Bayesian Inference of Survival Probabilities, Under Stochastic Ordering Constraints
Author(s): Elja Arjas and Dario Gasbarra
Source: *Journal of the American Statistical Association*, Vol. 91, No. 435 (Sep., 1996), pp. 1101-1109
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2291729
Accessed: 11/05/2010 06:54

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=astata.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association.*

http://www.jstor.org

Bayesian Inference of Survival Probabilities, Under Stochastic Ordering Constraints

Elja ARJAS and Dario GASBARRA

In the statistical analysis of survival data arising from two populations, it often happens that the analyst knows, a priori, that the life lengths in one population are stochastically shorter than those in the other. Nevertheless, survival probability estimates, if determined separately from the corresponding samples, may not be consistent with this prior assumption, because of inherent statistical variability in the observations. This problem has been considered in a number of papers during the past decade, by adopting a (generalized) maximum likelihood approach. Our approach is Bayesian and, in essence, nonparametric. The a priori assumption regarding stochastic ordering is formulated naturally in terms of a joint prior distribution defined for pairs of survival functions. Nonparametric specification of the model, based on hazard rates and using a few hyperparameters, allows for sufficient flexibility in practical applications. The numerical computations are based on a coupled version of the Metropolis–Hastings algorithm. The results from a statistical analysis are summarized nicely by a pair of predictive survival functions that are consistent with the assumed stochastic ordering.

KEY WORDS: Hazard rate; Markov chain Monte Carlo integration; Predictive distribution.

1. INTRODUCTION

A situation frequently encountered in the estimation of survival distributions from two samples is that the analyst knows, a priori, that the corresponding life lengths are stochastically ordered. For example, when considering survival data arising from a manufacturing process, one may know that of two production batches, one was made from cheaper and technically inferior raw materials than the other. Similarly, in an animal experiment it may be known in advance that exposure to some toxic substance can only shorten the lives of the animals, and thus the survival probabilities in the exposed group should be smaller than those in the control group. It is a natural requirement that the estimated survival functions should be consistent with such a stochastic ordering assumption. However, if the estimation is done separately from the two subsamples, and in particular if the samples are small, then this may not hold, because of inherent variability in the observations.

An obvious solution to this inferential problem is to apply constrained maximum likelihood estimation with the stochastic ordering condition being valid. It is commonly held, however, that parametric models may not offer sufficient flexibility for a realistic description of survival data. Therefore, nonparametric, or generalized maximum likelihood (ML), estimation methods may seem more appropriate. Brunk, Franck, Hanson, and Hogg (1966) considered the constrained estimation problem from complete (uncensored) data, thereby arriving at discrete estimators corresponding to two stochastically ordered empirical distributions functions. Dykstra (1982) extended these results to right-censored survival data and derived estimators that correspond to stochastically ordered Kaplan–Meier curves. An interpretation of Dykstra's estimators is that in cases where the ordinary Kaplan–Meier curves determined separately from the two subsamples would intersect, the corresponding risk sets are adjusted by the same amounts up and down at conveniently chosen time points, so that the Kaplan–Meier curves stay apart. The general case of N stochastically ordered survival functions was considered by Feltz and Dykstra (1985), and by Dykstra and Feltz (1989). The corresponding estimation problem under the stronger hypothesis of uniform conditional stochastic ordering (which for continuous distributions is equivalent to hazard rate ordering) was considered by Dykstra, Kochar, and Robertson (1991).

Our approach to this problem is Bayesian and, in essence, nonparametric. This means that we view the unknown survival functions as random elements in some large enough function space D and describe our (prior) knowledge by a probability model on that space. In the constrained estimation problem, we are then led to consider pairs of survival functions, say $(\overline{F_1}, \overline{F_2}) \in D \times D$, and specify the prior so that it is supported only by those $(\overline{F_1}, \overline{F_2})$ for which $\overline{F_1} \succcurlyeq_{st} \overline{F_2}$; that is, $\overline{F_1}(t) \geq \overline{F_2}(t)$ for all $t \geq 0$. Regardless of data, the posterior is then supported only by this part of the product space.

The model is parameterized in terms of hazard rates rather than distributions or survival functions. Hazard rates have a straightforward intuitive interpretation, their properties are well understood, and they are relatively easy to quantify in practice. These are important aspects, particularly in Bayesian inference, which involves an explicit specification of the prior. In principle, the only mathematical restriction is that hazard rates must be nonnegative functions of time. Our choice here is to also use a fixed time grid and assume that the hazard rates are piecewise constant on the grid intervals. Such a convention appears to be a reasonable compromise between sufficient flexibility and realistic description on the one hand and computational feasibility on the other. Each segment of the hazard rate on a finite interval is then parameterized in a finite manner, which enables us to use Markov chain Monte Carlo (MCMC) integration techniques in the numerical computations.

Elja Arjas is Professor and Dario Gasbarra is Research Assistant, Department of Mathematical Sciences, University of Oulu, SF-90570 Oulu, Finland. This study was supported by a grant from the Academy of Finland. The authors are grateful to the referees for their constructive comments.

© 1996 American Statistical Association
Journal of the American Statistical Association
September 1996, Vol. 91, No. 435, Theory and Methods

From a technical viewpoint, this article is a follow-up to earlier work (Arjas and Gasbarra 1994) where, in a one-sample situation, the hazard rate parameters had a similar piecewise constant structure except that the positions of the changepoints, and even their number, were also allowed to be random. The present method could be modified in a similar manner.

This article is organized as follows. Section 2 formulates our problem and the statistical model in exact terms. Section 3 describes in detail a coupled modification of the algorithm of Arjas and Gasbarra (1994), which is needed to generate the hazard rate processes from the constrained posterior distribution. Section 4 contains an example based on empirical data and a brief discussion on model checking. Finally, Section 5 provides some additional remarks. To make the article more self-contained, a short Appendix explains the ideas underlying our coupling method.

2. THE STATISTICAL MODEL

We consider simple right-censored survival data $\{(X'_j, \delta'_j); 1 \le j \le m'\} \cup \{(X''_k, \delta''_k); 1 \le k \le m''\}$ arising in a two-sample study; here X'_j (resp. X''_k) is the time that the jth individual in the first (the kth individual in the second) subsample was last seen and δ'_j (δ''_k) is the indicator of an observed failure at X'_j (X''_k).

All of the individuals in the same subsample are assumed to have a common hazard rate for failure, $\{\lambda'_t; t \ge 0\}$ and $\{\lambda''_t; t \ge 0\}$. These functions become the parameters of our model. Because a hazard rate cannot be directly observed, we shall treat it as a random function, or stochastic process, also assigning probabilities to its sample paths. It seems most natural to think about each sample path as corresponding to an individual's proneness to failure in an infinite collection of "similar" (i.e., exchangeable) individuals or objects. A probability on the space of hazard rate sample paths can be understood either subjectively, as a description of uncertainty regarding such proneness, or in terms of a superpopulation over different proneness classes.

