Bayesian Survival Curve Estimation for Multivariate Data

By

Ronald C. Pruitt

Technical Report No. 516

School of Statistics

University of Minnesota

19 July 1988

Abstract

This paper considers nonparametric estimation of a multivariate survival curve from incompletely observed data. The types of incomplete observations considered include, but are not limited to, observations which are censored on the right either individually or jointly. The posterior distribution of the survival curve is derived using a mixture of Dirichlet process priors assuming the available observations are known only to belong to particular measurable subsets of the observation space. For univariate data, this work extends the results of Susurla and Van Ryzin (1976) and verifies their conjecture that the posterior distribution is a mixture of Dirichlet processes. The form of the estimate for univariate data under weighted squared error loss simplifies to a product limit form which is an extension of the Kaplan-Meier estimate and which is easy to compute. For multivariate data, the estimate has a conceptually simple form but computation increases exponentially with the sample size.

1 Introduction.

Censored data often involves observations on more than a single variable. This may arise through grouping of the experimental units or through multiple observations on each experimental unit. Examples include matched pair and multivariable studies. Censoring may occur through many mechanisms which allow differential censoring in the variables. In pair studies, members of the pair may withdraw from the study at different times; in multivariable problems, the observations may be obtained in sequence and the study terminated at some point which only allows observation of some of the variables, for example. Much of this paper concentrates on bivariate right censoring, where the observable variables are X_1, X_2, D_1, and D_2, where $X_j = \min[T_j, C_j]$ and $D_j = 1[X_j = T_j]$. The distribution of the unobserved variable (T_1, T_2) is of interest, and the variable (C_1, C_2) is a nuisance censoring variable. We will discuss a nonparametric Bayesian estimate of the distribution of (T_1, T_2) based on sample data $(X_{1i}, X_{2i}, D_{1i}, D_{2i})$ for $i = 1, \ldots, n$. For sample data points subject to right censoring we will use the intuitive '+' notation: $(4+, 3)$ indicates a data point with T_1 censored at 4 and T_2 observed at 3, that is $(X_1, X_2, D_1, D_2) = (4, 3, 0, 1)$.

Many estimators have been proposed for bivariate survival curve data, including those of Muñoz (1980), Campbell (1981), Campbell and Földes (1982), Hanley and Parnes (1983), Tsai et al. (1986), and Dabrowska (1988). Muñoz (1980), Campbell (1981), and Hanley and Parnes (1983) all discuss the nonparametric maximum likelihood estimator (NMLE) for this problem. The main objection to this estimator which has been raised is that it is not uniquely defined. For example, if $(5.1+, 4.3)$ is observed, the NMLE does not specify how to distribute mass on this ray. This same situation arises in the univariate case when the largest observation is censored. The objection with multivariate data seems to be that the amount of indeterminate mass of this kind is not limited to n^{-1} although it does become negligible as the sample size increases under some conditions on the censoring distribution. If the support of (C_1, C_2) is all of the positive quadrant of $\Re^2$ (this rules out the univariate censoring of Leurgans, Tsai, and Crowley (1982)), this can be seen by noting that any ray which has a perpindicular ray cross it has no indeterminate mass, and eventually all rays will be crossed by perpindicular rays in any compact set bounded away from zero in either coordinate. As soon as all rays in a compact set have no indeterminate mass, the mass of

all doubly censored points will also be determined. These objections seem to have the same status in both the multivariate and univariate case. Another kind of nonuniqueness occurs when many rays cross, for example, if $(1+,2),(1+,3),(2,1+)$, and $(3,1+)$ are observed, the NMLE only specifies that the mass at $(2,2)$ and $(3,3)$ is γ and the mass at $(2,3)$ and $(3,2)$ is $0.5-\gamma$ where γ is between 0 and 0.5. The estimator is always a proper survival function, agrees with the empirical survival function when censoring is not present, and can be computed using the EM algorithm (Dempster, Laird, and Rubin (1977)). The connection between this estimate and the Bayesian estimate discussed in this paper will be explored in a further paper.

The other estimates all make some decomposition of the survival function based on the marginal distributions. The estimates of Campbell and Földes, and Dabrowska are not necessarily proper survival functions. Campbell and Földes decompose $\Pr\{T_1 > t_1, T_2 > t_2\}$ as $\Pr\{T_2 > t_2 | T_1 > t_1\} \Pr\{T_1 > t_1\}$ and estimate each term separately. Dabrowska estimates components of the bivariate cumulative hazard function separately and uses a product limit form for the survival curve as a function of these quantities. Characterization of Dabrowska's estimate as a bivariate Kaplan-Meier (1958) estimate does not seem appropriate (see Section 5). Tsai et al. decompose the survival curve in a seemingly more arbitrary manner and estimate several (sub)survival curves. A conditional subsurvival curve is estimated using nonparametric smoothing techniques which depend on smoothing parameters chosen by the practitioner and give a slow asymptotic rate of convergence. The resulting estimator is a step function which only assigns mass to uncensored points and rays which do not cross the $x_1 = x_2$ line. This estimator gives much different results than the others considered. Finally, another feature shared by only the NMLE and the Bayesian estimate among these five is that both are simple generalizations of the univariate case and involve no special consideration of singly and doubly censored variables.

