The Annals of Statistics
1990, Vol. 18, No. 4, 1556–1577

A WEIBULL MODEL FOR DEPENDENT CENSORING[1]

BY SHERRIE E. EMOTO AND PETER C. MATTHEWS

National Institutes of Health and University of Maryland Baltimore County

A bivariate Weibull model is proposed for censored survival data when there may be dependence between the survival and censoring random variables. The model is based on marginal transformations of the Pickands class of bivariate exponential distributions. Assuming the model to hold, the joint distribution of survival and censoring times is shown to be identifiable, and the maximum likelihood estimator of the parameters is shown to be consistent.

1. Introduction. The ability to estimate a survival distribution in the presence of censoring is important and has been studied extensively. If one is not willing to make parametric assumptions about the exact form of the underlying survival and censoring distributions but is willing to assume independence between these two mechanisms, Kaplan and Meier (1958) provided an estimator which is consistent, among other desirable properties. Many other estimators have been proposed which outperform the Kaplan–Meier estimator in various situations, but most still depend on the assumption of independence between the survival and censoring variables.

In recent years some work has been done, too, on estimation of the survival distribution without this independence assumption. Some of this work will be discussed in the following paragraph. Any work in this area must confront the problem of identifiability; without some assumptions it is impossible to infer the survival distribution from censored data or to judge whether the censoring and survival mechanisms are independent. Cox (1959) first discussed this problem. See Puri (1979) for a review and a list of references.

Most work in dependent censoring therefore relies on fairly strong assumptions on the form of the dependence between the survival and censoring mechanisms or is concerned only with bounds on the survival distribution. Fisher and Kanarek (1974) used stretching factors to produce a family of joint distributions with different levels of dependence. Moeschberger (1974) considered the bivariate Weibull distribution arising from the Marshall–Olkin bivariate exponential distribution. Peterson (1976) gave worst case bounds on the survival distribution without any assumptions. Williams and Lagakos (1977) provided conditions for the Kaplan–Meier estimator to be consistent even in the presence of dependence. Slud and Rubenstein (1983) gave bounds based on specific knowledge about hazard ratios. Robertson and Uppuluri (1984) ex-

Received July 1987; revised October 1989.

[1]Research partially supported by the National Security Agency under Grant MDA904-88-H-2014.

AMS 1980 *subject classifications*. Primary 62F10; secondary 62P10, 62N05.

Key words and phrases. Bivariate exponential distribution, dependent censoring, maximum likelihood estimation, survival analysis, Weibull distribution.

tended the Kaplan–Meier estimator to nonparametric estimates with dependent margins. Klein and Moeschberger (1984) assumed a particular form of dependence and gave bounds on the survival distribution in that case.

Here a model based on a bivariate Weibull distribution is presented. Identifiability assuming the model is shown and the maximum likelihood estimator (MLE) is shown to be consistent. An algorithm for fitting the model is presented and applied to a data set. The bivariate Weibull used is the one resulting from marginal transformations of the Pickands (1976) bivariate exponential. See Galambos [(1978), page 264] for a discussion of this distribution. This model places relatively weak nonparametric (parameterized by a measure) restrictions on the form of the dependence between the survival and censoring distributions. See Tawn (1988) for a discussion and references on the use of parametric subfamilies of this family of distributions in bivariate extreme value modeling. Tawn briefly discusses estimation for this model in the complete data case and dismisses it as troublesome. The methods of estimation given here apply to the complete data case as well, although the complete data problem appears to be no easier than the case of censored data.

The model consists of all bivariate Weibull distributions with unequal shape parameters that can arise as minimal extreme value distributions. Thus, if a minimal extreme value distribution is reasonable for the joint distribution of censoring and survival, then this model should be appropriate. The model contains all independent bivariate Weibull distributions with unequal shape parameters. Thus, if one would fit independent Weibull distributions to censored survival data, then one can fit this larger model and see how much the results deviate from independence. It is possible to formulate similar models with marginal distributions coming from a shape family different from the Weibull family. For many of these, similar results should hold.

The remainder of this article is organized as follows. Section 2 describes the model. The likelihood for a data set is derived in Section 3. Section 4 demonstrates that the model is identifiable. In Section 5 the maximum likelihood estimator is shown to be consistent. Section 6 gives an algorithm for fitting the maximum likelihood estimator. The algorithm is applied to a data set in Section 7. Finally, Section 8 is a discussion of the applicability of the model and possible paths for future investigation.

2. The model. The model is based on the bivariate exponential distribution of Pickands (1976), hereafter referred to simply as the bivariate exponential distribution. This distribution leads to a characterization of all bivariate minimal extreme value distributions, in that they can be derived from this distribution by univariate transformations of the two margins. See Galambos [(1978), pages 258–259] for a discussion. Here we will consider marginal transformations of this distribution of the form $x \mapsto x^{1/\alpha}$, $y \mapsto y^{1/\beta}$, $\alpha \neq \beta$, which is the class of all bivariate Weibull distributions with unequal shape parameters that are bivariate extreme value distributions.

The following notation will be used. A sample of n survival times $D_1, \ldots, D_n$ and a set of n potential censoring times $C_1, \ldots, C_n$ exists but is not observable. Within a pair, D_i and C_i may be dependent. The pairs (D_i, C_i),

$i = 1, \ldots, n$, are independent, each with joint distribution function $F_{D,C}$. The marginal distributions of D and C are F_D and F_C. The pairs are not observable, only (T_i, δ_i), $i = 1, \ldots, n$, where $T_i = \min(D_i, C_i)$ and $\delta_i = I_{D_i < C_i}$. In all the models considered $P(D_i = C_i) = 0$, so ties will be of no concern. Let $F_{T,\delta}(t, i) = P(T \leq t \cap \delta = i)$ for $0 \leq t < \infty$ and $i = 0, 1$. A survival function will be denoted by an overline, i.e., $\bar{F}_D(d) = P(D > d)$. Finally, a hazard function will be denoted by h,

$$h_D(x) = -\frac{d}{dx} \log \bar{F}_D(x).$$

The Pickands bivariate exponential distribution can be characterized in several ways. The best for the purposes of this article is to represent the bivariate random variable as a function of a Poisson process. See de Haan and Pickands (1986) for the best current results and a history of this technique. Their representations use a homogeneous Poisson process. Here it will be more convenient and intuitive to use a nonhomogeneous process. Consider the quadrant $R_2^+ = (0, \infty) \times (0, \infty)$, along with two rays R_D and R_C. Intuitively, these are horizontal and vertical rays at "infinity;" R_D goes from $(0, \infty)$ to (∞, ∞) and R_C goes from $(\infty, 0)$ to (∞, ∞). See Figure 1 for the intuitive placement of these rays. Consider a measure M on $\Omega = R_2^+ \cup R_D \cup R_C$ of the following form. For a segment in R_D or R_C, M is λ_d or λ_c times Lebesgue measure, respectively, where $\lambda_d \geq 0$ and $\lambda_c \geq 0$. On R_2^+, M is defined by a nonnegative measure μ on $(0, \pi/2)$ satisfying

$$\int_{(0, \pi/2)} \left(\frac{1}{\sin \theta} + \frac{1}{\cos \theta} \right) \mu(d\theta) < \infty.$$

M is most easily given on R_2^+ in polar coordinates. For a section of an annulus centered at 0, the set $\omega = \{(r, \theta) | 0 < R_1 < r < R_2, \theta_1 < \theta < \theta_2\}$, we have $M(\omega) = (R_2 - R_1) \times \int_{(\theta_1, \theta_2)} \mu(d\theta)$. This defines M for arbitrary Borel subsets of Ω. As an example, see Figure 1. There, for the segment A of R_D,

