_j(t)$. [In (2.13) we have suppressed the superscripts indexing the cycle to lighten the notation.] Since some of the A_k^*'s are singletons and some are intervals, use of $\overline{C}(t)$ provides a much more natural way to estimate the absolutely continuous and discrete components of the distribution of a future X given the data. This is illustrated in the example in Section 3.

2.3.3. *Uniform ergodicity.* Let π be the posterior distribution of $\mathbf{S}$ given `data` on the space $(\mathscr{R}^{2n} \times \Theta \times \Psi, \mathscr{B})$, where $\mathscr{B}$ is the collection of Borel sets on $\mathscr{R}^{2n} \times \Theta \times \Psi$. This π is also a stationary distribution for the chain $\{\mathbf{S}^{(jN)}\}_{j=0}^\infty$. We now show that the distribution of the Markov chain $\{\mathbf{S}^{(jN)}\}_j$ converges to π at a rate that is geometric and independent of the starting point. Before stating our theorem, we introduce some notation. Define

$$D = \{\mathbf{s} \colon x_i, y_i \geq 0 \text{ and } x_i \vee y_i = m_i \text{ for } i = 1, \ldots, n, \ \theta \in \Theta, \ \psi \in \Psi\},$$

and let $P^j(\mathbf{s}, C)$ be the j-step transition probabilities for the chain; that is,

$$P^j(\mathbf{s}, C) = P(\mathbf{S}^{(jN)} \in C \mid \mathbf{S}^{(0)} = \mathbf{s}).$$

Note that if ν is absolutely continuous with respect to Lebesgue measure, then so is $\nu_{\mathbf{x}}$. This implies $\nu_{\mathbf{x}}$ has a density with respect to Lebesgue measure, which we shall denote $\nu_{\mathbf{x}}'$. In the proof of Theorem 1, we shall need to find bounds on $\nu_{\mathbf{x}}'(\theta)$ for fixed θ as $\mathbf{x}$ varies over the compact set $[0, m_{(n)}]^n$. Here, $m_{(n)} = \max(m_1, \ldots, m_n)$. This would be straightforward if $\nu_{\mathbf{x}}'(\theta)$ were

continuous in $\mathbf{x}$, but inspection of (2.11) clearly shows this is not the case. For this reason, we introduce functions $f_i(\mathbf{x}, \theta)$ on $\mathscr{R}^i \times \Theta$, defined by

$$f_i(\mathbf{x}, \theta) = c_i(\mathbf{x})\left(\prod_{j=1}^i h_\theta(x_j)\right)\frac{(\alpha_\theta(\mathscr{R}))^i \Gamma(\alpha_\theta(\mathscr{R}))}{\Gamma(\alpha_\theta(\mathscr{R}) + n)}\nu'(\theta) \quad \text{for } i = 1, \ldots, n.$$

where the values $c_i(\mathbf{x})$ are constants such that, for each $\mathbf{x}$, $f_i(\mathbf{x}, \theta)$ is a density in θ. Note that

$$\nu'_{\mathbf{x}}(\theta) = f_{\#(\mathbf{x})}(\mathbf{x}^{\text{dist}}, \theta),$$

where $\mathbf{x}^{\text{dist}}$ is comprised of the distinct values of $\mathbf{x}$ arranged in any order.

THEOREM 1. *Consider the Markov chain* $\{\mathbf{S}^{(jN)}\}_j$ *resulting from applying the above algorithm to a parallel system of order* 2. *Suppose that the following hold*:

(i) *The observed maximum values* $m_1, \ldots, m_n$ *are distinct and positive.*

(ii) *There exists a compact set* $\Gamma_1 \subset \Theta$ *with* $\nu(\Gamma_1) > 0$ *such that* $\alpha'_\theta(x)$ *exists, is positive and is continuous in both* θ *and* x *for* $(\theta, x) \in \Gamma_1 \times [0, m_{(n)}]$; *similarly there exists a compact set* $\Gamma_2 \subset \Psi$ *corresponding to* β'_ψ.

(iii) *The prior* ν *is absolutely continuous with respect to Lebesgue measure and for each* $i, i = 1, \ldots, n, f_i(\cdot, \cdot)$ *is positive and continuous on* $[0, m_{(n)}]^i \times \Gamma_1$; *an analogous condition holds for the prior and posterior of* ψ.

Then there exists a value $\lambda > 0$ *such that*

$$(2.14) \qquad \sup_{C \in \mathscr{B}, \mathbf{s} \in D} |P^j(\mathbf{s}, C) - \pi(C)| < e^{-\lambda j} \quad \text{for all } j \geq 2.$$

Assumption (i) above is not essential. We impose it mainly to simplify the argument. The conclusion of our theorem remains true when there are ties among the values $m_1, \ldots, m_n$ so long as we make appropriate changes in our Gibbs sampling algorithm. Specifically, the systems having a common value of m_i must be treated as a group and updated simultaneously. The procedure for updating a "tied" group of systems is a straightforward generalization of the mixing procedure in Lemma 1.

It is not necessary to understand the proof of Theorem 1 in order to follow the rest of the paper, so the reader not interested in the proof can go directly to Section 2.4.

Recall that if $\{Z_k\}_{k=0}^\infty$ is a Markov chain on $(\mathscr{Z}, \mathscr{F})$ with stationary probability distribution π and n-step transition probabilities $P^n(z, C)$ which satisfy the *Doeblin condition*, that is, for some probability measure ρ on $(\mathscr{Z}, \mathscr{F})$, some positive integer n_0 and some $\varepsilon > 0$,

$$(2.15) \qquad P^{n_0}(z, C) \geq \varepsilon\rho(C) \quad \text{for all } z \in \mathscr{Z} \text{ and all } C \in \mathscr{F},$$

then

$$(2.16) \qquad \sup_{C \in \mathscr{F}} |P^n(z, C) - \pi(C)| \leq (1 - \varepsilon)^{\lfloor n/n_0 \rfloor} \quad \text{for all } z \in \mathscr{Z}.$$

The proof of this fact involves a coupling argument which we review below. We may write

$$(2.17) \qquad P^{n_0}(z, \cdot) = (1 - \varepsilon)\eta(z, \cdot) + \varepsilon\rho(\cdot) \quad \text{for all } z \in \mathscr{Z},$$

where $\eta(z, \cdot) = (P^{n_0}(z, \cdot) - \varepsilon\rho(\cdot))/(1 - \varepsilon)$ and we assume $\varepsilon < 1$. By the Doeblin condition (2.15), $\eta(z, \cdot)$ is a probability measure. Thus, $P^{n_0}(z, \cdot)$ is a convex combination of two measures, *the second of which does not involve z*. Therefore, the representation in (2.17) allows us to view each step in the evolution of the Markov chain $\{Z_{ln_0}\}_l$ as a coin-tossing experiment followed by a draw from either $\eta(z, \cdot)$ or $\rho(\cdot)$. This is the key that makes possible the coupling argument.

Now run two chains $\{W_l\}_l$ and $\{Y_l\}_l$ as follows. Let $W_0 = w_0$ and $Y_0 \sim \pi$. If the current state of the two chains is (w_{n-1}, y_{n-1}) with $w_{n-1} \neq y_{n-1}$, generate W_n and Y_n by first tossing a coin with probability of heads equal to ε. If the toss results in a head, select $V \sim \rho$, and set *both* W_n and Y_n equal to V. If the toss results in tails, independently select W_n from $\eta(w_{n-1}, \cdot)$ and Y_n from $\eta(y_{n-1}, \cdot)$. If $w_{n-1} = y_{n-1}$, then select $V \sim P^{n_0}(w_{n-1}, \cdot)$ and set $W_n = Y_n = V$. It is clear that $\{W_l\}_l$ and $\{Y_l\}_l$ are Markov chains with transition probabilities $P^{n_0 l}(z, C)$ and that $Y_l \sim \pi$ for all l. Moreover,

$$P\{W_l \neq Y_l \text{ for some } l \geq k\} \leq (1 - \epsilon)^k.$$

This argument shows that the Markov chain $\{Z_{ln_0}\}_l$ satisfies

$$\sup_{C \in \mathscr{F}} |P\{Z_{n_0 k} \in C\} - \pi(C)| \leq (1 - \varepsilon)^k,$$

from which (2.16) follows, since $\sup_{C \in \mathscr{F}} |P^n(z, C) - \pi(C)|$ is nonincreasing in n.