We now describe our statistical model. Because the treatment of $\{\lambda'_t; t \ge 0\}$ and $\{\lambda''_t; t \ge 0\}$ is symmetric, we often omit the primes from the notation. We fix a time grid by dividing the observation interval $[0, T_{\max}]$ into N equally long subintervals $0 = t_0 < t_1 < \cdots < t_N = T_{\max}, t_{i+1} - t_i = \Delta t = T_{\max}/N$. Then we assume (cf. Arjas and Gasbarra 1994) that both λ' and λ'' have the following structure (dropping the primes from the notation). The hazard rate sample paths $\{\lambda_t; t \ge 0\}$ are constant on the subintervals; therefore, they can be parameterized by vectors $(\lambda_1, \ldots, \lambda_N)$ belonging to $D = (\mathbb{R}^+)^N$ by setting $\lambda_i = \lambda(t_i)$. The unconstrained prior distributions are then specified by considering sequentially the values λ_i at the grid points:

a. The initial level λ_1 has distribution *gamma*(α_0, β_0) (with α_0 as the shape parameter and β_0 as the scale parameter).

b. Given $(\lambda_1, \ldots, \lambda_{i-1})$, λ_i has distribution *gamma*(α, β_i), where $\beta_i = \alpha/\lambda_{i-1}$.

Here α_0, β_0, and α are given hyperparameters controlling the initial level and constancy of the hazard rate. Obviously, they can be chosen differently for λ' and λ''. In penalized ML estimation they would be interpreted in terms of costs, with larger and more frequent oscillations involving a higher cost.

Note that the conditional expected values satisfy

$$E_{\text{prior}}(\lambda_i | \lambda_1, \ldots, \lambda_{i-1}) = \lambda_{i-1}, \quad i \ge 1. \tag{1}$$

We have thus made here the neutral (i.e., discrete time martingale) prior assumption that unconstrained hazard rates do not have upward or downward trends. We also see that the corresponding conditional standard deviation is given by $\sqrt{\text{Var}_{\text{prior}}(\lambda_i | \lambda_1, \ldots, \lambda_{i-1})} = \lambda_{i-1}/\sqrt{\alpha}$. In a sense, therefore, the prior variability of the hazard rate is proportional to its current value and inversely proportional to the square root of the hyperparameter α. Letting $\alpha \to \infty$ would correspond to fitting an exponential model to the data.

We have provided a more detailed discussion of the rationale behind this type of prior in earlier work (Arjas and Gasbarra 1994). We note, however, that only the algorithm used for numerical integration depends on this particular choice: If some other family of priors seems more appropriate in a concrete problem, then the algorithm could be adjusted accordingly.

For both subsamples the (unconstrained) posterior density of the hazard is proportional to the product of the prior density and the likelihood. The prior density of a segment $\{\lambda(t); 0 \le t \le T_{\max}\}$ of the hazard rate can now be written as $\gamma(\lambda_1; \alpha_0, \beta_0) \prod_{i=2}^{N} \gamma[\lambda_i; \alpha, \alpha/(\lambda_{i-1})]$, where we denote the density of the distribution gamma(α, β) by $\gamma(\cdot\,; \alpha, \beta)$. Assuming that the censoring mechanism is noninformative with regard to the hazard rate, the likelihood will be proportional to the usual product form

$$\prod_{j=1}^{m} \left[\lambda(X_j)^{\delta_j} \exp\left\{ -\int_0^{X_j} \lambda(s)\, ds \right\} \right]$$

$$= \prod_{j=1}^{m} (\lambda(X_j))^{\delta_j} \cdot \exp\left\{ -\int_0^{T_{\max}} Y(s)\lambda(s)\, ds \right\}, \tag{2}$$

where $T_{\max} = \max_{1 \le j \le n} X_j$ is the largest observation time and $Y(t) = n - \sum_{j=1}^{n} 1_{\{X_j < t\}}$ is the number of individuals at risk at time t.

Let us then return to our original constrained estimation problem. The hypothesis that the "true" hazard rates satisfy the (partial) ordering $\int_0^t \lambda'(s)\, ds \ge \int_0^t \lambda''(s)\, ds$ for all $t \in [0, T_{\max}]$ is henceforth denoted simply by $\lambda' \succcurlyeq \lambda''$. Our approach to solving the problem is constructive, and so we are led to consider the product space $D \times D$ or, more exactly, the subset

$$S = \{(\lambda', \lambda'') \in D \times D: \lambda' \succcurlyeq \lambda''\}$$

$$= \left\{(\lambda'_1, \ldots, \lambda'_N, \lambda''_1, \ldots, \lambda''_N) \in (\mathbb{R}^+)^{2N}: \sum_{i \le k} \lambda'_i \ge \sum_{i \le k} \lambda''_i, 1 \le k \le N \right\}. \quad (3)$$

It is a well-known consistency property of the Bayesian updating procedures that if the prior is supported completely by a subset of the parameter space, then so is the posterior. Thus a direct approach to solving the constrained estimation problem would start from the specification of a joint prior on S.

This could well be an elaborate task in practice. It would be natural to try to construct a joint prior supported by S from two given marginals on D. A first suggestion (which we in fact followed in an earlier version of this work) would be to form the product measure on $D \times D$ (corresponding to independence) and then simply restrict it to S and normalize to a probability measure. But although this procedure is simple, the resulting prior is rather unintuitive and has the drawback that its marginals do not coincide with the original univariate priors. Thus we have looked for a more satisfactory alternative. As a first guideline, we have the well-known theorem by Strassen (1965). Applied to the present context; this states that if π' and π'' are priors on D that are stochastically ordered (with respect to the partial order $\succcurlyeq$ on $(\mathbb{R}^+)^N$), then there is a joint prior on $D \times D$ supported completely by S and with marginals π' and π''. Although Strassen's theorem guarantees only existence, we actually have also an obvious construction, by stochastic monotonicity, if the shape parameters α of the level distributions in Condition b earlier are the same for λ' and λ''. Unfortunately, however, there is a further problem: If we make the simple prior postulate that $\pi' = \pi''$, then the constructed λ' and λ'' will coincide, thus leading to a singular joint prior π supported by the diagonal $\{(\lambda', \lambda'') \in D \times D: \lambda' = \lambda''\}$. Such a complete coupling is not desirable, as it would be inherited by the joint posterior regardless of data. Hence we look for yet another way of constructing joint (prior and posterior) distributions that would be supported by S.