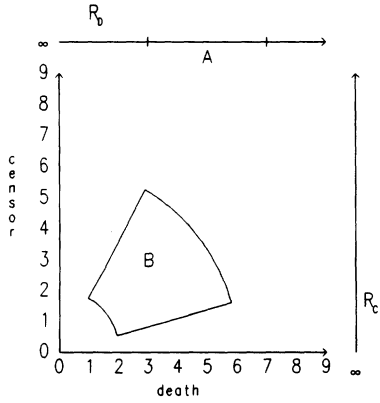

FIG. 1. *The set Ω along with two subsets A and B.*

$M(A) = 4\lambda_d$, and for the region B of R_2^+, $M(B) = (6 - 2)\int_{0.5}^{1.3}\mu(d\theta)$. To avoid degeneracy, we assume that either $\mu((0, \pi/2)) > 0$ or that both λ_d and λ_c are positive. The relevant Poisson process is then the Poisson process on Ω with intensity M; i.e., for $\omega \subset \Omega$, the number $N(\omega)$ of points in ω is Poisson with mean $M(\omega)$, etc. Let $(d_1, c_1), \ldots$ denote the points in a realization of this process. For this Poisson process a bivariate exponential distribution arises from considering $D^* = \inf(d_1, \ldots)$ and $C^* = \inf(c_1, \ldots)$. Under the preceding assumptions, the set of points is infinite almost surely and each of D^* and C^* is finite but nonzero almost surely. Then

$$P(D^* > d \cap C^* > c)$$
$$= P(N(\omega) = 0)$$
$$= \exp\left(-d\lambda_d - d\int_{[\arctan(c/d), \pi/2)} \frac{\mu(d\theta)}{\cos\theta} - c\lambda_c - c\int_{(0, \arctan(c/d))} \frac{\mu(d\theta)}{\sin\theta}\right),$$

where $\omega = \{(x, y) \in \Omega | x \le d \text{ or } y \le c\}$. The Pickands bivariate exponential has

$$P(D^* > d \cap C^* > c) = \exp\left(-\int_{[0,1]} \max(pd, (1-p)c)\, dT(p)\right),$$

where T is a positive measure on $[0, 1]$. Taking $p = (1 + \cot\theta)^{-1}$, calculus shows the representations to be equivalent.

Note that if μ assigns mass 0 to $(0, \pi/2)$, then D^* and C^* are independent, since deaths can only be caused by points on R_D and censorings can only be caused by points on R_C, which are independent Poisson processes. Loosely, more mass of μ near the center $\pi/4$ corresponds to stronger dependence, while more mass near the edges corresponds to weaker dependence.

As mentioned previously, the model involves marginal power transformations of this kind of bivariate exponential distribution. Let $F_\alpha(x) = 1 - \exp(-x^\alpha)$, the Weibull cumulative distribution function with scale parameter 1 and shape parameter α, for $\alpha > 0$. For any measure M as defined previously and the derived random variables D^* and C^*, let $D = F_\alpha^{-1}(1 - e^{-D^*}) = (D^*)^{1/\alpha}$ and $C = F_\beta^{-1}(1 - e^{-C^*}) = (C^*)^{1/\beta}$. Then D and C have a bivariate Weibull distribution with parameters α and β. In what follows, other shape families could be substituted for the Weibull and similar results would obtain. For simplicity we will discuss only the Weibull. For possible ease of application of the result to other shape families, formulas will be given in general terms involving the marginal distribution function rather than the special transformations for the Weibull.

The full model will now be formulated. The unknown parameters are λ_d, λ_c, α, β, and the measure μ. To make the model identifiable, the restriction $\alpha \neq \beta$ is placed on the model. If one is considering this model in practice, this is not likely to be too serious a restriction. Thus, the final model is that a sequence of independent replicates of $T = \min(D, C)$ and $\delta = 1_{D<C}$ are observable. Their joint distribution is assumed to arise from a bivariate Weibull distribution as before, with unknown parameters α, β, λ_d, λ_c, and μ, a

measure. Restrictions are that $\alpha \neq \beta$, either $\mu((0, \pi/2))$ is positive or both of λ_d and λ_c are positive, and

$$\int_{(0,\pi/2)} \left(\frac{1}{\sin\theta} + \frac{1}{\cos\theta} \right) \mu(d\theta) < \infty.$$

The set of all parameter values satisfying these restrictions will be denoted A.

3. The likelihood. To derive the likelihood of T and δ, recall that

$$P(D^* > d \cap C^* > c) = \exp\left[-d\left(\lambda_d + \int_{[r,\pi/2)} \frac{\mu(d\theta)}{\cos\theta} \right) - c\left(\lambda_c + \int_{(0,r)} \frac{\mu(d\theta)}{\sin\theta} \right) \right],$$

where $r = \arctan(c/d)$. In the preceding formulas a simple calculation shows that the value of the integral is the same whether any atom of μ at r is included in the first integral or the second. It has arbitrarily been placed in the first. Next, by transforming D^* and C^* to D and C, we see that

$$\begin{aligned} P(D > d \cap C > c) &= P\left(D^* > -\log\left(\bar{F}_\alpha(d)\right) \cap C^* > -\log\left(\bar{F}_\beta(c)\right)\right) \\ &= \exp\left[\log\left(\bar{F}_\alpha(d)\right)\left(\lambda_d + \int_{[R,\pi/2)} \frac{\mu(d\theta)}{\cos\theta} \right) \right. \\ &\qquad \left. + \log\left(\bar{F}_\beta(c)\right)\left(\lambda_c + \int_{(0,R)} \frac{\mu(d\theta)}{\sin\theta} \right) \right], \end{aligned} \tag{3.1}$$

where $R = \arctan(\log(\bar{F}_\beta(c))/\log(\bar{F}_\alpha(d)))$. For the Weibull, $-\log(\bar{F}_\alpha(x))$ is just x^α.

Define

$$R(t) = \arctan \frac{\log\left(\bar{F}_\beta(t)\right)}{\log\left(\bar{F}_\alpha(t)\right)}.$$

Then,

$$\begin{aligned} \bar{F}_T(t) = \exp\left[\log\left(\bar{F}_\alpha(t)\right)\left(\lambda_d + \int_{[R(t),\pi/2)} \frac{\mu(d\theta)}{\cos\theta} \right) \right. \\ \left. + \log\left(\bar{F}_\beta(t)\right)\left(\lambda_c + \int_{(0,R(t))} \frac{\mu(d\theta)}{\sin\theta} \right) \right]. \end{aligned} \tag{3.2}$$

Differentiating the negative of the logarithm of (3.2) with respect to t yields the hazard function for T,

$$\begin{aligned} h_T(t) = &\left[-\frac{d}{dt} \log\left(\bar{F}_\beta(t)\right) \right]\left[\lambda_c + \int_{(0,R(t))} \frac{\mu(d\theta)}{\sin\theta} \right] \\ &+ \left[-\frac{d}{dt} \log\left(\bar{F}_\alpha(t)\right) \right]\left[\lambda_d + \int_{(R(t),\pi/2)} \frac{\mu(d\theta)}{\cos\theta} \right] \\ &+ \mu\{R(t)\}\left[\log^2\left(\bar{F}_\alpha(t)\right) + \log^2\left(\bar{F}_\beta(t)\right) \right]^{1/2}. \end{aligned} \tag{3.3}$$

The first term in (3.3) is the hazard at time t due to censoring, the second is the hazard due to death, while the third is the hazard for simultaneous censoring and death.

PROPOSITION 3.4. $P(\mu\{R(T)\} \neq 0) = 0$.

PROOF. Consider the set of points $Q = \{(d, c) \in \Omega | d^\alpha = c^\beta\}$. For any atom θ of μ, Q intersects the ray $\tan(\theta) = d/c$ in only one point. For any such intersection point (d_0, c_0), $M(\{d = d_0 \cup c = c_0\}) = 0$. Since μ can have only a countable number of atoms, the union U of all such sets has M-measure 0 as well. However, $\mu\{R(T)\} \neq 0$ if and only if the point from the Poisson process leading to the observed death or censoring lies in U. Since $M(U) = 0$, the proposition is proven. □

Since $P(\mu\{R(T)\} \neq 0) = 0$, the last term in (3.3) can be dealt with arbitrarily in the joint likelihood of T and δ. It could be included in either of the previous two terms by closing the corresponding endpoint of the interval of integration. As we shall see later, in order that maximum likelihood estimates exist, this term must be included in each of the other two hazard terms by closing both intervals of integration. This will make both hazards, and hence the density, upper semicontinuous. The joint likelihood of T and δ can then be written, for $t > 0$, $i = 0$ or 1,

$$
\begin{aligned}
f(t,i) = \exp\Bigg(&\log\big(\bar{F}_\alpha(t)\big)\left[\lambda_d + \int_{[R(t),\pi/2)} \frac{\mu(d\theta)}{\cos\theta}\right] \\
&+\log\big(\bar{F}_\beta(t)\big)\left[\lambda_c + \int_{(0,R(t))} \frac{\mu(d\theta)}{\sin\theta}\right]\Bigg) \\
&\times\left(\left[-\frac{d}{dt}\log\big(\bar{F}_\beta(t)\big)\right]\left[\lambda_c + \int_{(0,R(t)]} \frac{\mu(d\theta)}{\sin\theta}\right]\right)^{1-i} \\
&\times\left(\left[-\frac{d}{dt}\log\big(\bar{F}_\alpha(t)\big)\right]\left[\lambda_d + \int_{[R(t),\pi/2)} \frac{\mu(d\theta)}{\cos\theta}\right]\right)^{i}.
\end{aligned} \tag{3.5}
$$

The first term is $\bar{F}_T(t)$. The last two terms are, respectively, the hazards of observable censoring and observable death $h_T(t,0)$ and $h_T(t,1)$ at time t.