Results of the form (2.14) are generally proved by verifying the Doeblin condition (2.15). However, we shall prove our theorem by dealing with the underlying coupling argument directly and explicitly because our arguments are then easier to follow.

PROOF OF THEOREM 1. Consider starting the Gibbs sampler from two different initial states $\mathbf{S}^{(0)}$ and $\tilde{\mathbf{S}}^{(0)}$ and producing two sequences $\{S^{(t)}\}$ and $\{\tilde{S}^{(t)}\}$. We shall "couple" these sequences by defining them on the same probability space in such a way that

$$(2.18) \qquad\qquad P\{\mathbf{S}^{(2N)} = \tilde{\mathbf{S}}^{(2N)}\} > \varepsilon$$

where ε is positive and can be chosen independently of the starting states $\mathbf{S}^{(0)}$ and $\tilde{\mathbf{S}}^{(0)}$. This clearly suffices to prove the theorem.

Condition (2.18) indicates that our proof requires two passes of the algorithm (or, equivalently, two "cycles" of the Gibbs sampler) to couple the two sequences. In the first pass, we show that there is a positive lower bound not depending on the starting states, for the probability that for each $i, i = 1, \ldots, n$, the minima $X_i \wedge Y_i$ and $\tilde{X}_i \wedge \tilde{Y}_i$ are not equal to any of the observed maxima $m_1, \ldots, m_n$. Then in the second pass, we show that there is also a

positive lower bound for the probability that for each i, $i = 1, \ldots, n$, $X_i = \tilde{X}_i$ and $Y_i = \tilde{Y}_i$; that is, $S_i = \tilde{S}_i$. The probability that $S_0 = \tilde{S}_0$ is handled in a manner similar to the proof of (2.16).

A simple example may help to identify the issues involved. Suppose that $n = 2$ and $m_1 < m_2$ with starting states $\mathbf{S}^{(0)} = (S_0^{(0)}, S_1^{(0)}, S_2^{(0)}) = ((\theta, \psi), (l_1, m_1), (m_1, m_2))$ and $\tilde{\mathbf{S}}^{(0)} = ((\tilde{\theta}, \tilde{\psi}), (m_1, l_2), (m_2, m_1))$, where $l_1, l_2 < m_1$. At time $t = 1$, we update θ, ψ and $\tilde{\theta}, \tilde{\psi}$, but this is not important. At time $t = 2$, we update (X_1, Y_1) and $(\tilde{X}_1, \tilde{Y}_1)$, and this is where the problem arises. For the sequence $\{\mathbf{S}^{(t)}\}$, $A_1^{(2)}$ [defined in (2.9)] has an atom at m_1, but $B_1^{(2)}$ does not, so $P(X_1^{(2)} = m_1, Y_1^{(2)} < m_1) = 1$. For the sequence $\{\tilde{\mathbf{S}}^{(t)}\}$, $\tilde{B}_1^{(2)}$ has an atom at m_1, but $\tilde{A}_1^{(2)}$ does not, so $P(\tilde{X}_1^{(2)} < m_1, \tilde{Y}_1^{(2)} = m_1) = 1$. Thus, with probability 1, the two sequences are *not* coupled during the first pass of the algorithm. In the proof, we will show that after one pass of the algorithm, the atoms at m_1 and m_2 are "out of the way" and thus there is a positive probability (that is independent of the starting states) that $\mathbf{S}^{(6)} = \tilde{\mathbf{S}}^{(6)}$, making (2.18) true.

The following will involve a specific implementation of the algorithm described by (2.6)–(2.11), strictly for use in our coupling argument. Let $\{U_{jk};\ j = 1, \ldots, \infty;\ k = 1, 2, 3, 4\}$ be an array of independent $\mathscr{U}(0, 1)$ [the uniform distribution on $(0, 1)$] random variables. We shall generate $\theta^{(t)}$, $\psi^{(t)}$, $X_i^{(t)}$ and $Y_i^{(t)}$ in terms of these uniform random variables using certain functions ρ_1, ρ_2 and ρ_3. These functions and their properties are described in detail later in this section. Specifically, we define the functions and show that using them produces the desired conditional distribution of $\mathbf{S}^{(t)}$ given $\mathbf{S}^{(t-1)}$.

For t with $(t - 1) \bmod N = 0$, define $S_0^{(t)} = (\theta^{(t)}, \psi^{(t)})$ by

$$(2.19) \quad \theta^{(t)} = \rho_1\big(\mathbf{X}^{(t-1)}, U_{t1}, U_{t2}\big) \quad \text{and} \quad \psi^{(t)} = \rho_2\big(\mathbf{Y}^{(t-1)}, U_{t3}, U_{t4}\big),$$

and for t with $(t - 1) \bmod N = K \neq 0$, define $S_K^{(t)} = (X_K^{(t)}, Y_K^{(t)})$ by

$$(2.20) \quad \big(X_K^{(t)}, Y_K^{(t)}\big) = \rho_3\big(\theta^{(t-1)}, \psi^{(t-1)}, m_K, \mathbf{X}_{(-K)}^{(t-1)}, \mathbf{Y}_{(-K)}^{(t-1)}, U_{t1}, U_{t2}, U_{t3}\big),$$

where $\mathbf{W}_{(-K)}$ is used to denote the $(n - 1)$-tuple obtained by deleting the Kth element of the n-tuple $\mathbf{W}$. $\square$

For the process $\{\tilde{\mathbf{S}}^{(t)}\}$, the quantities $\tilde{\theta}^{(t)}$, $\tilde{\psi}^{(t)}$, $\tilde{X}^{(t)}$ and $\tilde{Y}^{(t)}$ are defined in the same way except that θ, ψ, $\mathbf{X}$ and $\mathbf{Y}$ are everywhere replaced by $\tilde{\theta}$, $\tilde{\psi}$, $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$. The *same* uniform variates are used in generating both $\mathbf{S}^{(t)}$ and $\tilde{\mathbf{S}}^{(t)}$.

The essence of the proof consists of showing that there is a positive probability (which is independent of the starting states) of coupling the two processes if we implement the algorithm using ρ_1, ρ_2 and ρ_3 [i.e., (2.18) holds]. To this end, we introduce a sequence of events $A_1, A_2, \ldots, A_{2N}$ defined as follows. Let $\mathscr{M} = \{m_1, m_2, \ldots, m_n\}$. For $t = 1$ and $t = N + 1$, we define

$$A_t = \Big\{\theta^{(t)} = \tilde{\theta}^{(t)}, \psi^{(t)} = \tilde{\psi}^{(t)}, \theta^{(t)} \in \Gamma_1, \psi^{(t)} \in \Gamma_2\Big\}.$$

[The sets Γ_1 and Γ_2 are the compact sets given by assumption (ii) of Theorem 1.] For $2 \le t \le N$ and $K = (t - 1) \bmod N$ we define

$$A_t = \left\{ X_K^{(t)} \wedge Y_K^{(t)} \notin \mathcal{M}, \ \tilde{X}_K^{(t)} \wedge \tilde{Y}_K^{(t)} \notin \mathcal{M} \right\}.$$

For $N + 2 \le t \le 2N$ and $K = (t - 1) \bmod N$ we define

$$A_t = \left\{ X_K^{(t)} = \tilde{X}_K^{(t)}, \ Y_K^{(t)} = \tilde{Y}_K^{(t)}, \ X_K^{(t)} \wedge Y_K^{(t)} \notin \mathcal{M} \right\}.$$

Note that $\bigcap_{t=1}^{2N} A_t$ implies that $\mathbf{S}^{(2N)} = \tilde{\mathbf{S}}^{(2N)}$. We can write

$$P\left(\bigcap_{t=1}^{2N} A_t \right) = \prod_{t=1}^{2N} \zeta_t,$$

where $\zeta_1 = P(A_1)$ and $\zeta_t = P(A_t \mid \bigcap_{j<t} A_j)$ for $t > 1$. We shall show that each of the values ζ_t are bounded away from 0 by quantities which do not depend on $\mathbf{S}^{(0)}$ and $\tilde{\mathbf{S}}^{(0)}$. In the last part of this section we establish these bounds in Facts 1–6. Specifically, Fact 1 is used to bound ζ_1 and ζ_{N+1}; Facts 2 and 3 are used to bound $\zeta_2, \ldots, \zeta_N$. Now consider $N + 2 \le t \le 2N$ and $K = (t - 1) \bmod N$. Conditional on $\bigcap_{j<t} A_j$, the vectors $\mathbf{X}_{(-K)}^{(t-1)}$, $\tilde{\mathbf{X}}_{(-K)}^{(t-1)}$, $\mathbf{Y}_{(-K)}^{(t-1)}$ and $\tilde{\mathbf{Y}}_{(-K)}^{(t-1)}$ contain no entries equal to m_K. [Here we are using assumption (i) of Theorem 1; i.e., the values $m_1, \ldots, m_n$ are distinct.] This will allow us to use Fact 6 to bound $\zeta_{N+2}, \zeta_{N+3}, \ldots, \zeta_{2N}$, which will establish (2.18) and complete the proof of Theorem 1.

