SPLINE SMOOTH ESTIMATES OF SURVIVAL

Jerome Klotz

Ohio State University

1. Introduction

Let X be survival time with continuous distribution $F_X(x)$ and density $f_F(x)$. Similarly, let Y be time to censoring, independent of X, with continuous distribution $F_Y(y)$ and density $f_Y(y)$. We observe time on trial, T , and death or censoring indicator, D , where

$$T = \min(X,Y)$$
$$D = \begin{cases} 1 & \text{if } X \le Y \quad \text{(death)} \\ 0 & \text{if } X > Y \quad \text{(censoring)} \end{cases} .$$

Using a sample $\{(T_i,D_i);i=1,2,\ldots,n\}$ we wish to find a smooth estimate of the survival distribtion $1-F_X(x) = P[X>x]$.

Define the hazard function by

$$h_X(x) = f_X(x)/(1-F_X(x))$$

and the integrated hazard function by

$$H_X(x) = \int_0^x h_X(u)du = -\int_0^x d\,\ln(1-F_X(u)) \quad ,$$

which is related to survival by $1-F_X(x) = e^{-H_X(x)}$. Defining the indicator function I[A] (1 or 0 according as the event A holds or not), the sample cumulative distribution is

$$F_n(t) = \sum_{i=1}^{n} I[T_i \le t] / n .$$

We are concerned with a smooth approximation of the hazard function over subintervals using polynomials. We write the polynomials as linear combinations of B-splines defined on the knots or points defining the subintervals. The B-spline or order r or polynomial of degree r-1 is defined for the non-decreasing sequence of knots

(1) $$\tau_{-r+1},\ \tau_{-r+2},\ldots,\tau_0,\ \tau_1,\ldots,\tau_K,\ \tau_{K+1},\ldots,\tau_{K+r-1} \quad ,$$

using the following divided differences:

$$g_r(\tau_j;t) = (\tau_j - t)_+^{r-1} = [\max(0,\tau_j - t)]^{r-1}$$

$$g_r(\tau_j,\ \tau_{j+1};t) = [g_r(\tau_{j+1};t) - g_r(\tau_j;t)] / (\tau_{j+1} - \tau_j)$$

$$\vdots$$

$$g_r(\tau_j,\ \tau_{j+1},\ldots,\tau_{j+r},t) = \left[\frac{g_r(\tau_{j+1},\ldots,\tau_{j+r};t) - g_r(\tau_j,\ldots,\tau_{j+r-1};t)}{(\tau_{j+r} - \tau_j)}\right] .$$

Then the normalized B-spline is

$$N_{jr}(t) = (\tau_{j+r} - \tau_j)\ g_r(\tau_j,\ \tau_{j+1},\ldots,\tau_{j+r};t) \quad .$$

In case some knots coincide, continuity can be used for the definition. For a discussion of B-splines see de Boor (1976). Figure 1 gives graphs for $r = 2,3$ corresponding to linear and quadratic B-splines.

