# Logspline Density Estimation for Censored Data

CHARLES KOOPERBERG * AND CHARLES J. STONE †

Logspline density estimation is developed for data that may be right censored, left censored, or interval censored. A fully automatic method, which involves the maximum likelihood method and may involve stepwise knot deletion and either the Akaike information criterion (AIC) or Bayesian information criterion (BIC), is used to determine the estimate. In solving the maximum likelihood equations, the Newton–Raphson method is augmented by occasional searches in the direction of steepest ascent. Also, a user interface based on *S* is described for obtaining estimates of the density function, distribution function, and quantile function and for generating a random sample from the fitted distribution.

**Key Words:** AIC; BIC; Maximum likelihood; Polynomial splines; *S*; Stepwise knot deletion; User interface.

## 1. INTRODUCTION

Consider data that can be thought of as arising as a random sample from a distribution on a known open interval having an unknown positive, continuous density function on that interval. In practice the traditional way of modeling the unknown distribution is to assume a classical parametric model such as normal, lognormal, gamma, Weibull, Pareto, or beta. Alternatively, we can use a histogram, kernel, or other nonparametric estimate of the unknown density function.

We can think of histogram density estimation as modeling the unknown log-density function by a piecewise constant function and estimating the unknown coefficients of the model by the method of maximum likelihood. Similarly, we can model the log-density function by a linear spline (continuous, piecewise linear function), quadratic spline (continuously differentiable, piecewise quadratic polynomial), or cubic spline (twice continuously differentiable, piecewise cubic polynomial) and use the maximum likelihood method to estimate the unknown coefficients. The resulting methodology, known as logspline density estimation, has been studied in Stone and Koo (1986), Stone (1990), and

*Department of Statistics, University of Washington, Seattle, WA 98195
†Department of Statistics, University of California, Berkeley, CA 94720

Received February 1992; Revised October 1992

*Volume 1, Number 4, pp. 301–328*

Kooperberg and Stone (1991). (In these papers as well as the present article, cubic splines are employed.)

Here logspline density estimation will be further refined and studied and extended to handle data that may be right censored, left censored, or interval censored. In addition, we will describe a user interface that makes the extended procedure conveniently available within the *S* environment (see Becker, Chambers, and Wilks 1988). To evaluate the procedure in its present form, we will apply it to a number of simulated and real data sets. Finally, the main issues involved in the numerical implementation of the procedure will be discussed.

For work on kernel density estimation in the presence of right-censored data, see Marron and Padgett (1987) and the references cited therein. We are not aware of work on density estimation in the presence of right-censored, left-censored, and interval-censored data.

## 2. LOGSPLINE MODELS

In this section we will give a detailed description of logspline models and introduce some auxiliary notation that will be used later on. Given the integer $K \geq 3$, the numbers $L$ and $U$ with $-\infty \leq L < U \leq \infty$, and the sequence $t_1, \ldots, t_K$ with $L < t_1 < \cdots < t_K < U$, let $\mathcal{S}_0$ be the space of twice-continuously differentiable functions $s$ on $(L, U)$ such that the restriction of $s$ to each of the intervals $(L, t_1]$, $[t_1, t_2], \ldots, [t_{K-1}, t_K]$, $[t_K, U)$ is a cubic polynomial. The space $\mathcal{S}_0$ is $(K+4)$-dimensional, and the functions in this space are referred to as cubic splines having (simple) knots at $t_1, \ldots, t_K$. Let $\mathcal{S}$ be the subspace of $\mathcal{S}_0$ consisting of the functions in $\mathcal{S}$ that are linear on $(L, t_1]$ and on $[t_K, U)$. The space $\mathcal{S}$ is $K$-dimensional, and the functions in this space are referred to as natural (cubic) splines. Set $p = K - 1$. Then $\mathcal{S}$ has a basis of the form $1, B_1, \ldots, B_p$. We can choose $B_1, \ldots, B_p$ such that $B_1$ is linear with negative slope on $(L, t_1]$, $B_2, \ldots, B_p$ are constant on $(L, t_1]$, $B_p$ is linear with positive slope on $[t_K, U)$, and $B_1, \ldots, B_{p-1}$ are constant on $[t_K, U)$.

A column vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t \in I\!R^p$ is said to be *feasible* if

$$\int_L^U \exp(\theta_1 B_1(y) + \cdots + \theta_p B_p(y))dy < \infty$$

or, equivalently, if (1) either $L > -\infty$ or $\theta_1 < 0$ and (2) either $U < \infty$ or $\theta_p < 0$. Let $\Theta$ denote the collection of such feasible column vectors. Given $\boldsymbol{\theta} \in \Theta$, set

$$C(\boldsymbol{\theta}) = \log\left(\int_L^U \exp(\theta_1 B_1(y) + \cdots + \theta_p B_p(y))dy\right)$$

and

$$f(y; \boldsymbol{\theta}) = \exp(\theta_1 B_1(y) + \cdots + \theta_p B_p(y) - C(\boldsymbol{\theta})), \quad L < y < U.$$

Then $f(\cdot; \boldsymbol{\theta})$ is a positive density function on $(L, U)$ for $\boldsymbol{\theta} \in \Theta$. The corresponding distribution function $F(\cdot; \boldsymbol{\theta})$ and quantile function $Q(\cdot; \boldsymbol{\theta})$ are given by

$$F(y; \boldsymbol{\theta}) = \int_L^y f(z; \boldsymbol{\theta})dz, \quad L < y < U,$$

and

$$Q(p;\boldsymbol{\theta}) = F^{-1}(p;\boldsymbol{\theta}), \quad 0 < p < 1$$

(so that $F(Q(p;\boldsymbol{\theta});\boldsymbol{\theta}) = p$ for $0 < p < 1$ and $Q(F(y;\boldsymbol{\theta});\boldsymbol{\theta}) = y$ for $L < y < U$). If $U = \infty$, then the density function is exponential on $[t_K, \infty)$; if $L = -\infty$, then the density function is exponential on $(-\infty, t_1]$.

We now define various quantities that will appear in Section 3 in the formulas for the log-likelihood, score, and Hessian. First, given $y \in (L, U)$, set

$$\varphi(y;\boldsymbol{\theta}) = \log(f(y;\boldsymbol{\theta})) = \sum_j \theta_j B_j(y) - C(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta.$$

Next, to handle censoring, given a subinterval $A$ of $(L, U)$ having positive length, set

$$\varphi(A;\boldsymbol{\theta}) = \log\left(\int_A f(y;\boldsymbol{\theta})dy\right) = \log\left(\int_A e^{\varphi(y;\boldsymbol{\theta})}dy\right), \quad \boldsymbol{\theta} \in \Theta.$$

Given a function $g$ on $\Theta$, set

$$g_j(\boldsymbol{\theta}) = \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_j} \quad \text{and} \quad g_{jk}(\boldsymbol{\theta}) = \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}, \quad \boldsymbol{\theta} \in \Theta,$$

where $1 \le j, k \le p$. Then, in particular,

$$\varphi_j(y;\boldsymbol{\theta}) = B_j(y) - C_j(\boldsymbol{\theta}) \quad \text{and} \quad \varphi_{jk}(y;\boldsymbol{\theta}) = -C_{jk}(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta.$$

Moreover, when $A$ has positive length,

$$\varphi_j(A;\boldsymbol{\theta}) = \frac{\int_A \varphi_j(y;\boldsymbol{\theta}) f(y;\boldsymbol{\theta})dy}{\int_A f(y;\boldsymbol{\theta})dy}, \quad \boldsymbol{\theta} \in \Theta,$$

and

$$\begin{aligned}\varphi_{jk}(A;\boldsymbol{\theta}) \quad = \quad & \frac{\int_A \varphi_{jk}(y;\boldsymbol{\theta}) f(y;\boldsymbol{\theta})dy}{\int_A f(y;\boldsymbol{\theta})dy} + \frac{\int_A \varphi_j(y;\boldsymbol{\theta})\varphi_k(y;\boldsymbol{\theta}) f(y;\boldsymbol{\theta})dy}{\int_A f(y;\boldsymbol{\theta})dy} \\ & - \frac{\int_A \varphi_j(y;\boldsymbol{\theta}) f(y;\boldsymbol{\theta})dy \int_A \varphi_k(y;\boldsymbol{\theta}) f(y;\boldsymbol{\theta})dy}{\left(\int_A f(y;\boldsymbol{\theta})dy\right)^2}, \quad \boldsymbol{\theta} \in \Theta. \ .\end{aligned}$$

## 3. MAXIMUM LIKELIHOOD ESTIMATION

We will now discuss the implementation of the maximum likelihood method for estimating the unknown parameters $\theta_1, \ldots, \theta_p$ of the logspline model. For conceptual simplicity, we think of the method as being applied to a random sample $Y_1, \ldots, Y_n$ of size $n$ from a distribution on $(L, U)$ having density function $f$, distribution function $F$, and quantile function $Q$.

Let $A_1, \ldots, A_n$ be subintervals of $(L, U)$ such that it is known only that $Y_i \in A_i$ for $1 \le i \le n$. If $Y_i$ is uncensored, then $A_i = \{Y_i\}$. If $Y_i$ is right censored at $T_i < Y_i$, then $A_i = (T_i, U)$. If $Y_i$ is left censored at $T_i > Y_i$, then $A_i = (L, T_i)$. In either case we refer to $T_i$ as the censoring value of $Y_i$. If $Y_i$ is interval censored, then its censoring interval $A_i$ is a subinterval of $(L, U)$. Under the usual assumption that the random sample is independent of the censoring mechanism, the log-likelihood function corresponding to the logspline model is given by $l(\boldsymbol{\theta}) = \sum_i \varphi(A_i; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. Thus $l_j(\boldsymbol{\theta}) = \sum_i \varphi_j(A_i; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$, and $l_{jk}(\boldsymbol{\theta}) = \sum_i \varphi_{jk}(A_i; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$.

The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is given as usual by $l(\widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \Theta} l(\boldsymbol{\theta})$ and the log-likelihood of the model is given by $\hat{l} = l(\widehat{\boldsymbol{\theta}})$. The corresponding maximum likelihood estimates of $f$, $F$, and $Q$ are given by $\hat{f}(y) = f(y; \widehat{\boldsymbol{\theta}})$ for $L < y < U$, $\widehat{F}(y) = F(y; \widehat{\boldsymbol{\theta}})$ for $L < y < U$, and $\widehat{Q}(p) = Q(p; \widehat{\boldsymbol{\theta}})$ for $0 < p < 1$.

