6-2008

# Parametric Model Discrimination for Heavily Censored Survival Data

Lawrence Leemis
*William & Mary*, lmleem@wm.edu

A. D. Block



## Recommended Citation

Leemis, Lawrence and Block, A. D., Parametric Model Discrimination for Heavily Censored Survival Data (2008). *IEEE Transactions on Reliability*, 57(2), 248-259.
https://doi.org/10.1109/TR.2008.923488

# Parametric Model Discrimination for Heavily Censored Survival Data

A. Daniel Block and Lawrence M. Leemis

***Abstract*—Simultaneous discrimination among various parametric lifetime models is an important step in the parametric analysis of survival data. We consider a plot of the skewness versus the coefficient of variation for the purpose of discriminating among parametric survival models. We extend the method of Cox & Oakes from complete to censored data by developing an algorithm based on a competing risks model and kernel function estimation. A by-product of this algorithm is a nonparametric survival function estimate.**

***Index Terms*—Competing risks, distribution selection, kernel functions.**

### ACRONYM[1]

| | |
|---|---|
| APPL | a probability programming language |
| CDF | cumulative distribution function |
| HF | hazard function |
| MP | mercaptopurine |
| PDF | probability density function |
| PP | probability-probability |
| QQ | quantile-quantile |
| SF | survivor function |

### NOTATION

| | |
|---|---|
| $A$ | $\min\{S, R/1.34\}$ |
| $c_i$ | censoring time $i$ |
| $\delta_i$ | censoring indicator variable $i$ |
| $\hat{f}(y)$ | kernel density estimate |
| $f_{X_j}(t)$ | PDF associated with crude lifetime $j$ |
| $\gamma$ | coefficient of variation |
| $\hat{\gamma}$ | estimated coefficient of variation |
| $\gamma_3$ | skewness |
| $\hat{\gamma}_3$ | estimated skewness |
| $h$ | smoothing parameter |
| $h_{Y_j}(t)$ | HF associated with net lifetime $j$ |
| $\kappa$ | shape parameter for the Weibull distribution |
| $\hat{\lambda}_X$ | crude lifetime failure rates |
| $\mu$ | population mean |
| $n$ | sample size |
| $\pi_j$ | probability of failure from risk $j$ |
| $\hat{\pi}_j$ | estimated probability of failure from risk $j$ |
| $R$ | interquartile range |
| $\sigma$ | population standard deviation |
| $S$ | sample standard deviation |
| $*$ | denotes right-censored observation |
| $S_{X_j}(t)$ | SF associated with crude lifetime $j$ |
| $T$ | lifetime |
| $t_i$ | failure time $i$ |
| $w$ | kernel function |
| $x_i$ | $\min\{c_i, t_i\}$ |
| $X_j$ | crude lifetime associated with risk $j$ |
| $y_1, y_2, \ldots, y_n$ | data values |
| $Y_j$ | net lifetime associated with risk $j$ |

Manuscript received January 11, 2007; revised July 30, 2007 and October 17, 2007; accepted October 19, 2007. Associate Editor: M. Xie.

D. Block is a consultant with the Department of Mathematics, The College of William & Mary, Williamsburg, VA 23187 USA (e-mail: danielblock@cox.net).

L. Leemis is with the Department of Mathematics, The College of William & Mary, Williamsburg, VA 23187 USA (e-mail: leemis@math.wm.edu).

Digital Object Identifier 10.1109/TR.2008.923488

[1]The singular and plural of an acronym are always spelled the same.

## I. INTRODUCTION

A WELL-KNOWN technique for choosing a probability model to approximate survival data is to plot the standardized third moment (skewness) versus the coefficient of variation [6]. Their plot gives the trajectory of the population skewness versus population coefficient of variation for several popular parametric lifetime models (e.g., Weibull and log-logistic). The sample skewness, and the sample coefficient of variation can be plotted for a data set containing no censored data values. The proximity of the sample point to the population distribution values can be helpful in ruling in or ruling out particular parametric distributions as potential survival models.

This technique is superior to many of the standard exploratory data techniques for survival data (such as QQ plots, PP plots, probability plots, histograms, density function estimation via kernel functions, and empirical cumulative distribution function estimation via the Kaplan-Meier product-limit estimate) in the sense that several competing models are easily viewed simultaneously. The weaknesses of this technique are (a) it considers only second and third moments which, for example, might do a poor job of detecting appropriate tail behavior; and (b) it is not easily adaptable to right-censored data sets. The focus of this paper is overcoming the second weakness. We limit our discussion to survival models.

Cox & Oakes' [6] technique is extended here to allow the calculation of these statistics in the presence of heavy right censoring. Censored data is common in survival analysis. Recent work in the area includes [12], [13], [15], [17], [18], [21], [23], [25], [26]. While data sets with light censoring (as in Cox & Oakes' example with the cancer data from [2]) may be handled in a variety of ways (e.g., doubling censoring times or treating them as failures) without introducing significant error, heavily censored data leaves a data analyst with only heuristic methods that will significantly influence the position of the sample point. In this paper, we will present an analytical approach for plotting a point on the Cox & Oakes graph regardless of the fraction of censored observations in the data set.

Our approach assumes a random censoring scheme, where the censoring times and the failure times are independent random variables. Treating censoring and failing as risks allows us to analyze the data set using a competing risks framework. We then treat the censored observations as coming from one distribution, and the failures as coming from a second distribution. We use kernel estimation to create empirical density functions for these two data sets. Using a mathematically tractable distribution for the density function, the competing risks model is used to estimate the failure time distribution as if censoring were not present. Because the failed & censored data are observed in the presence of another risk, these random variables represent crude lifetimes. A crude lifetime [8] is the lifetime of an observation in the presence of other risks (also called a cause-specific distribution or cause-specific lifetime). Competing risks models are surveyed more recently in [7] and [19]. The goal of this paper is to find an estimate of the net lifetimes of the failure time distribution (i.e., when no censoring is present). The net lifetime is the lifetime of an observation only at risk of failing by one cause. We will use the crude lifetimes and statistical methods to "eliminate" the presence of the censoring risk. The result will be the distribution of the net lifetime of the time to failure, which can then be treated as any other distribution of observed failures. The sample skewness and sample coefficient of variation associated with this distribution will then be plotted on the graph. The proximity of this point to curves associated with well-known parametric survival models (such as the Weibull distribution) can be used to provide a list of potential models for data fitting purposes.

## II. Literature Review

Because several methodologies are used in the algorithm that plots the sample skewness versus the sample coefficient of variation, we discuss their literature in the following subsections: Cox & Oakes' Methodology, Kernel Functions, Competing Risks.

*Cox & Oakes' Methodology:* Section 2.4 of [6] outlines four methods for plotting or tabulating data to select a parametric survival distribution. One of these methods (illustrated on page 27) is a plot of the standardized third moment (skewness) $\gamma_3 = E[((T-\mu)/\sigma)^3]$ versus the coefficient of variation $\gamma = \sigma/\mu$, where $\mu$ and $\sigma$ are respectively the mean and standard deviation of the lifetime $T$. Cox & Oakes plot $\gamma_3$ versus $\gamma$ for several popular distributions. Their graph is replicated in Fig. 1 using APPL (A Probability Programming Language) code given in Appendix A. The code has been modified & augmented from [9]. APPL is described in [11].

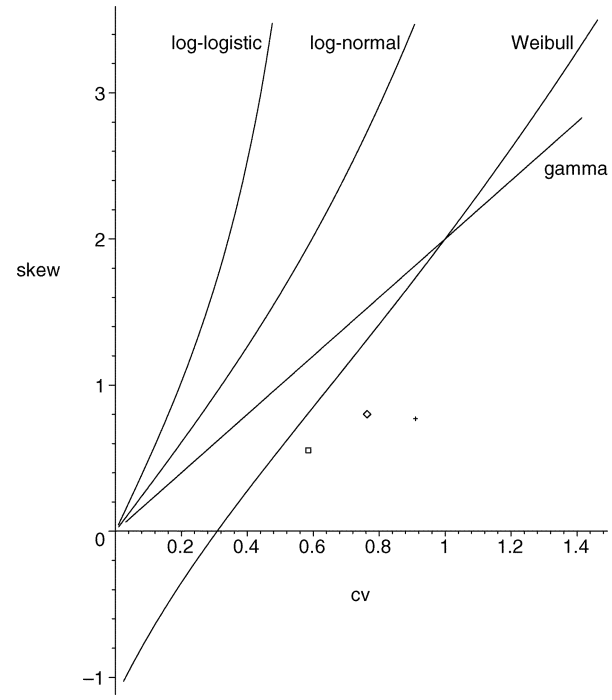


Fig. 1. Cox & Oakes graph with three heuristic methods employed for the 6-MP treatment group remission times.