2 Background and notation.

Let $\vec{T_1} = (T_{11}, \ldots, T_{1k}), \ldots, \vec{T_n}$ be independent, identically distributed random vectors with non-negative components and survival function $S(t_1, \ldots, t_k) = \Pr\{T_{11} > t_1, \ldots, T_{1k} > t_k\}$ which we wish to estimate. We observe $A_1, \ldots, A_n$ where each A_i is a measurable subset of

$\Re_+^k = [0,\infty) \times \cdots \times [0,\infty)$, and it is only known that $T_i \in A_i$. This is a generalization of the often considered right censoring. If $k = 1$, right censoring amounts to considering only A_i of the form $A_i = \{c_1\}$ or $A_i = [c_1, \infty)$. In two dimensions, right censored A_i are restricted to be one of the following four forms:

$$\begin{aligned} &\text{I.} \quad A_i = \{c_1\} \times \{c_2\} \\ &\text{II.} \quad A_i = [c_1,\infty) \times \{c_2\} \\ &\text{III.} \quad A_i = \{c_1\} \times [c_2,\infty) \\ &\text{IV.} \quad A_i = [c_1,\infty) \times [c_2,\infty) \end{aligned} \tag{2.1}$$

Sets of type I are referred to as uncensored, types II and III as singly censored, and type IV as doubly censored. We present theoretical results for the general form and then make some specific calculations for this specific type of censoring. Associated with each A_i are four random variables: X_{1i}, X_{2i}, D_{1i}, and D_{2i}. For a particular A_i these are given as follows: $X_{1i} = c_1$, $X_{2i} = c_2$, $D_{1i} = 1[A_i \text{ is type I or III}]$, and $D_{2i} = 1[A_i \text{ is type I or II}]$.

Other types of censoring are also covered by this construction. We mention double censoring encountered in bioassay and interval censoring as examples (see Kuo (1983) and Turnbull (1974)). Both of these types of censoring widen the possible observed sets A_i.

We next describe the prior distribution for the survival function. The survival function and probability distribution provide the same information and throughout we deal with the distribution through its probability measure. Basic familiarity with the Dirichlet process as a distribution on probability measures as described in Ferguson (1973) is assumed. The following material is based on Antoniak (1974) where mixtures of Dirichlet processes are described. Let $\mathcal{B}_+^k$ be the Borel σ-field on $\Re^k$ restricted to $\Re_+^k$. Also let $\mathcal{B}_A^k$ denote $\mathcal{B}_+^k$ restricted to the set A. We need the following definitions which are definitions 2, 3, and 4 of Antoniak (1974).

Definition 2.1 (Antoniak, def. 2) *Let $(\Theta, \mathcal{A})$ and $(U, \mathcal{B})$ be two measurable spaces. A* transition measure *on $U \times \mathcal{A}$ is a mapping α of $U \times \mathcal{A}$ into $[0,\infty)$ such that*

1. *For every $u \in U$, $\alpha(u, \cdot)$ is a finite, nonnegative, nonnull measure on $(\Theta, \mathcal{A})$.*
2. *For every $A \in \mathcal{A}$, $\alpha(\cdot, A)$ is measurable on $(U, \mathcal{B})$.*

Definition 2.2 (Antoniak, def. 3) *Let $(\Theta, \mathcal{A})$ be a measurable space, let $(U, \mathcal{B}, H)$ be a probability space called the index space, and let α be a transition measure on $U \times \mathcal{A}$. We say P is a mixture of Dirichlet processes on $(\Theta, \mathcal{A})$ with mixing distribution H on the index space $(U, \mathcal{B})$, and transition measure α, if for all $k = 1, \ldots$ and any measurable partition $A_1, \ldots, A_k$ of Θ we have*

$$\Pr\{P(A_1) \leq y_1, \ldots, P(A_k) \leq y_k\} = \int_U D(y_1, \ldots, y_k | \alpha(u, A_1), \ldots, \alpha(u, A_k)) dH(u),$$

where $D(\theta_1, \ldots, \theta_k | \alpha_1, \ldots, \alpha_k)$ denotes the distribution function of the Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_k)$.

Definition 2.3 (Antoniak, def. 4) *Let P be a mixture of Dirichlet processes on $(\Theta, \mathcal{A})$ with mixing distribution H on index space $(U, \mathcal{B})$, and transition measure α on $U \times \mathcal{A}$. We say that $\theta_1, \ldots, \theta_n$ is a sample of size n from P if for any $m = 1, \ldots$ and measurable sets $A_1, \ldots, A_m, C_1, \ldots, C_n$ we have:*

$$\Pr\{\theta_1 \in C_1, \ldots, \theta_n \in C_n | u, P(A_1), \ldots, P(A_m), P(C_1), \ldots, P(C_n)\} = \prod_{i=1}^{n} P(C_i) \text{ a.s.}$$