In parallel with this, we entertain the following idea. If the original hypothesis $\lambda' \succcurlyeq \lambda''$ is valid, then, if the one-sample posteriors in D are determined from the respective subsamples without constraining the estimation procedure, they should be (at least "almost") stochastically ordered in D. Conversely, if we find a way of coupling the one-sample posterior distributions such that the pairs (λ', λ'') are "almost only" in S, then this shows that the stochastic ordering hypothesis agrees well with the data. We return to this question at the end of Section 4.

In view of these remarks, to form such a coupling we build (given the data and by using a version of the Metropolis–Hastings algorithm) two ergodic Markov chains on the parameter space D with the respective one-sample posteriors as equilibrium distributions. Then, by coupling the Markov chains, we form a coupling of the equilibrium distributions. (For a more detailed explanation, see the Appendix.) Furthermore, conditioning the initial distribution and the transition probabilities of the coupled chain to the subset $S \subset D \times D$ gives a coupled constrained chain. In our case this chain is irreducible, with the constrained posterior as the equilibrium distribution, and thus it is ergodic with the same limit. A detailed description of the algorithm is given in Section 3.

3. A COUPLED AND CONSTRAINED METROPOLIS–HASTINGS ALGORITHM

The Metropolis–Hastings algorithm is a simulation method that enables one to draw random samples from a given target distribution (here the posterior). This happens by constructing an ergodic Markov chain whose equilibrium distribution is the given target. The algorithm is particularly useful in the computation of integrals when analytic or numerical integration, or straightforward Monte Carlo, are impossible to use in practice. Because of ergodicity, the terms of the chain, although dependent, can be used in the empirical approximation of integrals as if they were independent. (For applications of MCMC methods to Bayesian inferential problems, as well as for more references, see, e.g., Gelfand and Smith 1990, Roberts and Smith 1994 and Tierney 1993, for MCMC in order-restricted problems, see Gelfand, Smith, and Lee 1992.)

The general form of the Metropolis–Hastings algorithm can be sketched in the following way. Being currently in state $x^j \in X$, we choose a suitable "proposal transition density" $q(x^1, x^2)$; we then generate a new value x^* according to the distribution $q(x^j, \cdot)$ and compute the *Hastings ratio,*

$$R = \frac{p(x^*)}{p(x^j)} \frac{q(x^*, x^j)}{q(x^j, x^*)}. \quad (4)$$

The state x^{j+1} is now determined as follows:

with probability $\min(1, R)$, choose $x^{j+1} := x^*$;
otherwise, $x^{j+1} := x^j$.

Under some regularity conditions, the sequence of successive states $\{x^j\}$ will form an ergodic Markov chain with equilibrium distribution $p(x)$. The advantage of this algorithm is that it uses only the ratios of target densities, and thus such densities need to be specified only up to a proportionality constant. For example, in Bayesian computations the target is the posterior density $p(x|\text{data})$, but this is proportional in x to the joint density $p(x, \text{data})$.

In this version of the algorithm, the state space is $(\mathbb{R}^+)^N \times (\mathbb{R}^+)^N$ and a transition always concerns one coordinate λ_k of $(\lambda', \lambda'') \in D \times D$ at a time. (In this respect we mimic the Gibbs sampler algorithm, where the parameters are updated one by one from the full conditional distributions and the resulting transition is always accepted; however, because in our case the sampling from the full conditional is not straightforward, we choose another proposal distribution and perform the consequent Metropolis–Hastings acceptance–rejection step.)

The algorithm is symmetric for λ' and λ'', and thus we omit the primes from the notation, referring to the one currently being updated using the corresponding data subsample by λ.

Note that the full joint density $p(\lambda', \lambda'', \text{data})$ is proportional (in λ_k) to

$$\lambda_k^{(r_k-1)}\exp\left\{-\lambda_k\left(\beta_k + \int_{t_k}^{t_{k+1}} Y(s)\,ds\right)\right\} \times \exp\left\{-\frac{\alpha\lambda_{k+1}}{\lambda_k}\right\} = f_\zeta(\lambda_k)\cdot g_\zeta(\lambda_k), \quad (5)$$

where $Y(s)$ corresponds to the same data subsample as λ_k and the functions f_ζ (resp. g_ζ) are defined by

$$f_\zeta(\lambda) = \lambda^{(\zeta+r_k-1)}\exp\left\{-\lambda\left(\beta_k + \int_{t_k}^{t_{k+1}} Y(s)\,ds\right)\right\} \quad (6)$$

and

$$g_\zeta(\lambda) = \left(\frac{1}{\lambda}\right)^\zeta \exp\left\{-\frac{\alpha\lambda_{k+1}}{\lambda}\right\} \qquad (\zeta > 0). \quad (7)$$

Note also that f_ζ is proportional to the gamma density

$$\gamma\left(\cdot\,;\zeta + r_k, \beta_k + \int_{t_k}^{t_{k+1}} Y(s)\,ds\right)$$

(we use the same notation f_ζ after normalization), and for $\zeta > 1$, the function g_ζ is proportional to an inverse gamma density. Solving the equation

$$\frac{\zeta + r_k - 1}{\beta_k + \int_{t_k}^{t_{k+1}} Y(s)\,ds} = \frac{\alpha\lambda_{k+1}}{\zeta} \quad (8)$$

for ζ, we can find a value ζ^* such that f_{ζ^*} and g_{ζ^*} have their modes at the same point. Indeed, ζ^* coincides with the mode of the full conditional density of λ_k.

Finally we construct a Metropolis–Hastings algorithm by using the proposal distribution

$$q(x^1, x^2) = q(\lambda_k) = f_{\zeta^*}(\lambda_k). \quad (9)$$

The algorithm has three steps:

Step 1: Generate λ^* according to the distribution $q(\cdot)$:

Step 2: Compute the Hastings ratio

$$R = \frac{g_{\zeta^*}(\lambda^*)}{g_{\zeta^*}(\lambda_k^{\text{old}})}: \quad (10)$$

Step 3: Generate U uniform in [0, 1]. If $R > U$, then move to a new state with $\lambda_k^{\text{new}} = \lambda^*$; otherwise, stay at λ_k^{old}. (We have chosen $\zeta = \zeta^*$ to improve the acceptance probability for λ_k^{new}.)

At this stage we have built two independent Markov chains whose equilibrium distributions are the respective posteriors. Given the starting values, the chains can be coupled by coupling the transition probabilities. This actually is done by using the same sequence of pairs of [0, 1]-uniform random variables (i.e., one for generation of the proposals and another for the acceptance–rejection step) in the updating steps for both parameters. Given the data and the parameter coordinates λ_h' and λ_h'' for all $h \neq k$, denoting by $G'(\cdot)$ and $G''(\cdot)$ the univariate distribution functions of the respective proposal distributions for λ_k' and λ_k'', we first generate a random variable V from the uniform distribution in [0, 1] and independent from the other variables, and then take $\lambda'^* = G'^{-1}(V)$ and $\lambda''^* = G''^{-1}(V)$. We generate another random variable, U, also uniform in [0, 1], independently to perform both acceptance–rejection steps.