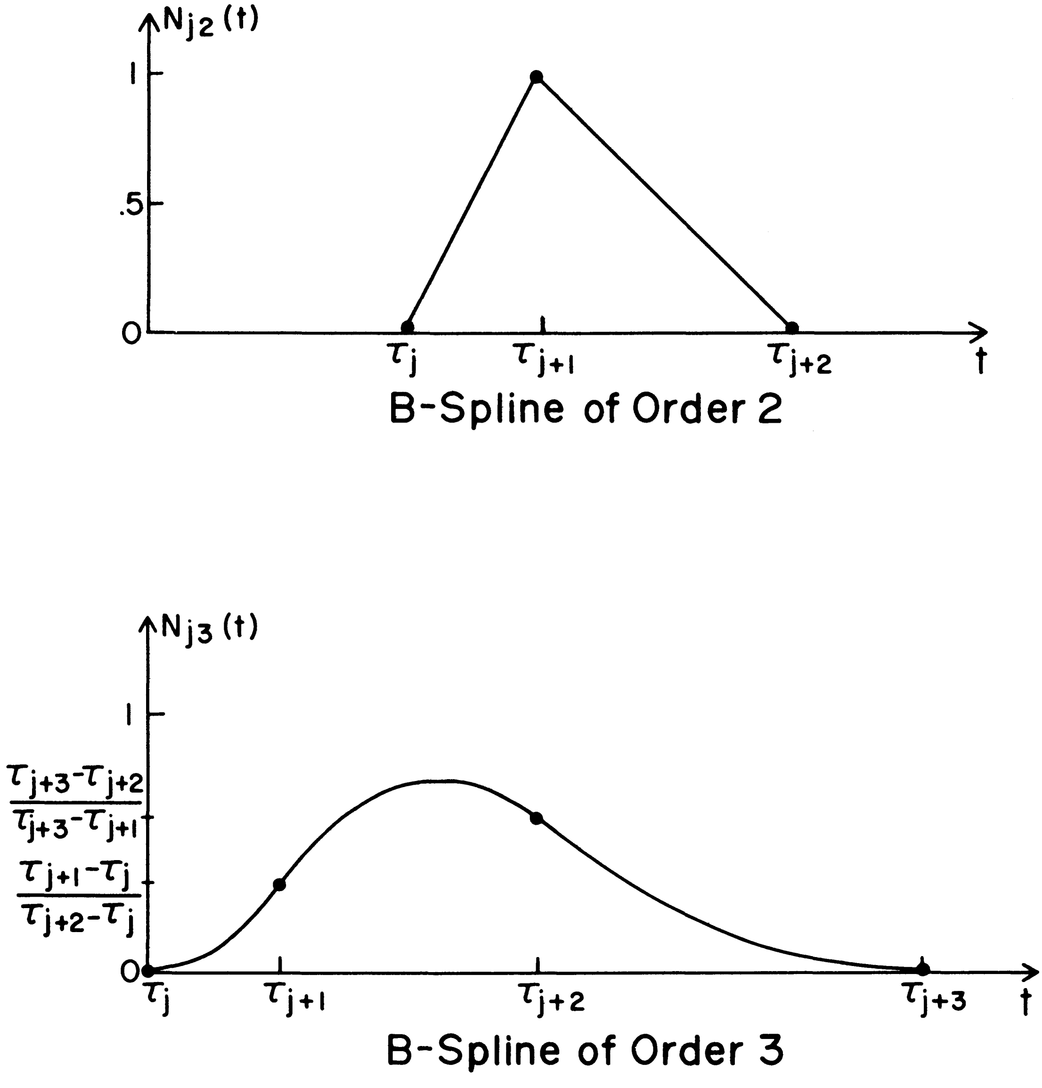

FIGURE 1. Linear and quadratic B-splines for knots $\{t_j\}$.

2. Hazard Approximation by Splines with Fixed Knots

We fit the model

$$h_X(x) = \sum_{j=-r+1}^{K-1} \theta_j N_{j,r}(x) \tag{2}$$

over the interval $0 \le x \le \tau_K$ by selecting knots

$$\tau_{-r+1} = \tau_{-r+2} = \cdots = \tau_0 = 0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_K = \tau_{K+1} = \cdots = \tau_{K+r-1} .$$

Although the model is parametric with parameters $\underset{\sim}{\theta} = (\theta_{-r+1}, \theta_{-r+2}, \ldots, \theta_{K-1})$, there is great flexibility through the choice of knots $\{\tau_k\}$ and spline order r.

We consider estimating $\underset{\sim}{\theta}$ by maximizing the likelihood. The joint continuous-discrete density under the random censorship model is

$$f_{T,D}(t,d) = [f_X(t)(1-F_y(t))]^d \, [f_Y(t)(1-F_X(t))]^{1-d}$$

$$= (h_X(t))^d \, (h_Y(t))^{1-d} \, (1-F_T(t)) \quad ,$$

where $1-F_T(t) = (1-F_X(t))\,(1-F_Y(t))$ by independence. The log-likelihood is then

$$\sum_{i=1}^{n} \ln f_{T,D}(t_i,d_i) = \sum_{i=1}^{n} [d_i \ln h_X(t_i) + \ln(1-F_X(t_i))]$$

$$+ \sum_{i=1}^{n} [(1-d_i) \ln h_Y(t_i) + \ln(1-F_Y(t_i))] \quad . \tag{3}$$

Differentiating (3) with respect to $\underset{\sim}{\theta}$ using (2) gives

$$\frac{\partial}{\partial \underset{\sim}{\theta}} \sum_{i=1}^{n} \ln f_{T,D}(t_i,d_i) = \sum_{i=1}^{n} [d_i \, \underset{\sim}{N}_r(t_i) \,/\, (\underset{\sim}{N}_r(t_i)\, \underset{\sim}{\theta}') - \int_0^{t_i} \underset{\sim}{N}_r(u)du] \quad ,$$

where $\underset{\sim}{N}_r(x) = (N_{-r+1,r}(x), N_{-r+2,r}(x), \ldots, N_{K-1,r}(x))$ and $\underset{\sim}{\theta}'$ is the transpose of $\underset{\sim}{\theta}$.

