Hazard Estimation with Flexible Tails

Charles Kooperberg*
University of Washington

Charles J. Stone†
University of California at Berkeley

Technical Report No. 388
April 23, 1993
Department of Statistics
University of California
Berkeley, California 94720

Abstract

Polynomial splines are used to estimate the log-hazard function based on possibly censored, positive data. Two additional log terms are incorporated into the fitted model for the log-hazard function to allow for greater flexibility in the extreme tails. A fully automatic procedure involving the maximum likelihood method, stepwise knot addition, stepwise knot deletion and BIC is used to select the final model. A user interface based on S is described for obtaining estimates of the hazard function, density function, distribution function and quantile function and for generating a random sample from the estimated distribution.

KEY WORDS: Maximum likelihood; Model selection; Polynomial splines.

*This research was supported in part by a grant from the Graduate School Fund of the University of Washington.

†This research was supported in part by National Science Foundation Grant DMS-9204247

1 Introduction

Consider positive data that may be (right-) censored. We think of the original, uncensored data as having arisen as a random sample from a distribution having an unknown positive density function on $(0, \infty)$. In this investigation, polynomial splines are used to estimate the logarithm of the unknown hazard function. In order to allow for greater flexibility in the extreme tails, two additional log terms are incorporated into the fitted model for the log-hazard function. The maximum likelihood method is used to estimate the unknown parameters of the model. We describe a fully automatic method for selecting the final model, which involves stepwise knot addition, stepwise knot deletion and BIC, and we also describe a user interface that makes the entire procedure conveniently available within the S environment (see Becker, Chambers and Wilks, 1988). In order to evaluate the procedure in its present form, we apply it to a number of simulated and real data sets. Finally, the technical issues involved in the numerical implementation of the procedure are discussed.

The discussion section in Abrahamowicz, Ciampi and Ramsay (1992) contains a good review of many of the papers on the use of splines to estimate density functions or hazard functions in the presence of censored data. These papers typically fall into two groups: those using smoothing splines or similar procedures, including Anderson and Senthilselvan (1980), Whittemore and Keller (1986), Senthilselvan (1987) and O'Sullivan (1988); those using polynomial splines, including Etezadi-Amoli and Campi (1987), Abrahamowicz, Ciampi and Ramsay (1992) and Kooperberg and Stone (1992). Among these papers, O'Sullivan (1988) is the only one that directly models the log-hazard function. Gu (1991) contains an asymptotic analysis of the hazard estimate in O'Sullivan (1988) that is different from the analysis of Cox and O'Sullivan (1990). Kooperberg and Stone (1992) model the log-density function. Most of the other papers model either the density function or the hazard function itself.

There are other methods for estimating the hazard function or the density function in the presence of right-censored data. In particular, Tanner and Wong (1983, 1984) and Marron and Padgett (1987) use kernel estimation, and Efron (1988) uses logistic regression to estimate the hazard function. In survival analysis, the classical approach to hazard estimation has typically been to use a parametric model for the hazard function (see, for example, Cox and Oakes (1984) and Miller (1981)). For further discussion, see Section IV.2 and the corresponding bibliographic remarks in Andersen, Borgan, Gill, and Keiding (1993).

2 Flexible Linear Models for the Log-Hazard Function

Let f be a positive density function on $(0, \infty)$, let $\varphi = \log f$ denote the corresponding log-density function, let F denote the distribution function, let $h = f/(1-F)$ denote the hazard

function, let $\lambda = \log h$ denote the log-hazard function, and let $Q = F^{-1}$ denote the quantile function (so that $Q(F(t)) = t$ for $t > 0$ and $F(Q(p)) = p$ for $0 < p < 1$). Then

$$1 - F(t) = \exp\left(-\int_0^T h(u)du\right) = \exp\left(-\int_0^T \exp(\lambda(u))du\right), \qquad t \geq 0.$$

Since $F(t) < 1$ for $0 < t < \infty$ and $\lim_{t\to\infty} F(t) = 1$, we conclude that $\int_0^T \exp(\lambda(u))du < \infty$ for $0 < t < \infty$, and $\int_0^\infty \exp(\lambda(t))dt = \infty$. Observe that $h = \exp\lambda$ and that $f(t) = \exp(\lambda(t))\exp(-\int_0^T \exp(\lambda(u))du)$ and $\varphi(t) = \lambda(t) - \int_0^T \exp(\lambda(u))du$ for $t > 0$.

Given the integer $K \geq 3$ and the sequence $t_1, \ldots, t_K$ with $0 < t_1 < \cdots < t_K < \infty$, let G_0 be the $(K-2)$-dimensional space of twice-continuously differentiable functions s on $[0, \infty)$ such that s is constant on $[0, t_1]$ and on $[t_K, \infty)$ and the restriction of s to each of the intervals $[t_1, t_2], \ldots, [t_{K-1}, t_K]$ is a cubic polynomial. The functions in G_0 are cubic splines having (simple) knots at $t_1, \ldots, t_K$. Let $B_1, \ldots, B_{K-2}$ be a basis of this space such that $B_{K-2} = 1$ on $[0, \infty)$ and $B_1, \ldots, B_{K-3}$ equal zero on $[t_K, \infty)$. (If $K = 3$, then G_0 is the space of constant functions on $[0, \infty)$ and $B_1 = 1$ on $[0, \infty)$.)

Given a positive number c (which will be defined in Section 3 in terms of the observed data in a simple manner), set

$$B_{-1}(t) = \log\frac{t}{t+c} \quad \text{and} \quad B_0(t) = \log(t+c) \qquad t > 0.$$

Also, set $p = K - 2$. Then $B_{-1}, B_0, B_1, \ldots, B_p$ is a basis of the linear space spanned by $G_0 \cup \{B_{-1}, B_0\}$.

Set

$$\lambda(\cdot;\boldsymbol{\theta}) = \theta_{-1}B_{-1} + \theta_0 B_0 + \theta_1 B_1 + \cdots + \theta_p B_p, \qquad \boldsymbol{\theta} = (\theta_{-1}, \theta_0, \theta_1, \ldots, \theta_p)^T \in \mathbb{R}^{p+2},$$

and

$$\begin{aligned}\Theta &= \left\{\boldsymbol{\theta} \in \mathbb{R}^{p+2} : \int_0^T \exp(\lambda(u;\boldsymbol{\theta}))du < \infty \text{ for } 0 < t < \infty \text{ and } \int_0^\infty \exp(\lambda(t;\boldsymbol{\theta}))dt = \infty\right\} \\ &= \left\{\boldsymbol{\theta} = (\theta_{-1}, \theta_0, \theta_1, \ldots, \theta_p)^T \in \mathbb{R}^{p+2} : \theta_{-1} > -1 \text{ and } \theta_0 \geq -1\right\}.\end{aligned}$$

We use $\lambda(\cdot;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, to model the log-hazard function. Given $\boldsymbol{\theta} \in \Theta$, the corresponding hazard function, density function and log-density function are given by $h(\cdot;\boldsymbol{\theta}) = \exp(\lambda(\cdot;\boldsymbol{\theta}))$,

$$f(t;\boldsymbol{\theta}) = \exp(\lambda(t;\boldsymbol{\theta}))\exp\left(-\int_0^T \exp(\lambda(u;\boldsymbol{\theta}))du\right), \qquad t > 0,$$

and

$$\varphi(t;\boldsymbol{\theta}) = \log f(t;\boldsymbol{\theta}) = \lambda(t;\boldsymbol{\theta}) - \int_0^T \exp(\lambda(u;\boldsymbol{\theta}))du, \qquad t > 0.$$