The exponential distribution occurs at the intersection of the Weibull and gamma curves at $\gamma = 1$, and $\gamma_3 = 2$. As stated earlier, Cox & Oakes "reasonably extrapolate" the light censoring in their example data set. Had the censoring been heavier, however, their analysis would have been unable to attain objective results. The square, diamond, and cross in Fig. 1 are from the heavily-censored 6-MP treatment group from [10]:

$$6, 6, 6, 6^*, 7, 9^*, 10, 10^*, 11^*, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^*, 32^*, 32^*, 34^*, 35^*.$$

In this data set, there are $n = 21$ individuals at risk, 12 of which are right-censored as indicated by an asterisk. The points on the graph, starting from the left, treat the censored observations heuristically as follows.

1) Censored observations are treated as failures (square).
2) Censored observations are doubled, then treated as failures (diamond).
3) Censored observations are quadrupled, then treated as failures (cross).

The considerable scatter in these three points gives sufficient impetus to search for a parameter-free technique for plotting skewness versus the coefficient of variation for a given data set. A goal of this paper is to devise an algorithm based on nonparametric methods that will handle randomly right-censored data, and not require parameters from an analyst.

*Kernel Functions/Density Estimation:* Bowman & Azzalini [4] give an excellent introductory framework for density estimation. They define a kernel density estimate as

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} w(y - y_i; h),$$

where $w$, known as a kernel function, is a probability density function whose variance is controlled by the smoothing parameter $h$; and $y_1, y_2, \ldots, y_n$ are the data values. The primary issue then becomes the choice of the kernel function. For their example kernel function, Bowman & Azzalini use the Gaussian distribution at $y - y_i$ with mean 0, and standard deviation $h$.

Section 3.4 of [22] gives comprehensive coverage on the topic of choosing the smoothing parameter $h$. Because density estimation is not a central topic of this paper, we will choose an $h$ for use in the remainder of this work. Silverman gives two "quick ways of choosing" the smoothing parameter based on the properties of the sample data. The first is $h_1 = 0.79Rn^{-1/5}$, where $R$ is the sample interquartile range; and the second is $h_2 = 0.9An^{-1/5}$, where $A = \min\{S, R/1.34\}$, and $S$ is the sample standard deviation. Based on the discussion in Silverman, we will choose $h_2$, although this choice will have little influence on our results. Biswas & Sundaram [1] apply kernel functions to survival analysis.

*Competing Risks:* Treatment of censoring as a risk for use in a competing risks model is mentioned in [8], [14], [20], and [24]. The articles of primary interest involve risk elimination, and net/crude relationships.

Williams & Lagakos [24] address a common assumption in many censoring models which states that the survival mechanism and the censoring mechanism are independent. They give examples when this is an invalid assumption, and then examine the consequences of survival times and censoring times being dependent. They also discuss the testability of a data set for censoring influences on the survival mechanism. Where applicable, they give statistical tests that may be performed. Framed in the context of this paper, they discuss situations where the crude lives of the observed data failing from different causes are dependent. Here, however, we will assume that the remaining longevity of an observation has no influence on its "surviving" censoring.

Prentice & Kalbfleisch [20] address three main issues: influence of regression coefficients on some, but not all failure causes; the interrelation between failure types; and the estimation of failure rates for one cause given the removal of some or all of the other causes. Starting from the assumption that the cause-specific hazard function is the only truly estimable portion of the competing risks model, they build models to address these three issues. While the first two issues are not relevant to this paper, the third is. They raise a list of concerns regarding cause-removal methods, and write about the most important concern: "In effect, the stochastic mechanism generating failures is assumed to continue beyond latent failure times for causes that have been removed until the smallest operative failure time is reached." The authors do not present a quantitative model. Their concerns were raised in response to [5], and do not affect our technique. Finally, the authors deal with censoring as a cause of failure. They abandon their previous concerns of cause removal stating "the marginal distribution that arises from the elimination of censoring is clearly the relevant target of estimation." This is, of course, the very goal of this paper. They do, however, cite Williams & Lagakos [24], and raise the issue of the independence of survival and censoring; we will avoid this issue.

Kalbfleisch & Mackay [14] extend the work of [24], and show that the constant-sum condition is equivalent to a simple relationship between hazard functions.

David & Moeschberger [8] formulate a mathematical model for competing risks in their monograph. They mention the notion of treating censored data as a risk in their first chapter. Their main work, as it pertains to this paper, is in the area of net versus crude lifetimes.

Leemis [16] also discusses the relationships between net and crude lifetimes. Section 5.1 steps through the various lifetimes, and ends with an equation that will be central in making the analytical connection between net and crude lifetimes. This equation is proved in Leemis' Appendix C, and states that, under the assumption of independent risks,

$$h_{Y_j}(t) = \frac{\pi_j f_{X_j}(t)}{\sum_{i=1}^{k} \pi_i S_{X_i}(t)}, \tag{1}$$

where $Y_j$ represents the net lifetime associated with risk $j$, $X_j$ represents the crude lifetime associated with risk $j$, $\pi_j$ is the probability of failure from risk $j$, and $f$, $S$, and $h$ denote the probability density, survivor, and hazard functions, respectively. The details on computing $h_{Y_j}(t)$ are given in Appendix B. More recent references in the competing risks literature are given by [3].

## III. A Parametric Example

Although the main emphasis of the work here is nonparametric, the following parametric example demonstrates the calculation of the net lifetime of a failure distribution. For this example, we make the assumption that both the failure times and censoring times are exponentially distributed random variables with different rates. For illustration, we will turn to the 6-MP treatment group data set discussed earlier [10]. Using competing risks terminology, risk 1 corresponds to censoring, and risk 2 corresponds to failure. In the 6-MP data set, there are $n = 21$ patients on test, and the number of observed failures (leukemia remissions) is $r = 9$. Using maximum likelihood, the failure rate associated with the first crude lifetime is

$$\hat{\lambda}_{X_1} = \frac{n - r}{\sum_{i|\delta_i=0} x_i} = \frac{12}{250},$$

where $\delta_i = 0$ denotes a censored observation, $x_i = \min\{c_i, t_i\}$, $c_i$ is the censoring time for patient $i$, and $t_i$ is the failure time for patient $i$ $(i = 1, 2, \ldots, n)$. The failure rate associated with the second crude lifetime is

$$\hat{\lambda}_{X_2} = \frac{r}{\sum_{i|\delta_i=1} x_i} = \frac{9}{109},$$

where $\delta_i = 1$ denotes an observed failure $(i = 1, 2, \ldots, n)$. We estimate $\pi_1$, and $\pi_2$ as

$$\hat{\pi}_1 = \frac{n-r}{n} = \frac{12}{21} \qquad \text{and} \qquad \hat{\pi}_2 = \frac{r}{n} = \frac{9}{21}.$$

Because of our exponential assumption

$$\hat{f}_{X_1}(t) = \hat{\lambda}_{X_1} e^{-\hat{\lambda}_{X_1} t}, \hat{S}_{X_1}(t) = e^{-\hat{\lambda}_{X_1} t},$$
$$\hat{f}_{X_2}(t) = \hat{\lambda}_{X_2} e^{-\hat{\lambda}_{X_2} t}, \hat{S}_{X_2}(t) = e^{-\hat{\lambda}_{X_2} t},$$

for $t > 0$. Using (1), we compose the hazard function for the net lifetime of the failure data

$$h_{Y_2}(t) = \frac{\pi_2 f_{X_2}(t)}{\pi_1 S_{X_1}(t) + \pi_2 S_{X_2}(t)},$$

for $t > 0$, which is estimated by

$$\hat{h}_{Y_2}(t) = \frac{\hat{\pi}_2 \hat{\lambda}_{X_2} e^{-\hat{\lambda}_{X_2} t}}{\hat{\pi}_1 e^{-\hat{\lambda}_{X_1} t} + \hat{\pi}_2 e^{-\hat{\lambda}_{X_2} t}} = \frac{\frac{27}{109} e^{-9t/109}}{4e^{-12t/250} + 3e^{-9t/109}},$$

for $t > 0$. The probability density function of the net lifetime associated with failure is

$$f_{Y_2}(t) = h_{Y_2}(t) e^{-\int_0^t h_{Y_2}(\tau) d\tau},$$

and is estimated by

$$\hat{f}_{Y_2}(t) = \hat{h}_{Y_2}(t) e^{-\int_0^t \hat{h}_{Y_2}(\tau) d\tau},$$

for $t > 0$. This estimated probability density function of the net lifetime of the failure data can be used to calculate the skewness, and coefficient of variation of the failure data. These statistics are $(\hat{\gamma}, \hat{\gamma}_3) = (1.40, 2.55)$. In the remainder of the paper, we develop a nonparametric method in order to avoid the assumption of exponentiality for the failure time and the censoring mechanism. Computation will be greatly simplified with the help of APPL.