For future reference note that

$$
F_{T,\delta}(t,1) = \int_0^t \bar{F}_T(s) h_T(s,1)\, ds \tag{3.6}
$$

is the subdistribution function of observable deaths.

In practice the survival distribution F_D is usually of foremost interest. In a model of this form it is a Weibull distribution with shape parameter α and scale parameter

$$
\Lambda_d = \lambda_d + \int_{(0,\pi/2)} \frac{\mu(d\theta)}{\cos\theta}.
$$

Similarly, the marginal censoring distribution is Weibull with shape parameter β and scale parameter

$$\Lambda_c = \lambda_c + \int_{(0,\pi/2)} \frac{\mu(d\theta)}{\sin\theta}.$$

For any dependent $F_{D,C}$ there is an independent bivariate distribution with the same $F_{T,\delta}$ [see, e.g., Cox (1959)]. For any bivariate Weibull (D, C) considered here, denote such a pair of independent random variables by D^i and C^i. Geometrically, it is easy to describe these variables: D^i is the leftmost point above or to the left of the curve $D = C$; C^i is similarly the lowest point below or to the right of this curve. The hazard functions for D^i and C^i are simply the same as the observable hazards, i.e., $h_{D^i}(t) = h_T(t, 1)$ and $h_{C^i}(t) = h_T(t, 0)$. Their cumulative distribution functions can be found by exponentiating the integrated hazards or, equivalently, by integrating M over the relevant regions of a graph like Figure 1 and using a Poisson process argument. One of D^i or C^i may be improper if λ_d or λ_c is zero.

LEMMA 3.7. *Fix α and β. Then the likelihood of a single observation $f(t, i)$ as a function of λ_d, λ_c, and μ is bounded by $(\beta/t)(1 - i) + (\alpha/t)i$. Further, if Λ_d (Λ_c) is bounded above by K, then for t near 0, $f(t, i)$ is bounded above by $\alpha K t^{\alpha-1} i + \beta K t^{\beta-1}(i - 1)$.*

PROOF. If $i = 0$, the maximum is attained by taking

$$\lambda_c + \int_{(0,R(t)]} \frac{\mu(d\theta)}{\sin\theta} = t^{-\beta} \quad \text{and} \quad \lambda_d + \int_{[R(t),\pi/2)} \frac{\mu(d\theta)}{\cos\theta} = 0.$$

The result follows, and a similar calculation proves the case $i = 1$. The second assertion can be proven similarly. □

4. Identifiability. Identifiability follows from a demonstration that the model parameters can be obtained from the distribution functions and likelihood. Dividing the likelihood $f(t, i)$ by $P(T > t)$ and setting i to 0 or 1 gives the individual hazards. $h_T(t, 0)/h_T(t, 1)$ is an increasing function of t if and only if $\beta > \alpha$ and is a decreasing function of t if and only if $\beta < \alpha$. Thus, the hazards determine which of these cases obtains. If $\beta > \alpha$, then

$$\beta = 1 + \lim_{t\to\infty} \frac{\log h_T(t, 0)}{\log t} \quad \text{and} \quad \alpha = 1 + \lim_{t\to 0} \frac{\log h_T(t, 1)}{\log t}.$$

The case $\beta < \alpha$ is similar. Once β and α are known, then all the integrals

$$\lambda_c + \int_{(0,t)} \frac{\mu(d\theta)}{\sin\theta} \quad \text{and} \quad \lambda_d + \int_{[t,\pi/2)} \frac{\mu(d\theta)}{\cos\theta}$$

are known. These clearly determine μ, λ_d, and λ_c.

Note that the same sort of identifiability proof will work for other families of marginal transformations besides the Weibull. Let $\{F_d, F_c\}$ be a family of

pairs of marginal transformations. The following proposition gives one set of circumstances under which the model will be identifiable. Its proof is an exact analog of the Weibull case.

PROPOSITION 4.1. *In order that the complete model* F_d, F_c, λ_d, λ_c, μ *be identifiable, it is sufficient that the following two conditions hold*:

(a) F_d *and* F_c *are identifiable from* $F_{T,\delta}$;
(b) *for any pair* F_d, F_c, *the range of* $R(t)$ *is the interval* $(0, \pi/2)$.

This makes clear why the case $\alpha = \beta$ must be excluded; condition (b) would be violated and the parameters λ_d, λ_c, and μ could not be identified. Further, although the shape parameter α of the survival distribution F_D can be identified, the scale parameter cannot be.

5. Consistency of the maximum likelihood estimator. The purpose of this section is to consider maximum likelihood estimation of the parameters of the joint distribution. It will be seen that maximum likelihood estimates exist with probability approaching 1 as the sample size grows. In fact, they will exist for practically any natural data set. The maximum likelihood estimates will in general be nonunique. However, any choice of MLE will be shown to be consistent for the true parameter. If interest is only in the marginal survival distribution, then much of the nonuniqueness will disappear.

THEOREM 5.1. *Let* A_n^* *denote the set of MLEs based on a sample of size* n. *Then* $P(A_n^* \text{ empty}) \to 0$ *as* $n \to \infty$ *and the set* A_n^* *converges in probability to the singleton* $\{a = (\alpha, \beta, \lambda_d, \lambda_c, \mu)\}$, *the true parameter value.*

The proof will follow a slight modification of a theorem of Bahadur (1967), as given in Grenander [(1981), pages 349–353]. The plan of the proof is to compactify the parameter space A by taking all weak limits of the distributions in the model and then to verify that all the conditions of Grenander (1981) are satisfied. A quirk of the Weibull family is that if X_n has a univariate Weibull distribution with parameters α_n and λ_n, where $\alpha_n \to \infty$ and $\lambda_n \sim x^{-\alpha_n}$, for some $x > 0$, then X_n converges in distribution to a point mass at x. It follows that for a single observation the Weibull likelihood is unbounded and the conditions for consistency of the MLE do not apply immediately. However, for two or more observations this problem evaporates. A similar pathology can occur for the bivariate Weibull if there is at most one death or at most one censored observation. For proving consistency in the bivariate situation, the following remedy will be used. First, prove that if the maximization is restricted to any subset of the parameter space of the form $\{a|\max(\alpha, \beta) \leq K\}$ and the true value of the parameter lies in this set, then this restricted MLE is consistent. Next, it will be shown that the unrestricted MLE must eventually be in a subset of the above form (the subset depends on

the true value of the parameters), so the unrestricted MLE will eventually be the same as the restricted MLE and will thus be consistent as well.

For convenience, the theorem of Bahadur, as in Grenander (1981), will be restated here. Let A be a metric space with distance $d(\cdot, \cdot)$. Let X denote the space on which the random variables of interest are defined. Assume each distribution in A has a density f with respect to a σ-finite measure ν.

DEFINITION 5.2. A compact metric space $\bar{A}$ is said to be a suitable compactification of A if the following hold:

(a) A is an everywhere dense subset of $\bar{A}$;

(b) for any $a_0 \in \bar{A}$, the function

$$f(x, a_0, \varepsilon) = \sup\{ f(x, a) | a \in A, d(a, a_0) < \varepsilon\}$$

is measurable in x for ε small enough;

(c) defining $f(x, a_0, 0) = \lim_{\varepsilon \to 0} f(x, a_0, \varepsilon)$, we have

$$\int_X f(x, a_0, 0) \nu(dx) \leq 1 \quad \text{for any } a_0 \in \bar{A};$$

for $B \subseteq A$ and h an extended real-valued function, write $h(B) = \sup\{h(a) | a \in B\}$.