Let $\mathbf{S}(\boldsymbol{\theta})$ denote the score at $\boldsymbol{\theta}$ (that is, the $p$-dimensional column vector with elements $l_j(\boldsymbol{\theta})$), and let $\mathbf{H}(\boldsymbol{\theta})$ denote the Hessian at $\boldsymbol{\theta}$ (that is, the $p \times p$ matrix with elements $l_{jk}(\boldsymbol{\theta})$). The Newton–Raphson method for computing $\widehat{\boldsymbol{\theta}}$ is to start with an initial guess $\widehat{\boldsymbol{\theta}}^{(\mathrm{o})}$ and iteratively determine $\widehat{\boldsymbol{\theta}}^{(m)}$ from the formula

$$\widehat{\boldsymbol{\theta}}^{(m+1)} = \widehat{\boldsymbol{\theta}}^{(m)} - \left[\mathbf{H}\left(\widehat{\boldsymbol{\theta}}^{(m)}\right)\right]^{-1} \mathbf{S}\left(\widehat{\boldsymbol{\theta}}^{(m)}\right).$$

If at some stage $l\left(\widehat{\boldsymbol{\theta}}^{(m+1)}\right) \le l\left(\widehat{\boldsymbol{\theta}}^{(m)}\right)$, then $\widehat{\boldsymbol{\theta}}^{(m+1)}$ should be replaced by

$$\widehat{\boldsymbol{\theta}}^{(m)} - \gamma \left[\mathbf{H}\left(\widehat{\boldsymbol{\theta}}^{(m)}\right)\right]^{-1} \mathbf{S}\left(\widehat{\boldsymbol{\theta}}^{(m)}\right)$$

for some constant $\gamma \in (0, 1)$. In our implementation we choose $\gamma = 2^{-\nu}$, where $\nu$ is the smallest nonnegative integer such that

$$l\left(\widehat{\boldsymbol{\theta}}^{(m)} - 2^{-\nu}\left[\mathbf{H}\left(\widehat{\boldsymbol{\theta}}^{(m)}\right)\right]^{-1} \mathbf{S}\left(\widehat{\boldsymbol{\theta}}^{(m)}\right)\right) > l\left(\widehat{\boldsymbol{\theta}}^{(m)}\right).$$

This procedure is referred to as the Newton–Raphson method with step-halving. We stop the iterations when

$$\frac{1}{2}\left[\mathbf{S}\left(\widehat{\boldsymbol{\theta}}^{(m)}\right)\right]^t \left[\mathbf{H}\left(\widehat{\boldsymbol{\theta}}^{(m)}\right)\right]^{-1} \mathbf{S}\left(\widehat{\boldsymbol{\theta}}^{(m)}\right) \ge -\epsilon,$$

where $\epsilon = 10^{-6}$; roughly speaking this corresponds to stopping the iterations when $l\left(\widehat{\boldsymbol{\theta}}^{(m+1)}\right) - l\left(\widehat{\boldsymbol{\theta}}^{(m)}\right) \le \epsilon$.

When there is no censoring the Hessian is globally negative definite, the log-likelihood function is strictly concave, and the maximum likelihood estimate of $\boldsymbol{\theta}$ is necessarily unique (see Stone 1990); moreover, the Newton–Raphson method with step-halving works quite well. When there is censoring, however, the Hessian need not be globally negative definite. Here we augment the Newton–Raphson method by occasional searches in the direction of steepest ascent; see the Appendix for details.

## 4. KNOT SELECTION

The knot selection methodology involves initial knot placement, stepwise knot deletion, and final model selection based on AIC or BIC. Let $n_u$ be the number of uncensored observations, $n_r$ the number of right-censored observations, $n_l$ the number of left-censored observations, and $n_i$ the number of interval-censored observations. Then

$n = n_u + n_r + n_l + n_i$. Set $n_c = n_u + n_i + .5(n_r + n_l)$, and let $N$ be the number of distinct sets among $A_1, \ldots, A_n$. As the default value for the initial number of knots $K$ we have used $K = \min(4n_c^{.2}, n_c/4, N, 25)$ when knot deletion is employed and $K = \min(2.5n_c^{.2}, n_c/4, N, 25)$ in the absence of knot deletion. (If the indicated value of $K$ is not an integer, it is rounded up to the next integer.) These choices are in reasonable agreement with conclusions reached about this issue in Kooperberg and Stone (1991). Observe that if $N \leq 25$ and $n_c$ is sufficiently large relative to $N$, then the number of parameters in the model is given by $p = K - 1 = N - 1$.

(Suppose all of the observations are interval censored and that the number $N$ of distinct intervals is not large. In principle it should be possible to fit a model with $N+1$ knots and hence $N$ free parameters. In practice, however, the numerical problem of obtaining the corresponding maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is likely to be ill-conditioned.)

We will now describe an initial knot placement rule, which is more complicated than the corresponding rule in Kooperberg and Stone (1991); the added complications are due to the need for handling censored observations. In order to place the $K$ initial knots in a reasonable manner, we first define a suitable smoothed empirical distribution on $(L, U)$. To this end, for $y \in (L, U)$, let $\#(y)$ denote the number of intervals $A_1, \ldots, A_n$ that correspond to uncensored, right-censored, or left-censored observations and that have $y$ as an endpoint, and set

$$H_1(y) = \sum_{z<y} \#(z) + \frac{1}{2} \sum_{z=y} \#(z).$$

Also, let $H_2(y)$ be the sum of

$$\frac{\text{length}(A_i \cap (L, y))}{\text{length}(A_i)}$$

over all values of $i$ that correspond to interval-censored observations, and set $H_3(y) = H_1(y) + H_2(y)$. Let $\mathcal{Y}$ denote the collection of finite endpoints of $A_1, \ldots, A_n$, and let $H_4$ be the function on $[\min(\mathcal{Y}), \max(\mathcal{Y})]$ obtained from $H_3(y)$, $y \in \mathcal{Y}$, by linear interpolation. Let $f_5$ be the function $H_4'$ normalized to have integral one, let $F_5$ be the distribution function of $f_5$, and let $Q_5 = F_5^{-1}$ be its quantile function.

When the $n$ observations are uncensored and distinct, $Q_5$ has a simple form. Let the observations be written in increasing order as $Y_{(1)}, \ldots, Y_{(n)}$. Then $\mathcal{Y} = \{Y_{(1)}, \ldots, Y_{(n)}\}$, $\min(\mathcal{Y}) = Y_{(1)}$, and $\max(\mathcal{Y}) = Y_{(n)}$. Also, $H_2(y) = 0$ for $y \in \mathcal{Y}$, so $H_3\left(Y_{(m)}\right) = m - \frac{1}{2}$ for $1 \leq m \leq n$. Thus

$$\int_{Y_{(1)}}^{Y_{(n)}} H_4'(y)dy = n - 1.$$

Consequently,

$$F_5\left(Y_{(m)}\right) = \frac{m-1}{n-1}, \quad 1 \leq m \leq n,$$

and $F_5$ is obtained for other values of $y \in \left[Y_{(1)}, Y_{(n)}\right]$ by linear interpolation. Therefore,

$$Q_5\left(\frac{m-1}{n-1}\right) = Y_{(m)}, \quad 1 \leq m \leq n,$$

and $Q_5(p)$ is obtained for other values of $p \in [0, 1]$ by linear interpolation.

The knot placement rule is governed by a sequence of numbers $r_2, \ldots, r_{K-1}$ such that $r_1 < r_2 < \cdots < r_{K-1} < r_K$, where $r_1 = 0$ and $r_K = 1$. Specifically, knots are placed initially at $\min(\mathcal{Y})$, $Q_5(r_k)$, $2 \le k \le K-1$, and $\max(\mathcal{Y})$.

Suppose first that $L = -\infty$ and $U = \infty$. The numbers $r_2, \ldots, r_{K-1}$ are then chosen to satisfy the symmetry condition $r_{K+1-k} = 1 - r_k$ for $1 \le k \le K$, which implies that

$$r_{K+1-k} - r_{K-k} = r_{k+1} - r_k, \quad 1 \le k \le K-1.$$

Motivated by the discussion in Section 5.1 of Kooperberg and Stone (1991), we require that

$$n(r_{k+1} - r_k) = 4[(4-\epsilon) \vee 1] \cdots [(4-(k-1)\epsilon) \vee 1] \tag{4.1}$$

for $1 \le k \le K/2$, where $\epsilon \in \mathbb{R}$; here $a \vee b = \max(a, b)$. The constant $\epsilon$ is determined as follows: If $K$ is an odd integer, then $r_{(K+1)/2} = 1/2$; if $K$ is an even integer, then $r_{K/2} + r_{K/2+1} = 1$.

For a numerical example of this rule, let $n = 100$. If $K = 11$, then $\epsilon \doteq 1.42$ and the numbers $r_1, \ldots, r_{10}$ are approximately as follows:

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_k$ | 0 | .04 | .14 | .26 | .38 | .50 | .62 | .74 | .86 | .96 | 1 . |

If $K = 10$, then $\epsilon \doteq 1.34$ and the numbers $r_1, \ldots, r_{10}$ are approximately as follows:

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $r_k$ | 0 | .04 | .15 | .29 | .43 | .57 | .71 | .85 | .96 | 1 . |

Suppose next that $L = -\infty$ and $U < \infty$. The numbers $r_2, \ldots, r_{K-1}$ are then chosen so that (4.1) holds for $1 \le k \le K-1$. Suppose instead that $L > -\infty$ and $U = \infty$. Then $r_2, \ldots, r_{K-1}$ are chosen so that

$$n(r_{K+1-k} - r_{K-k}) = 4[(4-\epsilon) \vee 1] \cdots [(4-(k-1)\epsilon) \vee 1], \quad 1 \le k \le K-1.$$

Suppose, finally, that $L > -\infty$ and $U < \infty$. Then $r_2, \ldots, r_{K-1}$ are chosen so that

$$r_{k+1} - r_k = r_k - r_{k-1}, \quad 2 \le k \le K-1;$$

thus

$$r_k = \frac{k-1}{K-1}, \quad 1 \le k \le K.$$

Generally, we have preferred stepwise knot deletion to using a fixed number of knots. In this procedure we place permanent knots at $\min(\mathcal{Y})$ and $\max(\mathcal{Y})$ and nonpermanent knots at $Q_5(r_k)$ for $2 \le k \le K-1$. Then we successively remove the least statistically significant among the remaining nonpermanent knots until there is only one nonpermanent knot. The statistical significance of a nonpermanent knot is measured by the absolute value of its Wald statistic $W_k = \widehat{\tau}_k/\mathrm{SE}(\widehat{\tau}_k)$. Here $\widehat{\tau}_k = \lambda_k^t \widehat{\boldsymbol{\theta}}$ is the jump of the third derivative of $\sum_j \widehat{\theta}_j B_j$ at the corresponding knot, where the $B_j'$'s are defined in terms of the remaining knots. Also,

$$\mathrm{SE}(\widehat{\tau}_k) = \sqrt{\lambda_k^t \hat{I}^{-1} \lambda_k},$$

where the estimate $\hat{I}$ of the information matrix is the negative of the Hessian matrix of the log-likelihood function at $\widehat{\boldsymbol{\theta}}$ for the model based on the remaining knots.

Using stepwise knot deletion we get a sequence of models indexed by $m \in \{0, \ldots, K-3\}$; the *m*th model has $K-1-m$ free parameters. Let $\hat{l}_m$ denote the log-likelihood of the *m*th model, and let $\mathrm{AIC}_{\alpha,m} = -2\hat{l}_m + \alpha(K-1-m)$ be the Akaike information criterion with penalty parameter $\alpha$ for this model. We choose the model corresponding to the value $\widehat{m}$ of $m$ that minimizes $\mathrm{AIC}_{\alpha,m}$. Traditionally, $\alpha = 2$. In Kooperberg and Stone (1991) we recommended choosing $\alpha = 3$ to reduce the chance of spurious modes in the density estimate. Based on our more recent experience, we now recommend choosing $\alpha = \log(n)$ as in the Bayesian information criterion (BIC) due to Schwarz (1978).

It would be worthwhile to extend the logspline methodology to allow for knot addition as in TURBO (Friedman and Silverman 1989) and MARS (Friedman 1991). This would obviate the need for the rather complicated knot placement rule described previously.

## 5. USER INTERFACE

A program for implementing logspline density estimation as it applies to possibly censored data has been written in *C* (see Appendix), and an interface based on *S* (see Becker, Chambers, and Wilks 1988, and Chambers and Hastie 1992) has also been developed. (The software is publicly available from StatLib. Send an electronic mail message with the body 'send logspline from S' to statlib@stat.cmu.edu to obtain the logspline density estimation program.) The interface has several purposes: (1) to facilitate the application of logspline density estimation to real data; (2) to facilitate the evaluation of the corresponding methodology and its comparison to kernel and other approaches to nonparametric density estimation; and (3) to explore the broader issue of the practical utility of nonparametric density estimation.