Observe that $f(\cdot;\boldsymbol{\theta})$ is a positive density function on $(0,\infty)$. The corresponding distribution function and quantile function are given, respectively, by $F(t;\boldsymbol{\theta}) = \int_0^T f(u;\boldsymbol{\theta})du$ for $t \geq 0$ and $Q(\cdot;\boldsymbol{\theta}) = F^{-1}(\cdot;\boldsymbol{\theta})$. Set $\varphi(\cdot,1;\boldsymbol{\theta}) = \varphi(\cdot;\boldsymbol{\theta})$ and $\varphi(\cdot,0;\boldsymbol{\theta}) = 1 - F(\cdot;\boldsymbol{\theta})$. Then

$$\varphi(y,\delta;\boldsymbol{\theta}) = \delta\lambda(y;\boldsymbol{\theta}) - \int_0^y \exp(\lambda(u;\boldsymbol{\theta}))du, \qquad y > 0 \text{ and } \delta \in \{0,1\},$$

$$\frac{\partial}{\partial\boldsymbol{\theta}_j}\varphi(y,\delta;\boldsymbol{\theta}) = \delta B_j(y) - \int_0^y B_j(u)\exp(\lambda(u;\boldsymbol{\theta}))du, \qquad -1 \leq j \leq p,\ \ y > 0 \text{ and } \delta \in \{0,1\},$$

and

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}\varphi_{jk}(y,\delta;\boldsymbol{\theta}) = -\int_0^y B_j(u)B_k(u)\exp(\lambda(u;\boldsymbol{\theta}))du,\ -1 \leq j,k \leq p,\ y > 0 \text{ and } \delta \in \{0,1\}.$$

It follows from the last result that $\varphi(y,\delta;\cdot)$ is a concave function on Θ for $y > 0$ and $\delta \in \{0,1\}$.

The two log terms in the model for the log-hazard function are easily motivated. Consider a positive density function f on $(0,\infty)$, and let F, h and λ denote, respectively, the associated distribution function, hazard function and log-hazard function. Suppose first that $f(t) \approx at^\gamma$ for $t \approx 0$, where $a > 0$ and $\gamma > -1$. Then $\log f(t) \approx \gamma \log t$ for $t \approx 0$. Since $1 - F(t) \approx 1$ for $t \approx 0$, we conclude that $\lambda(t) \approx \gamma \log t$ for $t \approx 0$. This motivates the inclusion of the term $\theta_{-1}B_{-1}(t)$ with $\theta_{-1} > -1$ in the model for the log-hazard function.

Suppose next that $f(t) \approx a\exp(-bt^\gamma)$ for $t \gg 1$, where $a > 0$, $b > 0$ and $\gamma > 0$. Then

$$1 - F(t) \approx \frac{a}{b\gamma t^{\gamma-1}}\exp(-bt^\gamma), \qquad t \gg 1,$$

so

$$h(t) \approx b\gamma t^{\gamma-1}, \qquad t \gg 1,$$

and hence $\lambda(t) \approx (\gamma - 1)\log t$ for $t \gg 1$. This motivates the inclusion of the term $\theta_0 B_0(t)$ with $\theta_0 > -1$ in the model for the log-hazard function.

Suppose instead that $f(t) \approx at^{-b-1}$ for $t \gg 1$, where $a, b > 0$. Then $1 - F(t) \approx ab^{-1}t^{-b}$ for $t \gg 1$, so $h(t) \approx bt^{-1}$ for $t \gg 1$ and hence $\lambda(t) \approx (-1)\log t$ for $t \gg 1$. This motivates allowing the possibility that $\theta_0 = -1$ in the model for the log-hazard function.

Suppose now that $K = 3$. Then $p = 1$ and $B_1 = 1$, so

$$\lambda(t;\boldsymbol{\theta}) = \theta_{-1}\log\frac{t}{t+c} + \theta_0\log(t+c) + \theta_1, \qquad t > 0.$$

This three-parameter model includes Weibull and Pareto distributions as special cases.

Consider first the Weibull density function f given by

$$f(t) = b\gamma t^{\gamma-1}\exp(-bt^\gamma), \qquad t > 0,$$

where $b > 0$ and $\gamma > 0$, whose distribution function is given by

$$F(t) = 1 - \exp(-bt^{\gamma}), \qquad t > 0. \tag{1}$$

The corresponding log-hazard function is given by $\lambda(t) = (\gamma - 1)\log t + \log b\gamma$ for $t > 0$. Thus $\lambda(\cdot) = \lambda(\cdot; \boldsymbol{\theta})$, where $\theta_{-1} = \theta_0 = \gamma - 1$ and $\theta_1 = \log b\gamma$. (Alternatively, we can get the Weibull model by setting $c = 0$, $\theta_{-1} = 0$, $\theta_0 = \gamma - 1$ and $\theta_1 = \log b\gamma$.)

Consider next the Pareto density function f given by

$$f(t) = \frac{bc^b}{(t+c)^{b+1}}, \qquad t > 0,$$

where $b > 0$ and $c > 0$, whose distribution function is given by

$$F(t) = 1 - \left(\frac{c}{t+c}\right)^b, \qquad t > 0. \tag{2}$$

The corresponding log-hazard function is given by $\lambda(t) = \log b - \log(t + c)$ for $t > 0$. Thus $\lambda(\cdot) = \lambda(\cdot; \boldsymbol{\theta})$, where $\theta_{-1} = 0$, $\theta_0 = -1$ and $\theta_1 = \log b$. (Here we have assumed that the parameter c of the three-parameter model coincides with the parameter c of the Pareto distribution; otherwise, the three-parameter model only provides an approximation to the Pareto distribution.)

3 Maximum Likelihood Estimation

Let $T_1, \ldots, T_n$ be a random sample of size n from the distribution on $(0, \infty)$ having density function f, and let $C_1, \ldots, C_n \in (0, \infty]$ be censoring times. It is assumed that $T_1, \ldots, T_n$, $(C_1, \ldots, C_n)$ are independent. For $1 \le i \le n$, set $Y_i = \min(T_i, C_i)$ and $\delta_i = \mathrm{ind}(T_i \le C_i)$; the observation Y_i is said to be uncensored or censored according as $\delta_i = 1$ or $\delta_i = 0$. We refer to (Y_i, δ_i), $1 \le i \le n$, as the observed data. As the default we chose the shift parameter c to be the upper quartile of the uncensored data; that is, of Y_i, $1 \le i \le n$, with $\delta_i = 1$.

The log-likelihood function corresponding to the observed data and the $(p+2)$-parameter model for the log-hazard function is given by

$$l(\boldsymbol{\theta}) = \sum_i \varphi(Y_i, \delta_i; \boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \Theta. \tag{3}$$

As noted by O'Sullivan (1988), the log-likelihood function is concave. Moreover,

$$\frac{\partial}{\partial \theta_j} l(\boldsymbol{\theta}) = \sum_i \frac{\partial}{\partial \theta_j} \varphi(Y_i, \delta_i; \boldsymbol{\theta}), \qquad -1 \le j \le p \text{ and } \boldsymbol{\theta} \in \Theta,$$

and

$$\frac{\partial^2}{\partial \theta_j \partial \theta_k} l(\boldsymbol{\theta}) = \sum_i \frac{\partial^2}{\partial \theta_j \partial \theta_k} \varphi(Y_i, \delta_i; \boldsymbol{\theta}), \qquad -1 \le j, k \le p \text{ and } \boldsymbol{\theta} \in \Theta.$$

The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is given as usual by $l(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta})$ and the log-likelihood of the model is given by $\hat{l} = l(\hat{\boldsymbol{\theta}})$. The corresponding maximum likelihood estimates of λ, h, f, F, Q are given by $\hat{\lambda}(t) = \lambda(t;\hat{\boldsymbol{\theta}})$, $\hat{h}(t) = h(t;\hat{\boldsymbol{\theta}})$ and so forth.