For a sample $x = (x_1, \ldots, x_n)$, consider the likelihood function

$$L_n(a; x) = \prod_{r=1}^{n} f(x_r, a).$$

The set of maximum likelihood solutions, possibly empty, is

$$\hat{A}_n = \{a | a \in A, L_n(a; x) = L_n(A; x)\}.$$

THEOREM 5.3 (Bahadur). *Assume that the following hold*:

(a) *There exists a suitable compactification* $\bar{A}$;

(b) $$E_a\left[\log \frac{f(x, A)}{f(x, a)}\right] < \infty \quad \textit{for any } a \in A;$$

(c) *for* $a \in A$ *and* $a_0 \in \bar{A}$ *with* $a \neq a_0$, *we have*

$$\nu\{x | f(x, a) \neq f(x, a_0, 0)\} > 0;$$

(d) A *is open in* $\bar{A}$;

(e) $f(x, a) = f(x, a, 0)$ *for all* x, *for any* $a \in A$.

Under conditions (a)–(e) *it is true with probability* 1 *that* $\hat{A}_n$ *is not empty for* n *large enough, and the set* $\hat{A}_n$ *converges to the one containing the single point* $a =$ *true value of the parameter*.

To apply Theorem 5.3, a suitable metric must be defined. This metric, denoted d_1, metrizes weak convergence of random vectors T, δ when mapped into $[0, 1] \times \{0, 1\}$. For technical convenience a second metric d_2 is defined as well. For $s \in [0, 1]$ and $a \in A$, let

$$G_a(s, i) = F_{T,\delta}\left(\frac{s}{1-s}, i\right) = P_a\left(\frac{T}{1+T} \le s, \delta = i\right) \quad \text{for } i = 0, 1.$$

Then, $G_a(s, 1) + G_a(s, 0) = P_a(T/(1 + T) \le s)$. Define a metric on A by

$$d_1(a, b) = \inf\{\varepsilon | G_a(s - \varepsilon, i) - \varepsilon \le G_b(s, i) \le G_a(s + \varepsilon, i) + \varepsilon \quad \text{for } i = 0, 1\}. \tag{5.4}$$

Convergence in this metric is equivalent to weak convergence of $(T/(1 + T), \delta)$. In terms of d_1 the assertion of Theorem 5.1 is

$$\lim_{n \to \infty} P\left(\sup_{a_1 \in A_n^*} d_1(a_1, a) > \varepsilon\right) = 0 \quad \text{for all } \varepsilon > 0. \tag{5.5}$$

Consider as well another metric d_2 defined on A. For $s \in (0, 1)$ and $a \in A$, let

$$j_d(a, s) = \lambda_d + \int_{[R(s/(1-s)),\, \pi/2)} \frac{\mu(d\theta)}{\cos\theta}$$

and

$$j_c(a, s) = \lambda_c + \int_{(0,\, R(s/(1-s))]} \frac{\mu(d\theta)}{\sin\theta}.$$

Define d_2 by

$$d_2(a_1, a_2) = \max\Big(|\alpha_1 - \alpha_2|, |\beta_1 - \beta_2|, \inf_{\varepsilon > 0}\{\varepsilon | \forall\, s \in (0, 1), \; j_d(a_1, s - \varepsilon) - \varepsilon \le j_d(a_2, s) \le j_d(a_1, s + \varepsilon) + \varepsilon \text{ and } j_c(a_1, s - \varepsilon) - \varepsilon \le j_c(a_2, s) \le j_c(a_1, s + \varepsilon) + \varepsilon\}\Big). \tag{5.6}$$

Later the parameter space will be compactified by adding some limit points in d_1. To prove Theorem 5.1, these must be identified. Consider a sequence of distributions $\{\alpha_n, \beta_n, \lambda_{dn}, \lambda_{cn}, \mu_n\}$ in A. By tightness this sequence must have a d_1 convergent subsequence. Then, without loss of generality, suppose the full sequence is convergent. To examine the possible limit points consider the following possibilities.

5.7.1. All of α_n, β_n, $|\alpha_n - \beta_n|$, Λ_{dn} and Λ_{cn} remain bounded away from 0 and ∞ as $n \to \infty$.

Then any subsequence has a further subsequence on which α_n, β_n, Λ_{dn} and Λ_{cn}, and hence μ, λ_d and λ_c as well, converge in d_2 to a point in A. Each of these further subsequences must converge in d_2 to the same limit point, or else the limit points will be indistinguishable, contradicting the identifiability

results of the last section. Thus, the full sequence must converge to this d_2 limit point, a point in A.

5.7.2. All of α_n, β_n, Λ_{dn} and Λ_{cn} remain bounded away from 0 and ∞ as $n \to \infty$ but $|\alpha_n - \beta_n| \to 0$. Then the limiting distribution of D, C is a bivariate Weibull. This distribution is not necessarily representable as the result of marginal transforms with the same shape parameters of a bivariate exponential unless a randomization scheme for deciding ties is introduced. The limiting cumulative distribution function of T,

$$F_T(t) = \lim_{n\to\infty} G\left(\frac{t}{1+t}, 0\right) + G\left(\frac{t}{1+t}, 1\right)$$

is a Weibull cumulative distribution function.

5.7.3. $\alpha_n \to 0$ or $\beta_n \to 0$. If $\alpha_n \to 0$ the limit of $G(s, 1)$ is flat except for jumps at 0 and 1, since the limiting distribution of deaths has point masses at 0 and "∞" (the masses depending on the behavior of Λ_{dn}). A similar situation arises if $\beta_n \to 0$.

5.7.4. Neither α_n nor β_n approaches zero or ∞, but Λ_{dn} or Λ_{cn} approaches 0 or ∞. Then as in 5.7.3 the limiting distribution of the deaths or censoring times will be a point mass at 0 or "∞," and hence either $G(s, 1)$ or $G(s, 0)$ will be flat except for a jump at 0 or 1.

5.7.5. $\alpha_n \to \infty$ or $\beta_n \to \infty$. If $\alpha_n \to \infty$ and $\Lambda_d \sim d_0^{-\alpha_n}$, then the limiting distribution of the deaths will have a point mass at d_0, hence $G(s, 1)$ will be flat except for a jump at $d_0/(1 + d_0)$. Similar situations can occur for the censoring variable. This situation will be ruled out in the proof and hence will be of no further concern.

LEMMA 5.8. *Consider a sequence of points $a_n \in A$ and another point $a \in A$. Then $d_1(a_n, a) \to 0$ if and only if $d_2(a_n, a) \to 0$.*

PROOF. Clearly, d_2 convergence implies d_1 convergence. For the reverse implication, suppose a_n has a subsequence that converges to a in d_1 but not in d_2. From the preceding discussion, d_1 convergence implies that α_n, β_n, $|\alpha_n - \beta_n|$, Λ_{dn} and Λ_{cn} all must eventually be bounded away from 0 and ∞. Thus, any subsequence has a d_2 convergent further subsequence. The convergence of any such subsequence to a point a_0 other than a would contradict identifiability; a_0 and a could not be distinguished. This implies that every subsequence has a further subsequence converging in d_2 to a, so the full sequence must converge to a. □

PROPOSITION 5.9. *Consider the following restriction of the original parameter space*

$$A_K = \{\alpha, \beta, \lambda_d, \lambda_c, \mu | \max(\alpha, \beta) \le K\}.$$

If the true parameter value a is in A_K and maximization is restricted to A_K, then Theorem 5.3 *applies and the MLE is consistent.*

PROOF. Here x is the pair (t, i). In what follows a will always refer to the true parameter value, while other points of the parameter space will be distinguished by subscripts. For any fixed K, compactify A_K by adding all $G(s,i)$ that are limit points of distributions in A_K in the metric d_1. Denote this space $\bar{A}_K$. This space is compact by Prohorov's theorem [see, e.g., Huber (1981), page 24]. To verify that this is a suitable compactification, notice that distributions of types 5.7.2, 5.7.3 and 5.7.4 only are in $\bar{A}_K \setminus A_K$.

The conditions in Bahadur's theorem will now be verified. First note that the compactification $\bar{A}_K$ is suitable because of the following:

(a) It is obtained by taking limit points of members of A_K, so A_K must be dense in it.