Details of the proof of Theorem 1. We shall now define in detail the functions ρ_1, ρ_2 and ρ_3 and demonstrate the properties of these functions that are used in the proof of our theorem. In this discussion we will use the following notation. If H is a distribution function, then $H^\dagger$ will denote a function with the property that $H^\dagger(U) \sim H$, when U is a $\mathcal{U}(0, 1)$ random variable. Such an $H^\dagger$ always exists. In the case of a distribution function on $\mathcal{R}^1$, the function $H^\dagger$ may be taken to be $H^\dagger(y) = \inf\{x: H(x) > y\}$. For economy of notation, if h is a density, then $h^\dagger$ is used if the distribution function associated with h has not been introduced. Finally, let U_1, U_2, U_3, U_4 be independent $\mathcal{U}(0, 1)$ random variables.

By (2.19), we generate θ and ψ from uniform random variables using functions ρ_1 and ρ_2 defined below. Define

$$\phi_1(u) = \inf_{\mathbf{x} \in E} \nu_\mathbf{x}'(u),$$

where $E = \{\mathbf{x}: 0 \le x_i \le m_i, 1 \le i \le n\}$,

$$\pi_1 = \int \phi_1(u)\, du,$$

and densities

$$\overline{\phi}_1(u) = \frac{\phi_1(u)}{\pi_1} \quad \text{and} \quad \overline{\nu}_\mathbf{x}(u) = \frac{\nu_\mathbf{x}'(u) - \phi_1(u)}{1 - \pi_1}.$$

Note that $\pi_1 > 0$. This is because by assumption (iii) of Theorem 1, for each i, the function f_i is bounded away from 0 over $[0, m_{(n)}]^i \times \Gamma_1$. Therefore, for each $u \in \Gamma_1$, $\phi_1(u)$ is positive. It follows that $\int_\Theta \phi_1(u)\, du \geq \int_{\Gamma_1} \phi_1(u)\, du > 0$.

Note that in the special case of the "exponential/gamma" setup used in our data analysis in Section 3, we have

$$\phi_1(u) = \inf_{\mathbf{x} \in E} \mathscr{G}(u \mid a(\mathbf{x}), b(\mathbf{x})),$$

where $a(\mathbf{x}) = a_1 + \#(\mathbf{x})$, and $b(\mathbf{x}) = b_1 + \Sigma^{\mathrm{dist}} x_i$. For $\mathbf{x}$ with $0 \leq x_i \leq m_i$, $i = 1, \ldots, n$, $(a(\mathbf{x}), b(\mathbf{x}))$ ranges over a compact set, so it is clear that $\phi_1(u) > 0$ for all $u > 0$.

Now we define the function ρ_1. For given values $\mathbf{x}, u_1, u_2$, we let

$$\rho_1(\mathbf{x}, u_1, u_2) = \begin{cases} \overline{\phi}_1^\dagger(u_2), & \text{if } u_1 \leq \pi_1, \\ \overline{\nu}_\mathbf{x}^\dagger(u_2), & \text{otherwise.} \end{cases}$$

It is easy to verify that $\rho_1(\mathbf{x}, U_1, U_2) \sim \nu_\mathbf{x}$.

For generating ψ we define a function ρ_2 in a similar manner. Define

$$\phi_2(u) = \inf_{\mathbf{y} \in E} \mu_\mathbf{y}'(u),$$

$$\pi_2 = \int \phi_2(u)\, du$$

and densities

$$\overline{\phi}_2(u) = \frac{\phi_2(u)}{\pi_2} \quad \text{and} \quad \overline{\mu}_\mathbf{y}(u) = \frac{\mu_\mathbf{y}'(u) - \phi_2(u)}{1 - \pi_2}.$$

Note that $\pi_2 > 0$. Now we define ρ_2. For given values $\mathbf{y}, u_1, u_2$, we let

$$\rho_2(\mathbf{y}, u_1, u_2) = \begin{cases} \overline{\phi}_2^\dagger(u_2), & \text{if } u_1 \leq \pi_2, \\ \overline{\mu}_\mathbf{y}^\dagger(u_2), & \text{otherwise.} \end{cases}$$

It is easy to verify that $\rho_2(\mathbf{y}, U_3, U_4) \sim \mu_\mathbf{y}$.

In the proof of our theorem, we need the following fact regarding ρ_1 and ρ_2. Let $\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{y}, \tilde{\mathbf{y}}$ be arbitrary fixed values in E. Define $\theta = \rho_1(\mathbf{x}, U_1, U_2)$, $\tilde{\theta} = \rho_1(\tilde{\mathbf{x}}, U_1, U_2)$, $\psi = \rho_2(\mathbf{y}, U_3, U_4)$ and $\tilde{\psi} = \rho_2(\tilde{\mathbf{y}}, U_3, U_4)$. From the definitions of ρ_1 and ρ_2 it is clear that

$$P\{\theta = \tilde{\theta},\, \theta \in \Gamma_1\} \geq \pi_1 \int_{\Gamma_1} \overline{\phi}_1(u)\, du = \int_{\Gamma_1} \phi_1(u)\, du > 0,$$

and similarly,

$$P\{\psi = \tilde{\psi},\, \psi \in \Gamma_2\} \geq \pi_2 \int_{\Gamma_2} \overline{\phi}_2(u)\, du = \int_{\Gamma_2} \phi_2(u)\, du > 0,$$

so that by independence we obtain the following.

FACT 1. We have

$$P\left\{\theta = \tilde{\theta},\, \psi = \tilde{\psi},\, \theta \in \Gamma_1,\, \psi \in \Gamma_2\right\} \geq \int_{\Gamma_1} \phi_1(u)\, du \int_{\Gamma_2} \phi_2(u)\, du.$$

Note that this bound does not involve $\mathbf{x}$, $\tilde{\mathbf{x}}$, $\mathbf{y}$ and $\tilde{\mathbf{y}}$.

By (2.20), the generation of $S_K = (X_K, Y_K)$ will be done using a function $\rho_3(\theta, \psi, m, \mathbf{x}, \mathbf{y}, u_1, u_2, u_3)$, where $m > 0$, $\mathbf{x} = (x_1, \ldots, x_{n-1})$ and $\mathbf{y} = (y_1, \ldots, y_{n-1})$ are $(n-1)$-tuples with $x_i, y_i \geq 0$ for all i. The function ρ_3 will be defined in the following discussion.

In Lemma 1, take

$$A = \frac{\alpha_\theta + N_\mathbf{x}}{\alpha_\theta(\mathscr{R}) + n - 1} \quad \text{and} \quad B = \frac{\beta_\psi + N_\mathbf{y}}{\beta_\psi(\mathscr{R}) + n - 1},$$

where $N_\mathbf{x} = \sum_{i=1}^{n-1} \delta_{x_i}$ and $N_\mathbf{y} = \sum_{i=1}^{n-1} \delta_{y_i}$. Define $p_1, \ldots, p_5$ and the distributions $\bar{A}_c, \bar{B}_c, \bar{A}_d, \bar{B}_d$ as in Lemma 1. These quantities are all (implicitly) functions of θ, ψ, $\mathbf{x}$, and $\mathbf{y}$.