Let $\mathbf{S}(\boldsymbol{\theta})$ denote the score at $\boldsymbol{\theta}$ (that is, the $(p+2)$-dimensional column vector with entries $\partial l(\boldsymbol{\theta})/\partial\theta_j$), and let $\mathbf{H}(\boldsymbol{\theta})$ denote the Hessian at $\boldsymbol{\theta}$ (that is, the $(p+2)\times(p+2)$ matrix with entries $\partial^2 l(\boldsymbol{\theta})/\partial\theta_j\partial\theta_k$). The Newton-Raphson method for computing $\hat{\boldsymbol{\theta}}$ is to start with an initial guess $\hat{\boldsymbol{\theta}}^{(0)}$ and iteratively determine $\hat{\theta}^{(m+1)}$ from $\hat{\boldsymbol{\theta}}^{(m)}$ according to the formula

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

Here we employ the Newton-Raphson method with step-halving, in which $\hat{\boldsymbol{\theta}}^{(m+1)}$ is determined from $\hat{\boldsymbol{\theta}}^{(m)}$ according to the formula

$$\hat{\boldsymbol{\theta}}^{(m+1)} = \hat{\boldsymbol{\theta}}^{(m)} - 2^{-\nu}[\mathbf{H}(\hat{\boldsymbol{\theta}}^{(m)})]^{-1}\mathbf{S}(\hat{\boldsymbol{\theta}}^{(m)}),$$

where ν is the smallest nonnegative integer such that

$$l(\hat{\boldsymbol{\theta}}^{(m)} - 2^{-\nu}[\mathbf{H}(\hat{\boldsymbol{\theta}}^{(m)})]^{-1}\mathbf{S}(\hat{\boldsymbol{\theta}}^{(m)})) \geq l(\hat{\boldsymbol{\theta}}^{(m)} - 2^{-\nu-1}[\mathbf{H}(\hat{\boldsymbol{\theta}}^{(m)})]^{-1}\mathbf{S}(\hat{\boldsymbol{\theta}}^{(m)})).$$

We stop the iterations when $l(\hat{\boldsymbol{\theta}}^{(m+1)}) - l(\hat{\boldsymbol{\theta}}^{(m)}) \leq \epsilon$, where $\epsilon = 10^{-6}$.

4 Knot Selection

Initially, we place knot at the quartiles of the uncensored data. Since $K = 3$, this corresponds to the three-parameter model discussed in Section 2. Then we successively add knots, at each step choosing a new knot by an heuristic search (described in Section 9.3) that is designed approximately to maximize the Rao statistic.

Given a model with $K-1$ knots and a potential additional knot, let $\hat{\boldsymbol{\theta}}_0$ be the maximum likelihood estimate of $\boldsymbol{\theta}$ based on the model with K knots subject to the constraint that the jump of the third derivative at the potential knot equals zero. Then the corresponding Rao statistic ((6e.3.6) of Rao, 1973) equals $[\mathbf{S}(\hat{\boldsymbol{\theta}}_0)]^T[\mathbf{I}(\hat{\boldsymbol{\theta}}_0)]^{-1}\mathbf{S}(\hat{\boldsymbol{\theta}}_0)$, where $\mathbf{I}(\hat{\boldsymbol{\theta}}_0) = -\mathbf{H}(\hat{\boldsymbol{\theta}}_0)$ with $\mathbf{S}(\cdot)$ and $\mathbf{H}(\cdot)$ corresponding to the model with K knots.

Upon stopping the stepwise knot addition (according to a rule that is described in Section 9.3), we proceed to stepwise knot deletion. Here we successively remove the least statistically significant among the K remaining knots until only three knots remain. The statistical significance of a remaining knot is measured by the absolute value of its Wald statistic $W = \hat{\tau}/\mathrm{SE}(\hat{\tau})$. Here $\hat{\tau} = \mathbf{c}^T\hat{\boldsymbol{\theta}}$ is the jump of the third derivative of $\sum_j \hat{\theta}_j B_j$ at the corresponding knot, where the $B_j's$ are defined in terms of the K remaining knots. Also, $\mathrm{SE}(\hat{\tau}) = \sqrt{\mathbf{c}^T[\mathbf{I}(\hat{\theta})]^{-1}\mathbf{c}}$, where $\mathbf{I}(\hat{\boldsymbol{\theta}}) = -\mathbf{H}(\hat{\boldsymbol{\theta}})$ with $\mathbf{H}(\cdot)$ corresponding to the model with K knots.

During the combination of stepwise knot addition and stepwise knot deletion, we get a sequence of models indexed by ν, with the νth model having p_ν parameters. Let $\hat{l}_\nu$ denote the log-likelihood of the νth model, and let $\text{AIC}_{\alpha,\nu} = -2\hat{l}_\nu + \alpha p_\nu$ be the Akaike Information Criterion with penalty parameter α for this model. We select the model corresponding to the value $\hat{\nu}$ of ν that minimizes $\text{AIC}_{\alpha,\nu}$. In light of Kooperberg and Stone (1992) and our experience in the present investigation, we recommend choosing $\alpha = \log(n)$ as in the Bayesian information criterion (BIC) due to Schwarz (1978).

5 User Interface

A program for implementing hazard estimation with flexible tails (HEFT) as described in this paper has been written in C (see Section 9), and an interface based on S (see Becker, Chambers and Wilks, 1988, and Chambers and Hastie, 1992) has also been developed[1]. The interface consists of nine S functions: `dheft`, `hheft`, `pheft`, `qheft`, `rheft`, `heft.fit`, `heft.summary`, and `heft.plot`. (Detailed documentation of each of these functions is included in the Appendix to this paper.) The functions `dheft`, `pheft`, `qheft`, `rheft` are analogous to the S functions `dnorm`, `pnorm`, `qnorm` and `rnorm`, respectively, and to similar four-tuples of S functions for t distributions, F distributions, gamma distributions, and so forth. Thus `dheft` gives the (estimated) density function, `pheft` gives the distribution function, `qheft` gives the quantile function, and `rheft` gives a random sample from the distribution. The function `hheft` gives the hazard function, `heft.fit` performs the model fitting and model selection tasks and supplies the modest output that is used as input to `dheft`, `hheft` and so forth. The function `heft.summary`, uses the output of `heft.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, `heft.plot` uses the output of `heft.fit` directly to produce a plot of the density, distribution, survival or hazard function.

6 Simulated Examples

In Section 2 we discussed how Pareto and Weibull distributions can be modeled using HEFT. To illustrate how HEFT can be used to approximate these distributions based on sample data, we generated a sample of size 200 from a Pareto distribution with parameters $b = 4$ and $c = 1$ (2). In the left side of Figure 1, we show the true density function (solid) corresponding to this distribution together with various estimates of the densty function based on the sample. The line with long dashes corresponds to the estimated densty function that was obtained from HEFT using the default parameters. As we noted in Section 2, HEFT can exactly fit a Pareto distribution if the shift parameter c in HEFT equals the parameter c in

[1]HEFT software is available from statlib. Send an email with the body **send heft from S** to statlib@stat.cmu.edu

the Pareto distribution. The default value for c in HEFT is the 75th percentile of the data, which was 0.4 for this sample. To examine the influence of the shift parameter, we also used `heft.fit` with the option `shift=1`, to make it possible for HEFT to fit the exact Pareto distribution. The third curve in the left side of Figure 1 that is based on HEFT uses the value for the shift parameter c that minimizes BIC. We determined that this was 2.7. The last curve is the logspline estimate of the density function (Kooperberg and Stone (1992)); in this and all other logspline estimates in the present paper we used the option `lbound=0`.

In the right side of Figure 1 we show the results of a similar set of computations. The data for this figure is a random sample of size 1000 from a Pareto distribution with parameters $b = 1$ and $c = 1$ (2). Again we show the estimate based on HEFT with the default choice for `shift` (which was 2.9), the theoretical optimal choice for `shift` ($c = 1$) and the value for `shift` that minimizes BIC ($c = 0.8$). As for the left hand side, the remaining curves are the density function corresponding to the logspline estimate and the true density function.