(b) Since the likelihood is upper semicontinuous, for any point in A_K the set where the likelihood is greater than a constant l is open, hence a similar set for $f(x, a_0, \varepsilon)$ is open and measurable.

(c) For limit points of types 5.7.3 and 5.7.4, it can be seen that $f((t, i), a_0, \varepsilon)$ converges to 0 almost everywhere as $\varepsilon \to 0$ for $i = 0$ or 1 or both. The other value of i can be handled as in type 5.7.1 below. For a type 5.7.2 limit point a_0 where both α_n and β_n converge to γ, choose a sequence that converges to a_0 in d_1 and choose a d_2 convergent subsequence. The marginal hazards of the limiting distribution are

$$h_T(t,1) = \gamma t^{\gamma-1} \lim_{n\to\infty} \lambda_{dn} + \int_{[R_n(t),\ \pi/2)} \frac{\mu_n(d\theta)}{\cos\theta}$$

and

$$h_T(t,0) = \gamma t^{\gamma-1} \lim_{n\to\infty} \lambda_{cn} + \int_{(0,\ R_n(t)]} \frac{\mu_n(d\theta)}{\sin\theta}.$$

As the limit of products of pairs of monotone functions, each $h_T(t,i)$ will be continuous almost everywhere. Further, any sequence that converges in d_1 to a_0 will have a d_2 convergent subsequence whose hazard functions converge to $h_T(t,i)$ at all continuity points of $h_T(t,i)$. Fix t_0, i_0 such that both the hazard functions of the limiting distribution are continuous at t. By Lemma 3.7 it is possible to choose a sequence of points a_n in A such that a_n is within a distance n^{-1} of a and

$$f((t_0,i_0),a_n) \geq f\big((t_0,i_0),a_0,n^{-1}\big) - \frac{1}{n}.$$

The sequence a_n contains a d_2 convergent subsequence, so along that subsequence the hazard functions for a_n must converge to those for a_0. However, by definition of the sequence a_n, the whole sequence $f((t_0,i_0),a_n)$ converges. Since this holds on a set of t values of full measure, condition (c) of Theorem 5.3 is satisfied for this type of limit point. For type 5.7.1 limit points, points in

A_K, a stronger result will be given in condition (e). Thus, condition (a) of Theorem 5.3 is satisfied.

Condition (b) follows since $f((t,i);A) = K/t$, and $\log t$ is integrable with respect to any of these distributions. Clearly, $E_a(\log f) < \infty$.

Condition (c) follows from the identifiability section and the discussion of the possible points in $\bar{A}_K$.

Condition (d) follows from the discussion of the points in $\bar{A}_K \setminus A_K$. From consideration of the marginal distribution of T for case 5.7.2 and $G(s,i)$ for cases 5.7.3 and 5.7.4, it is clear that no sequence of them could converge to a point in A_K.

Condition (e) is the reason for the definition of the joint likelihood given. As above, given $a \in A_K$ and (t_0, i_0), choose a sequence of points a_n converging to a in d_1 such that $f((t_0,i_0),a_n) > f((t_0,i_0),a,n^{-1}) - n^{-1}$. This sequence must converge in d_2 as well, by Lemma 5.8. The upper semicontinuity of the likelihood function implies condition (e) is satisfied. □

PROOF OF THEOREM 5.1. We must show that almost surely as $n \to \infty$, the supremum of the likelihood will be larger on some compact set A_K (depending on the true parameter value) than the supremum off the set, so almost surely as $n \to \infty$ the MLE on this compact set, when it exists, will in fact be the global MLE.

Fix the true parameter value a and consider the likelihood for some a_0 with $\beta_0 < \alpha_0$. Consider observations $\{x_1, \ldots, x_n\} = \{(T_1, \delta_1), \ldots, (T_n, \delta_n)\}$, without loss of generality written in order of increasing T_i. Let Δ_n denote $\sum_{i=1}^n \delta_i$. A portion of the likelihood at a_0 corresponding to the deaths is

$$\prod_{\delta_i=1} \alpha_0 \lambda_i T_i^{\alpha_0 - 1} e^{-\lambda_i T_i^{\alpha_0}}, \tag{5.10}$$

where

$$\lambda_i = \lambda_d + \int_{[R(T_i),\ \pi/2)} \frac{\mu(d\theta)}{\cos\theta}.$$

The remaining terms in the likelihood are bounded above by $\prod_{\delta_i=0} \alpha_0 / T_i$. Since $\beta_0 < \alpha_0$, $R(t)$ is a decreasing function of t; hence, the λ_i are nondecreasing. Subject to this restriction, the values of λ_i that maximize (5.10) are

$$\lambda_i \equiv \frac{1}{\Delta_n} \sum_{\delta_j=1} T_j^{\alpha_0}.$$

The product (5.10) is then less than

$$\prod_{\delta_i=1} \frac{\alpha T_i^{\alpha} \Delta_n}{T_i \sum_{\delta_i=1} T_i^{\alpha}}.$$

Thus, the whole log likelihood at α_0 is bounded by

$$n \log \alpha_0 + \sum_1^n \log T_i^{-1} + \sum_{\delta_i=1} \log\left\{ \frac{T_i^{\alpha_0}\Delta_n}{\sum_{\delta_i=1} T_i^{\alpha_0}} \right\}. \tag{5.11}$$

Note the following proposition, whose proof will be deferred.

PROPOSITION 5.12. *For any fixed $h > 0$, there exists an α_0 depending on a, such that*

$$\lim_{n\to\infty} n^{-1} \sup_{\alpha_1|\alpha_1=\alpha_0,\, \beta_1<\alpha_0} \log\left(\frac{L(\{x_1,\ldots,x_n\}, a_1)}{L(\{x_1,\ldots,x_n\}, a)} \right) < -h.$$

Next, choose $\alpha_0 > 1$ large enough to satisfy the conclusion of Proposition 5.12 with $h = 1$ and large enough that $\lim_{n\to\infty} n^{-1}H(\alpha_0, n) < -\log \alpha_0 - 1$. Then, once $n^{-1}H(\alpha_0, n) + \log \alpha_0 < 0$, which for almost all samples will hold from some n_0 on, (5.13) implies that

$$\sup_{\alpha_1 \geq \alpha_0} \log \alpha_1 + n^{-1}H(\alpha_1, n)$$

is attained at α_0. A similar argument with $\alpha_0 < \beta_0$ shows that almost surely as $n \to \infty$, the supremum of the likelihood off some set $A_K = \{a_0 | \max(\alpha, \beta) \leq K\}$ is less than the likelihood at the true parameter value. Thus, with probability going to 1, the MLE on some set A_K, when it exists, will be the global MLE. This combined with Proposition 5.9 proves Theorem 5.1. □

PROOF OF PROPOSITION 5.12. The second term of (5.11) and

$$\log L(\{x_1,\ldots,x_n\}, a)$$

grow linearly in n by the strong law of large numbers. It suffices to show that by making α_0 large enough the remaining two terms of (5.11), when divided by n, can be made to be asymptotically less than $-h$ for any $h > 0$. Let E_n denote averaging over the empirical distribution function of the deaths for a given sample of size n,

$$E_n g = \delta_n^{-1} \sum_{\delta_i=1} g(T_i).$$

Write the last term of (5.11) as

$$H(\alpha_0, n) = \Delta_n(\alpha_0 E_n \log T - \log E_n T^{\alpha_0}).$$

For any fixed α_0, note by the concavity of the logarithm that $H(\alpha_0, n) < 0$ with conditional probability 1 on the set $\Delta_n > 1$. Since $P(\Delta_1 = 0) < 1$ for any $a \in A$, the strong law of large numbers implies $H(\alpha_0, n)/n$ converges almost surely to some constant $c < 0$. Further, by the convexity of x^γ for $\gamma > 1$,

$$\log E_n T^{\alpha_0\gamma} \geq \gamma \log E_n T^{\alpha_0}.$$

This implies

$$|H(\alpha\gamma, n)| \geq |\gamma H(\alpha, n)| \quad \text{for } \gamma > 1. \tag{5.13}$$

This all says that by choosing α_0 large enough, $\lim_{n\to\infty} n^{-1}H(\alpha_0, n)$ can be made arbitrarily small. By (5.13), increasing α_0 will decrease the third term of (5.11) linearly in α_0, while it will increase the first only logarithmically. Thus, α_0 can be chosen large enough to satisfy the proposition. □

The convergence in Theorem 5.1 is weak convergence of the transformed distribution functions $G(s, i)$. This can be combined with Lemma 5.8 to give a more direct statement about the parameter values themselves. An argument by contradiction using Lemma 5.8 and (5.5) shows

$$\lim_{n\to\infty} P\left(\sup_{a_1 \in A_n^*} d_2(a_1, a) > \varepsilon \right) = 0 \quad \text{for all } \varepsilon > 0.$$

The definition (5.6) of d_2 then says that estimates of α and β converge in probability to the true values, while estimates of

$$\lambda_d + \int_{(r,\, \pi/2)} \frac{\mu(d\theta)}{\cos\theta} \quad \text{and} \quad \lambda_c + \int_{(0,r)} \frac{\mu(d\theta)}{\sin\theta}$$

converge weakly in probability to the true values.