If the solution of the derivative equation, $\partial \sum_i \ell n\, f_{T,D}(t_i, d_i)/\partial \underset{\sim}{\theta} = \underset{\sim}{0}$, gives a nonnegative function $\hat{h}_X(x) = \sum_{j=-r+1}^{K-1} \hat{\theta}_j N_{j,r}(x)$ then we propose the estimator $1 - \hat{F}_X(x) = \exp(-\hat{H}_X(x))$, where $\hat{H}_X(x) = \int_0^x \hat{h}_X(u)du$.

Because of the necessity of choosing the degree r as well as suitable knots $\{\tau_k\}$ and then solving a messy non-linear derivative equation which we can only hope has a non-negative solution $\tilde{h}_X$, we turn instead to a simplification.

3. An Ad Hoc Estimator

The model $h_X(x) = \underset{\sim}{N}_r(x)\ \underset{\sim}{\theta}'$ breaks down when the knots defining $\underset{\sim}{N}_r$ are random variables. Nevertheless, motivated by the success of the estimator of Breslow (1974) that uses constant splines over random death times, we propose a similar simplication using linear splines $(r = 2)$. Specifically, we replace the knots $0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_K$ in (1) by distinct death times $0 < T_{<1>} < T_{<2>} < \cdots < T_{<K>}$ which are different sorted values of T_i for which $D_i = 1$, $i = 1, 2, \ldots, n$. Using $N_{j,2}(T_{<j+1>}) = 1$, and 0 at other knots gives the minimizing solution $\tilde{\theta}_{-1} = 0$ and

$$\tilde{\theta}_j = m_{j+1} / \sum_{i=1}^{n} \int_0^{t_i} N_{j,2}(u)du \ , \quad \text{for } j = 0, 1, 2, \ldots, K-1 \ ,$$

where m_k is the number of death times equal $T_{<k>}$. Then the estimate of the integrated hazard is

$$\tilde{H}_X(x) = \sum_{k=1}^{K} \{m_k \int_0^x N_{k-1,2}(u)du \Big/ \sum_{i=1}^{n} \int_0^{t_i} N_{k-1,2}(t)dt\} \ .$$

From the identity

$$\int_0^x N_{k-1,2}(u)du = \left(\sum_{j \ge k-1} N_{j,3}(x)) \ (T_{<k+1>} - T_{<k>} \right) / \ 2 \ ,$$

we can cancel the nonzero factor $(T_{<k+1>} - T_{<k>})/2$ from both numerator and denominator to obtain

$$(4) \qquad \tilde{H}_X(x) = \sum_{k=1}^{K} \{m_k \sum_{j \geq k-1} N_{j,3}(x) \Big/ \sum_{i=1}^{n} \sum_{r \geq k-1} N_{r,3}(t_i)\} \quad .$$

For computing, with knots $\tau_{k-1} < \tau_k < \tau_{k+1}$, we have

$$\sum_{j \geq k-1} N_{j,3}(x) = \begin{cases} 0, \text{ for } x \leq \tau_{k-1} \\ (x - \tau_{k-1})^2/((\tau_k - \tau_{k-1})(\tau_{k+1} - \tau_{k-1})), \text{ for } \tau_{k-1} \leq x < \tau_k \\ 1 - (\tau_{k+1} - x)^2/((\tau_{k+1} - \tau_k)(\tau_{k+1} - \tau_{k-1})), \text{for } \tau_k \leq x \leq \tau_{k+1} \\ 1, \text{ for } x \geq \tau_{k+1} \quad . \end{cases}$$

The estimator (4) is a non-negative differentiable monotone increasing function of x on the interval $[0, T_{<k>}]$ and thus

$$1 - \tilde{F}_X(x) = e^{-\tilde{H}_X(x)}$$

is a differentiable monotone decreasing function on this interval. Figure 2 gives an example of the estimator contrasted with the Kaplan-Meier estimator (1958).