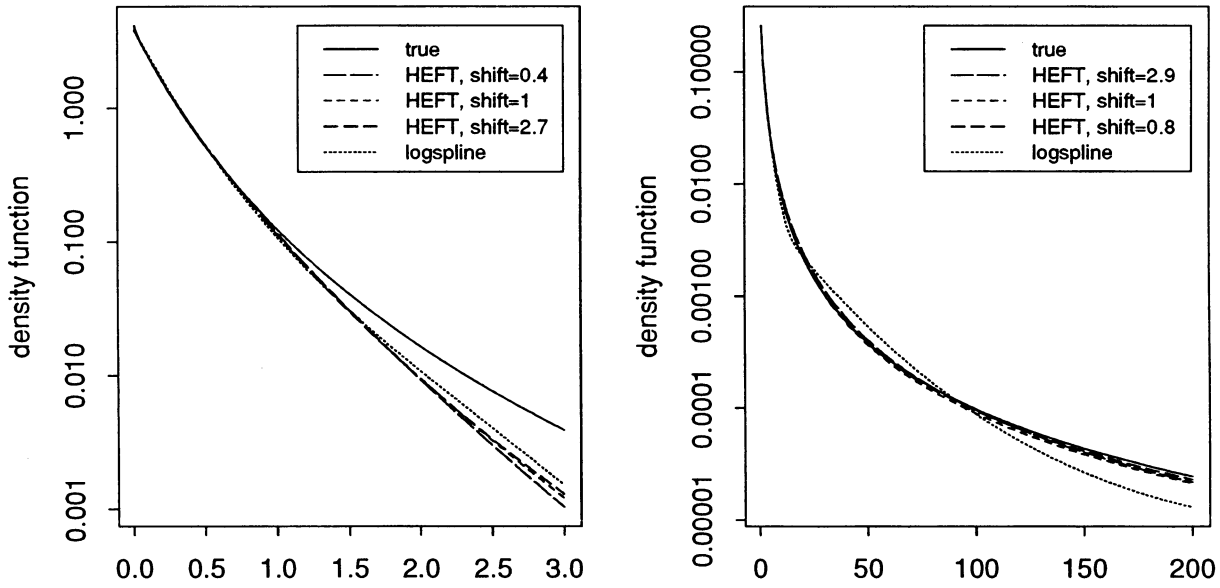

Fig. 1. Estimated density functions for Pareto distributions; left side: n=200, b=4, c=1; right side: n=1000, b=1, c=1.

From both of these examples (and from many more that we examined) we find that HEFT approximates Pareto distributions extremely well for sample sizes of 500 and larger, especially if `shift` is optimized but even with the default choice. It should be noted, though, that the HEFT estimate frequently does not coincide with the three parameter model described in Section 2. Often a few knots, close to the origin, remain. If the sample size is smaller, the HEFT estimate of the Pareto distribution typically has the form of the three parameter model in Section 2. Nevertheless, for these smaller sample sizes the HEFT estimate typically is not better than the logspline estimate.

Figure 2 is similar to Figure 1, but the underlying distributions for this figure are Weibull. The data for the left side of Figure 2 is a sample of size 200 from a Weibull distribution with parameters $b = 1$ and $\gamma = 0.25$ (1). In the figure we show the true density function corresponding to this Weibull distribution together with the estimate for this density function based upon HEFT using the default parameters and the estimate based upon the logspline density estimate for this data. In the right side of the Figure we show the results of similar calculations, based upon a sample of size 1000 from a Weibull distribution with parameters $b = 1$ and $\gamma = 4$ (1).

The HEFT fits to the Weibull distribution that are illustrated in Figure 2 turn out to be the three parameter model described in Section 2. The parameters $\hat{\theta}_{-1}$, $\hat{\theta}_0$ and $\hat{\theta}_1$ for the HEFT fit in the left side of Figure 2 are -0.755, -0.819 and -1.232 respectively while their theoretical values are -0.75, -0.75 and -1.386. Similarly the parameters for the HEFT fit in the right side of Figure 2 are 3.444, 2.383 and 2.182, while their theoretical values are 3, 3 and 1.386.

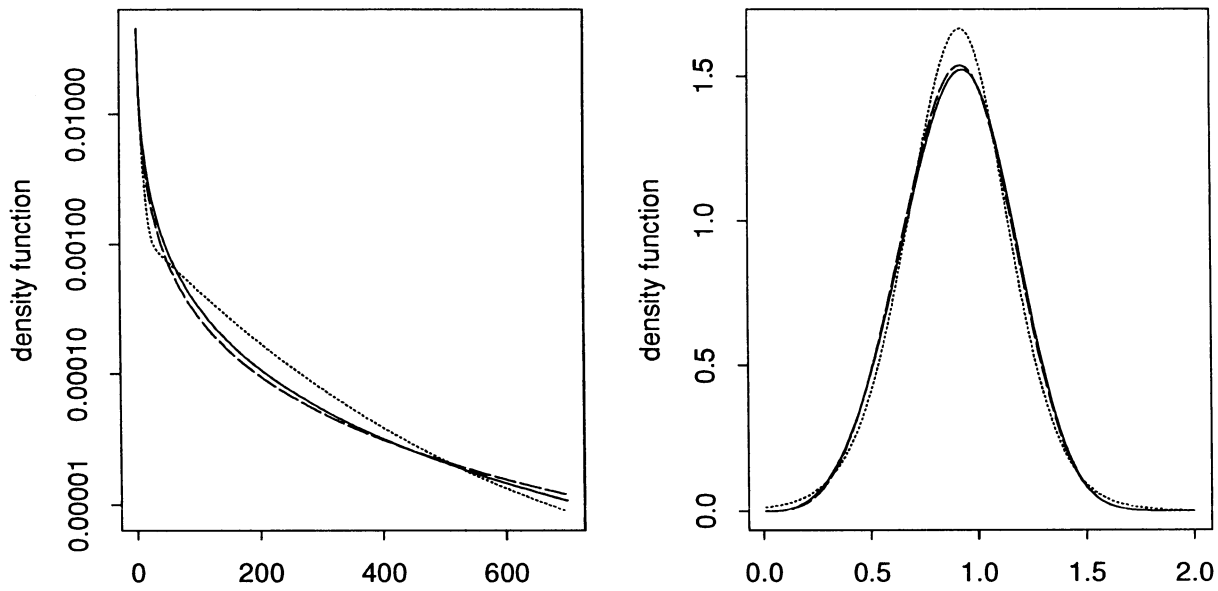

Fig. 2. Estimated density functions for Weibull distributions; left side: n=200, gamma=0.25; right side: n=1000, gamma=4; solid = truth, dashed = HEFT, dotted = logspline.

We carried out a small simulation study to determine how close these parameters typically are to their theoretical values. One hundred times we generated samples of size 200 and size 1000 from Weibull distributions with parameters $b = 1$ and $\gamma = 0.25$ and with parameters $b = 1$ and $\gamma = 4$. The large majority of HEFT fits are the three parameter model described in Section 2. The results are summarized in Table 1 below.