6. Maximum likelihood estimation in practice. In this section an algorithm for fitting the maximum likelihood estimator is presented. Two standard minimization programs are combined to give a slow, but effective, algorithm for estimation.

First, a more convenient notation for the data is introduced. Given α and β, let $(T_1, d_1, c_1), \ldots, (T_N, d_N, c_N)$ be the observed data, sorted so that $T_i^{\beta-\alpha}$ is increasing in i, where d_i is the number of failures occurring at time T_i and c_i is the corresponding number of observations newly censored at time T_i. If there are no ties in the observed data, then $d_i = \delta_i$ and $c_i = 1 - \delta_i$. Let $R_i = \arctan t_i^{\beta-\alpha}$.

First note that, given α and β, the search for the maximum can be narrowed to a discrete measure μ belonging to a fairly restricted set.

PROPOSITION 6.1. *Consider maximizing the likelihood for fixed α and β. The likelihood for any parameter value $(\lambda_d, \lambda_c, \mu)$ can be dominated by the likelihood of another parameter $(\hat{\lambda}_d, \hat{\lambda}_c, \nu)$, for which $\hat{\lambda}_d = \hat{\lambda}_c = 0$ and ν is supported on the points $R_1, \ldots, R_N$.*

PROOF. For any $(\lambda_d, \lambda_c, \mu)$, μ can be written as the sum of two measures, one μ^d with support $\cup_{i=1}^N \{R_i\}$ and the other μ^* assigning measure zero to this set. For $i = 1, \ldots, N-1$, let μ^i be the measure concentrated on R_i and R_{i+1} satisfying

$$\int_{[R_i, R_{i+1}]} \frac{1}{\cos(\theta)} (\mu^i(d\theta) - \mu^*(d\theta)) = 0$$

and

$$\int_{[R_i, R_{i+1}]} \frac{1}{\sin(\theta)} \left(\mu^i(d\theta) - \mu^*(d\theta)\right) = 0.$$

Further, let μ^0 be a point mass at R_1 satisfying

$$\frac{\mu^0\{R_1\}}{\sin(R_1)} = \lambda_c + \int_{(0, R_1]} \frac{1}{\sin(\theta)} \mu^*(d\theta),$$

and let μ^N be a point mass at R_N satisfying

$$\frac{\mu^N\{R_N\}}{\cos(R_N)} = \lambda_d + \int_{[R_N, \pi/2)} \frac{1}{\cos(\theta)} \mu^*(d\theta).$$

Let ν be the discrete measure $\mu^d + \sum_{i=0}^N \mu^i$. Then it is easy to check that for any observation T_i, $i = 1, \ldots, N$, the first term of the likelihood (3.5) is unchanged, while the second and third terms cannot be decreased. □

Thus, attention will be restricted to measures ν of this form, and λ_d and λ_c will be dropped for the remainder of this section.

Note that slightly more can be learned from the proof of Proposition 6.1. For fixed α and β, unless the optimal ν puts mass on both T_i and T_{i+1}, where $c_i = 0$ and $d_{i+1} = 0$, or on an endpoint T_1 if $d_1 = 0$ or T_N if $c_N = 0$, no measure μ can maximize the likelihood except one of the form ν. Thus, if $\hat{\nu}$ is the unique MLE of $(\lambda_d, \lambda_c, \mu)$ in the class of measures with support $\bigcup_{i=1}^N \{R_i\}$, and if $\hat{\nu}$ has support satisfying the above restriction, then $\hat{\nu}$ is unique in the class of all measures.

As a notational convenience, let $\nu_i = \nu\{R_i\}\sqrt{T_i^{2\alpha} + T_i^{2\beta}}$. Further, let

$$D_i = \int_{[R_i, \pi/2)} \frac{\nu(d\theta)}{\cos(\theta)} = \sum_{j=i}^N \frac{\nu_j}{T_j^\alpha}, \qquad D_{N+1} = 0,$$

and

$$C_i = \int_{(0, R_i]} \frac{\nu(d\theta)}{\sin(\theta)} = \sum_{j=1}^i \frac{\nu_j}{T_j^\beta},$$

and $C_0 = 0$.

Since the programs used to fit the MLE involve minimization rather than maximization, from here on we will work with the negative of the log likelihood, which can now be written

$$\begin{aligned} -\log L(\alpha, \beta, \nu(\alpha, \beta)) &= \sum_{i=1}^N (d_i + c_i)\left(T_i^\alpha D_i + T_i^\beta C_{i-1}\right) \\ &\quad - \sum_{i=1}^N d_i \log\left(\alpha T_i^{\alpha-1} D_i\right) - \sum_{i=1}^N c_i \log\left(\beta T_i^{\beta-1} C_i\right). \end{aligned} \tag{6.2}$$

Next note that, for fixed α and β, there is a unique value of ν that minimizes (6.2).

PROPOSITION 6.3. *For fixed α and β, the function* (6.2) *is a strictly convex function of ν on the set $V = \{\nu_i \geq 0,\ i = 1, \ldots, N\}$ and thus has a unique minimum in this set.*

PROOF. We calculate the Hessian of (6.2) and verify that it is positive definite where $-\log L$ is finite. This will prove the proposition. Less notation is involved in regarding (6.2) as a function of $\nu_1, \ldots, \nu_N$, so these will be used as the variables in differentiation. Let $\mathbf{A}$ and $\mathbf{B}$ be the matrices $\operatorname{diag}(T_1^{-\alpha}, \ldots, T_N^{-\alpha})$ and $\operatorname{diag}(T_1^{-\beta}, \ldots, T_N^{-\beta})$. Let $\mathbf{C}$ and $\mathbf{D}$ be the matrices $\operatorname{diag}(d_1/D_1^2, \ldots, d_N/D_N^2)$ and $\operatorname{diag}(c_1/C_1^2, \ldots, c_N/C_N^2)$, with the convention that $0/0 = 0$. Note that $\mathbf{C}$ and $\mathbf{D}$ have finite entries whenever (6.2) is finite. Finally, let Δ be an $N \times N$ lower triangular matrix of ones: $\Delta_{ij} = I(i \geq j)$. Let t denote transpose. Then the Hessian of (6.2) is

$$H = \mathbf{A}\Delta\mathbf{D}\Delta^t\mathbf{A} + \mathbf{B}\Delta^t\mathbf{C}\Delta\mathbf{B}.$$

This is clearly positive semidefinite. A simple calculation shows that for a vector x, $x^tHx = 0$ implies $x = 0$ where $-\log L$ is finite. Briefly, let E_k be the matrix with entries $E_{kij} = I(i \geq k,\ j \geq k)$. Then $\Delta\mathbf{D}\Delta^t = \sum_{k=1}^N \mathbf{D}_{kk}E_k$. Thus, for each k, either $\mathbf{D}_{kk} = 0$ or $\sum_{j \geq k}(\mathbf{A}x)_j = 0$. A similar set of equalities is obtained from $\mathbf{B}$ and $\mathbf{C}$. If the likelihood is finite, then for each k, at least one of $\mathbf{D}_{kk}$ and $\mathbf{C}_{kk}$ must be nonzero. This gives a set of at least N homogeneous linear equations that x must satisfy to have $x^tHx = 0$. It is straightforward to check that N of the equations are linearly independent, and thus x must be the zero vector. □

Thus, for fixed α and β, minimizing (6.2) over ν is relatively easy. It is a linearly constrained minimization problem in N variables with a convex objective function. The Fortran subroutine NPSOL [Gill, Murray, Saunders and Wright (1986)] is tailored for this kind of problem, and it is used in this part of our algorithm.