As a result of this procedure, we get a coupled Markov chain. In the Appendix we show that it is positive recurrent and that its invariant probability measure is a coupling of the marginal posteriors.

To constrain the coupled chain to $S \subset D \times D$, the algorithm is started from a pair with $\lambda' \succcurlyeq \lambda''$, and all of the transition probabilities are conditioned to stay in S. Before updating a coordinate λ_k', conditionally on $\lambda_h', h \neq k, \lambda''$ and the first data subsample, we compute the interval I such that $\lambda_k \in I$ iff $\lambda' \succcurlyeq \lambda''$; I is nonempty because the old value $\lambda_k'^{(\text{old})} \in I$, because we started from a configuration in S. Then, to constrain the transition probability, it is enough to condition the proposal gamma density f_{ζ^*} to the interval I. Because S is connected in $D \times D$, it is easy to see that the coupled chain is irreducible and that it will converge to the equilibrium distribution of the coupled chain, conditioned to S. Therefore, we are generating coupled and constrained Markov chains, whose equilibrium distributions are the constrained posteriors. Note that, provided we have started from an ordered pair $(\lambda'^{(0)}, \lambda''^{(0)})$, the equilibrium distributions will not depend on the choice of the initial distributions.

Remarks. At first look, our coupling of the posterior distributions of the parameters (Λ', Λ'') may seem quite ad hoc. We have not defined a joint prior for the parameters and then obtained the corresponding joint posterior using Bayes's formula. Instead, only the marginals of the prior are given, and the coupling of the posteriors is constructed by a Markov chain device a posteriori; that is, after observing the data (X', X'').

Nevertheless, by the same procedure we are able to construct also a compatible joint distribution for $(\Lambda', \Lambda'', X', X'')$.

Consider indeed a Markov chain $\{\Lambda_n, X_n\}$ with transition kernel density $k((\lambda, \xi) \to (l, x)) = p(x|\lambda)p(l|x)$. By construction, this Markov chain admits the invariant distribution that has density $p(\lambda, x) = p(\lambda)p(x|\lambda)$. If we construct in this way two independent chains, one for (Λ', X') and another for (Λ'', X''), then their joint invariant density will be the product of these invariant densities; that is, $p(\lambda')p(x'|\lambda')p(\lambda'')p(x''|\lambda'')$. A coupling of these distributions can be constructed by coupling the transitions of the two Markov chains. Given some initial values Λ_n', Λ_n'', we first sample independently the observations $X_{n+1}' \simeq p(x'|\Lambda_n')$ and $X_{n+1}'' \simeq p(x''|\Lambda_n'')$, and then sample jointly $(\Lambda_{n+1}', \Lambda_{n+1}'')$ from a coupling of $p(\lambda'|X_{n+1}')$ and $p(\lambda''|X_{n+1}'')$. In this way we obtain a (partially) coupled Markov chain. By using the fact that the marginal chains are positive recurrent, we show in the Appendix that this

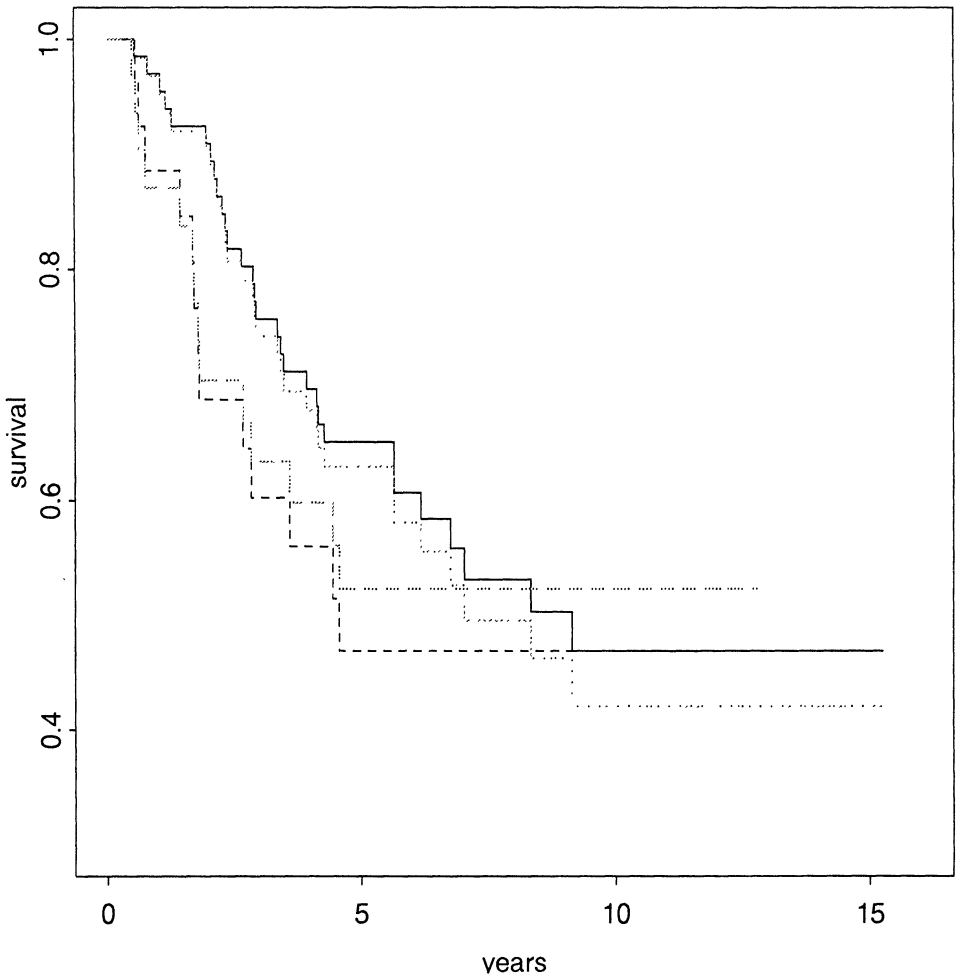

Figure 1. Kaplan–Meier and Constrained Dykstra Estimators of the Survival Functions of the Two Groups of Melanoma Patients after Surgery. ····· Kaplan–Meier estimator group 1; ∼∼∼∼∼ Kaplan–Meier estimator group 2; – – – Dykstra estimator group 1; —— Dykstra estimator group 2.

coupled chain is also positive recurrent. Therefore there is a joint invariant distribution for $(\Lambda', \Lambda'', X', X'')$, which is consistent with the coupling of the posteriors obtained by fixing any data in the algorithm and with the original (product) likelihood. On the other hand, it is tautological that, when it exists, the invariant distribution of the coupled chain must be a coupling of the invariant distributions of the marginal chains.