In Figure 3 logspline and HEFT are compared in a situation involving right-censoring and in which the true density function cannot be fitted exactly by HEFT. For this example we

TABLE 1. Coefficient estimates for HEFT fits of samples from Weibull distributions.

parameters	$b = 1,\ \gamma = 0.25$			$b = 1,\ \gamma = 4.00$		
sample size	200			200		
% of 3 parameter models	82%			89%		
coefficients	θ_{-1}	θ_0	θ_1	θ_{-1}	θ_0	θ_1
theoretical	−0.750	−0.750	−1.386	3.000	3.000	1.386
average	−0.735	−0.705	−1.327	2.951	3.194	1.213
standard deviation	0.060	0.156	0.716	1.426	2.369	2.701
parameters	$b = 1,\ \gamma = 0.25$			$b = 1,\ \gamma = 4.00$		
sample size	1000			1000		
% of 3 parameter models	76%			92%		
coefficients	θ_{-1}	θ_0	θ_1	θ_{-1}	θ_0	θ_1
theoretical	−0.750	−0.750	−1.386	3.000	3.000	1.386
average	−0.737	−0.746	−1.286	3.033	3.028	1.405
standard deviation	0.060	0.043	0.645	0.733	1.011	1.274

The averages and standard deviations for the coefficient estimates are based on all 100 samples.

generated a sample $T_i = \exp(Z_i)$, $1 \leq i \leq 1000$, where Z_i has a standard normal distribution, and an independent sample C_i, $1 \leq i \leq 1000$, from the same distribution. We then set $Y_i = \min(T_i, C_i)$ and $\delta_i = \text{ind}(T_i \leq C_i)$. This setup yields about 50% right-censoring. In the left side of Figure 3 we show the true lognormal density function (solid) together with the estimates obtained using HEFT (dashed) and logspline (dotted). In the right hand side of the figure we show the corresponding hazard estimates. As this figure illustrates, the HEFT estimate of the hazard function typically is more accurate than the logspline estimate in the right tail; otherwise the two estimates are comparable.

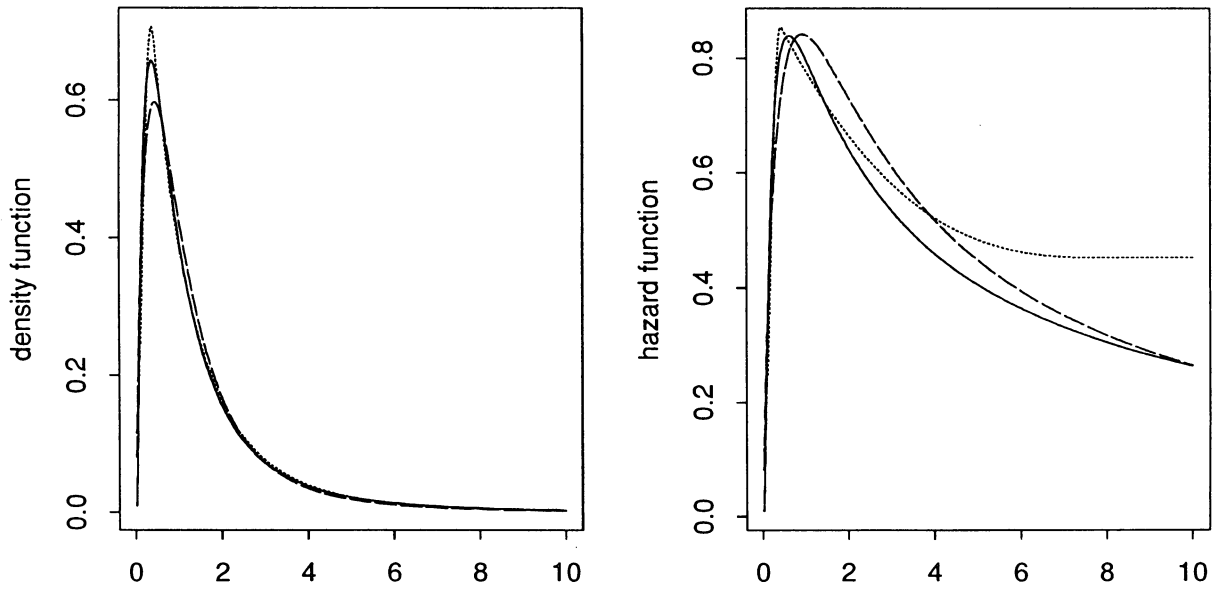

Fig. 3. Estimated density (left) and hazard (right) functions for lognormal distribution; n=1000, 50% right censored; solid = truth, dashed = HEFT, dotted = logspline.

In Section 6 of Kooperberg and Stone (1992) the logspline method is applied to samples from various distributions, a number of them which involve censoring. We compared the performance of logspline and HEFT on the examples in Figures 3 to 6 of the logspline paper, which involve only positive data, and found HEFT and logspline comparable in these examples. However, there are several circumstances in which HEFT has an advantage over logspline. In particular, the log-likelihood function (3) is concave for HEFT even when there is right censoring. This is not true for logspline, so that convergence to a global maximum cannot be guaranteed. (And indeed, we have seen cases with a large amount of right censoring for which HEFT converges to an acceptable solution, while the present implementation of logspline does not converge.) Another advantage of HEFT is that it is more flexible near the origin. In particular, it can deal without further adjustments with estimates for which $f(0) = 0$ and for which $f(0+) = \infty$, while this is not possible in logspline. A final advantage of HEFT over the present implementation of logspline is that HEFT employs knot addition and knot deletion, while logspline only employs knot deletion. Potentially this means that HEFT estimates are more flexible. Somewhat surprisingly, in spite of all these arguments, logspline is generally comparable to HEFT. We should mention that logspline has some advantages over HEFT too: it can deal with negative, left-censored and interval-censored data and density functions that are non-zero only on a known bounded interval. Finally, logspline often takes less cpu time than HEFT.

7 Real Examples

7.1 HEFT for Density Estimation

The data for Figure 4 consists of a random sample of size 7125 of annual net incomes in the United Kingdom (Family Expenditure Survey, 1968-1983)[2], which have been rescaled to have mean one. This data has also been used in Wand, Marron and Ruppert (1991) and in Kooperberg and Stone (1991, 1992). In the left side of Figure 4 we show the estimate of the density function that was obtained using HEFT together with the rather similar estimate obtained by using the logspline method (Kooperberg and Stone (1992)). In Kooperberg and Stone (1992) it was argued that the peak near 0.2 should have approximately the indicated height.

The logspline fit in Figure 4 uses 9 knots, while the HEFT fit uses 7 knots. Actually, 7 of the logspline knots are extremely close to the HEFT knots. The two extra logspline knots are at 2.03 and 11.46. Conceivably HEFT does not 'need' these knots because of the term $\log(t + 1.28)$ that is included in the model. (Note that 1.28 is the 75th percentile of the data.) The coefficient of this basis function was estimated by HEFT to be -1.00, the smallest value that yields a valid HEFT model. The coefficient of the other basis function $\log \frac{t}{t+1.28}$ involving a log-term turned out to be -0.14, with a standard error of 0.16. This led us to rerun HEFT but now using the options `leftlog=0` and `rightlog=-1`, which forces the coefficient of $\log \frac{t}{t+1.28}$ to be 0 and the coefficient of $\log(t + 1.28)$ to be -1. As noted in Section 2, if the coefficient of the basis function $\log(t + c)$ is 0 and the coefficient of the basis function $\log \frac{t}{t+c}$ is -1, the tail of the density is like that of a Pareto density. (Pareto distributions have been used to model the upper tail of income distributions.)

It is not surprising that the knots for the logspline and HEFT fits are fairly close in location. Since this dataset is fairly large, both logspline (which starts out with a large number of knots and then employs stepwise knot deletion) and HEFT (which uses knot addition and knot deletion) try a large number of knots, essentially covering the whole range of the data quite well.

In the right side of Figure 4 we show the same density estimate using HEFT that was shown in the left side of Figure 4 together with the HEFT estimate with options that force a "Pareto tail," as described above. It is hard to distinguish the two curves. Since one less parameter is estimated for the estimate with the Pareto tail, we prefer that estimate.