The interface consists of seven *S* functions: `dlogspline`, `plogspline`, `qlogspline`, `rlogspline`, `logspline.fit`, `logspline.summary`, and `logspline.plot`. The first four of these functions are analogous to the *S* functions `dnorm`, `pnorm`, `qnorm`, and `rnorm`, and to similar four-tuples of *S* functions for $t$ distributions, $F$ distributions, gamma distributions, and so forth. Thus `dlogspline` gives the density function corresponding to `logspline.fit`, `plogspline` gives the distribution function, `qlogspline` gives the quantile function, and `rlogspline` gives a random sample from the fitted distribution. The function `logspline.fit` performs the model fitting and model selection tasks and supplies the modest output that is used as input to `dlogspline`, `plogspline`, and so forth. This takes advantage of the feature of logspline density estimation that the estimate is determined by a moderate number of parameters, namely, $L$, $U$, $\alpha$ and the number and position of the initial knots. The function `logspline.fit` has defaults for all of these parameters.

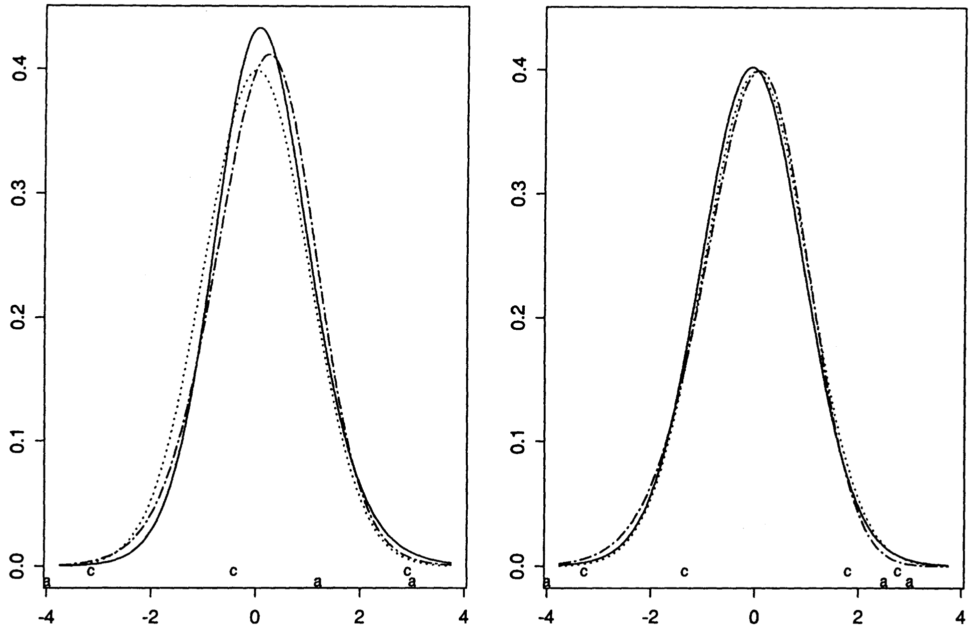

*Figure 1. Logspline Density Estimates for Interval-Censored Data From a Normal Density. —equals estimate using censored data, -— equals estimate using actual data, and ··· equals truth. Left, n = 100; right, n = 500.*

The function `logspline.summary` uses the output of `logspline.fit` to provide summary information about the fit and about the other fits that could be obtained by using alternative values of the penalty parameter. Finally, `logspline.plot` uses the output of `logspline.fit` directly to produce a plot of the density function, probability function, hazard function, or survival function.

## 6. SIMULATED EXAMPLES

An advantage of using simulated data for examples involving censoring is that we know not only the true density function from which the data was generated but also the actual values of the sample data before the censoring took place. In the figures in this section we indicate the position of the knots in the final estimate (after knot deletion): "**c**" indicates a knot for the density estimate based on the partly censored data; "**a**" indicates a knot for the estimate based upon the actual (uncensored) data.

Figure 1 contains examples involving the normal distribution. We generated a sample of size $n$ from the standard normal distribution and then grouped this data into the intervals $(-\infty, -3]$, $(-3, 2]$, $(-2, 1], \ldots, (2, 3]$, $(3, \infty)$. The logspline density estimate based on this interval-censored data (the counts in $(-3, 2], (-2, 1], \ldots, (2, 3])$, left-censored data (the count in $(-\infty, -3]$), and right-censored data (the count in $(3, \infty)$) is the solid line in Figure 1. We report here typical results for density estimates based on sample sizes of $n = 100$ and $n = 500$. The dashed line is the logspline density estimate based on the actual data. The dotted line is the true normal density.

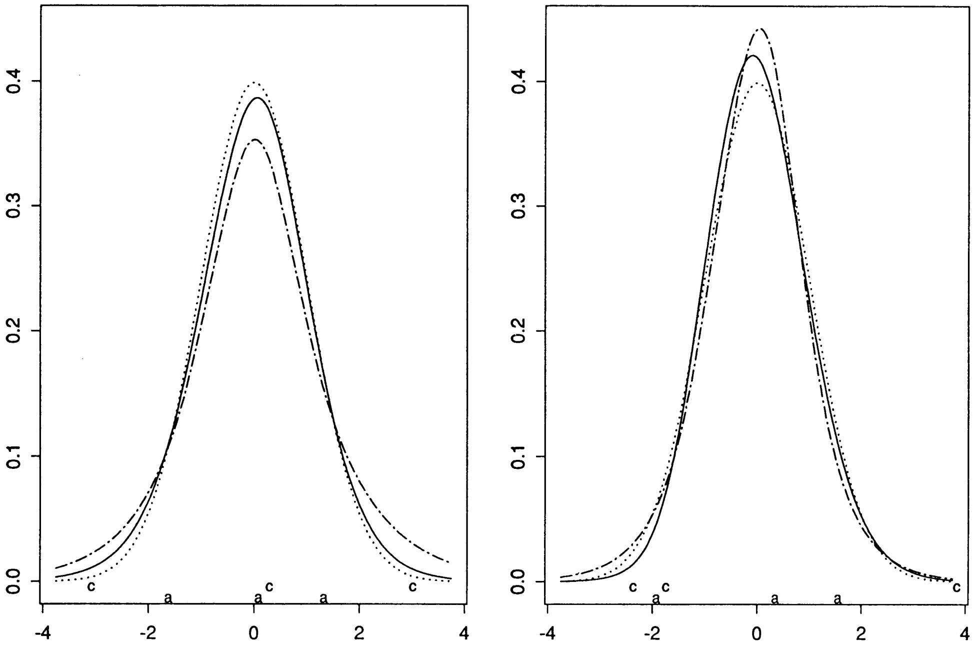

*Figure 2. Logspline Density Estimates for Normal Data, Left Censored, and Right Censored by Exponential Data. —equals estimate using partly censored data, —-— equals estimate using actual data, and · · · equals truth. Left, n = 100, 35% left censored, 33% right censored; right, n = 500, 29% left censored, 30% right censored.*

Figure 2 also contains examples involving the normal distribution. We generated a sample $Y_i$, $1 \le i \le n$, from the standard normal distribution. We also generated two independent samples, $L_i$, $1 \le i \le n$ and $R_i$, $1 \le i \le n$, from the exponential distribution with mean $\frac{2}{3}$. The solid line in Figure 2 is the density estimate based on $A_i$, $1 \le i \le n$, where

$$\begin{aligned} A_i &= (-\infty, -L_i) \text{ if } Y_i < -L_i \\ &= \{Y_i\} \text{ if } -L_i \le Y_i \le R_i \\ &= (R_i, \infty) \text{ if } Y_i > R_i. \end{aligned}$$

The dashed line is the logspline density estimate based on the actual data and the dotted line is the true normal density.

From Figures 1 and 2 we observe that the logspline procedure with censoring yields very good results. For each sample size the density estimate based on the partly censored data is just as good as the estimate based on the actual data. When the underlying density is smooth and unimodal the number of knots after knot deletion is typically 3 to 5, independent of the sample size (and thus of the initial number of knots).

Figures 3 and 4 contain logspline density estimates for a more traditional censoring scheme. We generated a sample $Y_i$, $1 \le i \le n$, from the gamma distribution with shape parameter 5 and scale parameter 1 and an independent sample $C_i$, $1 \le i \le n$, from the

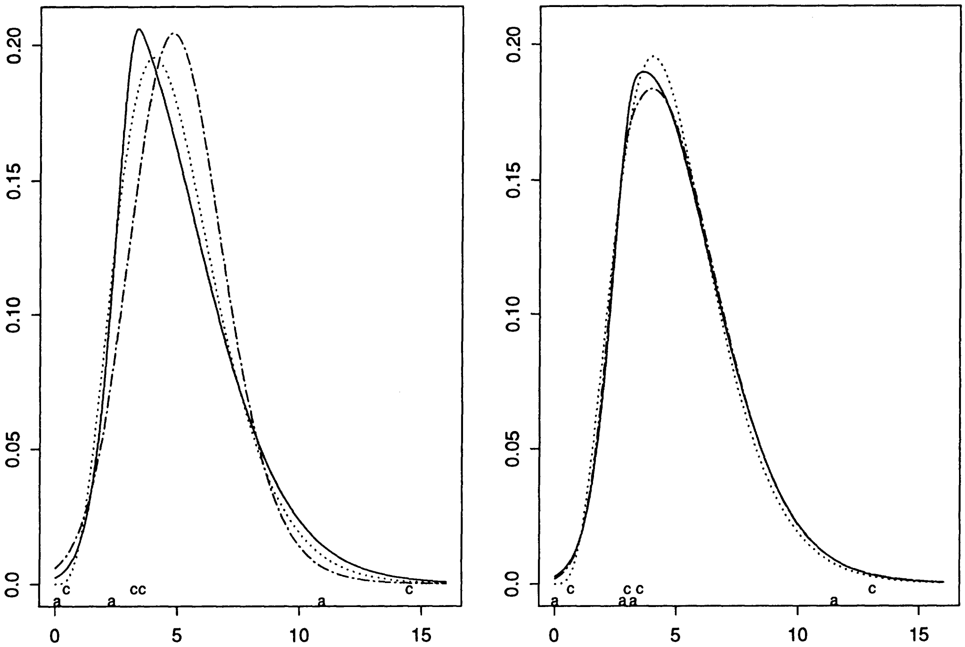

*Figure 3. Logspline Density Estimates for Gamma Data, Right Censored by Exponential Data. —equals estimate using partly censored data, —– equals estimate using actual data, and · · · equals truth. Left, n = 100, 58% right censored; right, n = 500, 54% right censored.*

exponential distribution with mean 6. (The mean was chosen so as to yield about 50% censoring.) The solid line in Figure 3 is the density estimate based on $A_i$, $1 \leq i \leq n$, where

$$\begin{aligned} A_i &= \{Y_i\} \text{ if } Y_i \leq C_i \\ &= (C_i, \infty) \text{ if } Y_i > C_i, \end{aligned}$$

and $L = 0$ and $U = \infty$.

The interpretation here is that $Y_i$ is the time of an event of interest for object $i$ and that we observe $Y_i$ unless it is greater than the censoring time $C_i$, in which case we know only that the event happened after the time $C_i$. The dashed line is the logspline density estimate based on the actual data and the dotted line is the true gamma density.