Now define ρ_3 by

$$\rho_3(\theta, \psi, m, \mathbf{x}, \mathbf{y}, u_1, u_2, u_3)$$

$$= \begin{cases} \left(\bar{A}_c^\dagger(u_3), m\right), & \text{if } 0 < u_1 \leq p_1 + p_2 \text{ and} \\ & \quad 0 < u_2 \leq p_1/(p_1 + p_2), \\ \left(m, \bar{B}_c^\dagger(u_3)\right), & \text{if } 0 < u_1 \leq p_1 + p_2 \text{ and} \\ & \quad p_1/(p_1 + p_2) < u_2 \leq 1, \\ \left(\bar{A}_d^\dagger(u_3), m\right), & \text{if } p_1 + p_2 < u_1 \leq p_1 + p_2 + p_3, \\ \left(m, \bar{B}_d^\dagger(u_3)\right), & \text{if } p_1 + p_2 + p_3 < u_1 \leq p_1 + p_2 + p_3 + p_4, \\ (m, m), & \text{if } p_1 + p_2 + p_3 + p_4 < u_1 \leq 1. \end{cases}$$

It is clear that the pair $(X, Y) = \rho_3(\theta, \psi, m, \mathbf{x}, \mathbf{y}, U_1, U_2, U_3)$ has the distribution given in (2.7) since the function ρ_3 merely describes a particular way to carry out the mixing procedure in Lemma 1.

The function ρ_3 has two properties which we use in the proof of our theorem. Pick arbitrary fixed values $\theta \in \Theta$, $\psi \in \Psi$, $m > 0$ and $\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{y}, \tilde{\mathbf{y}}$ in $\mathscr{R}^{n-1}$. Define

$$(X, Y) = \rho_3(\theta, \psi, m, \mathbf{x}, \mathbf{y}, U_1, U_2, U_3)$$

and

$$(\tilde{X}, \tilde{Y}) = \rho_3(\theta, \psi, m, \tilde{\mathbf{x}}, \tilde{\mathbf{y}}, U_1, U_2, U_3).$$

First note that $X \wedge Y$ is generated from a continuous distribution when $U_1 \leq p_1 + p_2$. Therefore we have the following fact.

FACT 2. If $U_1 \leq (p_1 + p_2) \wedge (\tilde{p}_1 + \tilde{p}_2)$, then both $X \wedge Y$ and $\tilde{X} \wedge \tilde{Y}$ are generated from continuous distributions.

In conjunction with this fact, we note that $p_1 + p_2 \geq f/(f+1)$, where

$$f = \frac{\alpha_\theta(m) \wedge \beta_\psi(m)}{n-1}.$$

To see this, we go back to the expressions for $p_1, \ldots, p_5$ given in Lemma 1. We find that

$$A_d(m)p_1 = A_c(m)(p_3 + p_5) \quad \text{and} \quad B_d(m)p_2 = B_c(m)(p_4 + p_5).$$

Since

$$A_d(m) \leq \frac{n-1}{\alpha_\theta(\mathscr{R}) + n - 1}, \qquad B_d(m) \leq \frac{n-1}{\beta_\psi(\mathscr{R}) + n - 1},$$

$$A_c(m) = \frac{\alpha_\theta(m)}{\alpha_\theta(\mathscr{R}) + n - 1}, \qquad B_c(m) = \frac{\beta_\psi(m)}{\beta_\psi(\mathscr{R}) + n - 1},$$

these relations imply

$$(n-1)p_1 \geq \alpha_\theta(m)(p_3 + p_5) \quad \text{and} \quad (n-1)p_2 \geq \beta_\psi(m)(p_4 + p_5).$$

Thus

$$p_1 + p_2 \geq f(p_3 + p_4 + p_5) = f(1 - p_1 - p_2),$$

so that $p_1 + p_2 \geq f/(1+f)$ as desired. From assumption (ii) of Theorem 1, we know that f is bounded below for $\theta \in \Gamma_1$ and $\psi \in \Gamma_2$. Therefore, we can state the following fact.

FACT 3.   For $\theta \in \Gamma_1$ and $\psi \in \Gamma_2$, the value $(p_1 + p_2)$ is bounded away from 0 by some quantity which does not depend on $\theta, \psi, \mathbf{x}$ or $\mathbf{y}$.

Now observe that A and $\tilde{A}$ have the same continuous part; that is, $A_c = \tilde{A}_c$. This implies that $\bar{A}_c = \bar{\tilde{A}}_c$ and thus $(\bar{A}_c)^\dagger = (\bar{\tilde{A}}_c)^\dagger$. Similarly, $(\bar{B}_c)^\dagger = (\bar{\tilde{B}}_c)^\dagger$. This leads to our next property.

FACT 4.   If $U_1 \leq (p_1 + p_2) \wedge (\tilde{p}_1 + \tilde{p}_2)$ and

$$U_2 \leq \frac{p_1}{p_1 + p_2} \wedge \frac{\tilde{p}_1}{\tilde{p}_1 + \tilde{p}_2} \quad \text{or} \quad U_2 > \frac{p_1}{p_1 + p_2} \vee \frac{\tilde{p}_1}{\tilde{p}_1 + \tilde{p}_2},$$

then $X = \tilde{X}$, $Y = \tilde{Y}$ and $X \wedge Y$ is generated from a continuous distribution.

Also note that when $N_{\mathbf{x}}(\{m\}) = N_{\mathbf{y}}(\{m\}) = 0$, we have

$$\frac{p_1}{p_1 + p_2} = \frac{\alpha_\theta(m)\beta_\psi'(m)}{\alpha_\theta(m)\beta_\psi'(m) + \alpha_\theta'(m)\beta_\psi(m)},$$

so that $p_1/(p_1 + p_2)$ does not depend on $\mathbf{x}$ and $\mathbf{y}$ in this situation. Now, using properties in assumption (ii) of Theorem 1, we obtain Fact 5.

FACT 5. If $N_x(\{m\}) = N_{\mathbf{y}}(\{m\}) = 0$, then for $\theta \in \Gamma_1$ and $\psi \in \Gamma_2$, the value $p_1/(p_1 + p_2)$ can be bounded away from 0 and 1 by quantities which do not depend on $\theta, \psi, \mathbf{x}$, and $\mathbf{y}$.

Combining Facts 3, 4, and 5 leads to our last property.

FACT 6. If $N_{\mathbf{x}}(\{m\}) = N_{\mathbf{y}}(\{m\}) = N_{\tilde{\mathbf{x}}}(\{m\}) = N_{\tilde{\mathbf{y}}}(\{m\}) = 0$, $\theta \in \Gamma_1$ and $\psi \in \Gamma_2$, then the probability of the event

$$\left\{ X = \tilde{X}, Y = \tilde{Y}, \text{ and } X \wedge Y \text{ is generated from a continuous distribution} \right\}$$

is bounded away from 0 by a quantity which does not depend on $\theta, \psi, \mathbf{x}, \mathbf{y}, \tilde{\mathbf{x}}$, or $\tilde{\mathbf{y}}$.

This concludes the details of the proof of Theorem 1. □

We note that the coupling argument in the proof of Theorem 1 also establishes uniform convergence of the chain with the added step (strictly speaking the chain that results if we append the extra step at the end of every two cycles, since our coupling argument requires two passes).

Note that

$$\mathscr{L}_{\text{data}}(F, G) = \int \left(\mathscr{D}_{\alpha_\theta + \sum_{i=1}^n \delta_{X_i}} \otimes \mathscr{D}_{\beta_\psi + \sum_{i=1}^n \delta_{Y_i}} \right) \mathscr{L}_{\text{data}}(d\theta, d\psi, d\mathbf{X}, d\mathbf{Y}).$$

(In particular, the marginal posterior distribution of F is a mixture of Dirichlets, and a similar statement holds for G.) Let

$$(F, G)^{(jN)} \sim \mathscr{D}_{\alpha_\theta(jN) + \sum_{i=1}^n \delta_{X_i}(jN)} \otimes \mathscr{D}_{\beta_\psi(jN) + \sum_{i=1}^n \delta_{Y_i}(jN)}.$$

COROLLARY 1. *Let $Q_{\mathbf{s}}^{(jN)}$ denote the distribution of $(F, G)^{(jN)}$ when the chain $\{\mathbf{S}^{(lN)}\}_l$ is started at $\mathbf{s}$. Then under the conditions of Theorem 1, we have*

$$(2.21) \quad \sup_{A \in \mathscr{A}} \left| Q_{\mathbf{s}}^{(jN)}\{(F, G)^{(jN)} \in A\} - P_{\text{data}}\{(F, G) \in A\} \right| \le e^{-\lambda j} \quad \text{for all } \mathbf{s},$$

where λ is the same λ that appears in the statement of Theorem 1.