[2]The calculations were made in collaboration with the Wirtschaftstheoretische Abteilung II, University of Bonn, Bonn, Germany

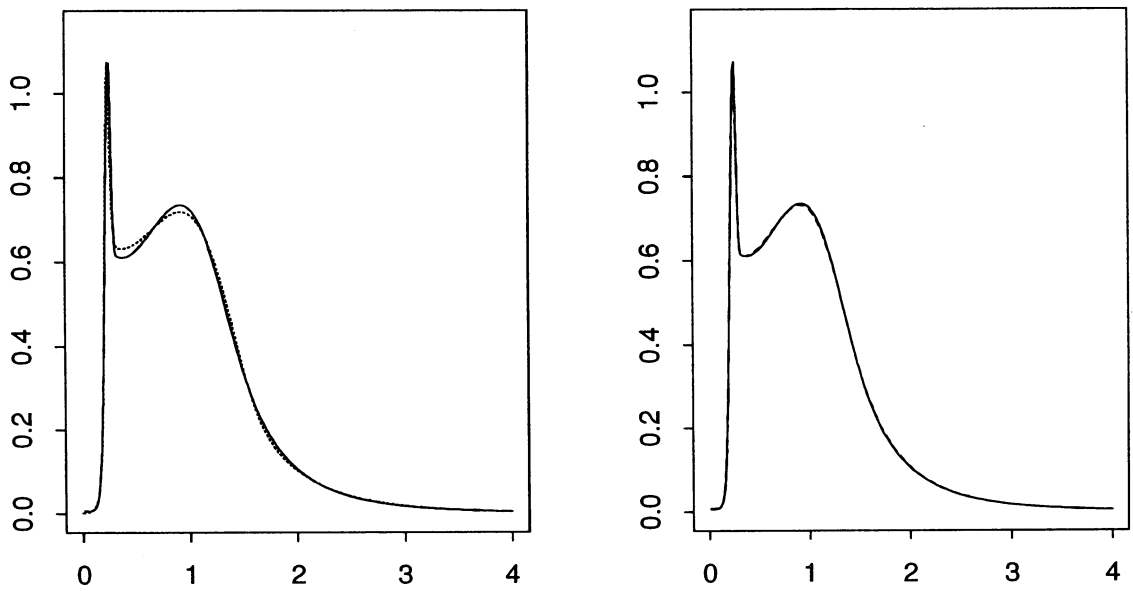

Fig. 4. Estimated density functions for the income data;
left side: solid = HEFT using defaults, dotted = logspline;
right side: solid = HEFT using defaults, dashed = HEFT forcing a Pareto tail.

7.2 HEFT as Pre-Processor for HARE

Even when there is a substantial percentage of right censoring, HEFT yields a reasonable estimate for the hazard function, as can be seen from Figures a-b in Section 6. As such, HEFT is a useful preprocessor for HARE (Kooperberg, Stone and Truong (1993)).

Hazard Regression (HARE) is a methodology for estimating the conditional log-hazard function based on possibly censored, positive response data and one or more covariates that has concurrently been developed. For HARE a MARS-like methodology (Friedman (1991)) is used to obtain a model for the conditional hazard function having the form $h(t|X_1,\ldots,X_p) = \exp\left(\sum_j \beta_j g_j(t,X_1,\ldots,X_p)\right)$; each g_j involves at most two of the variables $t, X_1,\ldots,X_p$ and has the form of a linear spline or tensor product of two linear splines with the linear splines in t being constant in the right tail and the knots selected by stepwise addition-deletion and BIC. Before applying HARE, HEFT can be used to transform time so that the transformed unconditional hazard function is approximately equal to one.

In particular, let T be the survival time, C the censoring time and $\boldsymbol{x}$ the vector of covariates for a randomly selected individual. It is assumed that T and C are conditionally independent and that T has conditional density function $f(\cdot|\boldsymbol{x})$ given $\boldsymbol{x}$. Set $Y = \min(T,C)$ and $\delta = \mathrm{ind}(T \le C)$; the random variable Y is said to be uncensored or censored according as $\delta = 1$ or $\delta = 0$. Consider n such individuals. For $1 \le i \le n$ let T_i be the survival time,

C_i the censoring time and $\boldsymbol{x}_i$ the vector of covariates for the ith such individual, and set $Y_i = \min(T_i, C_i)$ and $\delta_i = \text{ind}(T_i \leq C_i)$.

The HEFT methodology is now applied to (Y_i, δ_i), $1 \leq i \leq n$, to yield an estimate $\hat{h}_0$ of the unconditional hazard function. The HARE methodology is then applied to $(\hat{q}_0(Y_i), \delta_i, \boldsymbol{x}_i)$, yielding an estimate $\hat{h}_1$ of the conditional hazard function for the transformed data and the estimate $\hat{h}(t|\boldsymbol{x}) = \hat{h}_0(t)\hat{h}_1(\hat{q}_0(t), \boldsymbol{x})$ of the conditional hazard function for the untransformed data; here $\hat{q}_0 = -\log(1 - \hat{F}_0)$ with $\hat{F}_0$ being the distribution function corresponding to $\hat{h}_0$. It is easily seen that the unconditional hazard function of the transformed time is now approximately constant.

One of the examples that is used in Kooperberg, Stone and Truong (1993) involves data from a Veteran's Administration lung cancer trial (see Kalbfleisch and Prentice (1980)). The response is survival time, there are six predictors, and there are 137 cases of which 9 are censored. When HEFT is applied with the default options, the estimate for the hazard rate is the dotted line in the left side of Figure 5. The corresponding transformation $\hat{q}_0$ is shown in the right side of Figure 5. The estimated hazard function has no knots remaining and the coefficient of $\log \dfrac{t}{t + 145.75}$ is 0.0075, with a standard error of 0.1280, while the coefficient of $\log(t + 145.75)$ is -0.597 with a standard error of 0.321; the estimate of the intercept is -1.55. The BIC value for this model is 1508.73.

This leads us to use `heft.fit` with the option `leftlog=0`, forcing the coefficient of $\log \dfrac{t}{t + 145.75}$ to be 0. As expected, this hazard estimate again had no knots remaining. The coefficient for $\log(t + 145.75)$ is now -0.583 with a standard error of 0.211, and the intercept is -1.643, so that this model corresponds to

$$\hat{h}(t) \approx e^{-1.643}(t + 145.75)^{-0.583}. \tag{4}$$

The BIC value for this model is 1503.82 which is considerably smaller than the BIC value for the previous model since this model has one less parameter. The estimate for the unconditional hazard function and the corresponding transformation are the solid curves in Figure 5. These curves are hard to distinguish from the dotted ones corresponding to the previous fit.

Finally we applied `heft.fit` with the options `leftlog=0` and `rightlog=0`, forcing the coefficients of both log-based basis functions to be 0. This HEFT estimate has the form of a two parameter model involving four knots, and its BIC value is 1504.65. The estimate for the unconditional hazard function and the corresponding transformation for this fit are the dashed curves in Figure 5. Observe that this estimate differs considerably from the other two estimates. All in all, we like the solid curve corresponding to (4) best and use the resulting transformation in Kooperberg, Stone and Truong (1993).