Before discussing estimation of α and β, we note an interesting fact about estimation of ν. After deleting terms that depend only on α and β, (6.2) can be written

$$\sum_{i=1}^{N} (c_i + d_i)D_i - d_i \log(D_i) + \sum_{i=1}^{N} (c_i + d_i)C_{i-1} - c_i \log(C_i).$$

This is the sum of two objective functions in a form of isotonic regression [see Barlow, Bartholomew, Bremner and Brunk (1972), pages 43–45], except that the two sets of variables $(D_1, \ldots, D_N)$ and $(C_1, \ldots, C_N)$ are linked by the constraints $T_i^{\alpha}(D_i - D_{i+1}) = T_i^{\beta}(C_i - C_{i-1})$ for $i = 1, \ldots, N$. However, we have not been able to find a way to exploit exact algorithms for fitting isotonic regressions in this situation.

Next, we maximize the likelihood over α and β. From the preceding propositions, the function $l(\alpha, \beta) = \min_{\nu}(-\log L(\alpha, \beta, \nu))$ can be evaluated, though no explicit formula for the minimum is available. However, l is merely a function of two variables, and the IMSL derivative-free subroutine ZXMIN [IMSL (1982)] can be used to search for its minimum. This is the procedure we

used to maximize the likelihood, with the initial value for the iteration taken to be the MLEs of α and β, assuming independent Weibull survival and censoring distributions. ZXMIN minimizes the function $l(\alpha, \beta)$. Evaluation of the function $l(\alpha, \beta)$ is itself a minimization routine performed by NPSOL. This is fairly slow, but can undoubtedly be accelerated by more efficient minimization.

One outstanding point is whether $l(\alpha, \beta)$ has a unique minimum or whether the estimated parameters obtained depend on the initial values used in the iterative procedures. Lacking an explicit expression for $l(\alpha, \beta)$, we cannot prove that it is convex. However, we can offer some empirical and heuristic evidence for its convexity. First, on the empirical side, for every data set we have tried, the minimization procedure started from a variety of initial positions never converged to anything but the value obtained by starting at the MLEs assuming independence. On the heuristic side, consider a univariate Weibull distribution with shape parameter α and scale parameter λ, and hence density $f(x; \alpha, \lambda) = \alpha\lambda x^{\alpha-1} e^{-\lambda x^{\alpha}}$. The negative log likelihood of observations $x_1, \ldots, x_N$ is not a convex function. However, for fixed α it is convex in λ, and the function $\min_{\lambda}(-\log \text{likelihood})$ is convex in α. One can hope that this last property carries over to the bivariate Weibull.

In fitting the global MLE, the potential exists that the likelihood function may diverge to infinity as α or β goes to infinity. However, as long as the data set is reasonably large, it is unlikely that one of the estimated shape parameters will diverge to infinity. Indeed, examination of the likelihood shows that for $\beta > \alpha$ the only time the likelihood for a sample will be unbounded as $\beta \to \infty$ is when there is only one censored observation and it is the largest T_i. In this case, the maximum likelihood procedure will try to model the censoring distribution by a point mass at this observation. By letting $\beta \to \infty$, the joint distribution can approach this. A similar observation holds for deaths. Thus, if there is at least one observation larger than the smallest death and at least one observation larger than the smallest censoring time, then maximization of the likelihood will not lead to divergence. This will hold for practically any real data set.

The point that the MLE of $(\lambda_d, \lambda_c, \mu)$ may not be unique can cause some concern. However, much of the potential nonuniqueness of the MLE disappears if interest is only in the marginal survival distribution $F_D(t)$ or censoring distribution $F_C(t)$. Suppose α and β have unique estimates. Then, as noted after Proposition 6.1, there may be no other $(\lambda_d, \lambda_c, \mu)$ that has likelihood equal to that of ν. Even if this is not the case, the class of MLEs of $F_D(t)$ cannot be too varying, since each $(\lambda_d, \lambda_c, \mu)$ must agree with the uniquely determined ν in all of the values of D_i and C_i that appear in the likelihood. This may not completely determine the Weibull scale $\lambda_d + \int_0^{\pi/2} \mu(d\theta)/\cos(\theta)$, but, setting $I = \max\{i: d_i > 0\}$, it does determine $\lambda_d + \int_{(0, T_i]} \mu(d\theta)/\cos(\theta)$ for $i \leq I$, and similar integrals of $1/\sin(\theta)$, so there is not much variety in the class of MLEs of the marginal scales. In fact, as long as there is no mass at the angles corresponding to the first and last events, then the MLEs of the marginal scales are unique.

7. An example. In this section we demonstrate application of the model to an actual data set. Data were collected during a two-year pilot stroke data bank study initiated in 1980 by the National Institute of Neurological and Communicative Disorders and Stroke. The goal of the pilot study was to determine whether the collection of stroke data in this manner as a resource for future research was feasible. Four clinical centers collaborated to provide data on 1158 hospitalized stroke patients. A complete description of the study can be found in Kunitz et al. (1984). For our example we use survival and censoring data for the subset of 101 patients with diagnosed intracerebral hemorrhages. Forty deaths occurred among these 101 stroke patients, most during the first three months of study, while censorings were spread across the two years of observation.

In each of Figures 2 and 3, the step function plotted is the Kaplan–Meier estimator (KME) of the marginal survival and censoring distributions, respectively, assuming independence of the death and censoring times. The dashed curve in each figure represents a Weibull marginal fit under the same assumption. The estimated scale and shape parameters were 8.27×10^{-2} and 0.357 in Figure 2 and 9.21×10^{-4} and 1.26 in Figure 3. The solid curves in Figures 2 and 3 were generated under the bivariate Weibull assumptions of our model. The pictured marginal Weibull distributions appear to be very different from those produced under the independence assumption. Indeed, the scale and shape parameters for this full Weibull model were 5.19×10^{-2} and 0.59 in Figure 2 and 3.51×10^{-2} and 0.70 in Figure 3.

The $\hat{\mu}$ parameter in the full model put positive mass on four angles, which is typical of most data sets to which we have fit the model. In this case there