## IV. Methodology

This section details the steps involved in extracting a survival distribution estimate from the observed failure & censoring times. We will use the following eight-point data set in our examples.

$$x = [0.25^*, 0.25, 0.35^*, 0.45^*, 1.00, 1.15^*, 1.25, 1.35],$$

where an asterisk indicates a right-censored observation. The data could also be described as

$$x = [0.25, 0.25, 0.35, 0.45, 1.00, 1.15, 1.25, 1.35],$$
$$\delta = [0, 1, 0, 0, 1, 0, 1, 1],$$

where $x_i = \min\{t_i, c_i\}$, $t_i$ is the failure time, $c_i$ is the censoring time, and $\delta_i$ is 1 when $x_i = t_i$, and 0 when $x_i = c_i$. The random variables for failure times and censoring times are assumed to be mutually independent.

To verify the correctness of our method and the `CalcNetHaz` algorithm, we will work through two small examples.

The Gaussian distribution was an obvious choice for a kernel function. There are, however, several issues that preclude its use. The first issue is the range of support. Establishing a Gaussian kernel around any data value will include negative support values, which is an impossibility in lifetime data analysis because negative lifetimes never occur. The second issue is the intractability of the cumulative density function (CDF). While APPL was able to plot a (rather complicated) CDF, it was unable to calculate the coefficient of variation or the skewness.

We will instead use the PDF of the uniform, and triangular distributions as kernel functions. These distributions allow for simple, tractable CDF, and allow us to exploit APPL's piecewise function processing capability. We have avoided the problem of negative lifetimes with these distributions by carefully choosing our example data values and $h$ so as to avoid negative support for the kernel function.

### A. Uniform Kernel Function

Our method treats censored and observed data as coming from two different distributions. For simplicity in formulating, implementing, and testing our algorithm, we split the data into [0.25, 0.35, 0.45, 1.15], and [0.25, 1.0, 1.25, 1.35] where the first list contains the censoring times, and the second list contains the failure times. These data values were chosen arbitrarily with preference given to ease of verification by hand. According to the chosen method for calculating the bin width discussed in [22], we calculate $h$ as $h = 0.9sn^{-1/5}$, where $s$ is the sample standard deviation of the four values. For the first list, $h = 0.241$; and for the second, $h = 0.294$. We will, however, simplify by using an $h = 0.25$ for our calculations in order to ease arithmetic, and simplify verification. The first step is to create the two random variables with APPL:

```
h := 0.25;
C1 := UniformRV(0.25 − h, 0.25 + h);
C2 := UniformRV(0.35 − h, 0.35 + h);
C3 := UniformRV(0.45 − h, 0.45 + h);
C4 := UniformRV(1.15 − h, 1.15 + h);
X1 := Mixture([1/4, 1/4, 1/4, 1/4], [C1, C2, C3, C4]);
F1 := UniformRV(0.25 − h, 0.25 + h);
F2 := UniformRV(1.0 − h, 1.0 + h);
F3 := UniformRV(1.25 − h, 1.25 + h);
F4 := UniformRV(1.35 − h, 1.35 + h);
X2 := Mixture([1/4, 1/4, 1/4, 1/4], [F1, F2, F3, F4]);
```

At this point, APPL has created two density functions: $\hat{f}_{X_1}(t)$, and $\hat{f}_{X_2}(t)$. The kernel functions for these density functions are uniformly distributed, and centered at the observation with a width of $2h$. This APPL code returns the kernel

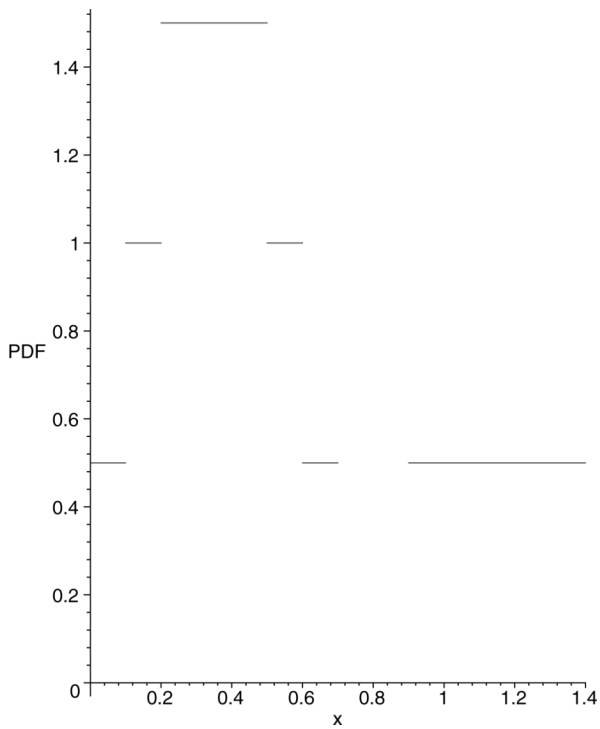


Fig. 2. Uniform kernel density estimate for the crude censoring time $X_1$.

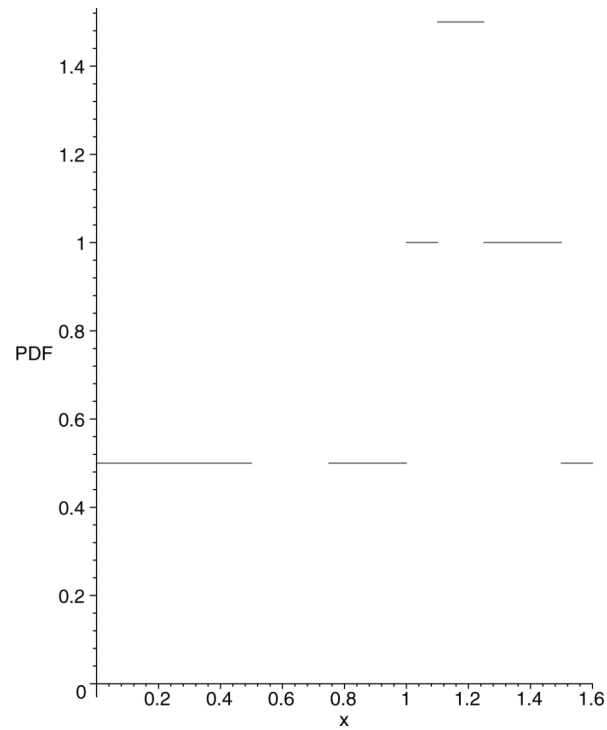


Fig. 3. Uniform kernel density estimate for the crude failure time $X_2$.

function estimate for the crude censoring time probability density function as

$$\hat{f}_{X_1}(t) = \begin{cases} 1/2 & 0 < t < 0.1 \\ 1 & 0.1 < t < 0.2 \\ 3/2 & 0.2 < t < 0.5 \\ 1 & 0.5 < t < 0.6 \\ 1/2 & 0.6 < t < 0.7 \\ 0 & 0.7 < t < 0.9 \\ 1/2 & 0.9 < t < 1.4. \end{cases}$$

Similarly, this APPL code returns the kernel function estimate for the crude failure time as

$$\hat{f}_{X_2}(t) = \begin{cases} 1/2 & 0 < t < 0.5 \\ 0 & 0.5 < t < 0.75 \\ 1/2 & 0.75 < t < 1.0 \\ 1 & 1.0 < t < 1.1 \\ 3/2 & 1.1 < t < 1.25 \\ 1 & 1.25 < t < 1.5 \\ 1/2 & 1.5 < t < 1.6. \end{cases}$$

Unless otherwise noted, all of the following plots were created using APPL's `PlotDist` command. Fig. 2 is a plot of $\hat{f}_{X_1}(t)$, and Fig. 3 is a plot of $\hat{f}_{X_2}(t)$.