[In (2.21), $\mathscr{A} = \mathscr{A}_0 \times \mathscr{A}_0$, where $\mathscr{A}_0$ is the smallest σ-field on the set of probability measures on $\mathscr{R}$ such that the map $P \mapsto P(B)$ is measurable for each Borel set $B \subset \mathscr{R}$.]

PROOF OF COROLLARY 1. Let $\pi_{\mathbf{s}}^{(jN)}$ be the distribution of $\mathbf{S}^{(jN)}$ when the Markov chain $\{\mathbf{S}^{(jN)}\}_j$ is started at $\mathbf{s}$. Fix $A \in \mathscr{A}$ and let

$$f_A(\mathbf{s}) = \left(\mathscr{D}_{\alpha_\theta + \sum_{i=1}^n \delta_{X_i}} \otimes \mathscr{D}_{\beta_\psi + \sum_{i=1}^n \delta_{Y_i}} \right)(A).$$

We know that

$$
\left| Q^{(jN)}_{\mathbf{s}^{(0)}}(A) - P_{\text{data}}\{(F,G) \in A\} \right|
$$

$$
= \left| \int f_A(\mathbf{s})\, d\pi^{(jN)}_{\mathbf{s}^{(0)}}(\mathbf{s}) - \int f_A(\mathbf{s})\, d\pi(\mathbf{s}) \right|
$$

(2.22)

$$
= \left| \int f_A(\mathbf{s})\, d\big(\pi^{(jN)}_{\mathbf{s}^{(0)}} - \pi\big)(\mathbf{s}) \right|
$$

$$
= \left| \int f_A(\mathbf{s})\, d\big(\pi^{(jN)}_{\mathbf{s}^{(0)}} - \pi\big)^{+}(\mathbf{s}) - \int f_A(\mathbf{s})\, d\big(\pi^{(jN)}_{\mathbf{s}^{(0)}} - \pi\big)^{-}(\mathbf{s}) \right|.
$$

(For a signed measure λ, the representation $\lambda = \lambda^{+} - \lambda^{-}$ is the standard Jordan decomposition of λ.) Since $f_A(\cdot)$ is a measurable function of $\mathbf{s}$ that satisfies $0 \le f_A(\mathbf{s}) \le 1$ for all $\mathbf{s}$, we see that each of the two integrals in the last line of (2.22) is bounded by $\exp(-\lambda j)$, and this proves Corollary 1. $\square$

Thus, if for example we want to estimate, for fixed v, the density of $\mathscr{L}_{\text{data}}(F(v))$, we would use the mixture

$$
\frac{1}{L} \sum_{j=1}^{L} \operatorname{beta}\left(\Big(\alpha_{\theta(jN)} + \sum_{i=1}^{n} \delta_{X_i^{(jN)}} \Big)(0,v], \Big(\alpha_{\theta(jN)} + \sum_{i=1}^{n} \delta_{X_i^{(jN)}} \Big)(v,\infty) \right),
$$

where $\operatorname{beta}(a, b)(\cdot)$ is the beta density.

2.4. *The algorithm for arbitrary systems.* Here we discuss the implementation and convergence of the algorithm in the general case. (Our discussion is in the context of the basic algorithm although all our comments pertain to the modified algorithm as well.) A discussion of the issues of identifiability and frequentist consistency appears in Lawson (1994).

2.4.1. *Implementation of the algorithm.* Let the autopsy statistics for system i, $i = 1, \ldots, n$, be (T_i, D_i), where T_i is the death time of the system and D_i is the set of components that are dead at time T_i. Recall that, after examining (T_i, D_i), each component in system i is put into exactly one of the categories C1, C2, C3 or C4 described in Section 1. For a component in category C1, one generates an observation according to the distribution $(A_i^{(t)})_{(T_i, \infty)}$, where $A_i^{(t)}$ is defined in (2.9) [i.e., the distribution $A_i^{(t)}$ restricted to (T_i, ∞) and renormalized to be a probability measure]. Similarly, for components in category C2, we generate an observation from $(A_i^{(t)})_{[0, T_i)}$. For a component in category C3, nothing needs to be done.

Suppose there are k components that fall into category C4. We then use an extension of Lemma 1 describing the conditional distribution of k independent random variables ($k > 2$), whose distributions have both absolutely continuous and discrete components, given the value of their maximum. The necessary formulas are easy to derive but require elaborate notation to write down explicitly, and so are not given here. We note, however, that the needed computer algorithm is relatively easy to implement.

We remark that the case of an arbitrary coherent system is no more difficult than that for a general parallel system if we note that the set of components in category C4 changes from system to system.

2.4.2. *Convergence of the algorithm.* When considering the case of an arbitrary system, it is helpful first to look at the situation when the prior distribution on each F_j is a single Dirichlet, that is, there is no mixing. In this case, the updating of the lifelength of component j in system i is based on (2.9), where $\alpha_{\theta^{(t)}}$ is replaced simply by α. When updating a component in category C1, the probability of drawing from the fixed probability measure proportional to $\alpha(\cdot \cap (T_i, \infty))$ is bounded below by $\alpha((T_i, \infty))/(\alpha((T_i, \infty)) + n - 1)$, independently of the current state of the chain. A similar statement holds for the lifelengths of components in category C2. We have already explained how to deal with the lifelengths for components in category C4 in Section 2.3. Thus a coupling argument along the lines of the proof of Theorem 1 gives convergence at a uniform geometric rate.

When the priors on the F_j's are mixtures of Dirichlets, a difficulty arises in that the distributions $\alpha_{\theta^{(t)}}$ in general need not have a uniform lower bound. For parallel systems we were able to find a uniform lower bound for the posterior distribution of θ given the lifelengths $\mathbf{X}$ only because $\mathbf{X}$ is known to lie in a compact set. Since the lifelengths of components in category C1 do not lie in a compact set, this argument no longer applies. For general systems convergence of the Markov chain can be established using the lower bounds established in Theorem 1 in conjunction with Theorem 1 of Athreya, Doss and Sethuraman (1996), which gives simple ergodicity (i.e., convergence, but not at a geometric rate).

3. Analysis of the U.S. Air Force C-17 fuel quantity computer data. We illustrate our algorithm on data involving survival times of the fuel quantity (FQ) computer system of the C-17 transport aircraft. The test program will eventually involve six aircraft being flown for approximately 10,000 cummulative hours. Our data set is taken relatively early in the test program, since only 2440 flight hours had been accumulated at the time of this writing. The data, listed below in Table 1, fall into one of three categories (we denote the failure times of the A-bus and B-bus as X and Y, respectively):

1. The FQ computer fails (both buses are dead) and the maximum survival time, say t_0, of the two buses is observed; that is, we have the usual autopsy statistics (system failure time and set of dead components). This type of observation has the form "$X \vee Y = t_0$."
2. The two components are checked at time t_1 and time t_2. Both buses are alive at t_1, but one of the buses, say B, is in a failed state at t_2. Even though the A-bus is alive at t_2, the FQ computer is replaced. This situation generates two observations, which have the form $Y \in (t_1, t_2]$ and $X \in (t_2, \infty)$.

TABLE 1
C-17 fuel quantity computer data: the first three lines contain observed FQ computer failure times; the next two categories (four observations) occurred when the FQ computer was replaced due to a failure in one bus; the remaining observations occurred because no more data is available

Aircraft Id	Event	Hours	Observation
P-1	FQ computer failed	43.4	$X_1 \vee Y_1 = 43.4$
P-1	FQ computer failed	236.8	$X_2 \vee Y_2 = 236.8$
P-2	FQ computer failed	244.0	$X_3 \vee Y_3 = 244.0$
T-1	A-bus and B-bus alive	11.9	
	B-bus dead	15.4	$Y_4 \in (11.9, 15.4]$
	FQ computer replaced	15.4	$X_4 \in (15.4, \infty)$
P-4	A-bus and B-bus alive	174.4	
	B-bus dead	181.8	$Y_5 \in (174.4, 181.8]$
	FQ computer replaced	181.8	$X_5 \in (181.8, \infty)$
T-1	A-bus and B-bus alive	819.6	$X_6 \in (819.6, \infty)$
	no more data available		$Y_6 \in (819.6, \infty)$
P-1	A-bus and B-bus alive	85.0	$X_7 \in (85.0, \infty)$
	no more data available		$Y_7 \in (85.0, \infty)$
P-2	A-bus and B-bus alive	476.4	$X_8 \in (476.4, \infty)$
	no more data available		$Y_8 \in (476.4, \infty)$
P-3	A-bus and B-bus alive	24.5	$X_9 \in (24.5, \infty)$
	independent software failure		$Y_9 \in (24.5, \infty)$
P-3	A-bus and B-bus alive	71.7	$X_{10} \in (71.7, \infty)$
	no more data available		$Y_{10} \in (71.7, \infty)$
P-4	A-bus and B-bus alive	68.4	$X_{11} \in (68.4, \infty)$
	no more data available		$Y_{11} \in (68.4, \infty)$
P-5	A-bus and B-bus alive	173.4	$X_{12} \in (173.4, \infty)$
	no more data available		$Y_{12} \in (173.4, \infty)$

3. Both components of the FQ computer are alive when the data are taken, but the aircraft had flown for t_3 hours. The failure times for both buses lie in the interval (t_3, ∞). Thus, two observations are generated: $X \in (t_3, \infty)$ and $Y \in (t_3, \infty)$.