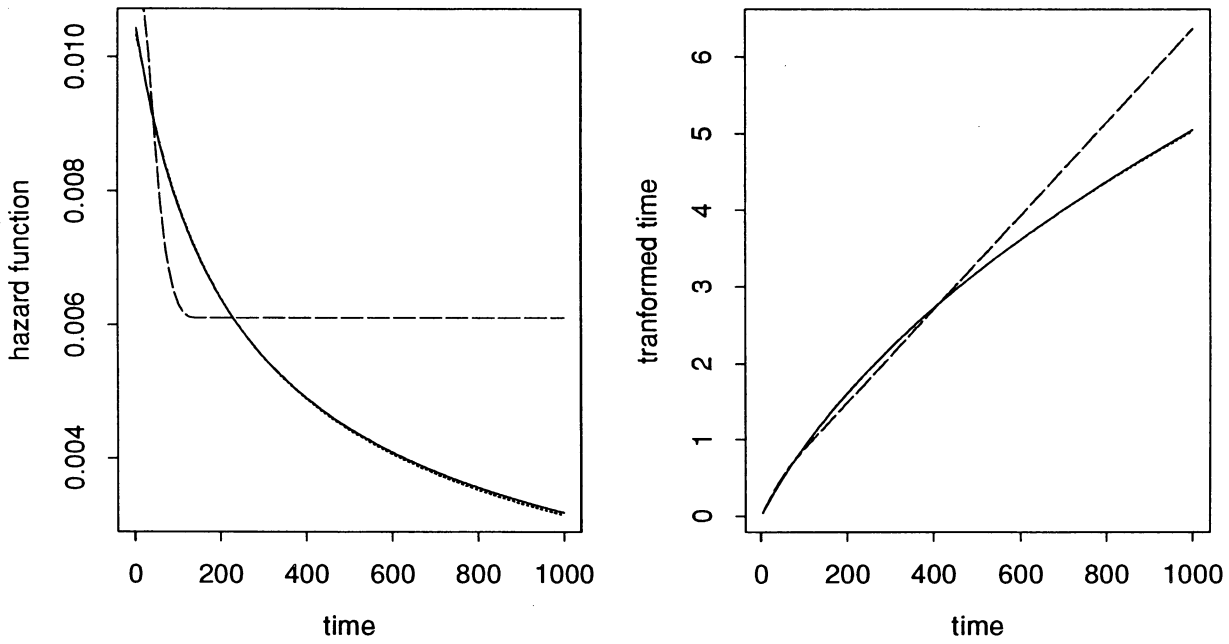

Fig. 5. Three estimates of the unconditional hazard function and the corresponding transformation of time using HEFT with various options for the lung cancer data.

Another example discussed in Kooperberg, Stone and Truong (1993) comes from 6 breast cancer trials conducted by the Eastern Cooperative Oncology Group[3]. There were 2404 patients in these studies. The response is survival time. Of the 2404 cases, 1116 were uncensored and 1288 were censored. The data is discussed in detail by Gray (1992).

In this dataset some of the survival times are exactly 0, which makes it impossible to include $\log \frac{t}{t+c}$ as a basis function in the HEFT fit. Thus, by giving the option `leftlin=T` to `heft.fit`, we let G_0 be the $(K-1)$-dimensional space of twice-continuously differentiable functions s on $[0,\infty)$ such that s is linear (as opposed to constant) on $[0,t_1]$ and constant on $[t_K,\infty)$ and the restriction of s to each of the intervals $[t_1,t_2],\ldots,[t_{K-1},t_K]$ is a cubic polynomial.

The estimated unconditional hazard function and the corresponding transformation of time are shown in Figure 6. This fit has four parameters since it is based on four knots and the term $\log(t+c)$.

[3]The data for this example was kindly provided by the Eastern Cooperative Oncology Group.

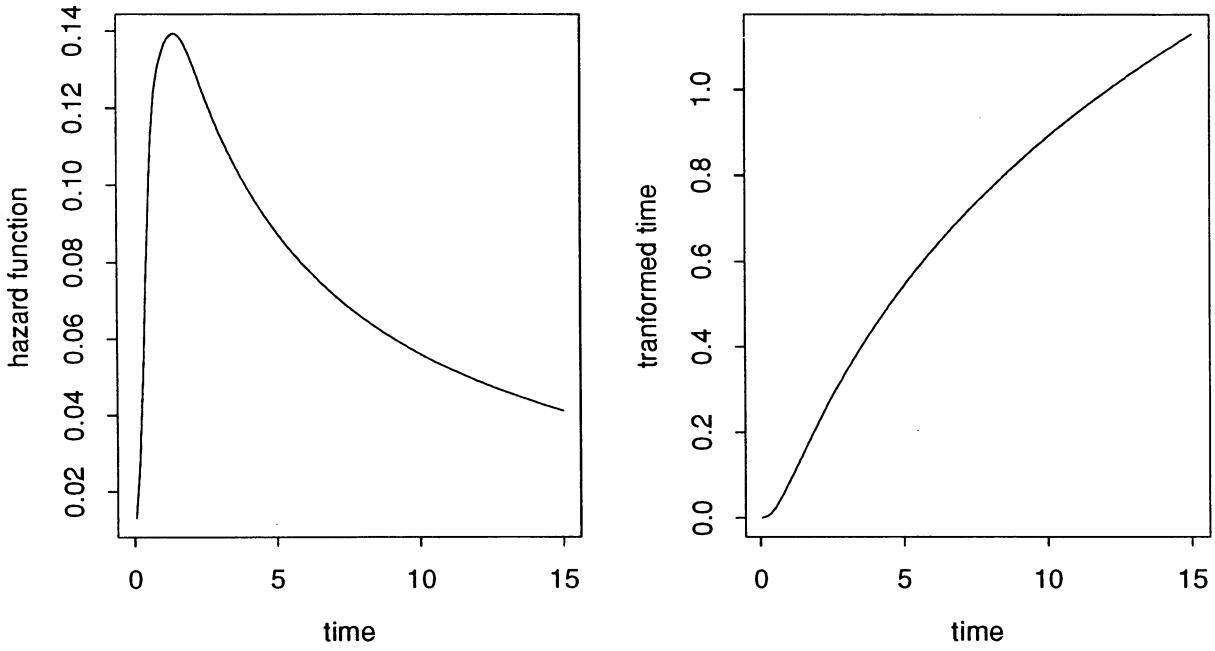

Fig. 6. Estimates of the unconditional hazard funtion and the corresponding transformation of time using HEFT for the breast cancer data.

8 Concluding Remarks

Hazard Estimation with Flexible Tails, as described in this paper, combines various features of the present implementation of logspline with several new innovations. In particular, HEFT combines automatic addition and deletion of knots; it has two extra log terms, which are specifically tailored to fit the tails of the underlying distribution; and, since the log-hazard function is directly modeled instead of the log-density function, the corresponding log-likelihood function is concave, guaranteeing convergence of the Newton-Raphson algorithm to the maximum likelihood estimate even in the presence of right-censored data.

An important improvement of HEFT over existing methodology is that it estimates the right tail of a distribution well even when there is a substantial amount of right-censoring while being just as good as other density estimates elsewhere. Moreover, it is an ideal pre-processor for HARE (Kooperberg, Stone and Truong (1993)).

The methodology in this paper is easily extended to handle random survival times having the form $T = \min(T_1, \ldots, T_M)$, where $T_1, \ldots, T_M, C$ are independent, positive random variables and T_m has log-hazard function λ_m for $1 \leq m \leq M$. Here we set $Y = \min(T, C)$, $\delta = 0$ if $C < T$ and $\delta = j \in \{1, \ldots, M\}$ if $Y = T_j$. Then the log-likelihood function has the

form $l(\boldsymbol{\theta}) = \sum_i \varphi(Y_i, \delta_i; \boldsymbol{\theta})$, where

$$\varphi(y, 0; \boldsymbol{\theta}) = -\sum_{m=1}^{M} \int_0^y \exp(\lambda_m(u; \boldsymbol{\theta}))du$$

and

$$\varphi(y, j; \boldsymbol{\theta}) = \lambda_j(y; \boldsymbol{\theta}) - \sum_{m=1}^{M} \int_0^y \exp(\lambda_m(u; \boldsymbol{\theta}))du, \qquad 1 \leq j \leq M.$$

On the other hand, it is not clear how to extend this methodology to handle multivariate survival distributions as treated in Section X.3 of Andersen, Borgan, Gill, and Keiding (1993).

Presumably, a large sample theory along the lines of Stone (1990) could be developed in the context of the present paper, but this has yet to be done.