It is useful to think of a mixture of Dirichlet processes in the following inuitive sense: think of the index u as a random variable with distribution H, and conditional given u, P is a Dirichlet process with parameter $\alpha(u, \cdot)$. This is completely general in the sense that the process P^* which chooses u according to H, and P from a Dirichlet process with parameter $\alpha(u, \cdot)$ is a mixture of Dirichlet processes as described in Definition 2.2. The class of Dirichlet mixtures is closed under sampling: if an observation is taken from a Dirichlet mixture, the resulting posterior distribution is again a mixture of Dirichlet processes with transition measure incremented by a point mass at the observation and mixing distribution given by the conditional distribution of u given the observation. This result is proved in Antoniak (1974, Corollary 3.2). We are interested in the case when the observations are not known exactly and a similar result holds in this case. We proved the following theorem before recognizing it as a special case of Theorem 3 of Antoniak (1974).

Theorem 2.1 (Antoniak (1974)) *Let P be a mixture of Dirichlet processes on $(\Re_+^k, \mathcal{B}_+^k)$ with standard Borel index space $(U, \mathcal{A})$, distribution H on $(U, \mathcal{A})$ and transition measure α on $U \times \mathcal{B}_+^k$. If θ is a sample of size 1 from P and A is a measurable subset of $\mathcal{B}_+^k$, then the*

distribution of P given $\theta \in A$ is a mixture of Dirichlet processes on $(\Re_+^k, \mathcal{B}_+^k)$, with index space $(A \times U, \mathcal{B}_A^k \times \mathcal{A})$, transition measure $\alpha_u + \delta_\theta$ on $(A \times U) \times \mathcal{B}_+^k$, and mixing distribution H_θ on the index space, where H_θ is the conditional distribution of (θ, u) given $\theta \in A$.

Proof: Apply Theorem 3 of Antoniak with $(\mathcal{X}, \mathcal{C}) = (\Re_+, \mathcal{B}_+)$ and $F(\theta, B) = 1[\theta \in A]$ for any $B \in \mathcal{B}_+$. □

This theorem of course immediately extends to a sample of size n.

Corollary 2.2 *Assume the conditions on P of Theorem 2.1. If $\vec{\theta}$ is a sample of size n from P, and $A_1, \ldots, A_n$ are measurable sets in $\mathcal{B}_+^k$, then the distribution of P given $\theta_1 \in A_1, \ldots, \theta_n \in A_n$ is a mixture of Dirichlet processes on $(\Re_+^k, \mathcal{B}_+^k)$, with index space $(A_1 \times \cdots \times A_n \times U, \mathcal{B}_{A_1}^k \times \cdots \times \mathcal{B}_{A_n}^k \times \mathcal{A})$, transition measure $\alpha_u + \sum \delta_{\theta_i}$ on $(A_1 \times \cdots \times A_n \times U) \times \mathcal{B}_+^k$, and mixing distribution $H_{\vec{\theta}}$ on the index space, where $H_{\vec{\theta}}$ is the conditional distribution of $(\theta_1, \ldots, \theta_n, u)$ given $\theta_1 \in A_1, \ldots, \theta_n \in A_n$.*

Theorem 2.1 and Corollary 2.2 simplify slightly when P is a simple Dirichlet process. In this case we have index space $(A_1 \times \cdots \times A_n, \mathcal{B}_{A_1}^k \times \cdots \times \mathcal{B}_{A_n}^k)$, transition measure $\alpha + \sum \delta_{\theta_i}$ on $(A_1 \times \cdots \times A_n) \times \mathcal{B}_+^k$, and mixing distribution $H_{\vec{\theta}}$ on the index space, where $H_{\vec{\theta}}$ is the conditional distribution of $(\theta_1, \ldots, \theta_n)$ given $\theta_1 \in A_1, \ldots, \theta_n \in A_n$. Intuitively this Corollary is reasonable, when the data is not known exactly the posterior is a mixture of distributions of exactly observed data with the mixing distribution the conditional distribution of the exactly observed data given the actual data. When Corollary 2.2 is applied to univariate data with a simple Dirichlet prior we get the result of Susurla and Van Ryzin. The details of the comparison are straightforward using Theorems 3.1 and 4.2.

3 Application to multivariate data.

In this section we consider how to make practical use of the result given in Corollary 2.2. For simplicity we consider only bivariate data and simple Dirichlet process priors. Development of the theory for higher dimensions and Dirichlet mixtures is routine, but computation of the estimates becomes much harder. We also illustrate the techniques for the usual single and double censoring given by (2.1). The implementation is specific to the type of

censoring involved, but some of the techniques developed here will be useful for other types of censoring as well. Make the following assumptions about the Dirichlet prior.

A0) Assume P is a Dirichlet process with parameter α.

A1) Assume the support of α is all of $\Re_+^2$.

A2) Assume that α has only finitely many discontinuities, that is

$$\alpha(A) = \sum_{j=1}^{p} \lambda_j 1[b_j \in A] + \int_A \alpha_0(s,t)dsdt$$

for some p and some $\lambda_j, b_j, j = 1, \ldots, p$.