Note that this data structure is a bit more complex than the data structure in the autopsy model; however, the required modifications to the algorithm involve no real difficulties.

The reader may wonder why there is a need for a computer, as opposed to a simple analogue gauge, to deal with fuel quantity. Indeed, this is not a frivolous issue. There is actually a need for a computer even during level flight, since the aircraft maintains its desired center of gravity via fuel transfer from one wing to another. This task is further complicated as the aircraft flies at different angles or possibly under turbulence. The FQ computer receives the current angle of flight from another computer and uses this information, along with readings from a series of probes in each fuel tank, to make accurate fuel quantity calculations. Also, the Mission Computer requires input from the FQ computer to make range calculations.

We analyzed the C-17 data using our proposed algorithm. We took our prior on both F and G to be (2.5), where H_θ is the exponential distribution with parameter θ (the mean is $1/\theta$). We assumed $\alpha_\theta(\mathscr{R})$ to be constant in θ and considered three cases: $\alpha_\theta(\mathscr{R}) = 1$, $\alpha_\theta(\mathscr{R}) = 10$ and $\alpha_\theta(\mathscr{R}) = 100$. We took $\nu = \mathscr{G}(a, b)$ (the gamma distribution with shape parameter a and scale parameter b), since we wished to center the prior around the family of exponential distributions, and the gamma is conjugate for this family. From (2.11), $\nu_{\mathbf{X}} = \mathscr{G}(a + \#(\mathbf{X}), b + \Sigma^{\mathrm{dist}} X_i)$, and $\mu_{\mathbf{Y}}$ is similarly defined.

We elicited the prior of a computer systems engineer from the C-17 Special Programs Office by asking his opinion about the FQ computer mean time before failure (MTBF) with respect to two reports. The first report, supplied by the C-17 manufacturer, provides target numbers for each "logical replaceable unit" (LRU), including the FQ computer. The manufacturer guarantees that the MTBF for each LRU, computed at the end of the acceptance testing period, will exceed that LRU's target number. The C-17 engineer thought it was highly likely (probability of 0.9) that the FQ computer MTBF would exceed the target number (which was 1300 hours). The second report, supplied to the C-17 Special Programs Office by the manufacturer's design group, contains a list of "mature" MTBF numbers for each LRU being evaluated. These numbers represent an average of MTBF's, by LRU, across many different aircraft which have similar LRU's. The data come from maintenance data accumulated following the acceptance testing periods for each aircraft (hence the word "mature"). Since these numbers come from mature aircraft, the C-17 engineer thought it was quite unlikely (probability of 0.1) that the FQ computer MTBF for the acceptance testing period would exceed the mature MTBF number (which was 3167 hours). Thus, we took 1300 hours and 3167 hours to be the 0.1 and 0.9 quantiles of the distribution of the MTBF for the FQ computer.

If $X \sim \mathscr{E}(\theta)$ (the exponential distribution with parameter θ) and $Y \sim \mathscr{E}(\psi)$, then MTBF $= E(X \vee Y) = (1/(\theta + \psi))[1 + \theta/\psi + \psi/\theta]$. If θ, ψ are iid $\sim \mathscr{G}(a, b)$, then the 0.1 and 0.9 quantiles of the distribution of the MTBF are equal to 1300 and 3167, respectively, when $a = 6.04424$ and $b = 6835.32$. Note that if the conditional distribution of X given θ is $\mathscr{E}(\theta)$ and θ is distributed as $G(a, b)$, then the unconditional cumulative distribution function of X is $F(t) = 1 - (b/(b + t))^a$, which is a "shifted Pareto" distribution with parameters a and b. The prior distributions of X and Y are each shifted Pareto distributions with parameters 6.04424 and 6835.32. The prior distribution of $X \vee Y$ is the product of two such distributions.