To call the `CalcNetHaz` procedure, we must first compute a mixture of these density functions. This mixture will serve as the denominator $\pi_1 S_{X_1}(t) + \pi_2 S_{X_2}(t)$ in the calculation of $h_{Y_2}(t)$. The $\pi_i$ values come from the competing risks discussion, and are the probability of failing due to the $i$th risk. In our example, four of the observations "fail" by censoring, and the other four are observed failures. Therefore, $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$. In general, for a data set with $n$ items on test, and $r$ observed failures, $\hat{\pi}_1 = ((n-r)/n)$, and $\hat{\pi}_2 = (r/n)$. With APPL's `Mixture` procedure, we compute this denominator as a mixture of density functions, and convert it to a survivor function before passing it to `CalcNetHaz`.

```
X12 := Mixture([1/2, 1/2], [X1, X2]);
```

This probability density function $\hat{f}_{X_{12}}(t)$ has the following mathematical form, and is plotted in Fig. 4.

$$\hat{f}_{X_{12}}(t) = 0.5\hat{f}_{X_1}(t) + 0.5\hat{f}_{X_2}(t) = \begin{cases} 1/2 & 0 < t < 0.1 \\ 3/4 & 0.1 < t < 0.2 \\ 1 & 0.2 < t < 0.5 \\ 1/2 & 0.5 < t < 0.6 \\ 1/4 & 0.6 < t < 0.7 \\ 0 & 0.7 < t < 0.75 \\ 1/4 & 0.75 < t < 0.9 \\ 1/2 & 0.9 < t < 1.0 \\ 3/4 & 1.0 < t < 1.1 \\ 1 & 1.1 < t < 1.25 \\ 3/4 & 1.25 < t < 1.4 \\ 1/2 & 1.4 < t < 1.5 \\ 1/4 & 1.5 < t < 1.6. \end{cases}$$

Referring back to (1) to calculate a net lifetime from the crude lifetimes, the hazard function estimate for the net lifetime $Y_2$ associated with the crude lifetime $X_2$ is

$$h_{Y_2}(t) = \frac{\pi_2 f_{X_2}(t)}{\pi_1 S_{X_1}(t) + \pi_2 S_{X_2}(t)}.$$

The denominator has been calculated in Maple in its density form, and can be converted to a survivor function with the APPL

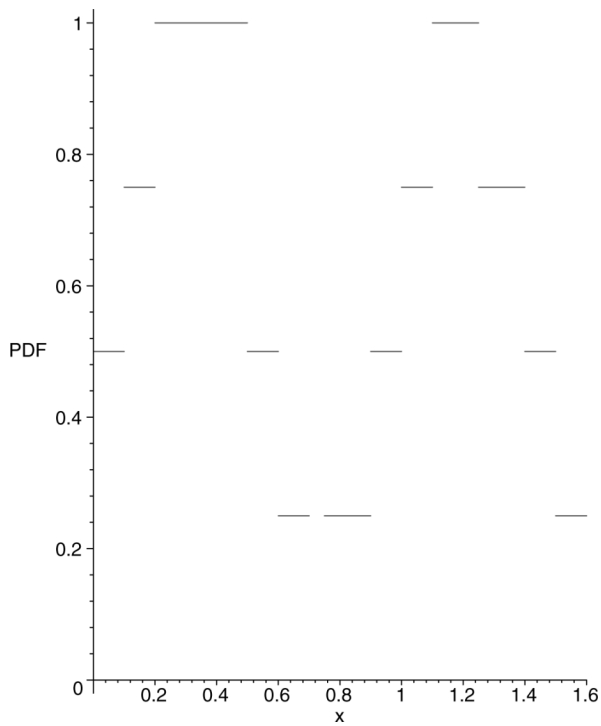


Fig. 4. Uniform kernel density estimate for $X_{12}$.

function SF. A call to CalcNetHaz with these three parameters gives the net lifetime hazard function for the observed failures.

```
Y2 := CalcNetHaz(X2, SF(X12), 0.5);
```

Because $X$ is the crude lifetime distribution, and $Y$ is the net lifetime distribution, we now have

$$\hat{h}_{Y_2}(t) = \frac{\pi_2 \hat{f}_{X_2}(t)}{\pi_1 \hat{S}_{X_1}(t) + \pi_2 \hat{S}_{X_2}(t)} = \begin{cases} \frac{1}{4-2t} & 0 < t < 0.1 \\ \frac{10}{41-30t} & 0.1 < t < 0.2 \\ \frac{10}{43-40t} & 0.2 < t < 0.5 \\ 0 & 0.5 < t < 0.75 \\ \frac{4}{11-4t} & 0.75 < t < 0.9 \\ \frac{20}{73-40t} & 0.9 < t < 1.0 \\ \frac{40}{3(31-20t)} & 1.0 < t < 1.1 \\ \frac{12}{23-16t} & 1.1 < t < 1.25 \\ \frac{4}{3(3-2t)} & 1.25 < t < 1.4 \\ \frac{20}{31-20t} & 1.4 < t < 1.5 \\ \frac{5}{8-5t} & 1.5 < t < 1.6. \end{cases}$$

The functional form of this hazard function, as computed by CalcNetHaz, has been verified by hand. Fig. 5 contains a plot of $\hat{h}_{Y_2}(t)$, which has a vertical asymptote at $t = 1.6$. The following additional APPL statements give the coordinates of a point that

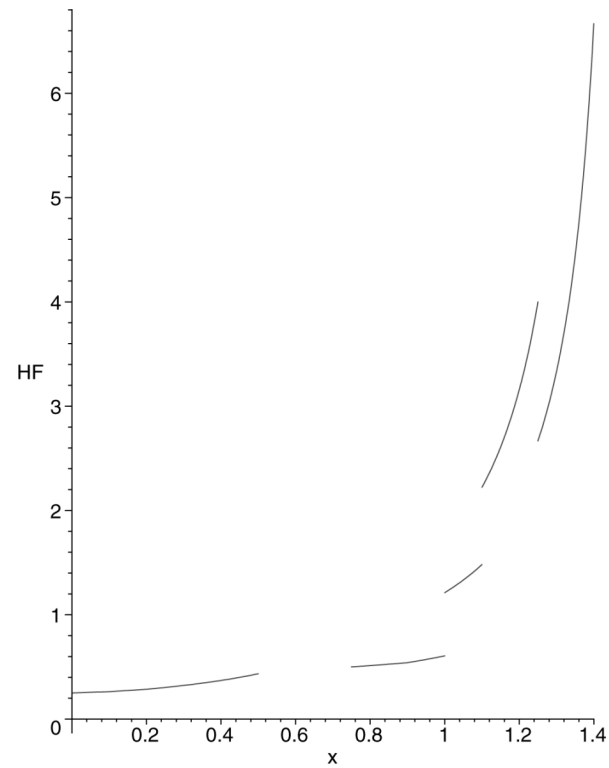


Fig. 5. Uniform kernel hazard function estimate for net lifetime of observed failures $Y_2$.

TABLE I
$\hat{\gamma}$, AND $\hat{\gamma}_3$ FOR DIFFERENT METHODS FOR HANDLING CENSORING

| Method | $\hat{\gamma}$ | $\hat{\gamma}_3$ |
|---|---|---|
| Increase censoring times by 0% | 0.58 | 0.55 |
| Increase censoring times by 50% | 0.69 | 0.76 |
| Increase censoring times by 100% | 0.76 | 0.80 |
| Parametric (exponential) analysis | 1.40 | 2.55 |
| Competing risks (uniform kernel) | 0.73 | 0.04 |

can be plotted on Cox & Oakes' parametric model discrimination plot:

```
cv := CoefOfVar(Y2);
skew := Skewness(Y2);
```

These statements yield the point $(\hat{\gamma}, \hat{\gamma}_3) = (0.3608, -1.2320)$. When this same technique is applied to the 6-MP treatment group data with $h_2 = 5.798$ for the censored data, and $h_2 = 3.970$ for the failure data, the point obtained is $(\hat{\gamma}, \hat{\gamma}_3) = (0.7338, 0.0425)$. Table I compares this point with the three points plotted in Fig. 1. There is considerable difference between the heuristic approaches, the parametric analysis, and the competing risks approach.