The solid line in Figure 4 is the density estimate based on $A_i$, $1 \leq i \leq n$, where

$$\begin{aligned} A_i &= (Y_i, \infty) \text{ if } Y_i \leq C_i \\ &= \{C_i\} \text{ if } Y_i > C_i, \end{aligned}$$

based on the same data as in Figure 3. The dashed line is the logspline density estimate based on $A_i = \{C_i\}$ for all $i$, and the dotted line is the true exponential density.

From Figures 3 and 4 we observe that it is possible in practice to recover both the underlying density of interest and the density of the censoring times using logspline density estimation from the information that is available in studies with right-censored data.

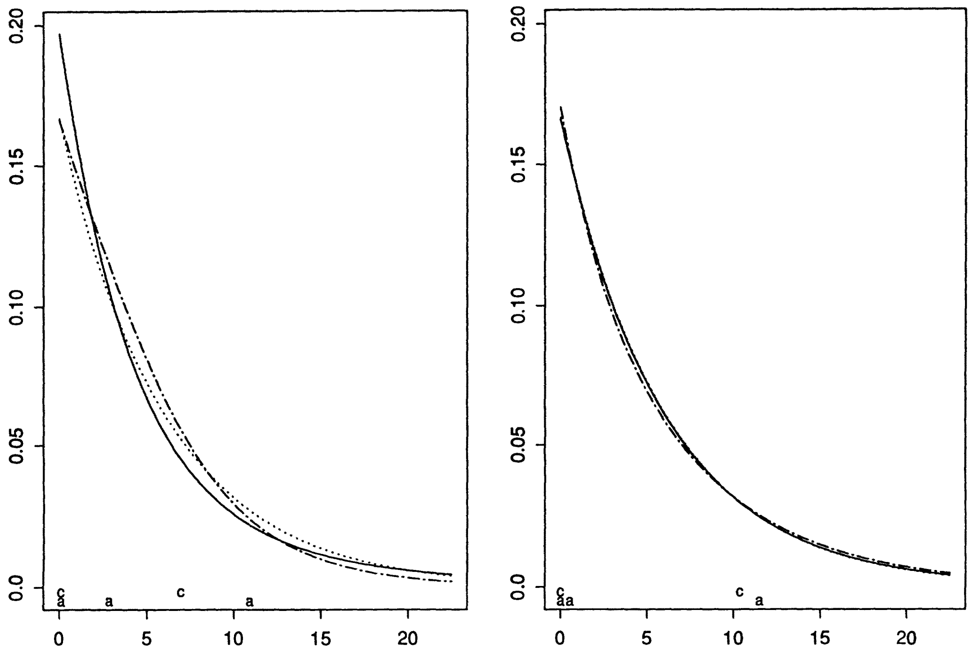

*Figure 4. Logspline Density Estimates for Exponential Data, Right Censored by Gamma Data. —equals estimate using partly censored data, —— equals estimate using actual data, and · · · equals truth. Left, n = 100, 42% right censored; right, n = 500, 46% right censored.*

The censoring scheme for Figure 5 is the same as the one for Figure 3. Here the $Y_i$'s were generated from a bimodal density $f$ that was used in Kooperberg and Stone (1991): $f = .8g + .2h$, where $g$ is the (lognormal) density of $\exp(Z/2)$, with $Z$ having the standard normal distribution, and $h$ is the normal density with mean 2 and standard deviation .17. The $C_i$'s were generated from the exponential distribution with mean 2.5. The solid line in Figure 5 is the density estimate based on $A_i$, $1 \le i \le n$, where

$$\begin{aligned} A_i &= \{Y_i\} \text{ if } Y_i \le C_i \\ &= (C_i, \infty) \text{ if } Y_i > C_i, \end{aligned}$$

and $L = 0$ and $U = \infty$. The dashed line is the logspline density estimate based on $A_i = \{Y_i\}$ for all $i$, $L = 0$ and $U = \infty$; the dotted line is the true bimodal density.

Even for this bimodal density, logspline density estimation does a decent job. For the smaller sample size, however, the estimate for the height of the second mode is not very accurate. This appears to be caused primarily by the sampling variation. (The number of data points close to the second mode, ignoring censoring, is binomial with parameters $n = 100$ and $p = .2$.) Actually, about one fifth of the time that we simulated this example, the number of data points near the second mode was so small that the density estimate missed the second mode completely. Although the percentage of censoring typically is less than 40%, about 55% of the cases in the range of the second mode get censored. Nevertheless, when the sample size is large this censoring has almost no influence on the fit.

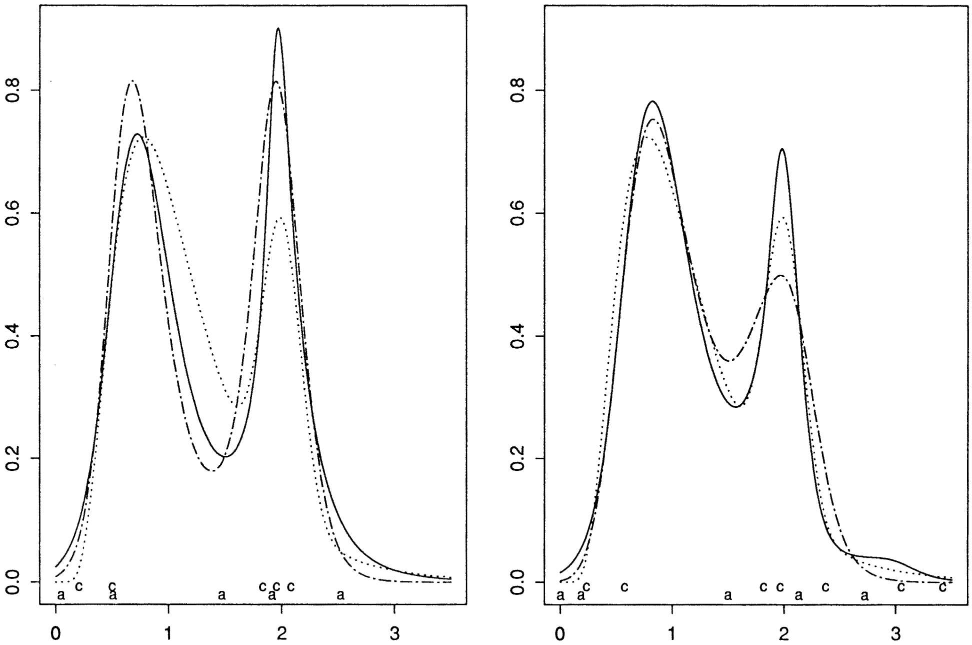

*Figure 5. Logspline Density Estimates for Bimodal Data, Right Censored by Exponential Data. — equals estimate using partly censored data, —— equals estimate using actual data, and · · · equals truth. Left, n = 100, 32% right censored; right, n = 500, 44% right censored.*

In Kooperberg and Stone (1991) many more examples of bimodal (uncensored) logspline density estimates are discussed. Based on the present version of the procedure, the estimates are at least as good as they were before.

For the left side of Figure 6 we generated a sample $Y_i$, $1 \leq i \leq 100$, from an exponential distribution with mean 1 and an independent sample $C_i$, $1 \leq i \leq 100$, from a uniform distribution on [0,1.5]. (The maximum of 1.5 was chosen so as to yield about 50% censoring.) The solid line in the left side of Figure 6 is the logspline estimate $\widehat{S}(y) = 1 - \widehat{F}(y)$ of the survival function based on $A_i$, $1 \leq i \leq 100$, where

$$\begin{aligned} A_i &= \{Y_i\} \text{ if } Y_i \leq C_i \\ &= (C_i, \infty) \text{ if } Y_i > C_i, \end{aligned}$$

The dashed line is the true survival function.

Because of the distribution of the censoring times in this example, all observations for which $Y_i > 1.5$ are censored. Logspline density estimates will have an exponential tail beyond the last knot. When the underlying density is (almost) exponential, as in the present example, this will be a useful approximation. (As usual, the user should be wary about extrapolation beyond the range of the data. In particular when all observations beyond a certain point are censored, as in Type I or Type II censoring, the reliability of conclusions about the right tail of the density may be severely limited.) To investigate the accuracy of logspline estimation of the survival function in the context of this example, we obtained 100 simulations of the data and recorded the estimated probability of survival

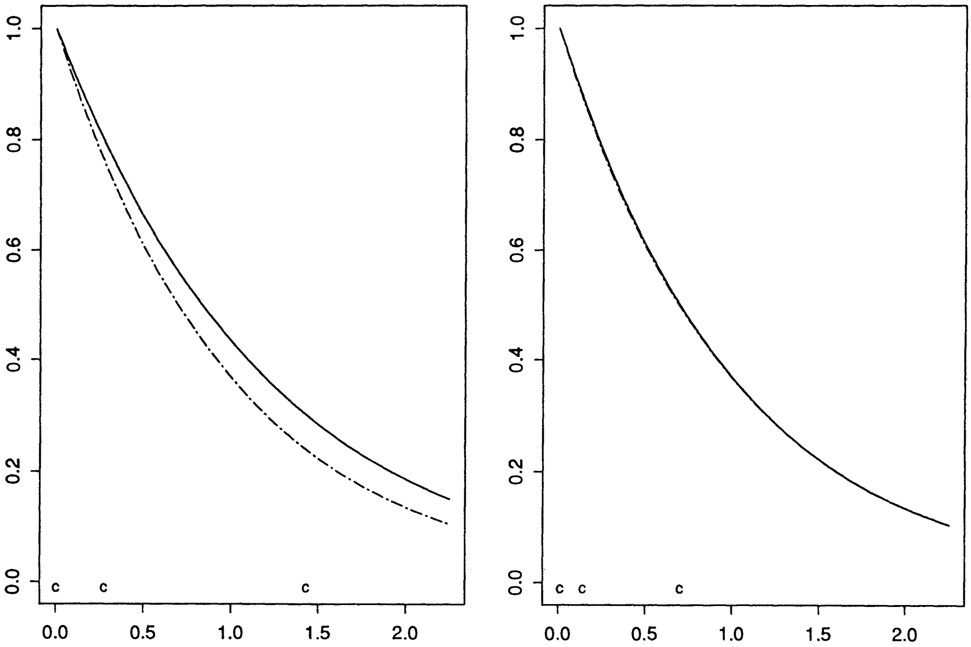

*Figure 6. Logspline Survival Estimates for Exponential Data. — equals estimate and —·— equals truth. Left, n = 100, 49% right censored; right, quantile sample, censored at .7.*

at 1.5 ("end of the data") and 2.0 ("extrapolation"). The averages and standard deviations of these 100 estimates are listed in the following table.

| | *Estimate* | | |
|---|---|---|---|
| *Probability* | *Average* | *Standard Deviation* | *Truth* |
| $1 - F(1.5)$ | .209 | .070 | .223 |
| $1 - F(2.0)$ | .127 | .065 | .135 |

It can be seen from this table that the estimates are biased slightly downward. (We also made density estimates using the 100 estimates of the probabilities of survival as the sample, which looked like normal density functions with the indicated means and standard deviations.)

We carried out identical calculations for deterministic censoring at .7 (i.e., $C_i = .7$ for $1 \le i \le 100$.) The results that we obtained were almost identical to those reported in the preceding table and in the left side of Figure 6.