A3) Assume that α_0 is uniformly continuous.

Assumption A1) is made for notational simplicity, and A2) and A3) are made so that the result is analytically tractable. Also let

$$\alpha_1(s,t) = \int_s^\infty \alpha_0(x,t)dx \qquad \text{and} \qquad \alpha_2(s,t) = \int_t^\infty \alpha_0(s,x)dx.$$

Finally define $\alpha^*(B,A) = \alpha(BA)$.

We first consider the case that the observed values θ_i fall in measurable sets A_i which each have positive α measure. In this case the conditional distribution of $\theta_1, \ldots, \theta_n$ given $\theta_1 \in A_1, \ldots, \theta_n \in A_n$ is

$$H_{\vec{\theta}}(B_1, \ldots, B_n) = \frac{P(\theta_1 \in B_1 A_1, \ldots, \theta_n \in B_n A_n)}{P(\theta_1 \in A_1, \ldots, \theta_n \in A_n)}.$$

These Dirichlet probabilities are most easily expressed using the law of total probability by conditioning on which of the θ_i are equal. For example

$$P(\theta_1 \in B_1 A_1, \theta_2 \in B_2 A_2, \theta_3 \in B_3 A_3, \theta_1 = \theta_2 = \theta_3) = \frac{\alpha^*(B_1 B_2 B_3, A_1 A_2 A_3)}{\alpha(\Re^2)} \frac{1}{\alpha(\Re^2)+1} \frac{2}{\alpha(\Re^2)+2}. \quad (3.1)$$

To complete this example we may write

$$\begin{aligned} \alpha(\Re^2)^{(3)} P(\theta_1 \in B_1 A_1, \theta_2 \in B_2 A_2, \theta_3 \in B_3 A_3) = & \\ & \alpha^*(B_1, A_1)\alpha^*(B_2, A_2)\alpha^*(B_3, A_3) + \alpha^*(B_1, A_1)\alpha^*(B_2 B_3, A_2 A_3) + \\ & \alpha^*(B_2, A_2)\alpha^*(B_1 B_3, A_1 A_3) + \alpha^*(B_3, A_3)\alpha^*(B_1 B_2, A_1 A_2) + \quad (3.2) \\ & 2\alpha^*(B_1 B_2 B_3, A_1 A_2 A_3) \\ \stackrel{\text{def}}{=} & \; Q_{3,\alpha}(B_1, B_2, B_3; A_1, A_2, A_3), \end{aligned}$$

where $a^{(n)} = a(a+1)\cdots(a+n-1)$ for real a and integer $n > 0$. In general when $\alpha(A_i) > 0$ for all i, define $Q_{n,\alpha}$ in this manner. That is

$$Q_{n,\alpha}(B_1,\ldots,B_n;A_1,\ldots,A_n) = \alpha(\Re^2)^{(n)}P(\theta_1 \in B_1A_1,\ldots,\theta_n \in B_nA_n). \qquad (3.3)$$

We have pursued this simple example because the general form of the summands in $Q_{n,\alpha}$ is difficult to write down with both precision and brevity, the following is a brief discussion of the form of the general summand. Each summand in $Q_{n,\alpha}$ arises from specified subsets of the observations being equal. The summand is a product of terms, one from each subset of equal observations. The term in the product is the α^* measure of the intersection of all the equal observations times the factorial of one less than the number of equal observations. How this arises is seen at (3.1). The α^* measure comes from the probability of getting an observation in the required set for the equalities to hold and the multiplier j arises from the probability that the $j+1^{st}$ observation is equal to the other j for $j = 1,\ldots,n-1$. We will extend the definition of α^* to get a definition of $Q_{n,\alpha}$ for all A_i at (3.5) and (3.9). Computation of these probabilities is straightforward, but without any known relationship among the A_i the number of terms in $Q_{n,\alpha}$ grows exponentially. We return to this in Section 4.

In general not all of the A_i will have positive α measure, and in this case we can use the methods of Pfanzagl (1979) to find a regular conditional distribution. If we let $A_i^\epsilon = \{\vec{x} : d(\vec{x},A_i) < \epsilon/2\}$ where $d(\cdot,\cdot)$ is supremum distance, and let

$$H^\epsilon_{\vec{\theta}}(B_1,\ldots,B_n) = \frac{P(\theta_1 \in B_1A_1^\epsilon,\ldots,\theta_n \in B_nA_n^\epsilon)}{P(\theta_1 \in A_1^\epsilon,\ldots,\theta_n \in A_n^\epsilon)}, \qquad (3.4)$$

then $H^\epsilon_{\vec{\theta}}$ is well defined since $\alpha(A_i^\epsilon) > 0$ by assumption, and by the theorem in Pfanzagl (1979), $H^\epsilon_{\vec{\theta}}$ converges weakly to $H_{\vec{\theta}}$, a probability measure which is a regular conditional distribution of $\theta_1,\ldots,\theta_n$ given $\theta_1 \in A_1,\ldots,\theta_n \in A_n$. If we restrict attention to A_i of the form given in (2.1), then the weak limit of the measures in (3.4) can be evaluated. This is because the class of sets given in (2.1) is closed under finite intersections, and it is straightforward to check $(A_iA_j)^\epsilon = A_i^\epsilon A_j^\epsilon$ for small enough ϵ. Under our assumptions on α_0