### B. Triangular Kernel Function

This section follows the technique and data set from Section IV-A, but uses the triangular distribution for the kernel

function in the density estimation method. The APPL code for calculating the kernel estimate for the PDF of the crude lifetimes is

```
h := 0.25;
C1 := TriangularRV(0.25 − h, 0.25, 0.25 + h);
C2 := TriangularRV(0.35 − h, 0.35, 0.35 + h);
C3 := TriangularRV(0.45 − h, 0.45, 0.45 + h);
C4 := TriangularRV(1.15 − h, 1.15, 1.15 + h);
X1 := Mixture([1/4, 1/4, 1/4, 1/4], [C1, C2, C3, C4]);
F1 := TriangularRV(0.25 − h, 0.25, 0.25 + h);
F2 := TriangularRV(1.0 − h, 1.0, 1.0 + h);
F3 := TriangularRV(1.25 − h, 1.25, 1.25 + h);
F4 := TriangularRV(1.35 − h, 1.35, 1.35 + h);
X2 := Mixture([1/4, 1/4, 1/4, 1/4], [F1, F2, F3, F4]);
```

The three parameters in TriangularRV are the minimum, mode, and maximum. Again, $\hat{f}_{X_1}(t)$, and $\hat{f}_{X_2}(t)$ are the density function estimates. Their kernel functions are now triangular distributions centered at the observation, and have a width of $2h$.

$$\hat{f}_{X_1}(t) = \begin{cases} 4t & 0 < t < 0.1 \\ 8t - \frac{2}{5} & 0.1 < t < 0.2 \\ 12t - \frac{6}{5} & 0.2 < t < 0.25 \\ 4t + \frac{4}{5} & 0.25 < t < 0.35 \\ -4t + \frac{18}{5} & 0.35 < t < 0.45 \\ -12t + \frac{36}{5} & 0.45 < t < 0.5 \\ -8t + \frac{26}{5} & 0.5 < t < 0.6 \\ -4t + \frac{14}{5} & 0.6 < t < 0.7 \\ 0 & 0.7 < t < 0.9 \\ 4t - \frac{18}{5} & 0.9 < t < 1.15 \\ -4t + \frac{28}{5} & 1.15 < t < 1.4. \end{cases}$$

$$\hat{f}_{X_2}(t) = \begin{cases} 4t & 0 < t < 0.25 \\ 2 - 4t & 0.25 < t < 0.5 \\ 0 & 0.5 < t < 0.75 \\ 4t - 3 & 0.75 < t < 1.0 \\ 1 & 1.0 < t < 1.10 \\ 4t - \frac{17}{5} & 1.10 < t < 1.25 \\ \frac{8}{5} & 1.25 < t < 1.35 \\ \frac{62}{5} - 8t & 1.35 < t < 1.5 \\ \frac{32}{5} - 4t & 1.5 < t < 1.6. \end{cases}$$

These probability density function estimates are plotted in Figs. 6 and 7.

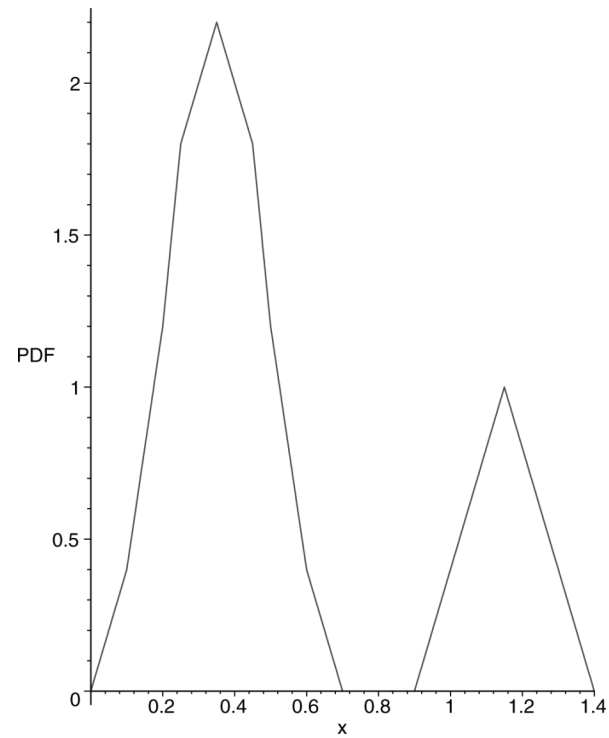


Fig. 6. Triangular kernel density estimate for crude censoring time $X_1$.

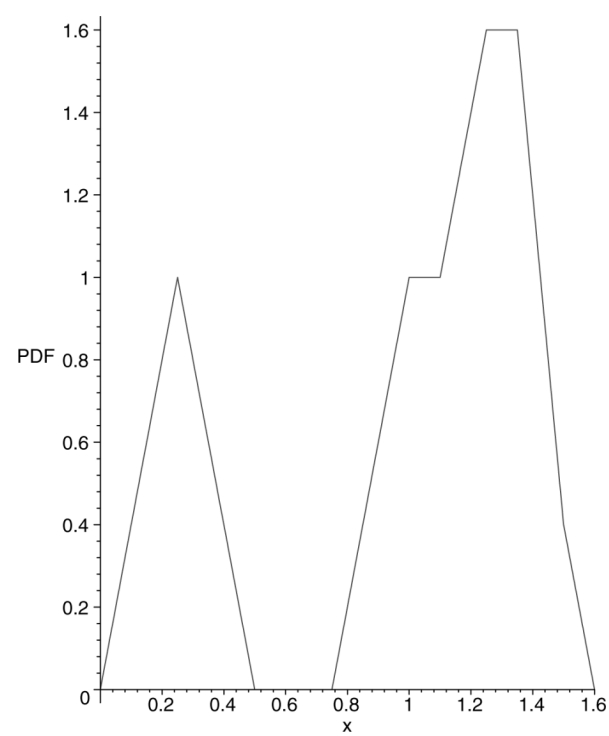


Fig. 7. Triangular kernel density estimate for crude failure time $X_2$.

As in the uniform case, we compute a mixture of these density functions to be the denominator of (1).

```
X12 := Mixture([1/2, 1/2], [X1, X2]);
```

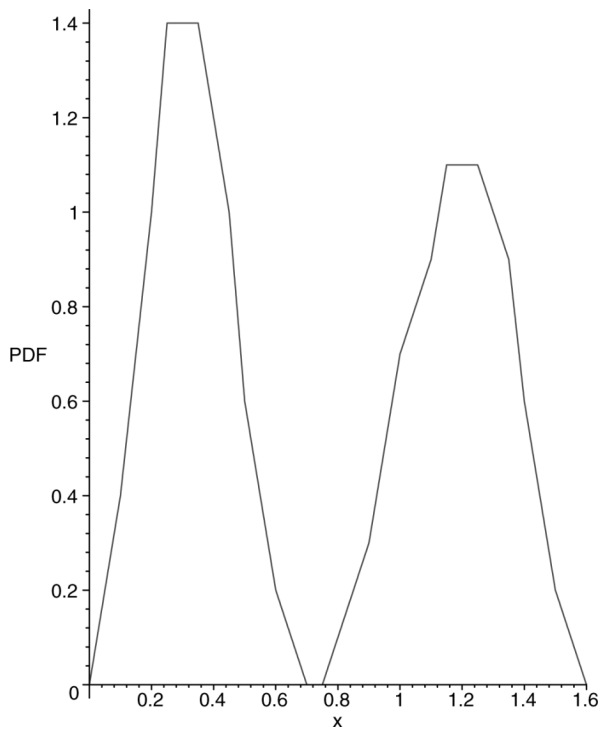


Fig. 8. Triangular kernel density estimate for $X_{12}$.