A corresponding prior for the parameters (Λ', Λ'') is obtained by marginalization, and again it is a coupling of the original priors. (Note that this coupling of the priors will generally depend on the dimension of the observed vector.)

The constrained distributions again are obtained by conditioning on $\{(\Lambda', \Lambda'') \in S\}$.

4. A NUMERICAL EXAMPLE

As mentioned previously, our numerical calculations are based on MCMC integration. An ergodic Markov chain $\{x^i = (\lambda', \lambda'')^i\}_{i\geq 0} \subset S$ has been constructed such that its invariant distribution is the constrained joint posterior $p(\lambda', \lambda''|\text{data}, \lambda' \succcurlyeq \lambda'')$. Consequently, for each real function Ψ: $S \to \mathbb{R}$ integrable with respect to the posterior, we have that almost surely

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Psi(x^i) = \int_S \Psi(x)\, dP(x|\text{data}). \tag{11}$$

Therefore, we can approximate numerically the posterior expectation of any integrable function Ψ by using hazard rates generated by the algorithm.

We obtain the constrained predictive survival functions $\Pr_S(X > t|\text{data})$ by considering in (11), for each time point t, the functional $\Psi(\lambda, t) = \exp\{-\int_0^t \lambda(\tau)\, d\tau\}$, where we choose $\lambda = \lambda'$ or λ''. Because the coupling was constrained to $S \subset D \times D$, it follows that $\Pr_S(X' > t|\text{data}) \leq \Pr_S(X'' > t|\text{data})$ for all $t > 0$. Using the generic shorthand notation $\bar{F}_{\text{pred}}(t)$ for such a survival function, the corresponding density $f_{\text{pred}}(t)$ is approximated in the simulation by considering in (11), for each fixed $t, \Psi(\lambda, t) = \lambda(t) \exp\{-\int_0^t \lambda(\tau)\, d\tau\}$. The natural notion of hazard, called here the *predictive hazard,* is defined as $\lambda_{\text{pred}}(t) = f_{\text{pred}}(t)/\bar{F}_{\text{pred}}(t)$ and corresponds to the hazard, according to the constrained posterior distribution, of some "new" individual (not in the data but "similar" in the sense of belonging to the same subpopulation) still alive at t. (Note that this is not the same as the posterior expecta-

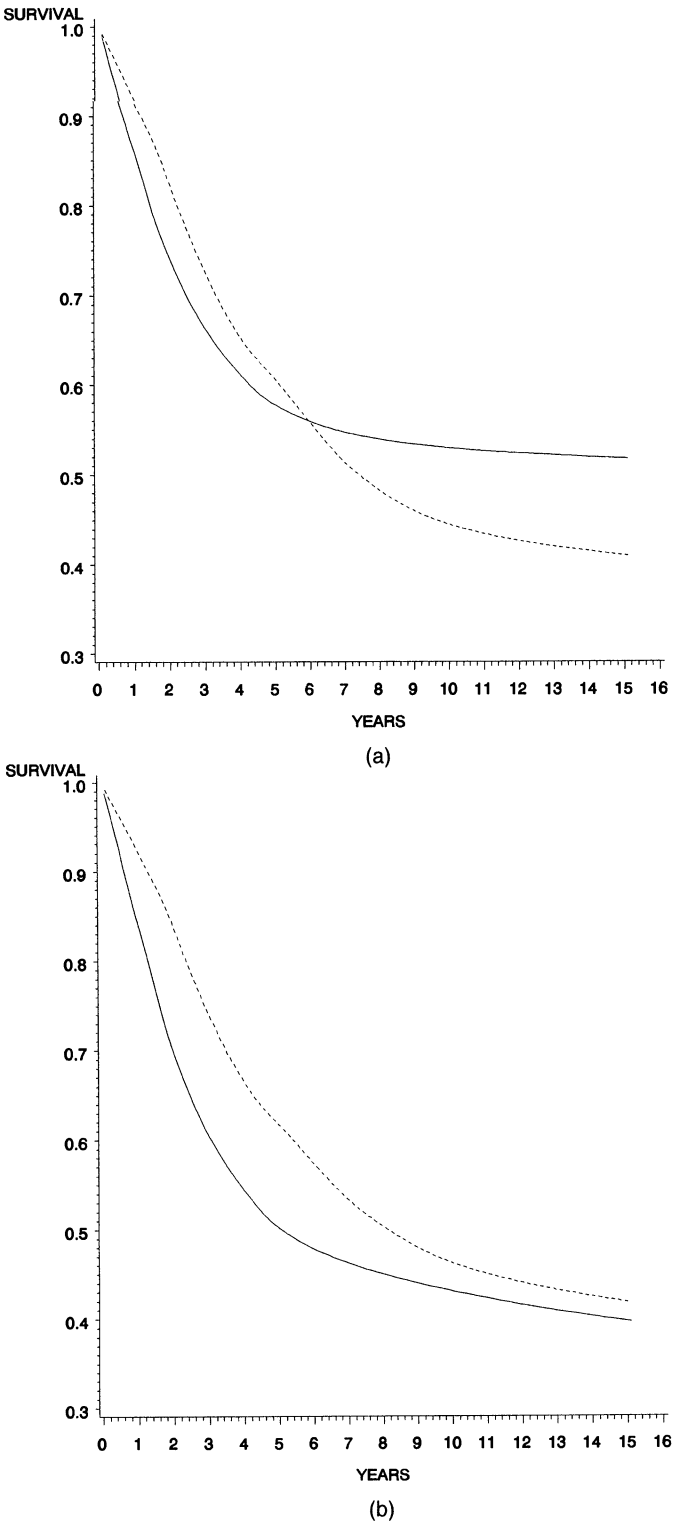

Figure 2. Predictive Survival Functions for the Two Groups (a) Under the Unconstrained Model and (b) Under the Constrained Model. The solid line represents group 1; the dashed line, group 2.

tion of the hazard rate, which would be approximated by $1/n\sum_{i=1}^{n}\lambda^i(t)$. The predictive hazard $\lambda_{\text{pred}}(t)$ is conditioned on both the data set and on the event that the new individual survives at least up to time t.)

As an illustration, we consider a data set studied in the monograph of Andersen, Borgan, Gill, and Keiding (1993) consisting of (eventually right-censored) survival times of patients with malignant melanoma after a surgery. Here we consider two groups: patients with tumor thickness > 5 mm (group 1, with 32 individuals) and patients with tumor thickness between 2 and 5 mm (group 2, with 64 individuals). The obvious constraining prior hypothesis is that the survival times are stochastically shorter in group 1 than in group 2. The corresponding Kaplan–Meier estimators, as well as the constrained Dykstra estimator (discussed in Sec. 1) are displayed in Figure 1.