To get some feeling for the portion of the error that was due to sampling variation (as opposed to bias in the estimate), we also "estimated" the survival function for a "deterministic sample." For the right side of Figure 6 we put $Y_i = -\log\left(1 - \frac{i}{101}\right)$, $1 \le i \le 100$ and $C_i = .7$, $1 \le i \le 100$. The solid line in the right side of Figure 6 is the logspline estimate of the survival function based on $A_i$, $1 \le i \le 100$, where

$$\begin{aligned} A_i &= \{Y_i\} \text{ if } Y_i \le C_i \\ &= (C_i, \infty) \text{ if } Y_i > C_i. \end{aligned}$$

The dashed line is the true, exponential, survival function. As can be seen from the right

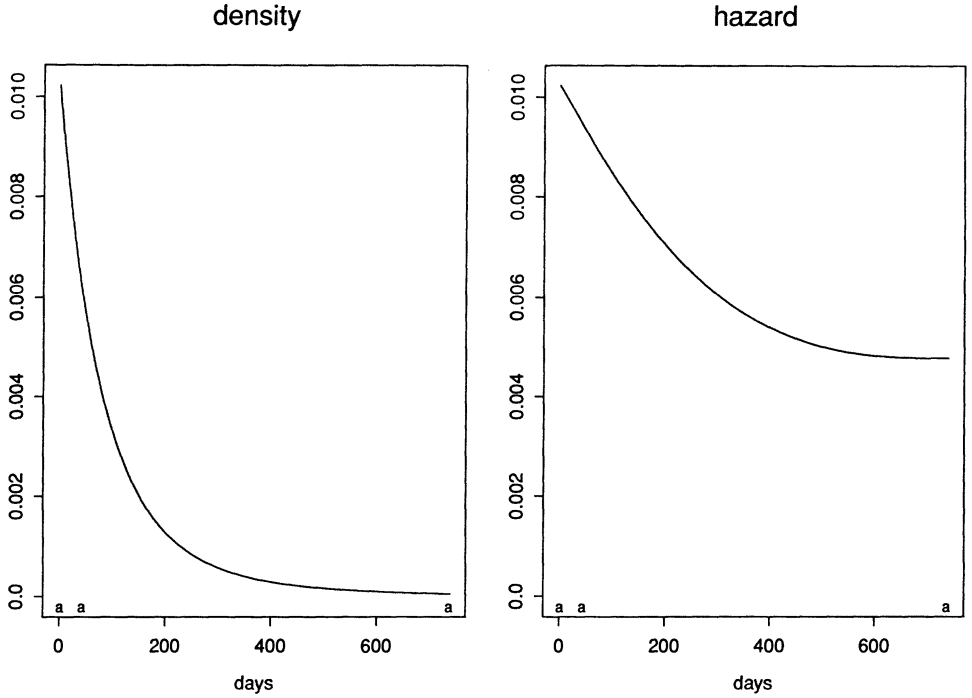

*Figure 7. Logspline Estimates for Suicide Data.*

side of Figure 6, the logspline method appears almost unbiased for this censoring scheme. We have found these "deterministic samples" to be very useful in getting a feeling about the behavior of logspline density estimation.

Figures 1–6 do not include any examples with Type I or Type II censoring. The simulations we have carried out suggest that the behavior of logspline density estimation with such censoring schemes is similar to that for the example shown in the left side of Figure 6.

## 7. REAL EXAMPLES

The data for Figure 7, which is labeled suicide, consist of 86 lengths of psychiatric treatment spells undergone by patients used as controls in a study of suicide risks reported by Copas and Freyer (1980). The data are used extensively in Silverman (1986), and they are also used in Wand, Marron, and Ruppert (1991) and Kooperberg and Stone (1991). The logspline estimate in the left side of Figure 7 was made using $L = 0$ and $U = \infty$. (In Kooperberg and Stone [1991], the data was first transformed to $(-\infty, \infty)$, after which a logspline density estimate with $L = -\infty$ was obtained.)

Because kernel density estimation handles finite boundaries in an awkward manner, there has been considerable discussion about the legitimacy of a peak around $y = 50$ in the kernel density estimate based on the suicide data. We are comfortable with the monotonic decreasing behavior of our estimate.

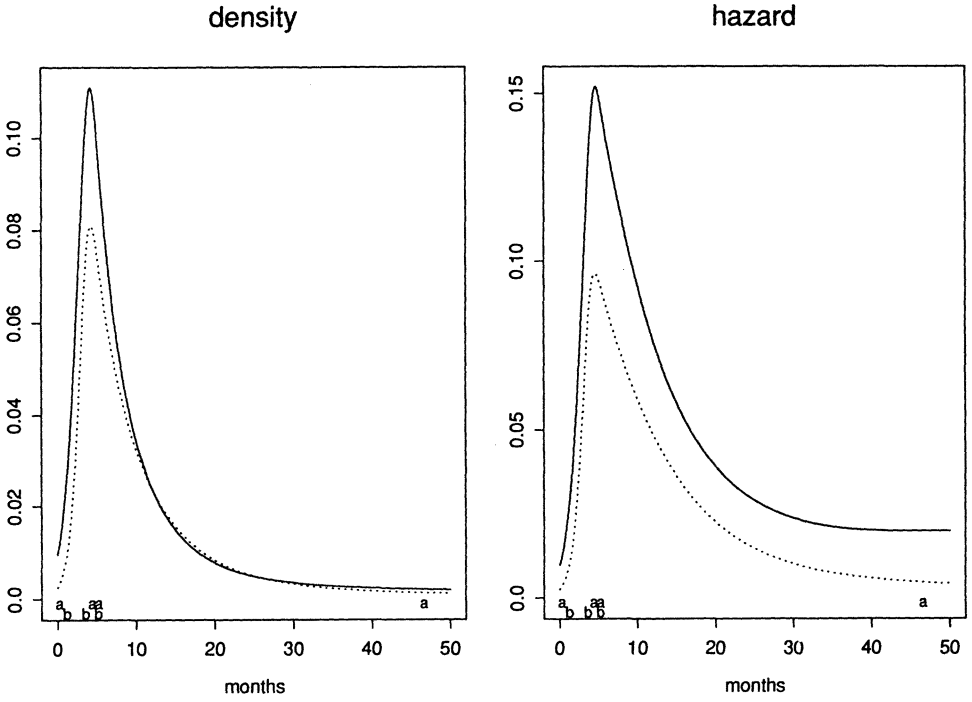

*Figure 8a. Logspline Estimates for Efron Data. —equals group A ($n = 51, 9$ censored) and $\cdots$ equals group B (n = 45, 14 censored).*

In the right side of Figure 7 we show the logspline hazard estimate $\hat{h}$ defined by

$$\hat{h}(y) = \frac{\hat{f}(y)}{1 - \widehat{F}(y)},$$

where $\hat{f}$ is the logspline density estimate and $\widehat{F}$ is the corresponding estimate of the distribution function. Because the logspline density estimate has very smooth tails, the estimate for the hazard rate is smooth too. It appears that kernel density estimation, unless transformations like those in Wand, Marron, and Ruppert (1991) are made, will yield density estimates that are too wiggly in the tails to provide smooth estimates of the hazard function. In this connection the reader may wish to compare our estimate with the one in Figure 6.5 in Silverman (1986, p. 150).

The data for Figure 8 comes from Efron (1988). He reports data on survival times for patients in a study of head-and-neck cancer. There were two treatments, A and B. In Group A there were 42 uncensored and 9 censored observations; in Group B there were 31 uncensored and 14 censored observations. For both groups we report logspline density estimate (the left side of Figure 8a) and the logspline hazard estimate (the right side of Figure 8a). The symbols "**a**" and "**b**" indicate the location of the knots (the right-most knots fall outside the plot). The hazard rates look very much like the ones in Efron (1988).

A possible use of `rlogspline` is to resample from a fitted logspline density. As in Figures 3 and 4 we estimated not only $f_A$ and $f_B$ but also the densities $g_A$ and $g_B$ of the

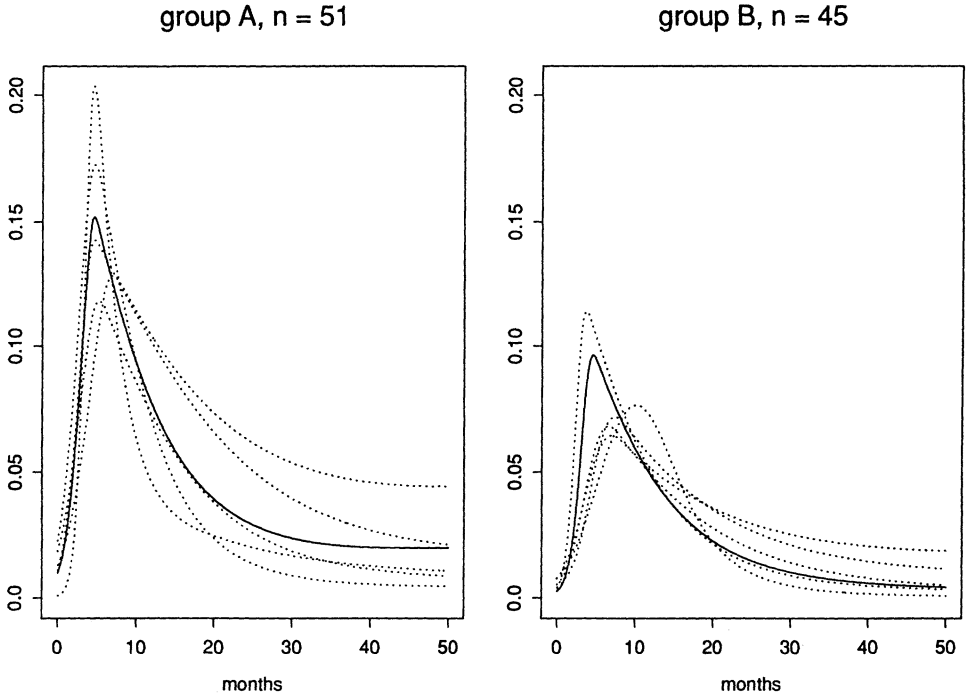

*Figure 8b. Logspline Hazard Estimates for the Efron Data (—) and Five Resampled Estimates (· · ·).*

censoring times. For both groups we then generated five samples of size 51 (for group A) or 45 (for group B) from $\hat{f}_A$, $\hat{g}_A$, $\hat{f}_B$, and $\hat{g}_B$. We then applied the same sampling scheme as in Figure 3, treating the samples from $\hat{f}_A$ and $\hat{f}_B$ as the actual times and the samples from $\hat{g}_A$ and $\hat{g}_B$ as the censoring times. The left side of Figure 8b shows $\hat{f}_A$ together with the five resampled logspline density estimates for group A. The right side of Figure 8b shows the same information for group B. Each of these figures gives an indication of the variability among logspline estimates based on different samples from a density like the one being studied. It remains to investigate the use of such figures in assessing the accuracy of logspline estimates.

Figure 9 involves the Stanford heart transplant data as taken from Kalbfleisch and Prentice (1980). There are 103 observations of which 75 are exact (deaths) and 28 are censored (survivals). In Figure 9a, left, we report the logspline density estimate (with $L = 0$ and $U = \infty$) and the logspline hazard estimate (Figure 9a, right). Because the density and hazard rate in this example are relatively high near the origin, we found it more useful to examine these estimates on a log-scale. Note that we have ignored the covariates in this set of data. It would be worthwhile to extend the logspline methodology to handle such covariates. In Figure 9b the logspline estimate $\widehat{S}(y) = 1 - \widehat{F}(y)$ of the survival function (solid line) based on the heart transplant data and the traditional product-limit (Kaplan–Meier) estimate (dotted line) are shown.