The probability density function $\hat{f}_{X_{12}}(t)$ has the following form, and is plotted in Fig. 8.

$$\hat{f}_{X_{12}}(t) = 0.5\hat{f}_{X_1}(t) + 0.5\hat{f}_{X_2}(t) = \begin{cases} 4t & 0 < t < 0.1 \\ 6t - \frac{1}{5} & 0.1 < t < 0.2 \\ 8t - \frac{3}{5} & 0.2 < t < 0.25 \\ \frac{7}{5} & 0.25 < t < 0.35 \\ -4t + \frac{14}{5} & 0.35 < t < 0.45 \\ -8t + \frac{23}{5} & 0.45 < t < 0.5 \\ -4t + \frac{13}{5} & 0.5 < t < 0.6 \\ -2t + \frac{7}{5} & 0.6 < t < 0.7 \\ 0 & 0.7 < t < 0.75 \\ 2t - \frac{3}{2} & 0.75 < t < 0.9 \\ 4t - \frac{33}{10} & 0.9 < t < 1.0 \\ 2t - \frac{13}{10} & 1.0 < t < 1.1 \\ 4t - \frac{7}{2} & 1.1 < t < 1.15 \\ \frac{11}{10} & 1.15 < t < 1.25 \\ -2t + \frac{18}{5} & 1.25 < t < 1.35 \\ -6t + 9 & 1.35 < t < 1.4 \\ -4t + \frac{31}{5} & 1.4 < t < 1.5 \\ -2t + \frac{31}{5} & 1.5 < t < 1.6. \end{cases}$$

Using (1) as before, we want to calculate the hazard function of the net lifetime for the observed failures. A call to `CalcNetHaz` with these three parameters gives the hazard function for the net lifetimes for the observed failures.

```
Y2 := CalcNetHaz(X2, SF(X12), 0.5);
```

If $X$ is the crude lifetime distribution, and $Y$ is the net lifetime distribution for a random variable, the following hazard function results.

$$\hat{h}_{Y_2}(t) = \frac{\pi_2 \hat{f}_{X_2}(t)}{\pi_1 \hat{S}_{X_1}(t) + \pi_2 \hat{S}_{X_2}(t)} = \begin{cases} \frac{2t}{-2t^2+1} & 0 < t < 0.1 \\ \frac{200t}{99+20t-300t^2} & 0.1 < t < 0.2 \\ \frac{40t}{19+12t-80t^2} & 0.2 < t < 0.25 \\ \frac{5(2t-1)}{-6+7t} & 0.25 < t < 0.35 \\ \frac{-200(2t-1)}{289-560t+400t^2} & 0.35 < t < 0.45 \\ \frac{-20(2t-1)}{37-92t+80t^2} & 0.45 < t < 0.5 \\ 0 & 0.5 < t < 0.75 \\ \frac{-8(4t-3)}{1-24t+16t^2} & 0.75 < t < 0.9 \\ \frac{-200(4t-3)}{349-1320t+800t^2} & 0.9 < t < 1.0 \\ \frac{200}{51+520t-400t^2} & 1.0 < t < 1.1 \\ \frac{-40(-17+20t)}{433-1400t+800t^2} & 1.1 < t < 1.15 \\ \frac{8(-17+20t)}{125-88t} & 1.15 < t < 1.25 \\ \frac{32}{125-144t+40t^2} & 1.25 < t < 1.35 \\ \frac{-20(-31+20t)}{677-900t+300t^2} & 1.35 < t < 1.4 \\ \frac{-20(-31+20t)}{481-620t+200t^2} & 1.4 < t < 1.5 \\ \frac{10}{8-5t} & 1.5 < t < 1.6. \end{cases}$$

The functional form of this hazard function has also been verified by hand. It is plotted in Fig. 9. As in the uniform case, $\hat{h}_{Y_2}(t)$ has a vertical asymptote at 1.6. Using APPL, the coordinates of this distribution on Cox & Oakes' parametric model discrimination plot are calculated as follows.

```
cv := CoefOfVar(Y2);
skew := Skewness(Y2);
```

This calculation yields the point $(\hat{\gamma}, \hat{\gamma}_3) = (0.3452, -1.3390)$. Not surprisingly, this point is in reasonable proximity to the point (0.3608, −1.2320) obtained with the uniform kernel.

## V. Monte Carlo Simulation Analysis

The Monte Carlo analysis of our algorithm begins with a plot of $\hat{\gamma}_3$ vs. $\hat{\gamma}$ for several parametric distributions which will serve as a baseline to assess how well our algorithm adapts to censored data values. The plot for various Weibull and log-logistic distributions appears in Fig. 10. The points plotted are associated with

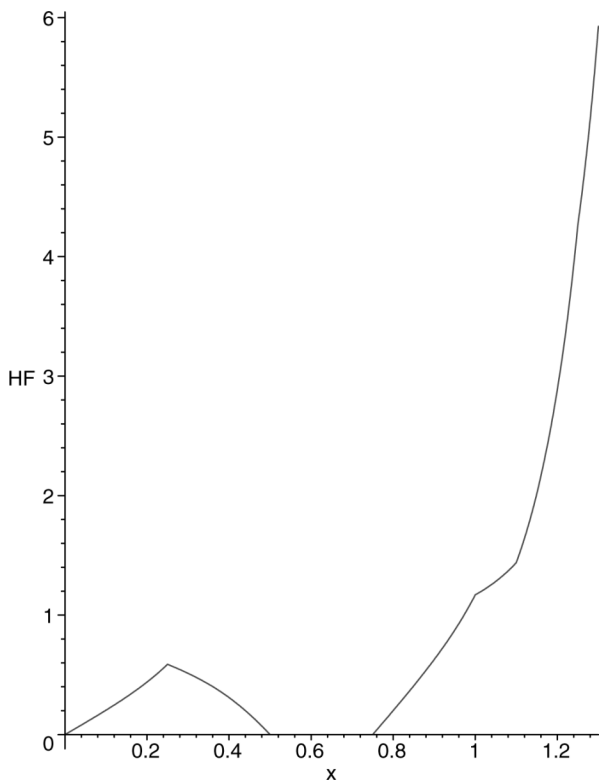


Fig. 9. Triangular kernel hazard function estimate for the net lifetime of the observed failures $Y_2$.

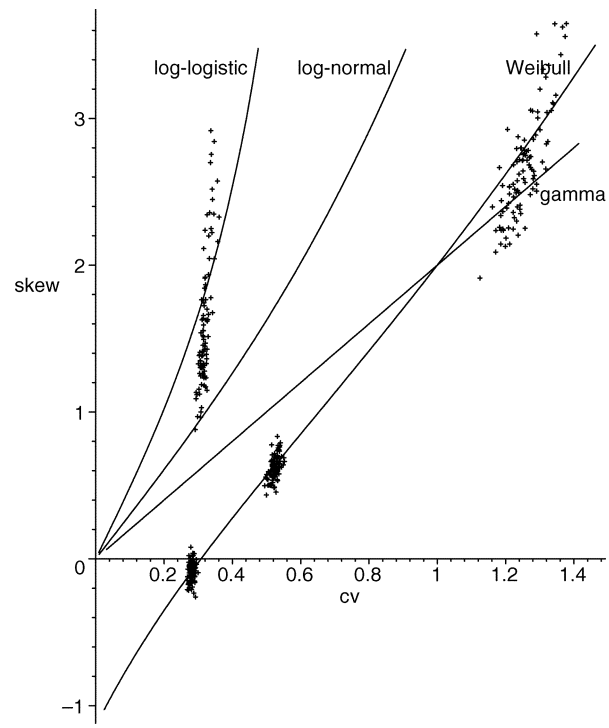


Fig. 10. Complete data sets for Weibull with $\kappa = 4, 2, 0.8$; and log-logistic with $\kappa = 6$.