Here we generated, using the Metropolis–Hastings algorithm, sample paths of the hazard rate processes λ' (group 1) and λ'' (group 2) from the posterior under the constraint $\lambda' \succcurlyeq \lambda''$. To assess the effect of the constraint, we built also an unconstrained version of the algorithm, as described in Section 3.

Figures 2 and 3 plot some results from this analysis: the constrained and unconstrained predictive survival functions and predictive hazard rates for both groups. In the parameterization of the hazard rates, we used a time grid of 100 evenly spaced points.

To assess the convergence of the algorithms (constrained and unconstrained), we ran for both algorithms additional simulations of different lengths (up to 200,000 iterations), also starting the Markov chains at different points (chosen randomly from the respective priors). The empirical results remained virtually unchanged and were stable already after 1,000 iterations. The posterior estimates in the figures are obtained from samples of 5,000 iterations, discarding the first 1,000 iterations. On a mainframe IBM ES/9121 these runs took about 32 minutes for the constrained algorithm and 11 minutes for the unconstrained algorithm. (The constrained algorithm is slower, because at each parameter update the ordering must be checked.)

In the absence of further knowledge of the values of λ' and λ'', we used the same unconstrained priors for both groups. In Figures 2 and 3 we used the hyperparameter values $\alpha_0 = 5$ and $\beta_0 = 16.4$, corresponding to an initial hazard λ_1 with prior mean $E_{\text{prior}}(\lambda_1) = \alpha_0/\beta_0 = .13$ and coefficient of variation $\sqrt{\text{Var}_{\text{prior}}(\lambda_1)}/E_{\text{prior}}(\lambda_1) = 1/\sqrt{\alpha_0} = 1/\sqrt{5}$, and the value $\alpha = 15$ corresponding to conditional standard deviation $\sqrt{\text{Var}_{\text{prior}}(\lambda_i|\lambda_1,\ldots,\lambda_{i-1})} = \lambda_{i-1}/\sqrt{15}$. With these hyperparameter choices, the Metropolis–Hastings sampler accepted 78% of the proposals.

In the figures the unconstrained predictive survival curves closely follow the corresponding Kaplan–Meier curves, also intersecting each other. The effect of the constraint on the predictive distributions is more evident in group 1 than in group 2. This is a consequence of the fact the sample size in group 1 is smaller than in group 2, and consequently its posterior is more easily adjusted to satisfy the stochastic ordering hypothesis.

To study the effect of the choice of the hyperparameters, we changed the value of β_0 to $\beta_0 = 152.4$ and that of α to $\alpha = 35$, leaving the other values as they were in Figure 2. The lowering of the prior mean $E_{\text{prior}}(\lambda_1)$ from .13 to .032, using a different scale parameter, obviously increases all predictive survival probabilities. However, the change was relatively small (consistently less than .03). Effectively, the

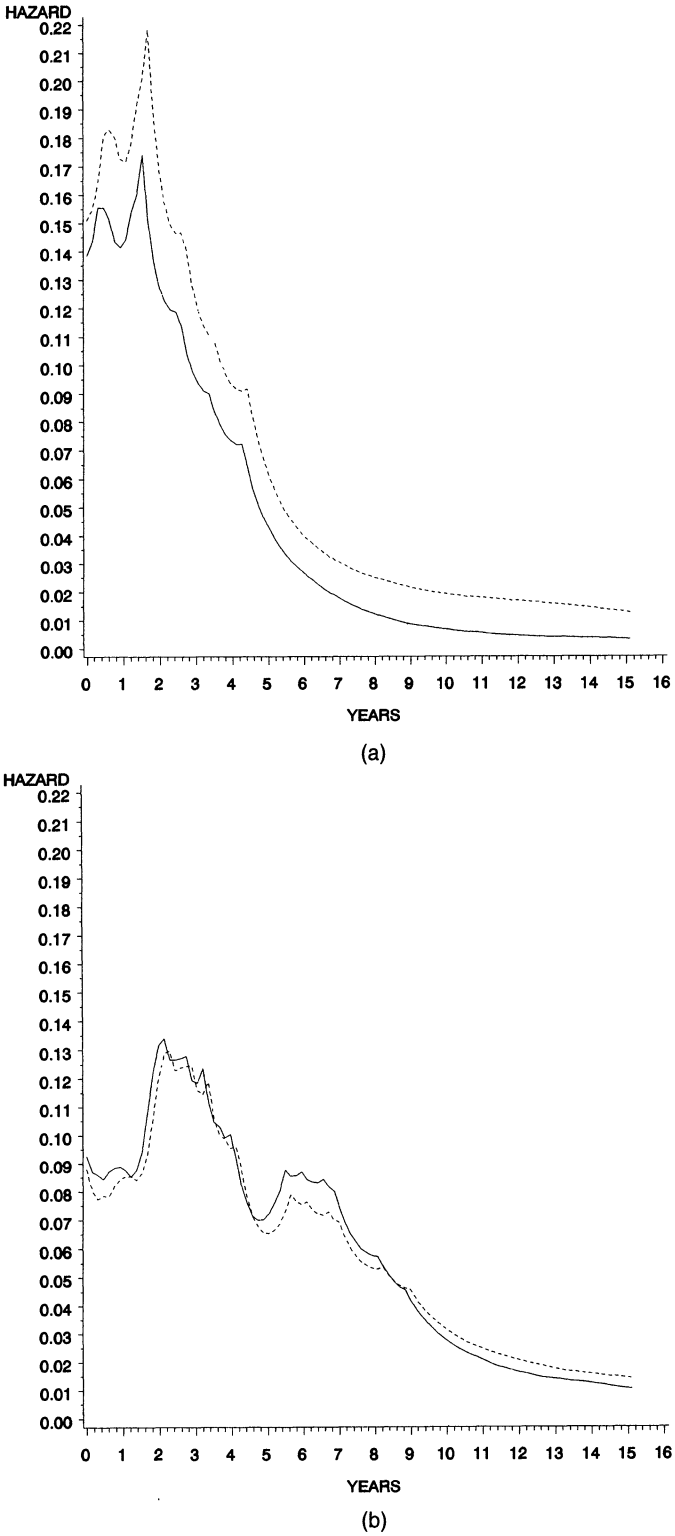

Figure 3. Constrained (Dashed Line) and Unconstrained (Solid Line) Predictive Hazards (a) for Group 1 and (b) for Group 2.

predictive hazards reflect this change only during the first 2 years of the follow-up. Giving a higher value to the hyperparameter α can be expected to reduce the oscillations of the hazard rate, as explained in Section 2. This actually happened, while all four predictive survival curves remained virtually unchanged.