Of particular interest to the C-17 engineers is the question of how the lifelength of a future FQ computer, as well as the lifelengths of a future A-bus and B-bus, would be distributed. The Bayes approach is especially well-suited to answer such a question. Figures 1, 2 and 3 give the prior density and a representation of the posterior distribution for the future lifelengths of the C-17 FQ computer (maximum lifelength of the A-bus and B-bus), the A-bus and the B-bus, for the cases $\alpha_\theta(\mathscr{R}) = 1$, $\alpha_\theta(\mathscr{R}) = 10$ and $\alpha_\theta(\mathscr{R}) = 100$. Recall from Table 1 that three system failures were observed, at times 43.4, 236.8
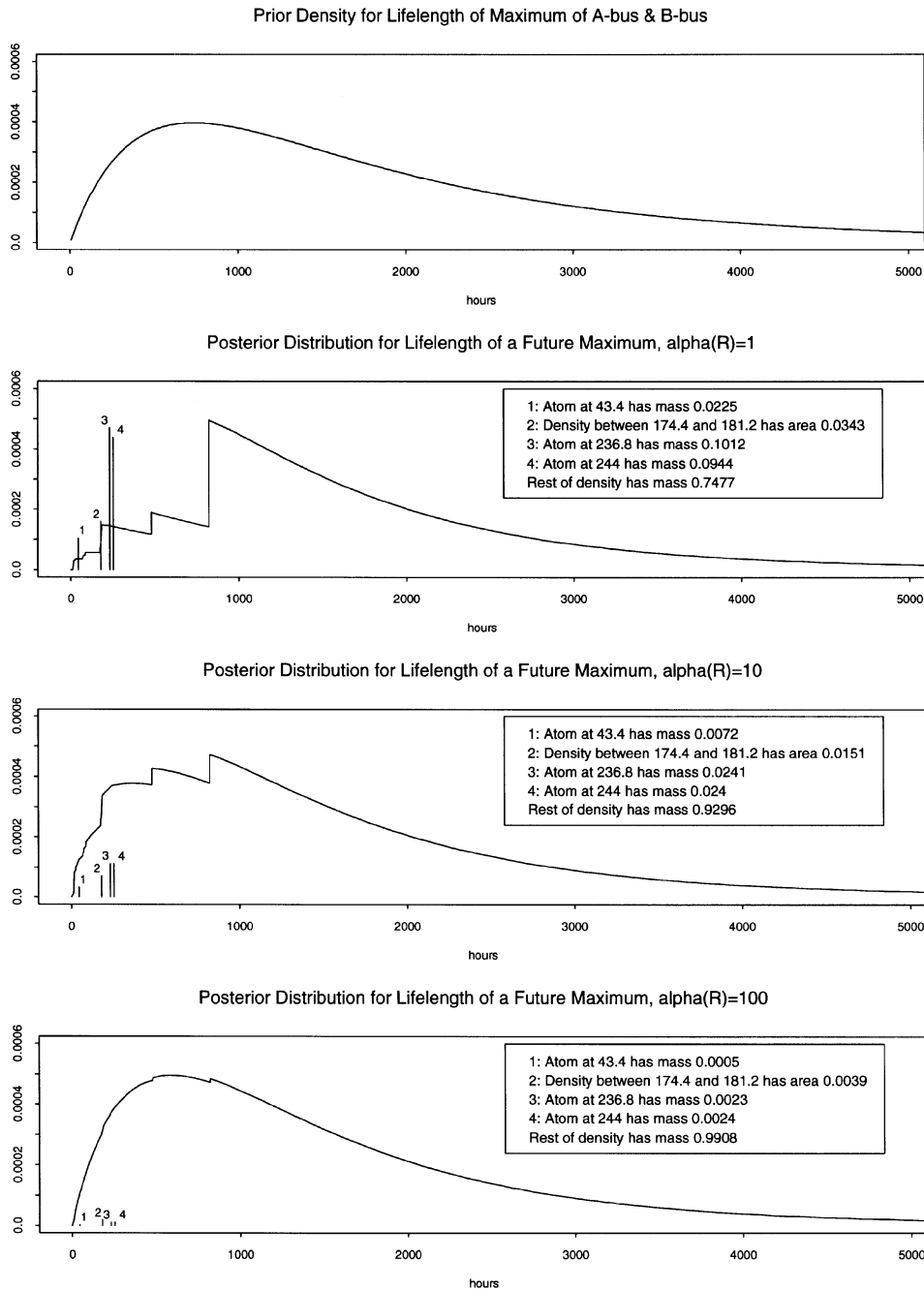
FIG. 1. *Prior density and estimates of the posterior distributions for the lifelength of the FQ computer, $\alpha(R) = 1$, 10 and 100; posterior distributions have both continuous and discrete components. Very sharp peaks in the density have been removed and replaced by spikes.*
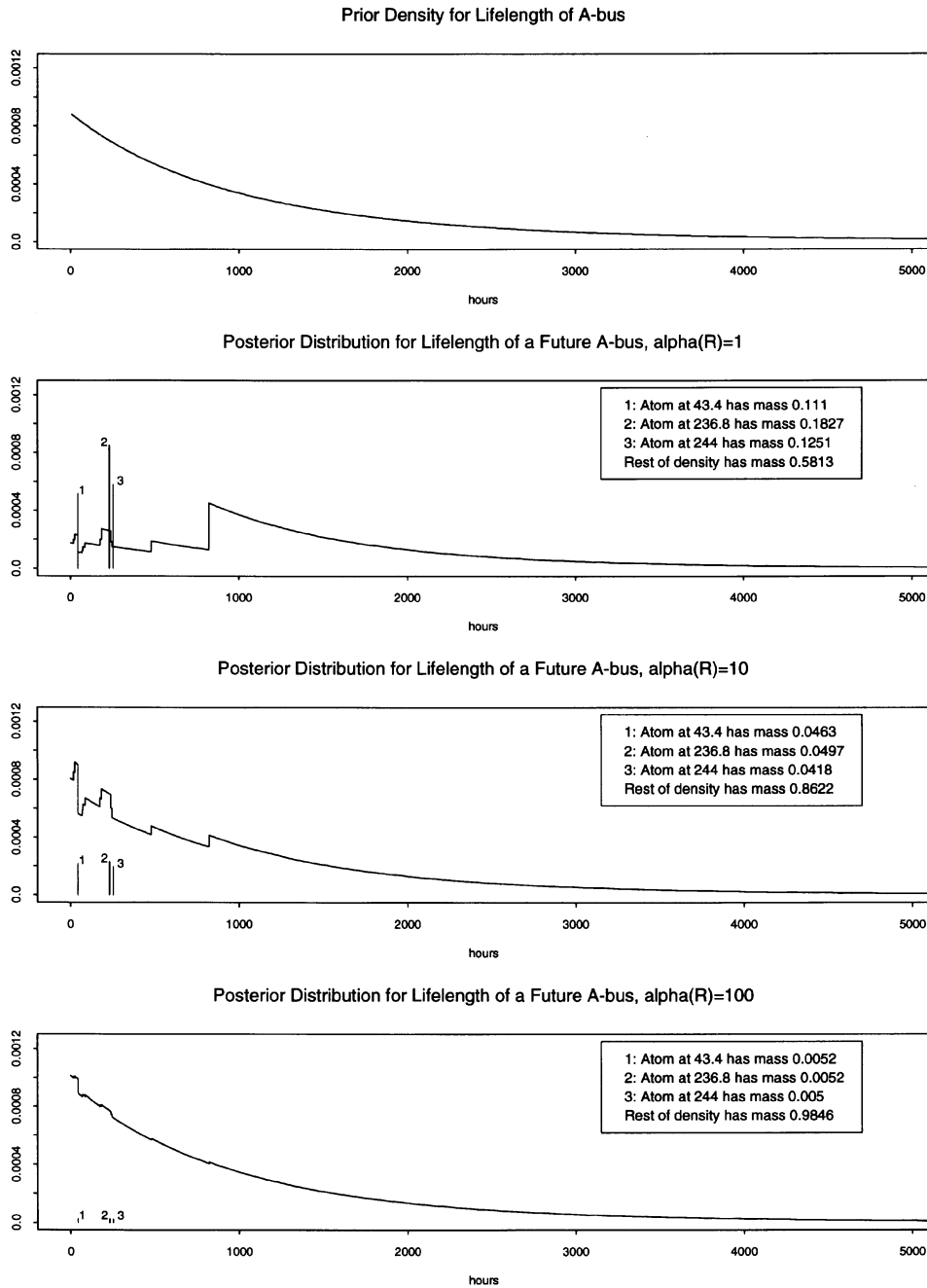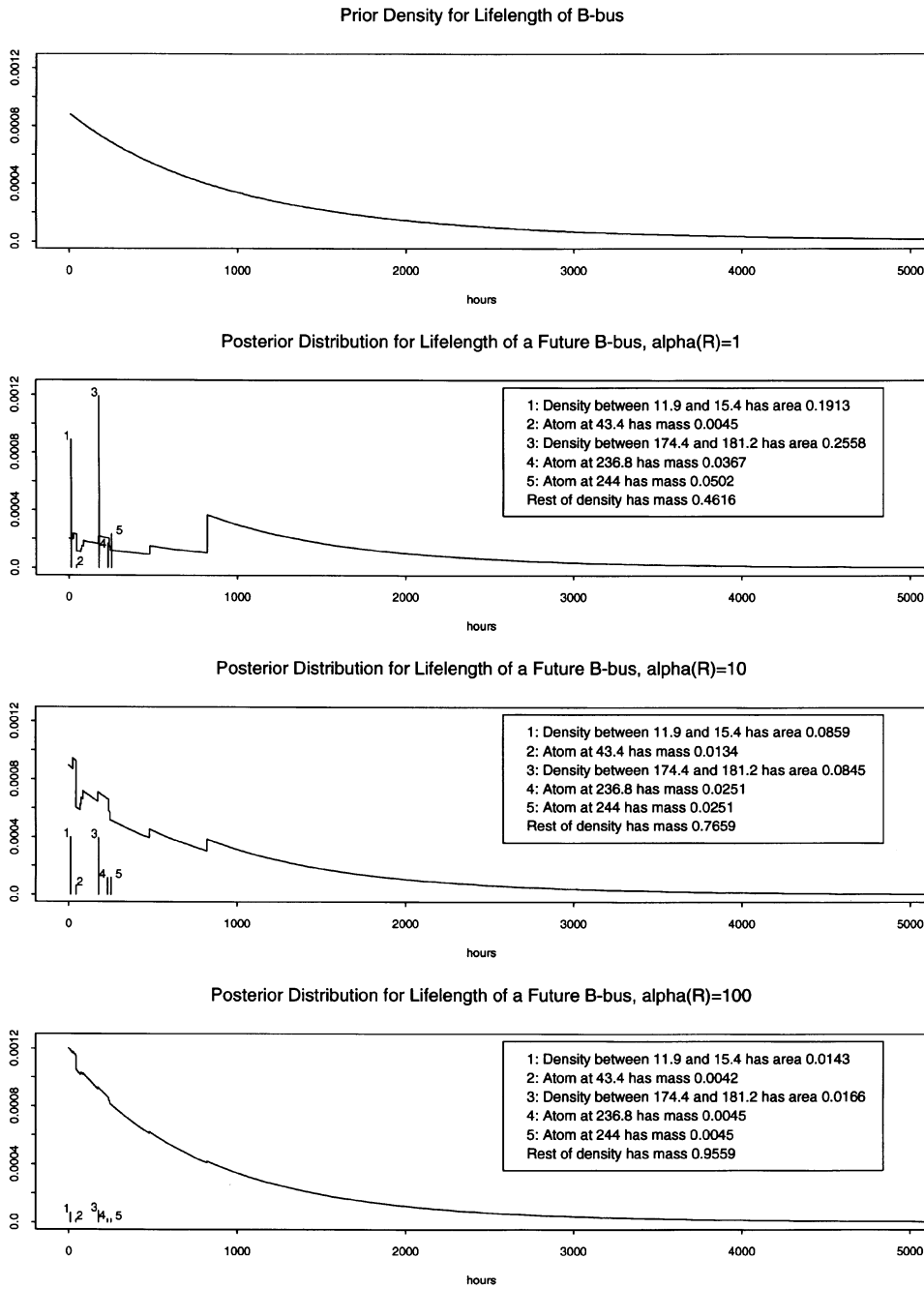
FIG. 2. *Prior density and estimates of the posterior distributions for the lifelength of the FQ computer's A-bus,* $\alpha(R) = 1,\ 10$ *and* 100; *posterior distributions have both continuous and discrete components. Very sharp peaks in the density have been removed and replaced by spikes.*