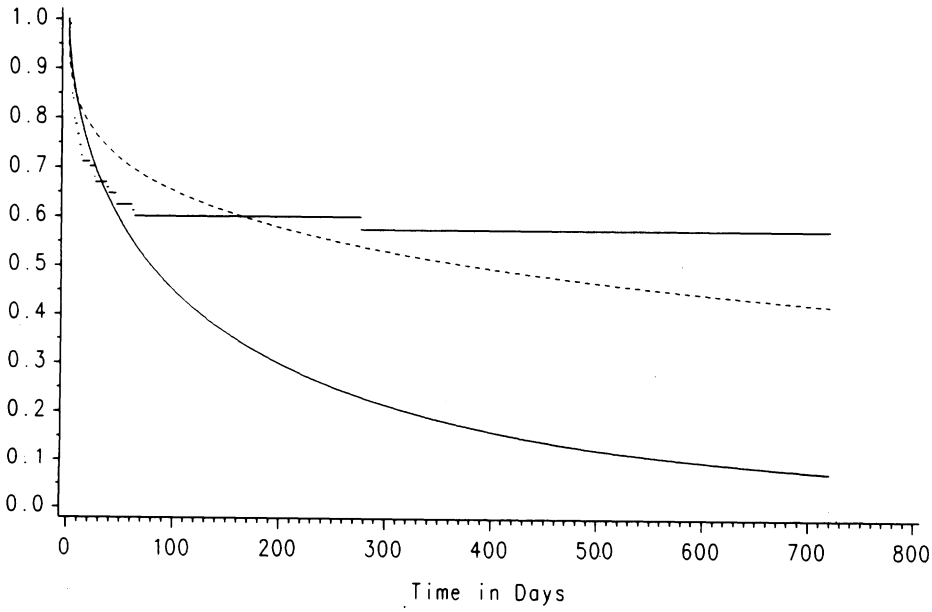

FIG. 2. *Survival function estimates for failure time distribution.*

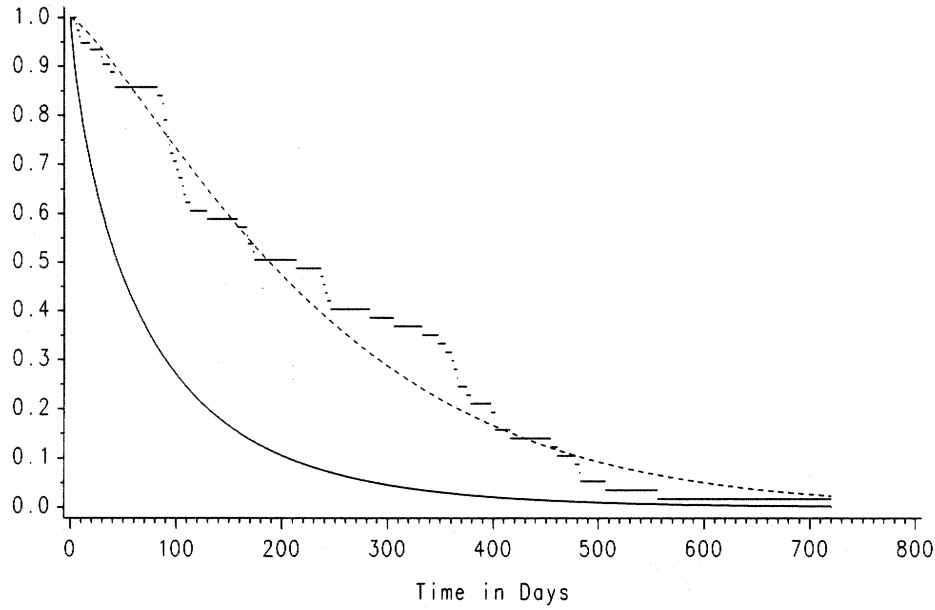

FIG. 3. *Survival function estimates for censoring distribution.*

are no masses at the first and last observations, and no masses at adjacent observations, so $\hat{\mu}$ is unique, and the estimates of marginal scales are unique as well. The distribution of the masses is as given in Table 1.

Since $\hat{\beta} > \hat{\alpha}$, $\arctan T^{\hat{\beta}-\hat{\alpha}}$ is an increasing function in T, so as time increases, the point masses of $\hat{\mu}$ will change from causing observed deaths to causing observed censoring. Thus, the model allows the hazard rate for observable deaths to decrease relative to the corresponding hazard rate for a Weibull distribution with the same shape parameter in the case of independent censoring. These discrete drops come at days 5, 29, 82 and 283. The time of the first censored observation is 5 days while the time of the last death is 283 days. The model fit puts probability 0 on observable censorings before 5 days or observable deaths after 283 days. This must be taken with a grain of salt; just as point masses of the KME are not taken literally to be atoms of the survival distribution, these point masses should not be taken as anything more

TABLE 1
Distribution of masses

Mass	Angle	Time (days)
7.55×10^{-3}	0.871	5
3.17×10^{-3}	0.962	29
1.69×10^{-2}	1.013	82
1.22×10^{-2}	1.071	283

than an estimate of a much smoother distribution. The most important things are the marginal distributions.

Our bivariate Weibull model seems reasonable for this data set partly because the Kaplan–Meier estimators plotted in both figures appear fairly Weibull in shape. The KME and dependent Weibull curves should not be compared directly, as they are fit under essentially different sets of assumptions. As mentioned, there is no way to test the appropriateness of the model; however, in the case of independence, it is worthwhile to check that the Weibull is a reasonable model. Next, the Weibull is a rich family of distributions and the dependence we allow is fairly general, even including the case of independence of the survival and censoring mechanisms. Finally, an argument can be made for positive dependence in that a patient who has remained stroke free is probably more likely to be happily involved in the study and less likely to leave than one who has suffered a stroke.

Recall from Figures 2 and 3 and previous discussion that assuming independence of the marginal Weibull distributions leads to very different results than operating under the assumptions of our bivariate Weibull model. For the case of survival, each of the marginal Weibull scale and shape parameter estimates under independence are within the same order of magnitude as that produced under the full model. However, values in the tails of the distributions, where interest often lies, can vary greatly, as seen in Figure 2. For censoring, the parameter estimates under the two models obviously differ, even by orders of magnitude. In fact, one leads to an increasing hazard rate and the other to a decreasing rate. The two curves in Figure 3 appear quite disparate, especially in the non-tail sections. For both survival and censoring, the independent Weibull marginals tend to overestimate these probabilities as compared to the dependent Weibulls of our model. Thus, by using our model in this study, one may have a more pessimistic, but possibly more realistic, view of survival after intracerebral hemorrhage.

How are these estimates to be interpreted? Of course, no model for dependent censoring can work magic and tell us whether censoring is independent of failure or not. What they can do is give an indication, under minimal assumptions, of the possible consequences of dependent censoring. If a model for dependent censoring gives an estimated survival distribution that is significantly different from the distribution estimated under independence, then it may give an investigator cause to think about assumptions more carefully.

Acknowledgments. We would like to thank Jerzy Filar for recommending the package NPSOL and Jim Dambrosia for the intracerebral hemorrhage data set in our example.

REFERENCES

BAHADUR, R. R. (1967). Rates of convergence of estimates and some test statistics. *Ann. Math. Statist.* **38** 303–324.

BARLOW, R. E., BARTHOLOMEW, D. S., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

COX, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *J. Roy. Statist. Soc. Ser. B* **21** 411–421.

DE HAAN, L. and PICKANDS, J., III (1986). Stationary min-stable stochastic processes. *Probab. Theory Related Fields* **72** 477–492.

FISHER, L. and KANAREK, P. (1974). Presenting censored data when censoring and survival times may not be independent. In *Reliability and Biometry* (F. Proschan and R. Serfling, eds.) 303–326. SIAM, Philadelphia.

GALAMBOS, J. (1978). *The Asymptotic Theory of Extreme Order Statistics* 258–265. Wiley, New York.

GILL, P. E., MURRAY, W., SAUNDERS, M. A. and WRIGHT, M. H. (1986). Users guide for NPSOL. Systems Optimization Laboratory, Technical Report SOL 86-2, Stanford Univ.

GRENANDER, U. (1981). *Abstract Inference* 349–352. Wiley, New York.

HUBER, P. J. (1981). *Robust Statistics* **24**. Wiley, New York.

IMSL (1982). *IMSL Library Reference Manual*. IMSL Inc., Houston.

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

KLEIN, J. P. and MOESCHBERGER, M. L. (1984). Bounds on net survival probabilities for dependent competing risks. Technical Report 291, Dept. Statist., Ohio State Univ.

KUNITZ, S. C., GROSS, C. R., HEYMAN, A., KASE, C. S., MOHR, J. P., PRICE, T. R. and WOLF, P. A. (1984). The pilot stroke data bank: Definition, design, and data. *Stroke* **15** 740–746.

MOESCHBERGER, M. L. (1974). Life tests under dependent competing causes of failure. *Technometrics* **16** 39–47.

PETERSON, A. V. (1976). Bounds for a joint distribution function with fixed subdistribution functions: Application to competing risks. *Proc. Nat. Acad. Sci. U.S.A.* **73** 11–13.

PICKANDS, J., III (1976). A class of multivariate negative exponential distributions. Preprint. Dept. Statist., Univ. Pennsylvania, Philadelphia.

PURI, P. S. (1979). On certain problems involving non-identifiability of distributions arising in stochastic models. In *Optimization Methods in Statistics* (J. S. Rustagi, ed.) 403–417. Academic, New York.

ROBERTSON, J. B. and UPPULURI, V. R. R. (1984). A generalized Kaplan–Meier estimator. *Ann. Statist.* **12** 366–371.

SLUD, E. V. and RUBENSTEIN, L. V. (1983). Dependent competing risks and summary survival curves. *Biometrika* **70** 643–649.

TAWN, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika* **75** 397–415.

WILLIAMS, J. S. and LAGAKOS, S. W. (1977). Models for censored survival analysis: Constant-sum and variable-sum models. *Biometrika* **64** 215–224.

NATIONAL INSTITUTE OF NEUROLOGICAL
DISORDERS AND STROKE
7550 WISCONSIN AVENUE, ROOM 7C02
BETHESDA, MARYLAND 20892

DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF MARYLAND BALTIMORE COUNTY
BALTIMORE, MARYLAND 21228