As mentioned in Section 2, we also can use our construction for a posterior assessment of the concordance or discordance between the original stochastic ordering hypothesis and the data. A possible criterion is to compute the posterior probability $P(S|\text{data}) = P(\lambda' \succcurlyeq \lambda''|\text{data})$ under the coupled unconstrained model. Ideally, of course, we would like to have a coupling that is maximally supported by S. Unfortunately, it is extremely difficult to find such a coupling in practice. Using the simulation results of the coupled unconstrained Metropolis–Hastings Markov chain described in Section 4, we can approximate this probability directly by the corresponding relative frequency

$$\frac{1}{n}\sum_{i=1}^{n} 1_S(\lambda'^{(i)}, \lambda''^{(i)}). \tag{12}$$

Choosing again the hyperparameters $\alpha_0 = 5$, $\beta_0 = 16.4$, and $\alpha = 15$, we obtained the approximation $P(\lambda' \succcurlyeq \lambda''|\text{data}) \simeq .03$. Under the coupled prior, the corresponding probability was $P(\lambda' \succcurlyeq \lambda'') \simeq .14$. In other words, the data have the effect of reducing the probability of S to less than one-fourth of its prior value. On the other hand, if we still believe that the stochastic ordering hypothesis is correct, then it is essential to apply constrained estimation in inference, because the constraint will have a strong modifying effect on the empirical results.

For a comparison, we tried our algorithm also to data that were obviously consistent with the stochastic ordering hypothesis. We collapsed the earlier groups 1 and 2 into a single group with tumor thickness at least 2 mm, and then formed a second group from the remaining patients in the original data, with tumor at most 2 mm thick. For these two subsamples, the Kaplan–Meier estimators stay nicely apart, and thus they coincide with the Dykstra estimators. Choosing again the same hyperparameters for the coupled prior, we obtained the approximation $P(\lambda' \succcurlyeq \lambda''|\text{data}) \simeq .93$.

5. CONCLUDING REMARKS

Our approach to the present two-sample problem can be viewed as having the following ingredients: We use hazard rates in setting up the statistical model; this is attractive, because their qualitative properties are well understood and backed by a strong probabilistic intuition and a long tradition in the modeling and analysis of survival data. We then introduce a hierarchical Bayesian model structure, using as conventions piecewise constant sample paths of the hazard rate and gamma distributions (depending on three hyperparameters, controlling the initial level of hazard and its variability over time) to set up a prior on the space D of distribution functions. We then express the postulated stochastic ordering in the two-sample problem by forming a coupling of the models for the two samples and restricting it to the relevant subset of $D \times D$. We use MCMC methods for the coupling and in the practical numerical computations.

As the foregoing example shows, our method seems to work well in practice. An obvious attraction of the Bayesian approach is that we can summarize the results from empirical study in terms of predictive distributions, or the corresponding predictive hazards, without the need to use confi-

dence bands (typically based on asymptotic considerations) to support the derived point estimates. Integration with respect to the posterior gives smooth versions for these functions, despite the fact that the hazard rate sample paths (model parameters) were, for simplicity, assumed to have a simple piecewise constant structure. In particular, predictive survival functions, which involve a further integration over a time interval, seem to be relatively insensitive to the local fluctuations of the hazard rates.

APPENDIX: SOME IDEAS ON COUPLING AND MARKOV CHAIN COUPLING

The following definitions are from the monograph of Lindvall (1992).

Definition. A coupling of the probability measures P' and P'' on a measurable space $(D, \mathcal{D})$ is a probability measure on $\tilde{P}$ on $D \times D$ such that $P'(A) = \tilde{P}(A \times D)$ and $P''(A) = \tilde{P}(D \times A)$ for each $A \in \mathcal{D}$.

For example, the product measure $P' \times P''$ is always a (trivial) coupling of P' and P''. Under this coupling, the first component of a random element in $D \times D$ is independent from the second and vice-versa.

Defining a random element in $(D, \mathcal{D})$ as a quadruple $(\Omega, \mathcal{F}, P, Z)$, where $(\Omega, \mathcal{F}, P)$ is an underlying probability space and Z is an $\mathcal{F}/\mathcal{D}$-measurable mapping from Ω to D, we get an equivalent and more practical definition.

Definition. A coupling of two random elements $(\Omega', \mathcal{F}', P', Z')$ and $(\Omega'', \mathcal{F}'', P'', Z'')$ in $(D, \mathcal{D})$ is a random element $(\tilde{\Omega}, \mathcal{F}, \tilde{P}, (\tilde{Z}', \tilde{Z}''))$ in $(D \times D, \mathcal{D} \times \mathcal{D})$ such that $Z' \stackrel{d}{=} \tilde{Z}'$ and $Z'' \stackrel{d}{=} \tilde{Z}''$. We denote the couple $(\tilde{Z}', \tilde{Z}'')$ by $\tilde{Z}$.

For example, if F and G are continuous distribution functions on $\mathbb{R}, \Omega' = \Omega'' = [0,1], P' = P'' =$ Lebesgue measure, $\mathcal{F}' = \mathcal{F}'' =$ the σ algebra of Lebesgue measurable sets, and $Z'(\omega') = F^{-1}(\omega'), Z''(\omega'') = G^{-1}(\omega'')$, then we form a coupling by taking $\tilde{\Omega} = \Omega' = \Omega'' = [0,1]$ with the same measure and σ algebra and define $(\tilde{Z}', \tilde{Z}'')(\tilde{\omega}) = (F^{-1}(\tilde{\omega}), G^{-1}(\tilde{\omega}))$. (Note that there are many possible ways to couple a given pair of distributions; here the distribution on $\mathbb{R} \times \mathbb{R}$ arising from the coupling $(\tilde{Z}', \tilde{Z}'')$ is singular with respect to the product measure $dF \times dG$, which corresponds to the trivial coupling.)

We now consider a time-homogeneous Markov chain $\{Z_n\}_{n \in \mathbb{N}}$ with state space $(D, \mathcal{D})$. A sample from this Markov chain will be a random element $(Z_n)_{n \in \mathbb{N}} \in (D)^\infty$, and a distribution can be specified by giving a transition probability kernel $K(z, A)$ for $z \in D, A \in \mathcal{D}$, and a distribution P_0 on D for the initial state Z_0. If two such chains are given, say $\{Z'_n\}_{n \in \mathbb{N}}$ and $\{Z''_n\}_{n \in \mathbb{N}}$, with the same state space $(D, \mathcal{D})$, then we can couple them together as random elements in $(D)^\infty$, looking for a measure on $(D)^\infty \times (D)^\infty$ with the given marginals.