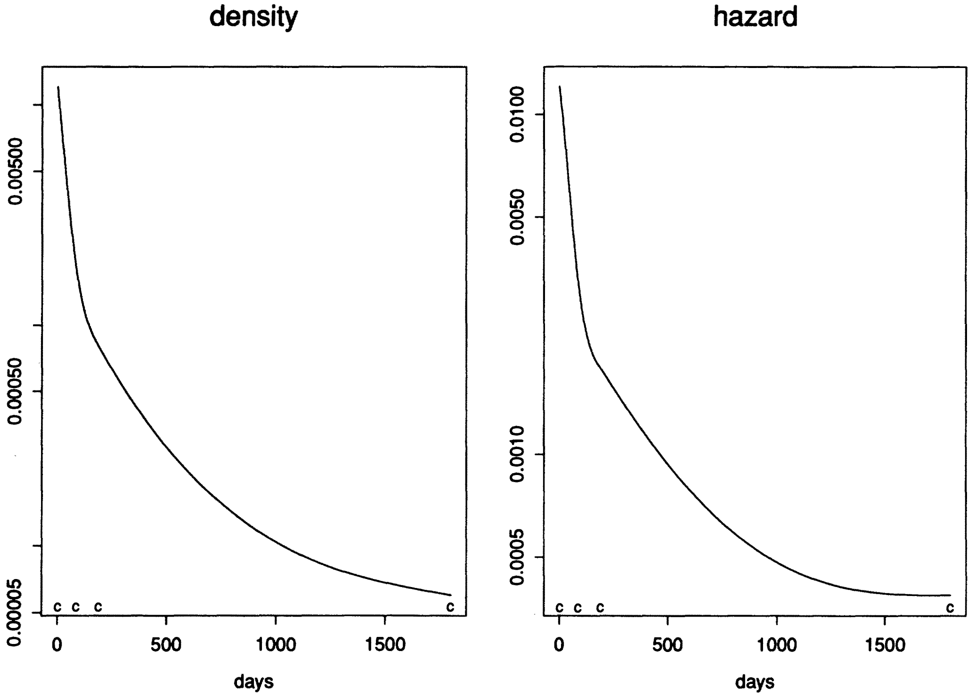

*Figure 9a. Logspline Estimates for the Stanford Heart Transplant Data. n = 103, 28 censored.*

The data for Figure 10, which is labeled income, consist of a random sample of size 7,125 annual net incomes in the United Kingdom (Family Expenditure Survey 1968–1983). (The data have been rescaled to have mean one as in Wand, Marron, and Ruppert, 1991.) The logspline density estimate in the left side of Figure 10a is essentially the same as in Kooperberg and Stone (1991). The sharpness of the peak near .24 is remarkable.

For a check on the height of our estimate of the peak we selected those incomes that were between .19 and .27. The dotted line in the right side of Figure 10a is the logspline density estimate based on the 452 cases (with $L = .19$ and $U = .27$), rescaled to integrate to $\frac{452}{7125}$; the solid line is part of the logspline density estimate based on all the data. It is clear from this figure that the peak is "real." Although position of the peak has shifted a bit, the height of the peak is essentially unchanged.

Alternatively, we can use `plogspline` to get that $\widehat{F}(.27) - \widehat{F}(.19) \approx .063 \approx \frac{449}{7125}$. Thus the estimated probability in this interval almost coincides with the proportion of cases between .19 and .27. As a further check, in Figure 10b we plotted $\widehat{F}\left(Y_{(i)}\right)$ against $\frac{i}{n+1}$ on the logit scale. Interpretation of the figure shows that the fit is excellent, even far out in the tails.

Wand, Marron, and Ruppert (1991) compute a fixed-width kernel density estimate after an initial transformation of the data. While their estimate has a smooth tail, the height of the first peak in their estimate is considerably smaller than that of the logspline estimate.

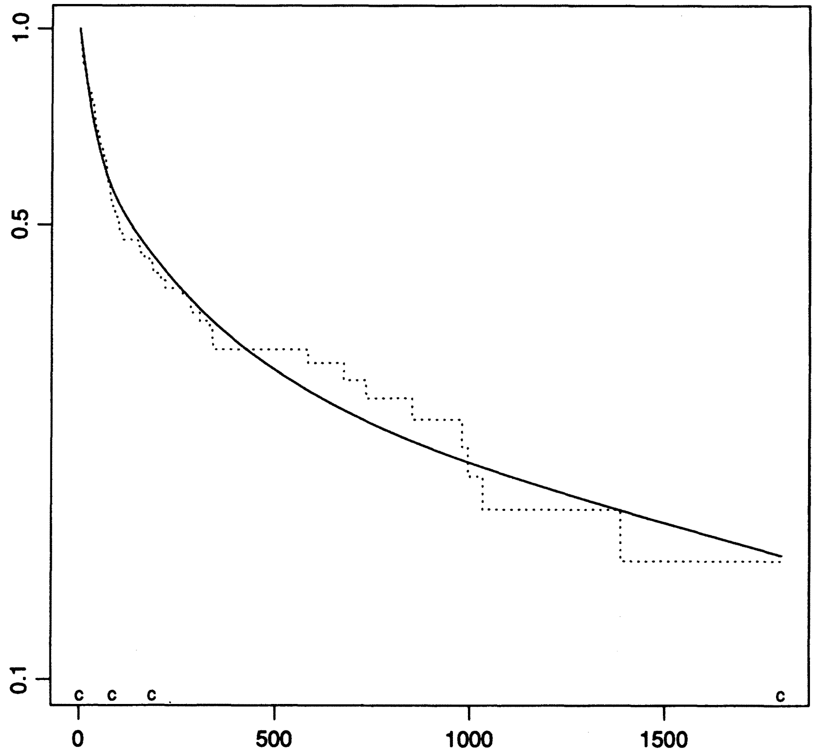

*Figure 9b. Logspline and Product-Limit Estimates of the Survival Curve for the Heart Transplant Data.*

How much of the income data does the logspline density estimate need to detect the sharp spike? To investigate this question we repeatedly sampled 500 cases from the complete set of data and obtained the density estimates based on the sample of size 500. In Figure 10c we show the part of the logspline density estimate that contains both modes for five of these samples, together with the estimate based on the complete set of data. The five density estimates are typical of the many more we have examined; for samples of size 500 the sharp peak is nearly always detected.

Generally, if the true density is unimodal, then the logspline density estimate is insensitive to the position of the initial knots, provided that they are reasonably spread out. When applied to random samples of size 500 from the income data, the estimate needs two initial knots near the spike, but is otherwise insensitive to the initial knot placement. If there is only one initial knot near the spike, then the estimate may substantially understate the height of the spike or even fail to detect it. Fortunately, the knot placement rule described in Section 4 almost always places two initial knots near the spike.

To study the effect of the penalty term $\alpha$, we focused on the sample of the income data that resulted in the logspline density estimate indicated with two bullets in Figure 10c. The function `logspline.summary` for this fit gave the following results:

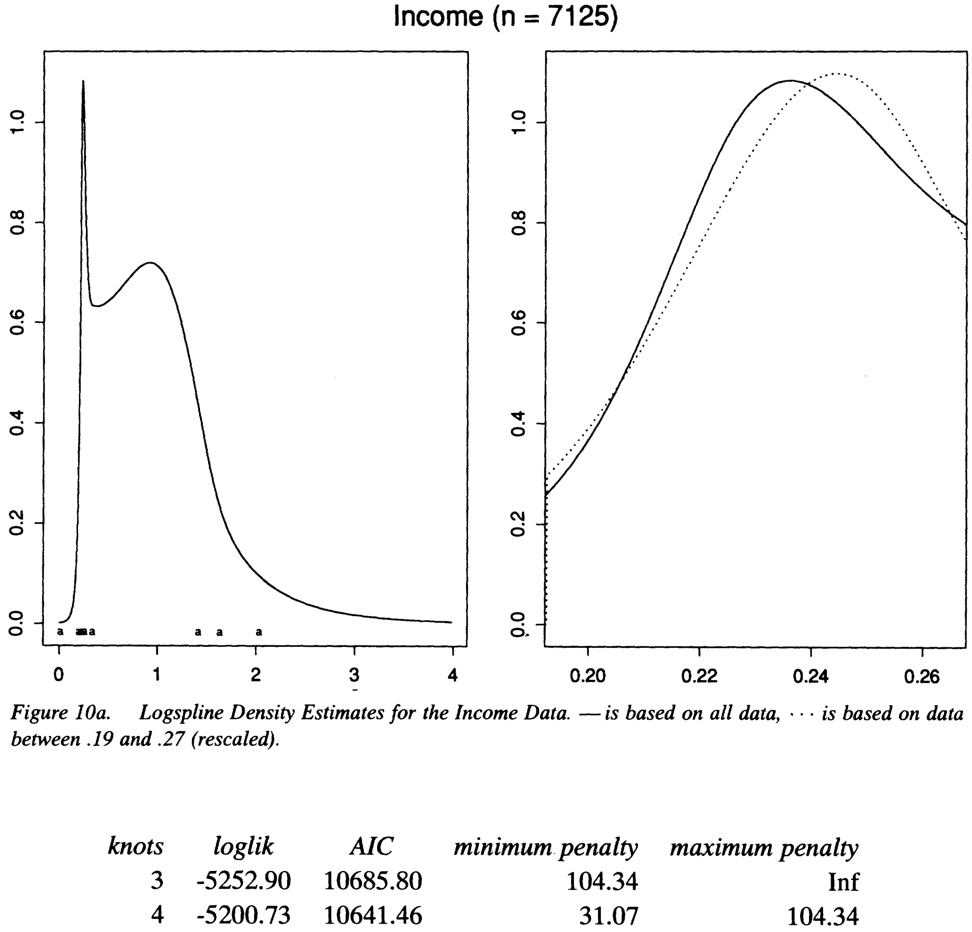

*Figure 10a. Logspline Density Estimates for the Income Data. — is based on all data, · · · is based on data between .19 and .27 (rescaled).*

```
knots    loglik      AIC    minimum penalty   maximum penalty
    3  -5252.90  10685.80            104.34               Inf
    4  -5200.73  10641.46             31.07            104.34
    5  -5197.78  10695.55                NA                NA
    6  -5169.66  10699.33              2.93             31.07
    7  -5168.20  10756.39              0.79              2.93
    8  -5167.80  10815.60              0.45              0.79
    9  -5167.58  10875.15              0.40              0.45
   10  -5167.38  10934.75              0.06              0.40
   11  -5167.36  10994.72                NA                NA
   12  -5167.33  11054.66                NA                NA
   13  -5167.28  11114.56              0.00              0.06
   14  -5167.28  11174.56              0.00              0.00
The present optimal number of knots is 6.
Penalty was the default: log(samplesize)=log(500)=6.21.
```

According to the table, if we had chosen the penalty $\alpha$ to be any number between 2.93 and 31.07, we would have obtained the same fit, while if we had chosen $\alpha$ to be between .79 and 2.93, we would have obtained a fit with 7 knots. Note also that for no value of $\alpha$ would we have obtained a fit with 5 knots. Figure 10d shows the logspline density estimate with 6 knots based on the default penalty; it is shown here as a solid