Fig. 3. *Prior density and estimates of the posterior distributions for the lifelength of the FQ computer's B-bus,* $\alpha(R) = 1$, 10 *and* 100; *posterior distributions have both continuous and discrete components. Very sharp peaks in the density have been removed and replaced by spikes.*

and 244.0. This causes all the posterior distributions to have atoms at these three failure times. The masses at these three failure times have been removed and plotted as distinct spikes, with their masses labeled separately. It is interesting to note that the posterior distribution for the B-bus gives large probability to the intervals (11.9, 15.4) and (174.4, 181.8). Even though the distribution has a density in those intervals, the mass there has been removed and plotted as spikes, in order to keep the scale on the vertical axis reasonable. The large mass there can be attributed to the two pairs of observations (X_4, Y_4) and (X_5, Y_5) (see Table 1), which arose from a failure in the B-bus while the A-bus was still functioning. For low values of $\alpha_\theta(\mathscr{R})$ [the cases $\alpha_\theta(\mathscr{R}) = 1$ and $\alpha_\theta(\mathscr{R}) = 10$], these two observations cause the algorithm to assign a fairly high value to the conditional probability that a future Y lies in one of these two intervals. As a consequence, the posterior distribution of the B-bus has much more mass in the lower tail than that of the A-bus. For example, when $\alpha_\theta(\mathscr{R}) = 1$, for the B-bus the mass to the left of 236.8 hours is 0.49 versus 0.16 for the A-bus, and when $\alpha_\theta(\mathscr{R}) = 10$ these two numbers are 0.34 and 0.21, respectively. From Figure 1 it can be seen that the masses at the three observed FQ computer failure times account for much of the mass in the conditional distribution of a future FQ computer for the cases $\alpha_\theta(\mathscr{R}) = 1$ and $\alpha_\theta(\mathscr{R}) = 10$ (0.22 and 0.06).

The means of the posterior distributions of the lifelengths of the FQ computer, for the cases $\alpha_\theta(\mathscr{R}) = 1$, $\alpha_\theta(\mathscr{R}) = 10$ and $\alpha_\theta(\mathscr{R}) = 100$, are 1443, 1536, and 1568 hours, respectively. At this relatively early stage of the study, we conclude that the performance of the FQ computer is not as good as the Air Force would like to see, as the means are only slightly above the minimum acceptable MTBF. However, we caution that the study will continue for an additional two years beyond the close of our data set, and that some of the early failures experienced can be directly attributed to ongoing design changes. It will be interesting to rerun the algorithm on the updated data set at the close of the acceptance testing period.

The plots were obtained by using the algorithm with the extra step and the method of Rao–Blackwellization discussed in Section 2.3 [see (2.13)]. The algorithm was run for 1,000,000 cycles, but because Rao–Blackwellization here consumes substantially more time than it takes to carry out one cycle, we Rao–Blackwellized only every five cycles, so our final estimates are based on averages of 200,000 replicates.

To see how much of an improvement results from the inclusion of the extra step and the use of Rao–Blackwellization, we computed estimates of the posterior mean of a future X value by each of four methods: we ran the basic algorithm and the algorithm with the extra step, and for each chain we formed the estimates of the posterior mean of a future X value without and then with Rao–Blackwellization. [Note that for the basic algorithm, Rao–Blackwellization is the method described immediately following (2.12) while, for the chain with the extra step, Rao–Blackwellization involves using the conditional distributions in (2.13).]

TABLE 2
Means and variances of estimates of the posterior expectation of the lifelength of a future A-bus

		$\alpha(\mathscr{R}) = 1$		$\alpha(\mathscr{R}) = 10$		$\alpha(\mathscr{R}) = 100$	
		Non-RB	**RB**	**Non-RB**	**RB**	**Non-RB**	**RB**
Basic algorithm	Mean	1008.0	1007.5	1093.5	1091.3	1133.3	1132.7
	Var	1968.5	1563.7	1325.1	404.2	1119.7	334.1
Algorithm with	Mean	1006.8	1009.8	1092.5	1089.2	1134.6	1134.2
extra step	Var	782.4	136.1	1052.2	202.6	937.1	301.4

Each chain was run for 400,000 complete cycles (after an initial burn-in period of 2000 cycles), and 200 batches of size 2000 were formed. Means and variances of the batch means were computed. Table 2 summarizes our results for three values of $\alpha(\mathscr{R})$. From the table we see that the variance of the estimates always decreases if we use the algorithm with the extra step or Rao–Blackwellize, and as expected, the gains from using the modified algorithm are greatest for small values of $\alpha(\mathscr{R})$.

Of course, Rao–Blackwellization may be hard to implement in some problems (e.g., when dealing with the posterior mean of the maximum of an A-bus and B-bus) because of the difficulty of computing the relevant conditional distributions. The comparisons in Table 2 do not address this issue, nor the fact that, when Rao–Blackwellizing, the gains in statistical efficiency may be outweighed by the increased computational and programming effort required.

REFERENCES

ANTOINE, R., DOSS, H. and HOLLANDER, M. (1993). On the identifiability in the autopsy model of reliability theory. *J. Appl. Probab.* **30** 913–930.

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

ARJAS, E. and GASBARRA, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statist. Sinica* **4** 505–524.

ATHREYA, K. B., DOSS, H. and SETHURAMAN, J. (1996). On the convergence of the Markov chain simulation method. *Ann. Statist.* **24** 69–100.

BARLOW, R. E. and PROSCHAN, F. (1981). *Statistical Theory of Reliability and Life Testing*. To Begin With, Silver Spring, Md.

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

BUSH, C. and MACEACHERN, S. (1996). A semi-parametric Bayesian model for randomized block designs. *Biometrika* **83** 275–285.

DOSS, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22** 1763–1786.

DOSS, H., FREITAG, S. and PROSCHAN, F. (1989). Estimating jointly system and component reliabilities using a mutual censorship approach. *Ann. Statist.* **17** 764–782.

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

GEYER, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statist. Sci.* **7** 473–511.

GILKS, W. R. and WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41** 337–348.

LAWSON, K. L. (1994). Bayesian nonparametric estimation via Gibbs sampling for coherent systems with redundancy. Ph.D. dissertation, Dept. Statistics, Florida State Univ.

MEILIJSON, I. (1981). Estimation of the lifetime distribution of the parts from the autopsy statistics of the machine. *J. Appl. Probab.* **18** 829–838.

MEILIJSON, I. (1994). Competing risks on coherent reliability systems: estimation in the parametric case. *J. Amer. Statist. Assoc.* **89** 1459–1464.

NOWIK, S. (1990). Identifiability problems in coherent systems. *J. Appl. Probab.* **27** 862–872.

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

TANNER, M. A. and WONG, W. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550.

WATELET, L. (1990). Nonparametric estimation of component life distributions in Meilijson's competing risks model. Ph.D. dissertation, Dept. Statistics, Univ. Washington.

WEST, M., MÜLLER, P. and ESCOBAR, M. D. (1994). Heirarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty*: *A Tribute to D. V. Lindley* (P. Freeman and A. F. M. Smith, eds.) 363–386. Wiley, New York.

HANI DOSS
DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210
E-MAIL: doss@stat.ohio-state.edu

FRED W. HUFFER
DEPARTMENT OF STATISTICS
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FLORIDA 32306
E-MAIL: huffer@stat.fsu.edu

KEVIN L. LAWSON
ABBOTT DIAGNOSTICS DIVISION
ABBOTT LABORATORIES
200 ABBOTT PARK ROAD
DEPT. 87F, AP31
ABBOTT PARK, ILLINOIS 60064