the α measure of the sets A_i can be easily approximated. Extend the definition of α^* by

$$\alpha^*(B,A) = \begin{cases} \alpha(BA) & \text{if } \alpha(A) > 0 \text{ or } A \text{ is type IV} \\ \alpha_1(BA) & \text{if } \alpha(A) = 0 \text{ and } A \text{ is type II} \\ \alpha_2(BA) & \text{if } \alpha(A) = 0 \text{ and } A \text{ is type III} . \\ \alpha_0(BA) & \text{if } \alpha(A) = 0 \text{ and } A \text{ is type I} \\ 0 & \text{if } A = \emptyset \end{cases} \tag{3.5}$$

Then by a standard argument,

$$\alpha(BA^\epsilon) = \epsilon^\gamma(\alpha^*(B,A) + o(1)), \tag{3.6}$$

where the $o(1)$ term is uniform for all sets A of the form (2.1) since α_0 is uniformly continuous. Here γ is 0 if α^* is given by line 1 or 5 of (3.5), 1 if α^* is given by line 2 or 3 of (3.5), and 2 if α^* is given by line 4 of (3.5). This allows us to take the limit in (3.4). The procedure for finding the limit in (3.4) is clear: we rewrite (3.4) as

$$H^\epsilon_{\vec{\beta}}(B_1,\ldots,B_n) = \frac{Q_{n,\alpha}(B_1,\ldots,B_n;A_1^\epsilon,\ldots,A_n^\epsilon)}{Q_{n,\alpha}(A_1^\epsilon,\ldots,A_n^\epsilon;A_1^\epsilon,\ldots,A_n^\epsilon)}, \tag{3.7}$$

using (3.3), and then take term by term limits in numerator and denominator. In general the power of ϵ will not be the same for each summand in $Q_{n,\alpha}$, but (3.6) and the uniformity of the $o(1)$ term guarantee the existence of a unique integer $\Gamma[= \Gamma(A_1,\ldots,A_n)]$ such that

$$0 < \lim_{\epsilon\to 0} \epsilon^{-\Gamma} Q_{n,\alpha}(A_1^\epsilon,\ldots,A_n^\epsilon;A_1^\epsilon,\ldots,A_n^\epsilon) < \infty. \tag{3.8}$$

This Γ will be the lowest power of ϵ for any summand in $Q_{n,\alpha}(A_1^\epsilon,\ldots,A_n^\epsilon;A_1^\epsilon,\ldots,A_n^\epsilon)$ given at (3.2). With Γ defined in this manner, let

$$Q_{n,\alpha}(B_1,\ldots,B_n;A_1,\ldots,A_n) = \lim_{\epsilon\to 0} \epsilon^{-\Gamma} Q_{n,\alpha}(B_1,\ldots,B_n;A_1^\epsilon,\ldots,A_n^\epsilon). \tag{3.9}$$

This limit exists by (3.6) and (3.8), it may equal zero. We first discuss the case when all powers of ϵ are the same. In this case, $Q_{n,\alpha}$ is given by both (3.2) and (3.9), since by (3.6) if all powers of ϵ are the same α may be replaced by α^*.

Theorem 3.1 *Assume A0) – A3). Relabel the A_i so that $A_{r+1}, \ldots, A_n$ are the sets of type I. Let $\beta = \alpha + \sum_{r+1}^{n} \delta_{A_i}$. Assume that no two sets of type II with β measure zero intersect, and also no two sets of type III with β measure zero intersect. Also assume if A_i of type II and A_j of type III have $\beta(A_i A_j) = 0$, then $\beta(A_i) = 0$ and $\beta(A_j) = 0$. Then the posterior distribution of P given $\theta_1 \in A_1, \ldots, \theta_n \in A_n$ is a mixture of Dirichlet processes on $(\Re_+^2, \mathcal{B}_+^2)$, with index space $(A_1 \times \cdots \times A_r, \mathcal{B}_{A_1}^2 \times \cdots \times \mathcal{B}_{A_r}^2)$, transition measure $\beta + \sum_1^r \delta_{\theta_i}$ on $(A_1 \times \cdots \times A_r) \times \mathcal{B}_+^2$ and mixing distribution*

$$H_{\vec{\theta}}(B_1, \ldots, B_r) = \frac{Q_{r,\beta}(B_1, \ldots, B_r; A_1, \ldots, A_r)}{Q_{r,\beta}(A_1, \ldots, A_r; A_1, \ldots, A_r)}, \tag{3.10}$$

with $Q_{r,\beta}$ given as in (3.2).