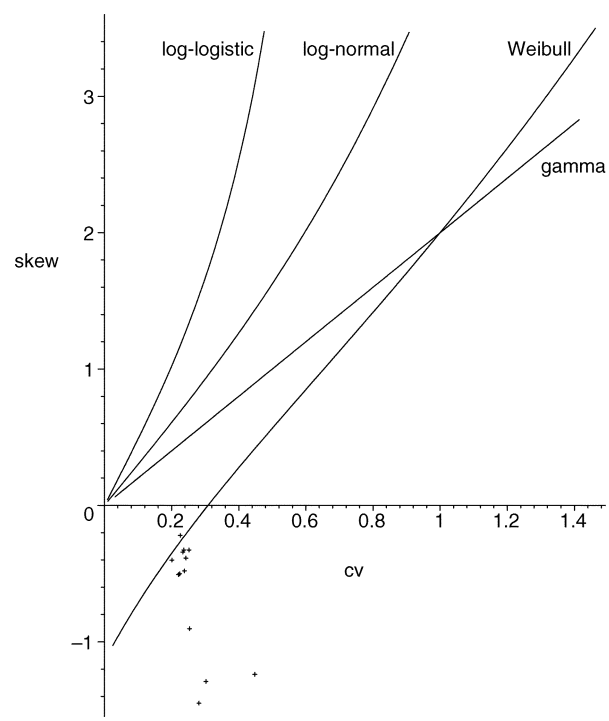


Fig. 11. Censored data sets for Weibull with shape parameter 5 with sample size $n = 200$.

a sample size of $n = 1000$ for each of four parametric distributions (one log-logistic, and three Weibull with shape parameters 4, 2, and 0.8 from left to right). The plotted values cluster around the correct point on the chart, but have considerable dispersion for such a large sample size. Not surprisingly, the dispersion associated with the skewness consistently exceeds the dispersion associated with the coefficient of variation. Also, as the shape parameter in the Weibull distribution decreases, the points spread. Some of the parameter choices exhibit a positive correlation between $\hat{\gamma}$ and $\hat{\gamma}_3$.

Appendix C contains a Maple implementation of a Monte Carlo simulation of the method described here. Fig. 11 displays points on the Cox & Oakes' graph where the lifetimes are drawn from a Weibull population with a shape parameter of 5, and were calculated from data sets of $n = 200$ values. Computational time requirements prevented the plotting of as many values as in Fig. 10. The points cluster around the appropriate point on the Weibull curve, although there are several points that fall significantly to the southeast of the target. The smaller number of items on test relative to the simulation illustrated in Fig. 10 results in a wider dispersion of the plotted values.

## VI. Conclusions, and Further Work

Cox & Oakes' parametric survival model discrimination plot has been extended to the case of a right-censored survival data set using kernel functions to estimate probability density functions of the crude lifetime PDF, and a competing risks framework to adjust for the effect of censoring. APPL and Maple were used for numerical and analytical computations, as well as for bookkeeping piecewise functions. This research work has a practical significance in that it allows a modeler to simultaneously assess competing parametric models for their fit to a data set consisting of censored observations.

Future work on this topic could proceed in several directions. Other kernel functions could be used to obtain a scatter-plot of skewness versus a coefficient of variation. Using the Weibull

distribution with $\kappa = [1/(1 - \ln(2))] \cong 3.25889$ (where the mode equals the median to give a bell-shaped kernel density function estimate) would be a reasonable approximation to the Gaussian distribution, and would prevent negative lifetime support. In addition, because the Weibull distribution has positive support, there will be fewer piecewise segments to account for. The influence that $h$ has on a plotted point could also be investigated.

## Appendix A
## APPL Code to Create Cox & Oakes' Graph

```
unassign('kappa'):
lambda := 1:
X := GammaRV(lambda, kappa):
c := CoefOfVar(X):
s := Skewness(X):
GammaPlot := plot([c, s, kappa = 0.5..999],
labels = ["cv", "skew"]):

unassign('kappa'):
lambda := 1:
X := WeibullRV(lambda, kappa):
c := CoefOfVar(X):
s := Skewness(X):
WeibullPlot := plot([c, s, kappa = 0.7..50.7]):

unassign('kappa'):
lambda := 1:
Y := LogNormalRV(lambda, kappa):
c := CoefOfVar(Y):
s := Skewness(Y):
LogNormalPlot := plot([c, s, kappa = 0.01..0.775]):

unassign('kappa'):
lambda := 1:
Y := LogLogisticRV(lambda, kappa):
c := CoefOfVar(Y):
s := Skewness(Y):
LogLogisticPlot := plot([c, s, kappa = 4.3..200.5]):

cnsrgrp := plot([[0.5849304, 0.5531863]],
style = point, symbol = box):
cnsrgrp15 := plot([[0.6883908, 0.760566]],
style = point, symbol = cross):
cnsrgrp20 := plot([[0.7633863, 0.8009897]],
style = point, symbol = diamond):
with(plots):

lll := textplot([0.17, 3.3, "log − logistic"],
'align = {ABOVE, RIGHT}'):
lnl := textplot([0.59, 3.3, "log − normal"],
'align = {ABOVE, RIGHT}'):
wbl := textplot([1.2, 3.3, "Weibull"], 'align =
{ABOVE, RIGHT}'):
gml := textplot([1.3, 2.44, "gamma"], 'align =
{ABOVE, RIGHT}'):

plots[display]({lll, lnl, wbl, gml,
GammaPlot, WeibullPlot, LogNormalPlot,
    LogLogisticPlot, cnsrgrp, cnsrgrp15, cnsrgrp20},
    scaling = unconstrained);
```

## Appendix B
## Computational Issues

With our kernel density estimates being mixtures of large numbers of random variables, even small data sets could result in piecewise functions with an unmanageable number of segments. To assist in the computation of these functions, we turned to the APPL programming language [11]. APPL is a Maple-based programming language that allows for statistical manipulations of piecewise random variables. In addition, APPL allows for the creation, and combination of all types of standard random variables (uniform, Gaussian, triangular, Weibull, etc.) we use in our kernel functions. The flexibility of APPL will allow for the efficient manipulation of many random variables.

Despite APPL's comprehensive random variable handling ability, the equation at the core of our analysis (1) has not been implemented. This necessitated our devising an algorithm (using the APPL language as a platform) that could perform the implementation of (1) for random variables defined in a piecewise manner. The Maple function `CalcNetHaz` calculates $h_{Y_j}(t)$ for crude lifetimes defined in a piecewise manner. It must be passed the APPL probability density function (PDF) subprogram for the numerator $f_{X_j}(t)$, a mixture of APPL survival functions for the denominator, and the numerator's $\pi_j$. The procedure `CalcNetHaz` returns the hazard function of the time to failure using (1). The code used to check for

- the correct number of arguments;
- the correct format for the PDF of the numerator, and the mixture of survivor functions in the denominator;
- the correct type (continuous) of random variables $X_1, X_2, \ldots, X_k$;
- the numerator given as a PDF, and the denominator as a SF; and
- $0 < \pi_j < 1$

is suppressed for brevity. Because the kernel estimate for the failure & censoring distributions may be defined in a piecewise fashion (e.g., for a uniform or triangular kernel), the procedure accommodates piecewise distributions.

```
CalcNetHaz := proc(num :: list(list),
denom :: list(list), NumPI :: float)
```

```
    local retval, nsegn, i, j:
    retval := []:
    nsegn := nops(num[2]):
    i := 1:
    for j from 2 by 1 to nsegn do
        while denom[2][i] < num[2][j] do
            retval := [op(retval), unapply(simplify(
                (NumPI*num[1][j - 1])/denom[1][i])(x), x)]:
            i := i + 1:
        end do:
    end do:
    return([retval, denom[2], ["Continuous", "HF"]]):
end:
```

The first two arguments to CalcNetHaz are lists of three lists; the last argument, $\pi_j$, is a scalar. The first list in the first parameter's three lists is of the numerator's $n-1$ density functions. These correspond to the $n$ breakpoints in the numerator. The first list in the second parameter's three lists is of the denominator's $m-1$ density functions. These correspond to the $m$ breakpoints in the denominator. These breakpoints are found in the second of the three lists. The third list contains the strings "Continuous", and either "PDF" or "SF" to denote the type of distribution representation. For each of the segments, the algorithm calculates a hazard function for the current segment based on (1). The algorithm assumes that the denominator is a mixture distribution involving the term in the numerator. We make this assumption because, in (1), $S_{X_j}$ in the denominator is derived from $f_{X_j}$ in the numerator (or vice-versa), and results in denominator segment breaks that are a superset of those in the numerator. After looping through each of the segments, the algorithm returns the list of hazard functions along with the segment breaks of the denominator.