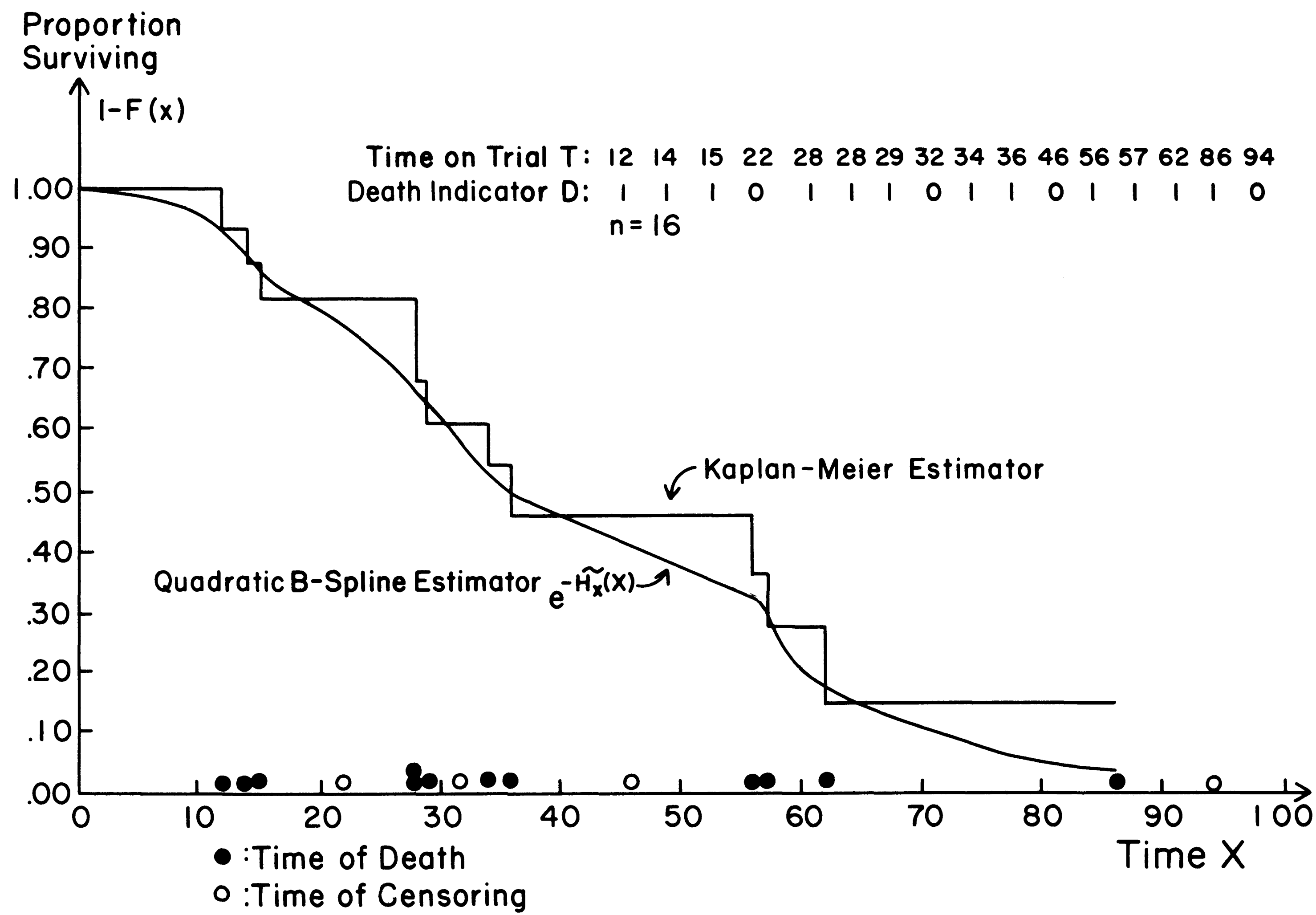

FIGURE 2. Comparison of the Kaplan-Meier estimator and the quadratic B-spline estimator.

4. Consistency of $\tilde{H}_X(x)$

The following theorem gives consistency of $\tilde{H}_X(x)$ and consequently $1-\tilde{F}_X(x)$ under some assumptions.