Proof: First note that the posterior distribution of P given $\theta_{r+1} \in A_{r+1}, \ldots, \theta_n \in A_n$ is Dirichlet with parameter β. We need only check that the power of ϵ is the same for all the non-zero summands in $Q_{r,\beta}$, the result then follows by (3.6) and (3.7) using Corollary 2.2. By the closure of the class of sets in (2.1) and induction we need only check that for any two sets B_1 and B_2, and any two sets A_1 and A_2 of type II-IV either 1) $\beta^*(B_1 B_2, (A_1 A_2)^c) = 0$ or 2) $\beta^*(B_1 B_2, (A_1 A_2)^c)$ and $\beta^*(B_1, A_1^c)\beta^*(B_2, A_2^c)$ have the same power of ϵ. This may be readily checked by examining all the relevant possibilities. By symmetry we need only check the type pairs II-II, II-III, II-IV, and IV-IV. The IV-IV pair is obvious since the intersection of two type IV sets is again a type IV set. The II-II pair is similarly obvious since if two type II sets intersect they each have positive β measure and so does their intersection. Hence 2) holds. For the II-III pair, 1) holds unless $B_1 B_2 A_1 A_2 \neq \emptyset$. In this case, either β of the intersection point is non-zero in which case 2) holds with zero as the power of ϵ, or if β of the intersection point is zero then the power of ϵ is two using the assumption about II-III intersections. The II-IV argument is similar, let A_1 be the set of type II. Again 1) holds unless $B_1 B_2 A_1 A_2 \neq \emptyset$. In this case, either β of the intersection set is non-zero in which case 2) holds with zero as the power of ϵ, or if β of the intersection set is zero then the power of ϵ is one since A_1 and $A_1 A_2$ are both sets of type II. □

Clearly Theorem 3.1 holds with the assumptions about II-II (III-III) disjoint and II-III intersections removed if the conclusion stops at (3.10), but the formula for $Q_{r,\beta}$ becomes notationally more cumbersome than (3.2). Instead of trying to state this form, we indicate the changes. The powers of ϵ are now not all the same for each summand of

$Q_{r,\beta}(A_1^\epsilon, \ldots, A_r^\epsilon; A_1^\epsilon, \ldots, A_r^\epsilon)$, and only terms with the lowest power need to be considered. If multiple sets of type II with zero β measure intersect, the conditional probability of these observations being equal is one and only terms in (3.2) which include all these observations being equal need be considered. Also if sets of type II and III with non-zero β measure intersect in a point with zero β measure, then the conditional probability that the observations are equal is zero. In this case only summands which have all of these observations being distinct from each other need to be considered.

4 Computation.

In this section we provide methods to compute the Bayes estimate. To ease the notation somewhat define

$$\tilde{Q}_{n,\alpha}(A_1, \ldots, A_n) = Q_{n,\alpha}(A_1, \ldots, A_n; A_1, \ldots, A_n), \quad \text{and} \quad \tilde{\alpha}^*(A_i) = \alpha^*(A_i, A_i).$$

We first get the following theorem on the mean value of the posterior distribution.

Theorem 4.1 *Under the assumptions of Theorem 3.1,*

$$E[P(B|\theta_1 \in A_1, \ldots, \theta_n \in A_n)] = \frac{Q_{r+1,\beta}(B, A_1, \ldots, A_r; \Re^2, A_1, \ldots, A_r)}{(\alpha(\Re^2) + n)\tilde{Q}_{r,\beta}(A_1, \ldots, A_r)}.$$

Proof: Both sides of the equation are equal to $\Pr\{\theta_{n+1} \in B | \theta_1 \in A_1, \ldots, \theta_n \in A_n\}$. □

This gives us a formula for the Bayes estimate under weighted squared error loss which depends only on Q. When combined with the fact that the easiest way to compute Q is recursively, this gives us a method to compute the estimate. To illustrate the recursive computation of Q, return to the example of (3.2) and suppose we wish to compute $Q_{4,\alpha}$. Each of the summands in $Q_{3,\alpha}$ will split into $p+1$ summands where p is the number of multipliers in the summand. For example, the summand $\tilde{\alpha}^*(A_1)\tilde{\alpha}^*(A_2A_3)$ which arises from the event $\theta_1 \neq \theta_2 = \theta_3$ splits into 3 summands based on whether θ_4 is unequal to θ_1 or θ_2, is equal to θ_1, or is equal to θ_2. These three summands are respectively given by $\tilde{\alpha}^*(A_1)\tilde{\alpha}^*(A_2A_3)\tilde{\alpha}^*(A_4)$, $\tilde{\alpha}^*(A_1A_4)\tilde{\alpha}^*(A_2A_3)$, and $2\tilde{\alpha}^*(A_1)\tilde{\alpha}^*(A_2A_3A_4)$. The factor two occurs in the final term since $P(\theta_4 = \theta_2 | \theta_1 \neq \theta_2 = \theta_3) = 2/(\alpha(\Re^2) + 3)$. The computations are straightforward but time consuming as the number of terms grows exponentially. Some simplification is possible for two reasons: 1) summands become zero as soon as the specified

set intersection is empty, and 2) if the sets have nested subsequences the following theorem may be used.