A *Markovian coupling* of two Markov chains is a coupling on $(D)^\infty \times (D)^\infty$, which itself is a Markov chain. This means that the Markov chains are coupled step by step on $D \times D$, by defining couplings for the initial distributions P'_0, P''_0 and then for all possible transition probabilities $K'(z', \cdot), K''(z'', \cdot)$ of the respective chains.

Let $\{\tilde{Z}_n\}$ be a Markovian coupling of two positive recurrent Markov chains $\{Z'_n\}$ and $\{Z''_n\}$, both defined on countable or complete separable metric spaces. We now show that, under mild conditions, $\{\tilde{Z}_n\}$ also will be positive recurrent. First, however, we need some terminology.

Definition. A set $A \subset D$ is called *uniformly transient* if there is $M < \infty$ such that $\sum_1^\infty K^n(z, A) < M$, for all $z \in A$.

Definition. A Markov chain is *weakly Feller* if, for any open set $O, K(\cdot, O)$ is a lower semicontinuous function; that is, $\{z \in D: K(z, O) > u\}$ is open for all $u \in [0,1]$. (This condition means that the transition kernel is compatible with the topology.)

The next lemma follows by combining propositions 6.2.8 (ii) and 8.3.5 of Meyn and Tweedie (1993).

Lemma. Let $\{Z_n\}$ be a weakly Feller, φ-irreducible, and transient Markov chain, where φ is a positive measure with support containing an open set. Then every compact set is uniformly transient.

We can now prove the following theorem.

Theorem. Let $\{Z'_n\}$ and $\{Z''_n\}$ be positive recurrent Markov chains on complete separable metric spaces D', D''. If $\{\tilde{Z}_n\}$ is a φ-irreducible weakly Feller Markovian coupling of $\{Z'_n\}$ and $\{Z''_n\}$, and if supp$\{\varphi\}$ contains an open set, then $\{\tilde{Z}_n\}$ is positive recurrent.

Proof. A φ-irreducible Markov chain must be either transient or recurrent. If $\{\tilde{Z}_n\}$ is recurrent, then it must be positive recurrent, because the marginals are positive recurrent. We show that it cannot be transient.

Fix a starting point $\tilde{z} = (z', z'')$. For each n, the joint n-step transition kernel $\tilde{K}^n(\tilde{z}, \cdot)$ is a coupling of the marginal n-step transition kernels $K'^n(z', \cdot)$ and $K''^n(z'', \cdot)$. Because the marginal chains are positive recurrent, it follows that the sequences of probability measures $\{K'^n(z', \cdot), n \in \mathbb{N}\}$ and $\{K''^n(z'', \cdot), n \in \mathbb{N}\}$ are tight (see, e.g., Meyn and Tweedie 1993). This implies tightness of the sequence of couplings $\{\tilde{K}^n(\tilde{z}, \cdot), n \in \mathbb{N}\}$.

In particular, there is a compact set C such that $\tilde{K}^n(\tilde{z}, C) > 1/2$, for all n. It is not a restriction to take $\tilde{z} \in C$ (add $\tilde{z}$ to C if necessary). Because $\sum_1^\infty \tilde{K}^n(\tilde{z}, C) = \infty, C$ is a compact set that is not uniformly transient. By the previous lemma, it follows that the Markov chain is not transient.

Notes. It is straightforward to see that the (unique) invariant probability measure for the coupled chain $\{\tilde{Z}_n\}$ is necessarily a coupling of the invariant distributions of the marginal chains $\{Z'_n\}, \{Z''_n\}$.

We can now apply the theorem to our MCMC algorithms by choosing $Z'_n = (\Lambda'_n, X'_n)$ and $Z''_n = (\Lambda''_n, X''_n)$.

It is easy to see that the resulting coupled chain $\{\tilde{Z}_n\} = \{\tilde{\Lambda}'_n, \tilde{X}'_n, \tilde{\Lambda}''_n, \tilde{X}''_n\}$ is *Lebesgue* irreducible on the product space, because in a single cycle we can reach any open neighborhood of the product space. In fact, the conditional density of the observation (X', X''), given any value of the parameters Λ', Λ'', is absolutely continuous with an everywhere positive density on $(\mathbb{R}^+)^{m'} \times (\mathbb{R}^+)^{m''}$. Also, starting from any initial values λ', λ'', the marginal transition kernel $K((\lambda', \lambda'') \to \cdot)$ is absolutely continuous with respect to the Lebesgue measure and has a positive density on $D \times D$. The weak Feller property and aperiodicity follow from the absolute continuity and positivity of $p(x', x''|\lambda', \lambda'')$.

From the same arguments, it follows that we can apply the theorem to the coupled Markov chain $\{\tilde{\Lambda}'_n, \tilde{\Lambda}''_n\}$, obtained from $\{\tilde{Z}_n\}$ by fixing X' and X'' at their observed values and updating only the parameters. Therefore, this coupled Markov chain is ergodic, and its equilibrium distribution is a coupling of the marginal posteriors.

[Received November 1993. Revised November 1995.]

REFERENCES

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer.

Arjas, E., and Gasbarra, D. (1994), "Nonparametric Bayesian Inference from Right-Censored Survival Data, Using the Gibbs Sampler," *Statistica Sinica*, 4, 505–524.

Brunk, H. D., Franck, W. E., Hanson, D. L., and Hogg, R. V. (1966), "Maximum Likelihood Estimation of the Distribution of Two Stochastically Ordered Random Variables," *Journal of the American Statistical Association*, 61, 1067–1080.

Dykstra, R. L. (1982), "Maximum Likelihood Estimation of the Survival Functions of Stochastically Ordered Random Variables," *Journal of the American Statistical Association*, 77, 621–628.

Dykstra, R. L., and Felz, C. J. (1989), "Nonparametric Estimation of Partially Stochastically Ordered Distributions and the Dual Problem," *Biometrika*, 76, 331–341.

Dykstra, R. L., Kochar, S., and Robertson, T. (1991), "Statistical Inference for Uniform Stochastic Ordering in Several Populations," *The Annals of Statistics*, 19, 870–888.

Felz, C. J., and Dykstra, R. L. (1985), "Maximum Likelihood Estimation of the Survival Functions of N Stochastically Ordered Random Variables," *Journal of the American Statistical Association*, 80, 1012–1019.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelfand, A. E., Smith, A. F. M., and Lee, T.-M. (1992), "Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using the Gibbs Sampler," *Journal of the American Statistical Association*, 87, 523–532.

Lindvall, T. (1992), *Lectures on the Coupling Method*, New York: John Wiley.

Meyn, S. P., and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, New York: Springer.

Roberts, G. O., and Smith, A. F. M. (1994), "Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis–Hastings Algorithms," *Stochastic Processes and Their Applications*, 49, 207–216.

Strassen, V. (1965), "The Existence of Probability Measures With Given Marginals," *Annals of Mathematical Statistics*, 36, 423–439.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1762.