## Appendix C
## Monte Carlo Simulation

```
n := 1000:
kappa := 5:
for i from 1 to 80 do
    r := 0:
    X1 := []:
    X2 := []:
    for k from 1 to n do
        f := -log(UniformVariate())^(1/kappa):
        c := -log(UniformVariate())^(1/kappa):
        iff < c then
            r := r + 1:
            X2 := [op(X2), f]:
        else
            X1 := [op(X1), c]:
        end if:
end do:
if r < n and r > 0 then
    dists := []:
    weights := []:
    R := describe[quartile[3]](X1) - describe[quartile[1]](X1):
    h := 0.79 * R * (n - r)^(-0.2):
    for j from 1 to n - r do
        weights := [op(weights), 1/(n - r)]:
        if h > sort(X1)[1] then
            fd := fopen("mapsim2", APPEND):
            fprintf(fd, "The following line was padded: h = %g min = %g \ n",
            h, sort(X1)[1]):
            fclose(fd):
            h := sort(X1)[1]:
        end if:
        dists := [op(dists), UniformRV(X1[j] - h, X1[j] + h)]:
    od:
    f_X1 := Mixture(weights, dists):
    dists := []:
    weights := []:
    R := describe[quartile[3]](X2) - describe[quartile[1]](X2):
    h := 0.79 * R * (r^(-0.2)):
    for j from 1 to r do
        weights := [op(weights), 1/r]:
        if h > sort(X2)[1] then
            fd := fopen("mapsim2", APPEND):
            fprintf(fd, "The following line was padded:h = %g min = %g \ n",
            h, sort(X2)[1]):
            fclose(fd):
            h := sort(X2)[1]:
```

```
            end if:
            dists := [op(dists), UniformRV(X2[j] -
            h, X2[j] + h)]:
        od:
        f_X2 := Mixture(weights, dists):
        f_X12 := Mixture([(n-r)/n, r/n], [f_X1, f_X2]):
        h_Y2  :=
        CalcNetHaz(f_X2, SF(f_X12), evalf(r/n)):
        f_Y2 := PDF(h_Y2):
        mu := Mean(f_Y2):
        ExpValueXSqrd := ExpectedValue(f_Y2, x- >
        x^2):
        sigma := sqrt(ExpValueXSqrd - mu^2):
        Term1 := ExpectedValue(f_Y2, x- > x^3):
        Term2 := 3 * mu * ExpValueXSqrd:
        Term3 := 2 * mu^3:
        skew := (Term1 - Term2 + Term3)/sigma^3:
        cov := sigma/mu:
        fd := fopen("mapsim2", APPEND):
        fprintf(fd, "[[%g, %g]], \n", Re(cov),
        Re(skew)):
        fclose(fd):
    elif r = n then
        skew := describe[skewness](X2):
        cov  :=
        describe[coefficientofvariation](X2):
        fd := fopen("mapsim4", APPEND):
        fprintf(fd, "[[%g, %g]], \n", Re(cov),
        Re(skew)):
        fclose(fd):
    end if:
end do:
```


## ACKNOWLEDGMENT

The authors thank Michael Trosset for his help with references on kernel functions. They also thank Diane Evans and Andy Glen for the ability to use the APPL programming language. They also thank the associate editor, and two referees for their insightful comments and suggestions. The second author thanks The College of William & Mary for the research leave to complete this research.


## REFERENCES


[1] A. Biswas and R. Sundaram, "Kernel survival function estimation based on doubly censored data," *Communications in Statistics: Theory and Methods*, vol. 35, pp. 1293–1307, 2006.

[2] J. W. Boag, "Maximum likelihood estimates of the proportion of patients cured by cancer therapy," *Journal of the Royal Statistical Society*, vol. 11, pp. 15–53, 1949.

[3] N. Bousquet, H. Bertholon, and G. Celeux, "An alternative competing risk model to the Weibull distribution for modeling aging in lifetime data analysis," *Lifetime Data Analysis*, vol. 12, pp. 481–504, 2006.

[4] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press, 1997.

[5] C. L. Chiang, *Introduction to Stochastic Processes in Biostatistics*. New York: Wiley, 1968.

[6] D. R. Cox and D. Oakes, *Analysis of Survival Data*. New York: Chapman and Hall, 1984.

[7] M. J. Crowder, *Classical Competing Risks*. Boca Raton, Fla: Chapman and Hall/CRC Press, 2001.

[8] H. A. David and M. L. Moeschberger, *The Theory of Competing Risks*. New York: Macmillan Publishing Co., 1978, Griffin's Statistical Monographs & Courses No. 39.

[9] D. L. Evans and L. M. Leemis, "Input modeling using a computer algebra system," in *Proceedings of the 2000 Winter Simulation Conference*, J. Joines, R. Barton, P. Fishwick, and K. Kang, Eds., 2000, pp. 577–586.

[10] E. A. Gehan, "A generalized Wilcoxon test for comparing arbitrarily singly-censored samples," *Biometrika*, vol. 52, pp. 203–223, 1965.

[11] A. Glen, D. Evans, and L. M. Leemis, "APPL: A probability programming language," *The American Statistician*, vol. 55, pp. 156–166, 2001.

[12] R. Jiang and A. K. S. Jardine, "Composite scale modeling in the presence of censored data," *Reliability Engineering & System Safety*, vol. 91, no. 7, pp. 756–764, 2006.

[13] S. T. Jiang, T. L. Landers, and T. R. Rhoads, "Semi-parametric proportional intensity models robustness for right-censored recurrent failure data," *Reliability Engineering & System Safety*, vol. 90, no. 1, pp. 91–98, 2005.

[14] J. D. Kalbfleisch and R. J. MacKay, "On constant-sum models for censored survival data," *Biometrika*, vol. 66, pp. 87–90, 1979.

[15] D. Kundu and A. M. Sarhan, "Analysis of incomplete data in presence of competing risks among several groups," *IEEE Trans. on Reliability*, vol. 55, no. 2, pp. 262–269, 2006.

[16] L. M. Leemis, *Reliability: Probabilistic Models and Statistical Methods*. Englewood Cliffs, New Jersey: Prentice-Hall, 1995.

[17] C. H. Li and N. Fard, "Optimum bivariate step-stress accelerated life test for censored data," *IEEE Trans. on Reliability*, vol. 56, no. 1, pp. 77–84, 2007.

[18] C. Park, "Parameter estimation of incomplete data in competing risks using the EM algorithm," *IEEE Trans. on Reliability*, vol. 54, no. 2, pp. 282–290, 2005.

[19] M. Pintilie, *Competing Risks: A Practical Perspective*. Hoboken, N.J.: John Wiley & Sons Inc., 2006.

[20] R. L. Prentice and J. D. Kalbfleisch, "The analysis of failure times in the presence of competing risks," *Biometrics*, vol. 34, pp. 541–554, 1978.

[21] A. M. Sarhan, "Analysis of incomplete, censored data in competing risks models with generalized exponential distributions," *IEEE Trans. on Reliability*, vol. 56, no. 1, pp. 132–138, 2007.

[22] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall, 1986.

[23] A. A. Soliman, "Estimation of parameters of life from progressively censored data using Burr-XII model," *IEEE Trans. on Reliability*, vol. 54, no. 1, pp. 34–42, 2005.

[24] J. S. Williams and S. W. Lagakos, "Models for censored survival analysis: Constant-sum and variable-sum models," *Biometrika*, vol. 64, pp. 215–224, 1977.

[25] L. F. Zhang, M. Xie, and L. C. Tang, "Robust regression using probability plots for estimating the Weibull shape parameter," *Quality and Reliability Engineering International*, vol. 22, no. 8, pp. 905–917, 2006.

[26] L. F. Zhang, M. Xie, and L. C. Tang, "Bias correction for the least squares estimator of Weibull shape parameter with complete and censored data," *Reliability Engineering & System Safety*, vol. 91, no. 8, pp. 930–939, 2006.



**A. Daniel Block** received his B.A. degree in economics, and his M.S. degree in Computational Operations Research from The College of William & Mary. He is presently teaching mathematics, and doing consulting work in price optimization. His research interests are in practical implementation of statistical models.

**Lawrence M. Leemis** received his B.S. and M.S. degrees in mathematics; and his Ph.D. in Industrial Engineering from Purdue University. He is presently teaching Operations Research at The College of William & Mary. His research interests are in reliability, simulation, and computational probability.