Theorem 4.2 *If α satisfies A0) – A3) and $A_1 \subseteq \cdots \subseteq A_n$ then*

$$\tilde{Q}_{n,\alpha}(A_1, \ldots, A_n) = \tilde{\alpha}^*(A_1)(\alpha(A_2) + 1) \cdots (\alpha(A_n) + n - 1).$$

Proof: The proof is by induction. The theorem is true for $m = 1$. Assume the theorem is true for $m = 1, \ldots, n$. Note that for any set A_{n+1},

$$\Pr\{\theta_{n+1} \in A_{n+1} | \theta_1 \in A_1, \ldots \theta_n \in A_n\} = \frac{\alpha(A_{n+1}) + n}{\alpha(\Re^2) + n}.$$

Using (3.3) and (3.9), this is enough to show the theorem. □

Even with these simplifying theorems, the estimate is still difficult to compute for moderately sized samples, and approximation methods will usually be needed. The straightforward method outlined above to compute the estimate requires both computing time and memory which increase exponentially with the sample size. The number of terms in $\tilde{Q}_{n,\alpha}$ is already 115975 for $n = 10$. The usual Monte Carlo methods for Dirichlet processes are not applicable here: the conditional distribution of θ_2 given θ_1 and $\theta_2 \in A_2$ does not have a simple form. However there is a Monte Carlo method which still uses exponential time but only quadratic memory. This is an improvement, but the computation is still prohibitive. We have developed an approximation not using Monte Carlo methods which runs much faster and where computation increases as n^5 where n is the sample size. The approximation seems reasonable although we have not yet proved any properties of it. The Monte Carlo and approximation methods will be reported on elsewhere.

5 Examples.

Theorem 4.2 also provides a logical connection to the univariate problem. Using the partial ordering induced by set inclusion, if we have sets $A_1, \ldots, A_n$ which satisfy the hypotheses of Theorem 4.2, the problem can be considered as only one dimensional. It seems a much more natural way to embed the one dimensional problem in two dimensions than by looking

Table 1: Estimates from data: $(1, 2.5), (2+, 0.5+), (3, 1.5)$.

	amount of mass			
Estimate	$(1, 1.5)$	$(1, 2.5)$	$(3, 1.5)$	$(3, 2.5)$
Bayes	0	1/3	2/3	0
NMLE	0	1/3	2/3	0
Redistribution	0	1/3	1/2	1/6
Dabrowska	-1/6	1/2	2/3	0
Tsai et al.	0	1/2	1/2	0

at the marginal distributions. This is closely related to the notion of homogeneous censoring considered by Hanley and Parnes (1983). Here we present the next simplest example possible beyond completely nested sets, and examine the estimates proposed by the various techniques. Suppose we observe three data points: $(1, 2.5), (2+, 0.5+)$, and $(3, 1.5)$. Some of the estimates are given in Table 1. It can be seen that Dabrowska's estimate is not a proper survival function, and that the Tsai et al. estimate does not agree with the likelihood based procedures. The Tsai et al. procedure does not depend on the smoothing parameters in this case. The Bayes estimate listed is the limit obtained as $\alpha(\Re^2) \to 0$. This limit usually depends on α but does not in this case. There is another sensible procedure which seems to be closely related to Dabrowska's method. I will refer to this method as the redistribution procedure. In this procedure with n data points, we start with mass $1/n$ at each data point, and then redistribute mass twice using Efron's (1967) redistribute to the right algorithm for univariate data. First, the second variable is ignored and the data is considered as n univariate observations. The mass of the observations with first variable censored is redistributed according to the redistribute to the right algorithm but keeping the value of the second variable fixed for each observation. The resulting mass is then redistributed by the same procedure interchanging the roles of the first and second variables. This estimate can

Table 2: Estimates from data: $(1,5.5),(2+,3.5),(3+,2.5+),(4+,1.5),(5,0.5+),(6,4.5)$.

	amount of mass				
Estimate	$(1,5.5)$	$(6,4.5)$	$(5,1.5)$	$(5,3.5)$	other
Bayes-1	0.167	0.243	0.172	0.217	0.201
Bayes-2	0.125	0.173	0.105	0.119	0.478
NMLE	0.167	0.230	0.230	0.373	0.000
Redistribution	0.167	0.194	0.125	0.153	0.361
Dabrowska	0.264	0.154	0.107	0.155	0.320
Tsai et al.	0.200	0.600	0.000	0.000	0.200

be shown to be well-defined and is given the final row in Table 1. This estimate is easy to compute, has marginals which agree with the usual Kaplan-Meier estimates, and is a proper survival function. We assert that any sensible procedure of the empirical distribution function type, as opposed to the methods pursued by Berliner and Hill (1986), would begin by assigning mass 1/3 to each of the observations. The only question is how to assign the mass of the doubly censored point. It seems clear that it is desirable to assign this mass within the region $(2+,0.5+)$ as is done by the NMLE, Bayes, and redistribution procedures. The Dabrowska procedure follows this prescription at the expense of not being a survival function, and the Tsai et al. procedure does not follow this prescription.