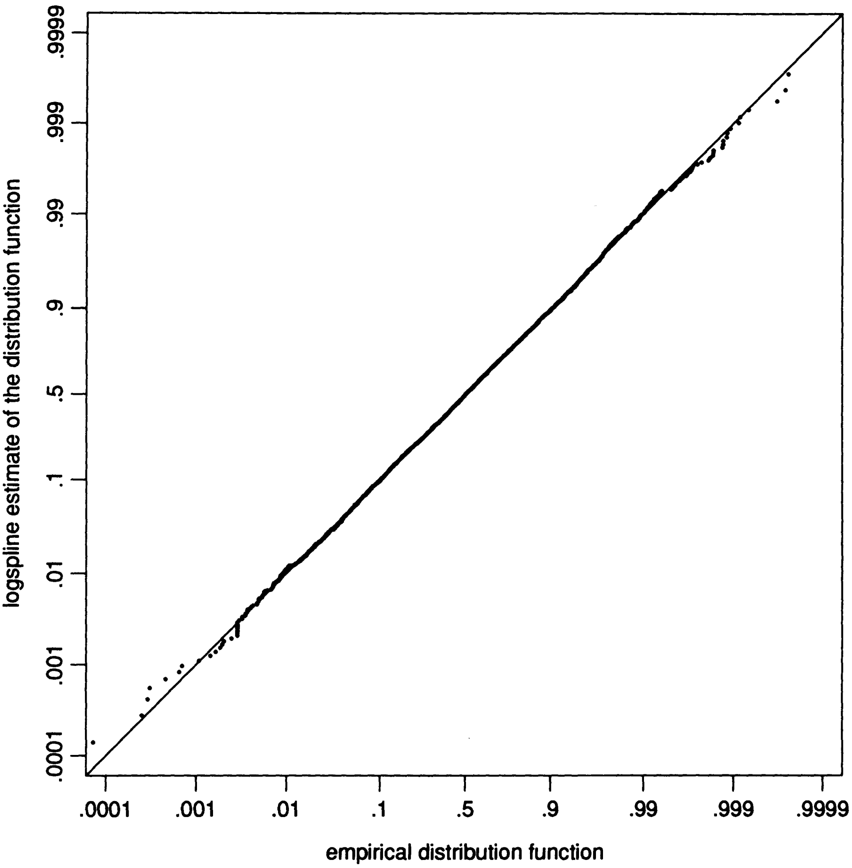

*Figure 10b. P-P plot for the Income Data.*

curve and was shown with bullets in Figure 10c. Also shown in Figure 10d is the fit with 7 knots (dashed curve) obtained by choosing $\alpha = 2$ and the fit with 4 knots (dashed curve) obtained by choosing $\alpha = 40$. The fit with 7 knots is qualitatively similar to the default, but the fit with 4 knots is nearly constant between the two modes based on all data.

As the previous plots and summary table illustrate, it is rather typical in logspline density estimation for the estimate based on the default penalty to be quite reasonable and to correspond to a fairly wide range of values of $\alpha$. If we make $\alpha$ too large, however, we get too much bias and may miss some genuine modes; if we make it too small, we get too much variance and possibly one or more spurious modes.

## 8. CONCLUDING REMARKS

In light of the examples in Sections 6 and 7 and much additional experience with logspline density estimation and its user interface, we are convinced that the current im-

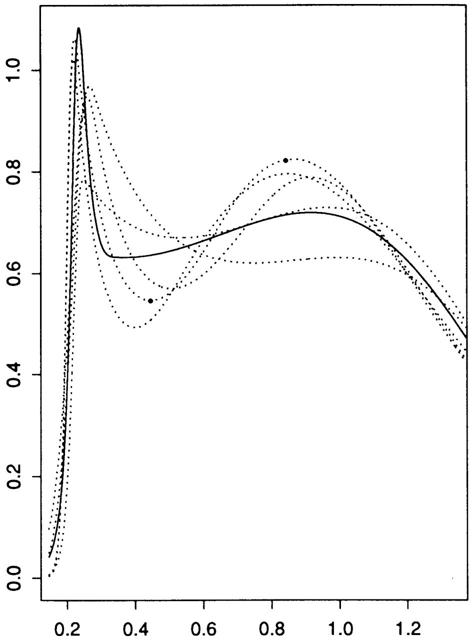

*Figure 10c. Logspline Density Estimates for Samples of the Income Data. — is based on all data and · · · is based on samples of size 500.*

plementation is of considerable practical value in data analysis. It is sufficiently accurate and flexible to handle sharp peaks in the middle of the data (see Figures 5 and 10), and it also works well far out in the tails (see Figure 10b) without using preliminary transformations as described in Kooperberg and Stone (1991). Moreover, a moderate amount of censoring has very little effect on the accuracy of estimation and the procedure can deal effectively with a high proportion of censoring. The estimation procedure is rather insensitive to the initial knot placement and choice of penalty parameter.

In Section 7 we have compared logspline fits to the suicide and income data with published results for kernel fits to these data sets. Based in part on such comparisons, it is our tentative belief that, for kernel density estimation to be as effective as logspline density estimation in handling both regular tails and sharp spikes, it would need to combine variable kernel width as in Silverman (1986) and transformations as in Wand, Marron, and Ruppert (1991). It would certainly be worthwhile to have publicly available user-friendly software that implements kernel density estimation with these two refinements.

It is noteworthy and perhaps a little surprising that, with proper algorithmic development and programming, maximum likelihood can be tailored to logspline density estimation so as to work well in the context of right-, left-, and interval-censored data, where the log-likelihood function can fail to be concave. This suggests that the logspline methodology could be modified to handle deconvolution in a reasonable manner.

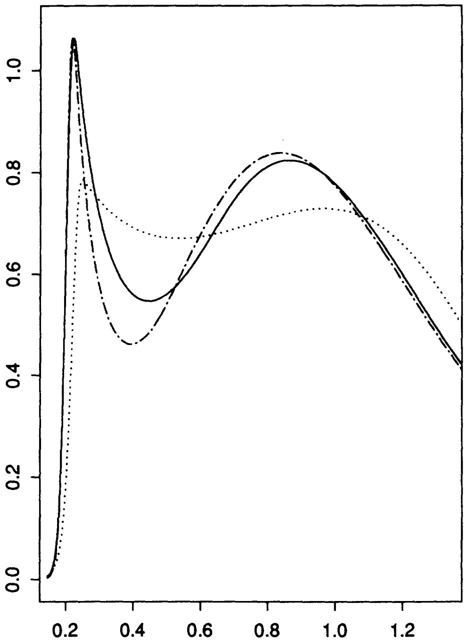

*Figure 10d. Effect of Penalty Choice on Logspline Density Estimates based on a sample of size 500. —equals default, · · · equals larger penalty, and —— equals smaller penalty.*

# APPENDIX: NUMERICAL IMPLEMENTATION

The program for logspline density estimation is written in *C*, with an interface to *S*. In the following sections we describe some of the more interesting features of the program.

## A.1 Starting Values

Newton–Raphson approximations to a maximum converge extremely fast to the maximum when the starting values are sufficiently good, but the method can behave poorly when the starting values are not very good. The following idea for the starting values is due to Jin (1992).

Set

$$f'(y) = \frac{df(y)}{dy},$$

$$B_j'(y) = \frac{dB_j(y)}{dy},$$

and

$$B_j''(y) = \frac{d^2B_j(y)}{dy^2}.$$

The basic idea is to minimize the $L_2$ distance between the score function, $d\log(f(y))$ $dy$, and its logspline approximation relative to the density; that is,

$$\widehat{\boldsymbol{\theta}}^{(\mathrm{o})} = \arg\min_{\boldsymbol{\theta}} \int_L^U \left( \sum_j \theta_j B_j'(y) - \frac{f'(y)}{f(y)} \right)^2 f(y)dy.$$

Setting the derivatives with respect to $\theta_k$ for $1 \le k \le p$ equal to 0, we get that

$$\int_L^U \left( \sum_j \widehat{\theta}_j^{(0)} B_j'(y) - \frac{f'(y)}{f(y)} \right) B_k'(y)f(y)dy = 0, \quad 1 \le l \le p.$$

This can be written as

$$\int_L^U \sum_j \widehat{\theta}_j^{(0)} B_j'(y)B_k'(y)f(y)dy = \int_L^U B_k'(y)f'(y)dy, \quad 1 \le k \le p.$$

Applying integration by parts to the right side of this equation, we get that

$$\begin{aligned} &\int_L^U \sum_j \widehat{\theta}_j^{(0)} B_j'(y)B_k'(y)f(y)dy \\ = &\ -\int_L^U B_k''(y)f(y)dy - B_k'(L)f(L) + B_k'(U)f(U) \\ = &\ -\int_L^U B_k''(y)f(y)dy, \quad 1 \le k \le p, \end{aligned}$$

where we have assumed that $f(L) = f(U) = 0$.

Because $f$ is unknown, the integrals are replaced by sums over the observations. Thus solving

$$\mathbf{A}\widehat{\boldsymbol{\theta}}^{(\mathrm{o})} = D,$$

where $\mathbf{A}$ is a $p \times p$ matrix with elements

$$a_{jk} = \sum_{i=1}^n B_j'(Y_i)B_k'(Y_i)$$

and $D$ is a column vector of length $p$ with elements

$$d_j = -\sum_{i=1}^n B_j''(Y_i),$$

we get starting values $\widehat{\boldsymbol{\theta}}^{(\mathrm{o})}$. Because the $B_j$'s are (almost) B-splines (de Boor 1978), $\mathbf{A}$ is a band-matrix with seven diagonals.

The previous algorithm is used to compute the starting values if there is no censoring. If an observation is left censored or right censored, we replace it by its censoring value; if the observation is interval censored, we replace it by the midpoint of its censoring interval.

## A.2 COMPUTATION OF NORMALIZING CONSTANT, LOG-LIKELIHOOD FUNCTION, SCORE FUNCTION AND HESSIAN.

The main numerical task of the algorithm is the computation of the normalizing constant $C(\boldsymbol{\theta})$, the log-likelihood function $l(\boldsymbol{\theta})$, the score function $\mathbf{S}(\boldsymbol{\theta})$, and the Hessian $\mathbf{H}(\boldsymbol{\theta})$ for various values of $\boldsymbol{\theta}$. In the absence of censoring, the computation of $l(\boldsymbol{\theta})$ is trivial and the computation of $C(\boldsymbol{\theta})$, $\mathbf{S}(\boldsymbol{\theta})$, and $\mathbf{H}(\boldsymbol{\theta})$ amounts to the computation of

$$\begin{aligned} a_{ij} &= \int_{t_i}^{t_{i+1}} y^j f(y;\boldsymbol{\theta})dy \\ &= \int_{t_i}^{t_{i+1}} y^j e^{b_0+b_1x+b_2x^2+b_3x^3} dy, \quad 1 \leq i < K \text{ and } 0 \leq j \leq 6, \\ a_{0j} &= \int_{L}^{t_1} y^j f(y;\boldsymbol{\theta})dy = \int_{L}^{t_1} y^j e^{b_0+b_1y} dy, \quad 0 \leq j \leq 2, \end{aligned}$$

and

$$a_{Kj} = \int_{t_K}^{U} y^j f(y;\boldsymbol{\theta})dy = \int_{t_K}^{U} y^j e^{b_0+b_1y} dy, \quad 0 \leq j \leq 2.$$

It is possible to compute $a_{0j}$ and $a_{Kj}$ analytically. We compute the other $a_{ij}$'s using Gaussian quadrature (Abramowitz and Stegun 1970), with 6 points between knots, unless some preliminary diagnostics suggest using a larger number of points.

If the $i$th observation is censored, it is also necessary to compute

$$b_{ij} = \int_{A_i} x^j f(y;\boldsymbol{\theta})dy, \quad 0 \leq j \leq 6.$$

Potentially, this increases the number of numerical integrations from the order of magnitude of $7K$ to $7(K+n)$ per iteration. Therefore, in the presence of censoring, we first compute the maximum likelihood estimator for an approximate problem:

$$\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_i \varphi(B_i;\boldsymbol{\theta}),$$

where $B_i$ is defined by

$$\begin{aligned} B_i &= A_i \text{ if } A_i = \{Y_i\}, \\ &= \{M_i\} \text{ if } A_i \text{ is a bounded interval of } (L,U), \\ &= (L,t_j) \text{ if } A_i = (L,T_i), \text{ and} \\ &= (t_j,U)\text{if}A_i = (T_i,U). \end{aligned}$$

Here $M_i$ is the midpoint of the bounded interval and $t_j$ is the $j$th knot, determined by the condition that $|T_i - t_j| \leq |T_i - t_k|$ for all $j \neq k$.