THEOREM 1:

If $f_X(t)>0$ a.e. on the interval of t values for which $F_T(t)<1$, $\tilde{H}_X(x) \xrightarrow{P} H_X(x)$ as $n\to\infty$ for x in the interior of the interval.

PROOF:

From equation (5) we obtain the inequalities

$$\text{(6)} \qquad I[\tau_{k+1} \le x] \le \sum_{j\ge k-1} N_{j,3}(x) \le I[\tau_{k-1} < x] \quad .$$

By the continuity of F_X and F_Y, F_T is continuous and the ordered times on trial $0<T_{(1)}<T_{(2)}<\cdots<T_{(n)}$ are distinct with probability 1. Consequently the ordered death times $0<T_{[1]}<T_{[2]}<\cdots<T_{[M]}$ where $M=\sum_{i=1}^{n} D_i$ are distinct with probability one and $K=M$. Thus, $T_{<k>}=T_{[k]}$ and we have

$$L_n(x) \le \tilde{H}_X(x) \le U_n(x) \quad ,$$

where the upper bound

$$\text{(7)} \qquad U_n(x) = \sum_{k=1}^{M} I[T_{[k-1]} < x] \,/(n(1-F_n(T^-_{[k+1]})))$$

is obtained by replacing the numerator and denominator of (4) by upper and lower bounds in (6) with knots $\{T_{[k]}\}$. Similarly

$$\text{(8)} \qquad L_n(x) = \sum_{k=1}^{M} I[T_{[k+1]} \le x] \,/\, (n(1-F_n(T_{[k-1]})))$$

is obtained from lower and upper bounds in the numerator and denominator of (4). Here we use the conventions $T_{[0]} = 0$ and $T_{[M+1]} = T_{[M]}$ in (7) and (8) so that (6) holds for $k = 1$ and $k = M$. Intuitively, the bounds $U_n(x)$ and $L_n(x)$ will be close to the empirical integrated hazard function

$$H_n(x) = \sum_{k=1}^{M} I[T_{[k]} < x] \,/\, (n(1 - F_n(T_{[k]}^-)))$$

$$= \sum_{i=1}^{n} \{D_i \; I[T_i < x] \,/\, (n(1 - F_n(T_i -)))\}$$

shown by Breslow and Crowley (1974) to converge weakly to

$$H_x(x) = E\{D \; I[T < x] \,/\, (1 - F_T(t))\}$$

using methods of Billingsly (1968). Consistency will follow by showing $U_n(x) - H_n(x) \overset{P}{\to} 0$ and $H_n(x) - L_n(x) \overset{P}{\to} 0$. We show convergence for the upper bound; the argument for the lower bound is similar. Write

$$U_n(x) - H_n(x) = \frac{1}{n} \sum_{k=1}^{M} \left\{ \frac{I[T_{[k-1]} < x]}{1 - F_n(T_{[k+1]}^-)} - \frac{I[T_{[k]} < x]}{1 - F_n(T_{[k]}^-)} \right\}$$

$$= \sum_{k=1}^{M} \frac{I[T_{[k-1]} < x \leq T_{[k]}]}{n(1 - F_n(T_{[k-1]}^-) - w_{nk} - w_{nk+1})} + \tag{9}$$

$$\sum_{k=1}^{M} \frac{I[T_{[k]} < x] \, w_{nk+1}}{n(1 - F_n(T_{[k]}^-))(1 - F_n(T_{[k]}^-) - w_{nk+1})} \; ,$$

where $w_{nk} = F_n(T_{[k]}^-) - F_n(T_{[k-1]}^-)$. The expression (9) is in turn bounded by

$$[n(1 - F_n(x) - 2w_n^*)]^{-1} + H_n(x)\, w_n^* \,/\, (1 - F_n(x) - w_n^*) \; ,$$

where $w_n^* = \max\limits_{1 \le k \le M} w_{nk}$. Since $\bar{F}_n(x) \xrightarrow{P} F_T(x) < 1$ and $H_n(x) \xrightarrow{P} H_X(x) < \infty$ we complete the proof by showing $w_n^* \xrightarrow{P} 0$. We bound w_{nk} by

$$
\begin{aligned}
w_{nk} &= F_n(T_{[k]}) - F_n(T_{[k-1]}) \\
&= F_T(T_{[k]}) - F_T(T_{[k-1]}) + \\
&\qquad (F_n(T_{[k]}) - F_T(T_{[k]})) - (F_n(T_{[k-1]}) - F_T(T_{[k-1]})) \\
&\le F_T(T_{[k]}) - F_T(T_{[k-1]}) + 2\sup(F_n(t) - F_T(t)) \quad .
\end{aligned}
$$