We give a second example which illustrates some of the differences between the NMLE and the Bayesian estimate. The data are given in Table 2. The other estimates are provided for comparison. The estimate of Dabrowska is again not a proper distribution giving mass -0.045 to the points $(1,3.5)$ and $(1,4.5)$, and mass -0.006 to the point $(1,1.5)$. The Tsai et al. estimate depends on smoothing parameters only in the 0.200 mass along the ray $(4+,1.5)$. The Bayes estimates are for α a multiple of the probability measure given by independent exponential distributions each with mean 10/3. The Bayes-1 estimate is the

limit as $\alpha(\Re^2) \to 0$, and the Bayes-2 estimate is for $\alpha(\Re^2) = 2$. Of these estimates, the NMLE, Bayes, and redistribution estimates are the only ones which assign mass $1/n$ to each of the observed sets A_i in the absence of prior information. Considering this redistribution of mass within the sets A_i more carefully, the Bayesian estimate seems to be preferable to the NMLE for the following reason. Examine what happens with the mass associated with the point $(4+, 1.5)$ under the NMLE and Bayes schemes. For the NMLE, the mass is all concentrated at the point $(5, 1.5)$ regardless of whether the observation $(5, 0.5+)$ is equal to $(5, 1.5)$ or not. If the observation $(5, 0.5+)$ is not equal to $(5, 1.5)$, there seems to be no reason to prefer that the observation $(4+, 1.5)$ be equal to $(5, 1.5)$ rather than some other point. This is what happens with the Bayes estimate. With the Bayes estimate, the mass of the observation $(4+, 1.5)$ is split as follows: mass 0.086 is assigned to the point $(5, 1.5)$ corresponding to when observations $(4+, 1.5)$ and $(5, 0.5+)$ are equal, and the remaining mass 0.080 is distributed along the ray $(4, 1.5)$ to $(\infty, 1.5)$ according to the mean of the prior distribution since nothing else is known. The redistribution algorithm goes to the other extreme and gives the ray crossings no special significance. Mass is redistributed based on a global estimate of the distribution: two rays $(10+, 11.5)$ and $(10+, 20.5)$ would always have identical mass distributions in the first coordinate. It seems strange that observations such as $(11, 19.5+)$ should cause identical mass redistribution in both of the rays $(10+, 11.5)$ and $(10+, 20.5)$. This is a point that Tsai et al. try to address by making their estimate depend on a local rather than global estimate of the distribution.

The global estimation of the distribution by the redistribution algorithm also has the effect that stochastically ordered data does not necessarily remain so under the redistribution process. This was pointed out in Hanley and Parnes (1983) as a reason not to use univariate analyses on bivariate data, but the same objection can be made for not using the redistribution method or the method of Dabrowska. To use their example suppose the observations are: $(1.8, 5.2), (4.7, 11.8), (6.2+, 6.2+), (13.6+, 13.6+)$, and $(17.0, 21.7)$. Then the second variable is larger than the first in this sample, but the redistribution algorithm assigns mass 1/15 to the point $(17.0, 11.8)$.

These examples point out some of the weaknesses of the estimates which have been proposed for bivariate survival data. The objections are of four types: first, the estimates of Campbell and Földes, and Dabrowska are not necessarily survival functions and should

not be used; second, the method of Tsai et al. does not redistribute mass within each of the observed sets A_i; third, the NMLE gives special prominence to the intersection points of observed rays while the redistribution algorithm does not give them enough emphasis; and fourth, the redistribution method fails to retain stochastic ordering when it occurs in a sample. The Bayes estimate is the only one of these which does not have any of these weaknesses.

References

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

BERLINER, L.M., and HILL, B.M. (1986). Bayesian nonparametric survival analysis. *Technical Report* **351**, Ohio State University.

CAMPBELL, G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika* **68** 417-422.

CAMPBELL, G., and FÖLDES, A. (1982). Large sample properties of nonparametric bivariate estimators with censored data. In *Nonparametric Statistical Inference, Colloquia Mathematica-Societatis János Bolyai* (B.V. Gnedenko, M.L. Puri, and I. Vincze, eds.). North Holland, Amsterdam.

DABROWSKA, D.M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statist.* To appear.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1-38.

EFRON, B. (1967). The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics, IV.* New York, Prentice-Hall 831-853.

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

HANLEY, J.A., and PARNES, M.N. (1983). Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics* **39** 129-139.

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

KUO, L. (1983). Bayesian bioassay design. *Ann. Statist.* **11** 886-895.

LEURGANS, S., TSAI, W.-Y., and CROWLEY, J. (1982). Freund's bivariate exponential distribution and censoring. In *Survival Analysis* (J. Crowley and R.A. Johnson, eds.) 230-242. IMS, Hayward, Calif.

MUÑOZ, A. (1980). Nonparametric estimation from censored bivariate observations. *Technical Report* **60**, Stanford University.

PFANZAGL, J. (1979). Conditional distributions as derivatives. *Ann. Probab.* **7** 1046-1050.

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

TSAI, W.Y., LEURGANS, S., and CROWLEY, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *Ann. Statist.* **14** 1351-1365.

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