No extra integrals, other than the $a_{ij}$'s, have to be computed to determine $\widetilde{\boldsymbol{\theta}}$. Starting at $\widetilde{\boldsymbol{\theta}}$ the algorithm typically converges in at most three steps; this limits the number of times we have to compute $7n$ integrals.

## A.3 Alternating Between Newton–Raphson and Steepest Ascent

In the absence of censored data the Hessian of the log-likelihood function at $\boldsymbol{\theta}$ is $n$ times the Hessian of $C(\boldsymbol{\theta})$, so the log-likelihood function is strictly concave. Therefore, if a maximum of the log-likelihood function exists, it is unique. If some of the observations are censored, however, the log-likelihood function need not be concave. As O'Sullivan (1988) pointed out, logspline hazard estimation, in which the log-hazard function is modeled by a spline, does result in a concave log-likelihood function if all the observations are either uncensored or right censored. If some of the observations are left censored or interval censored, the log-likelihood function is not necessarily concave for this estimation problem either. For this reason, and because logspline hazard estimation is not invariant under multiplication by a negative number, we did not pursue logspline hazard estimation in this project.

If the log-likelihood function is not concave we are not guaranteed that there is a unique maximum. A modified version of the Newton–Raphson algorithm, which alternates between Newton–Raphson approximations and steepest ascent searches, invariably seems to converge to the global maximum in our context. After computing $\mathbf{S}(\boldsymbol{\theta})$ and $\mathbf{H}(\boldsymbol{\theta})$, as described in the previous section, we use a Linpack routine (Dongarra, Bunch, Moler, and Stewart 1979) to determine whether or not the eigenvalues of $\mathbf{H}(\boldsymbol{\theta})$ are all positive. If they are all positive $l(\boldsymbol{\theta})$ is locally concave and we proceed with a Newton–Raphson approximation. Otherwise we carry out a steepest ascent search in the direction of $\mathbf{S}(\boldsymbol{\theta})$. To increase the speed of the procedure, we do one more steepest ascent search in the direction of (the new) $\mathbf{S}(\boldsymbol{\theta})$ and conclude the steepest ascent cycle with a steepest ascent search in the direction of $\widehat{\boldsymbol{\theta}}^{(m)} - \widehat{\boldsymbol{\theta}}^{(m-2)}$.

After two steepest ascent cycles we do one Newton–Raphson approximation to avoid the known slow convergence of the steepest ascent algorithm. Our experience is that, after at most a few steepest ascent searches, the log-likelihood function is locally concave and a maximum is found by Newton–Raphson approximations.

This algorithm will converge to a local maximum of the log-likelihood function. When that function is not concave, there is no guarantee that such a local maximum is unique or that it is a global maximum. However, we have not found a (nonpathological) example where $\widehat{\boldsymbol{\theta}}$, as found by the algorithm, does not correspond to a reasonable estimate of the unknown density function.

## A.4 Management of L and U

For a vector $\widehat{\boldsymbol{\theta}} \in I\!R^p$ to be feasible, it is necessary that (1) $\widehat{\theta}_1 < 0$ or $L > -\infty$ and (2) $\widehat{\theta}_p < 0$ or $U < \infty$. If $L = -\infty$ or $U = \infty$, however, the Newton–Raphson algorithm does not prevent $\widehat{\boldsymbol{\theta}}^{(m)} - [\mathbf{H}(\widehat{\boldsymbol{\theta}}^{(m)})]^{-1}\mathbf{S}(\widehat{\boldsymbol{\theta}}^{(m)})$ from yielding intermediate estimates for $\widehat{\boldsymbol{\theta}}$ that are not feasible. (An almost identical situation arises during a linear search in a steepest ascent direction.) We found the obvious solution, to choose $\alpha$ in $\widehat{\boldsymbol{\theta}}^{(m+1)} = \widehat{\boldsymbol{\theta}}^{(m)} - \alpha[\mathbf{H}(\widehat{\boldsymbol{\theta}}^{(m)})]^{-1}\mathbf{S}(\widehat{\boldsymbol{\theta}}^{(m)})$ such that $\widehat{\theta}_1^{(m+1)} < 0$ and $\widehat{\theta}_p^{(m+1)} < 0$, not practical because it leads to extremely slow convergence. Instead, we have used the following algorithm when $L = -\infty$ and $U = \infty$ (when $L = -\infty$ and $U < \infty$ or when $L > -\infty$ and $U = \infty$ we proceed similarly):

Table 1. Approximate CPU Time (Seconds) for Simulated Examples

| *Distribution* | *n = 100* | *n = 500* |
|---|---|---|
| Examples with no censoring<br>(the actual data in Figures 1–6) | 1.5 | 4.0 |
| Figure 1—interval censored data. | 1.5 | 5.0 |
| Partly left and right censored<br>(the partly censored data in Figures 2–6) | 5.5 | 29.5 |

1. Set $L_{\text{tmp}} = 2t_1 - t_2$ and $U_{\text{tmp}} = 2t_K - t_{K-1}$.
2. If $\widehat{\theta}_1^{(0)} < 0$ and $\widehat{\theta}_p^{(0)} < 0$ go to step 5; else go to step 3.
3. Instead of using

$$C(\boldsymbol{\theta}) = \log\left(\int_{I\!R} \exp\left(\theta_1 B_1(y) + \cdots + \theta_p B_p(y)\right) dy\right),$$

use

$$\tilde{c}(\boldsymbol{\theta}) = \log\left(\int_{L_{\text{tmp}}}^{U_{\text{tmp}}} \exp\left(\theta_1 B_1(y) + \cdots + \theta_p B_p(y)\right) dy\right).$$

Compute $\widehat{\boldsymbol{\theta}}$ using the algorithm as described.

4. If $\widehat{\theta}_1 < 0$ and $\widehat{\theta}_p < 0$ after the previous step,
   (a) then carry out Step 5, using $\widehat{\boldsymbol{\theta}}$ of Step 3 as the starting values for Step 5.
   (b) else, if either $\widehat{\theta}_1 \geq 0$ or $\widehat{\theta}_p \geq 0$ set $L_{\text{new}} = 2L_{\text{tmp}} - t_1$ and $U_{\text{new}} = 2U_{\text{tmp}} - t_K$. Go back to Step 3, while using $L_{\text{new}}$ and $U_{\text{new}}$ instead of $L_{\text{tmp}}$ and $U_{\text{tmp}}$.
5. Compute $\widehat{\boldsymbol{\theta}}$, integrating from $-\infty$ to $\infty$. If at some intermediate stage $\widehat{\theta}_1^{(m)} \geq 0$ or $\widehat{\theta}_p^{(m)} \geq 0$, go back to Step 4b, thereby using for $L_{\text{tmp}}$ and $U_{\text{tmp}}$ the values last used during Steps 1 or 3. If the algorithm converges without this happening, we have obtained $\widehat{\boldsymbol{\theta}}$.

Experience has shown that one does not have to use Step 3 very often, but if one does have to use it, one pass of Step 3 is typically sufficient and only for very heavy tailed distributions are more than two passes required. It almost never happens that one has to go back from Step 5 to Step 3. For extremely heavy tailed densities, such as the Cauchy density, $\widehat{\theta}_1$ or $\widehat{\theta}_p$ is sometimes numerically indistinguishable from 0. In this case even if $L = -\infty$ or $U = \infty$ the algorithm increases $L_{\text{tmp}}$ and $U_{\text{tmp}}$ only until maximum values $L_{max}$ and $U_{max}$ ($t_1 - L_{max} = U_{max} - t_K = 25\times$(range of the data)) are reached. In this situation the algorithm will not fit a model with $K$ or fewer knots.

## A.5 CPU TIME

Logspline density estimation, as currently implemented, is computer intensive, especially in the presence of censored data or if $n$ is very large. In Tables 1 and 2 we report CPU time (in seconds) on a SPARCstation 2 for computing the logspline density estimates that were shown in Figures 1–10.

Table 2. CPU Time (Seconds) for Real Examples

| | |
|---|---|
| Suicide data (Figure 7) | 1.5 |
| Efron—Group A (Figure 8) | 2.5 |
| Efron—Group B (Figure 8) | 2.0 |
| Stanford heart data (Figure 9) | 5.5 |
| Income data (Figure 10) | 73.0 |

It should be noted that for the interval censored data in Figure 1, the program makes use of the fact that $A_i$ takes on only a few distinct values and hence that the number of distinct integrals that have to be computed is very limited.

## ACKNOWLEDGMENTS

The authors wish to thank the associate editor and referees for a number of helpful suggestions that improved the organization and readability of the article. This research was supported in part by National Science Foundation Grant DMS-9100723. The calculations for Figure 10 were made in collaboration with the Wirtschaftstheoretische Abteilung II, University of Bonn, Bonn, West Germany.

## REFERENCES

Abramowitz, M., and Stegun, I. A. (1964), *Handbook of Mathematical Functions*, Washington, DC: National Bureau of Standards.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S Language*, Pacific Grove, CA: Wadsworth.

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag.

Chambers, J. M., and Hastie, T. J. (1992), *Statistical Models in S*, Pacific Grove, CA: Wadsworth.

Copas, J. B., and Freyer, M. J. (1980), "Density Estimation and Suicide Risk in Psychiatric Treatment," *Journal of the Royal Statistical Society*, Ser. A, 143, 167–176.

Dongarra, J. J., Bunch, J. R., Moler, C. B., and Stewart, G. W. (1979), *Linpack User's Guide*, Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1988), "Logistic Regression, Survival Analysis, and the Kaplan–Meier Curve," *Journal of the American Statistical Association*, 83, 414–425.

Family Expenditure Survey (1968–1983), *Annual Base Tapes and Reports (1968–1983),* London: Department of Employment, Statistics Division, Her Majesty's Stationary Office. (The data used in this paper were made available by the ESRC Data Archive at the University of Essex.)

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines" (with discussion), *The Annals of Statistics,* 19, 1–141.

Friedman, J. H., and Silverman, B. W. (1989), "Flexible Parsimonious Smoothing and Additive Modeling" (with discussion), *Technometrics*, 31, 3–39.

Jin, K. (1992), "Empirical Smoothing Parameter Selection in Adaptive Estimation," *The Annals of Statistics,* 20, 1844–1874.

Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley.

Kooperberg, C., and Stone, C. J. (1991), "A Study of Logspline Density Estimation," *Computational Statistics and Data Analysis*, 12, 327–347.

Marron, J. S., and Padgett, W. J. (1987), "Asymptotically Optimal Bandwidth Selection for Kernel Density Estimators From Randomly Right-Censored Samples," *The Annals of Statistics*, 15, 1520–1535.

O'Sullivan, F. (1988), "Fast Computation of Fully Automated Log-Density and Log-Hazard Estimators," *SIAM Journal of Scientific and Statistical Computing*, 9, 363–379.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Smith, P. L. (1982), "Curve Fitting and Modeling With Splines Using Statistical Variable Selection Methods," Report 166034, NASA, Langley Research Center, Hampton, VA.

Stone, C. J. (1990), "Large-Sample Inference for Log-Spline Models," *The Annals of Statistics*, 18, 717–741.

Stone, C. J., and Koo, C.-Y. (1986), "Logspline Density Estimation," in *AMS Contemporary Mathematics*, Series 29, Providence, RI: American Mathematical Society, pp. 1–15.

Wand, M. P., Marron, S. J., and Ruppert D. (1991), "Transformations in Density Estimation" (with discussion), *Journal of the American Statistical Association*, 86, 343–361.