Using the Glevenko Cantelli Theorem, $\sup(F_n(t) - F_T(t)) \xrightarrow{P} 0$, and so we show $\max\limits_{1 \le k \le M} F_T(T_{[k]}) - F_T(T_{[k-1]}) \xrightarrow{P} 0$. Now for $\varepsilon,\ \delta > 0$,

$$
\begin{aligned}
& P[\max_{1 \le k \le M} (F_T(T_{[k]}) - F_T(T_{[k-1]})) > \varepsilon] \\
\le & \sum_{n(p-\delta) \le m \le n(p+\delta)} P[\max_{1 \le k \le M} (F_T(T_{[k]}) - F_T(T_{[k-1]})) > \varepsilon \mid M = m]\ P[M=m] \\
& \qquad + P[|M - np| > n\delta] \qquad (10) \\
\le & \max_{n(p-\delta) \le m \le n(p+\delta)} P[\max_{1 \le k \le m} (F_T(T^*_{(k)}) - F_T(T^*_{(k-1)})) > \varepsilon] \\
& \qquad + P[|M - np| > n\delta] \quad ,
\end{aligned}
$$

where M has a binomial (n,p) distribution, $p = P[X \le Y]$, and $T^*_{(1)}, \ldots, T^*_{(m)}$ are order statistics for an independent sample of size m from the distribution

$$F_*(t) = F_{T|D}(t|1)$$

with density

$$f_*(t) = f_{T|D}(t|1) = f_X(t)\,(1 - F_Y(t))/p \quad . \tag{11}$$

The 2nd term in (10) goes to zero as $n \to \infty$ and so it remains to show $\max_{1 \le k \le m} (F_T(T^*_{(k)}) - F_T(T^*_{(k-1)})) \overset{P}{\to} 0$, as $m \to \infty$.

By the assumptions, we see from (11) that $F_*(t)$ is continuous in addition to $F_T(t)$ and we can write

$$\max_{1 \le k \le m} [F_T(T^*_{(k)}) - F_T(T^*_{(k-1)})]$$

$$= \max_{1 \le k \le m} [F_T(F_*^{-1}(F_*(T^*_{(k)}))) - F_T(F_*^{-1}(F_*(T^*_{(k-1)})))]$$

$$= \max_{1 \le k \le m} [F_T(F_*^{-1}(U_{(k)})) - F_T\, F_*^{-1}\,(U_{(k-1)})] \; ,$$

where $U_{(k)}$ are order statistics from a uniform (0,1) distribution. Thus, by continuity it remains to show

$$\max_{1 \le k \le m} (U_{(k)} - U_{(k-1)}) \overset{P}{\to} 0 \quad \text{as } m \to \infty.$$

Finally, by properties of uniform order statistics,

$$P[\max_{1 \le k \le m} (U_{(k)} - U_{(k-1)}) > \varepsilon] \le \sum_{k=1}^{m} P[U_{(k)} - U_{(k-1)} > \varepsilon]$$

$$= m\, P[U_{(1)} > \varepsilon] = m(1 - \varepsilon)^m \to 0 \quad \text{as } m \to \infty \quad .$$

It is likely that these strong assumptions can be weakened for proving consistency. However, some control on the spacings of adjacent death times may be required around the point x as $\tilde{H}_X(x)$ is a function of the order statistics and cannot be written as a counting process.

ACKNOWLEDGEMENT

Research was supported in part by the National Institutes of Health Grant No. 2-R01-CA-18332-07.

REFERENCES

Billingsly, P. (1968). Weak Convergence of Probability Measures. J. Wiley, New York.

Breslow, N. (1974). Covariance analysis of censored survival data. Biometrics **30**, 89-99.

Breslow, N. and Crowley, J. (1974). A large sample study of the life tables and product limit estimates under random censorship. Annals of Statistics **3**, 437-453.

de Boor, C. (1976). Splines as linear combinations of B-splines. A survey. University of Wisconsin - Madison Mathematics Research Center Technical Report 1667.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association **53**, 457-481.