9 Numerical Implementation

9.1 Starting Values

As the starting value for the maximum likelihood estimate of the log-hazard function in the linear space corresponding to the three-parameter model, we use the maximum likelihood constant estimate $\hat{\lambda} = \log(\sum_i \delta_i / \sum_i Y_i)$ of this function. In the context of stepwise addition, the starting value for the next step is the exact maximum likelihood estimate from the previous step, which is possible since the new linear space contains the previous one as a proper subspace.

In the context of stepwise deletion, let $\hat{\theta}_{-1}B_{-1} + \hat{\theta}_0 B_0 + \hat{\theta}_1 B_1 \cdots + \hat{\theta}_p B_p$ be the maximum likelihood estimate of the log-hazard function corresponding to a model with $K = p$ knots, and let $B_{-1}, B_0, \tilde{B}_1, \ldots, \tilde{B}_{p-1}$ be the basis corresponding to the deletion of one of the K knots. Also, for $1 \leq j \leq p$, let $\sum_{k=1}^{p-1} a_{jk}\tilde{B}_k$ be the orthogonal projection of B_j onto the span of $\tilde{B}_1, \ldots, \tilde{B}_{p-1}$ relative to the inner product $\langle h_1, h_2 \rangle = \sum_i h_1(Y_i)h_2(Y_i)$. As the starting value for the maximum likelihood estimate of the log-hazard function corresponding to the model with $K - 1$ knots, we use

$$\hat{\theta}_{-1}B_{-1} + \hat{\theta}_0 B_0 + \sum_{j=1}^{p} \hat{\theta}_j \left(\sum_{k=1}^{p-1} a_{jk}\tilde{B}_k \right) = \hat{\theta}_{-1}B_{-1} + \hat{\theta}_0 B_0 + \sum_{k=1}^{p-1} \left(\sum_{j=1}^{p} a_{jk}\hat{\theta}_j \right) \tilde{B}_k.$$

9.2 Computation of the Log-Likelihood Function, Score Function and Hessian

The main numerical task of the algorithm is the computation of the log-likelihood $l(\boldsymbol{\theta})$, the score $\mathbf{S}(\boldsymbol{\theta})$ and the Hessian $\mathbf{H}(\boldsymbol{\theta})$ for various models and values of $\boldsymbol{\theta}$. The time consuming

aspect of this computation involves the numerical approximation of

$$\sum_i \int_0^{Y_i} \psi(u)du = \int_0^{\infty} N(u)\psi(u)du, \qquad N(u) = \#(\{i : Y_i \geq u\}),$$

for many functions ψ, each of which is twice continuously differentiable on $(0, \infty)$ and three times continuously differentiable on each of the intervals

$$(0, t_1], [t_1, t_2], \ldots, [t_{K-1}, t_K], [t_K, \infty).$$

Note that the function $N(\cdot)$ is piecewise constant, has jumps at the observations $Y_1, \ldots, Y_n$, and equals zero to the right of the maximum observation $Y_{(n)}$.

Let $J_1, \ldots, J_M$ be a partition of $(0, Y_{(n)}]$ into disjoint intervals whose endpoints contain all of the initial knots. Then

$$\int_0^{\infty} N(u)\psi(u)du = \sum_{\nu} \int_{J_\nu} N(u)\psi(u)du.$$

Thus the time consuming aspect of the computation involves the computation of

$$\int_J N(u)\psi(u)du,$$

where J is a bounded interval and ψ is a three times continuously differentiable function on a bounded interval J_0 containing J. Let b_1, b_2, b_3 and b_4 be distinct points in J_0, and let P be the cubic polynomial that interpolates the values of ψ at these points. We approximate $\int_J N(u)\psi(u)du$ by $\int_J N(u)P(u)du$. According to the Lagrange interpolation formula, $P(u) = \sum_l \psi(b_l)P_l(u)$, where $P_l(u) = \prod_{m \neq l}(u - b_m)/\prod_{m \neq l}(b_l - b_m)$. Observe that

$$\int_J N(u)P(u)du = \int_J N(u) \sum_l \psi(b_l)P_l(u) = \sum_l \psi(b_l) \int_J N(u)P_l(u)du,$$

where the quantities $\int_J N(u)P_l(u)du$ (which can be evaluated analytically) need only be obtained once, right after the partition $J_1, \ldots, J_M$ and the four interpolation points corresponding to each of these intervals is determined.

Suppose one or more of the uncensored observations equals zero. If the coefficient θ_{-1} of the basis function B_{-1} is negative, then the log-likelihood function is infinite at zero. In order to avoid this difficulty, we omit the basis function B_{-1} and let G_0 be the $(K-1)$-dimensional space of twice-continuously differentiable functions s on $[0, \infty)$ such that s is *linear* on $[0, t_1]$ and constant on $[t_K, \infty)$ and the restriction of s to each of the intervals $[t_1, t_2], \ldots, [t_{K-1}, t_K]$ is a cubic polynomial.

9.3 Stepwise Knot Addition

Let $t_1 < t_2 < \cdots < t_K$ be the knots presently in the model, to which we want to add one more knot, and let $T_{(1)}, \ldots, T_{(n_u)}$ be those observations Y_i, $1 \le i \le n$ and $\delta_i = 1$, written in nondecreasing order. Define l_i and u_i by

$$
\begin{aligned}
l_i &= 6 + \arg \max_{1 \le j \le n_u} T_{(j)} \le t_i, \qquad i = 1, \ldots, k, && (5)\\
u_i &= -6 + \arg \min_{1 \le j \le n_u} T_{(j)} \ge t_{i+1}, \qquad i = 0, \ldots, k-1, && (6)\\
l_0 &= 1 \text{ and}\\
u_k &= n_u.
\end{aligned}
$$

For those $i = 0, \ldots, K$ for which $u_i \ge l_i$ we compute r_i, the Rao statistic as described in Section 4, for the model with knots at $t_1, t_2, \ldots, t_k$ and a potential knot at $T_{(m_i)}$, where $m_i = [(l_i + u_i)/2]$. Because of the 6 and -6 in (5) and (6) it is possible that $u_i < l_i$ for some i; if so, then no knot can be added between t_i and t_{i+1}. This forces knots in a model to be at least 6 order statistics apart, which improves the numerical stability. If there is no i for which $u_i \ge l_i$ no knots can be added to the model.

We place the new knot in the interval $[T_{(l_{i^*})}, T_{(u_{i^*})}]$, where $i^* = \arg\max r_i$. We proceed by computing the Rao statistic r_l for the model with knots at $t_1, t_2, \ldots, t_k$ and a potential knot at $T_{(l)}$, where $l = [(l_{i^*} + m_{i^*})/2]$, and r_u for the model with knots at $t_1, t_2, \ldots, t_k$ and a potential knot at $T_{(u)}$, where $u = [(m_{i^*} + u_{i^*})/2]$. If $r_{i^*} \ge r_l$ and $r_{i^*} \ge r_u$ we place the new knot at $T_{(m_{i^*})}$; if $r_{i^*} < r_l$ and $r_l \ge r_u$ we continue searching for a knot location in the interval $[T_{(l_{i^*})}, T_{(m_{i^*})}]$; if $r_{i^*} < r_u$ and $r_l < r_u$ we continue searching for a knot location on the interval $[T_{(m_{i^*})}, T_{(u_{i^*})}]$.

Note that for each candidate for the new knot only one column of $\mathbf{H}(\cdot)$ and one element of $S(\cdot)$ have to be computed, all other elements having already been computed during the most recent set of iterations.

We stop knot addition when one of the following three conditions is satisfied:

- the number of knots K is equal to K_{max}, where $K_{max} = \min(4n^{.2}, n/4, 30)$;
- $\hat{l}_K - \hat{l}_k < \frac{1}{2}(K - k) - 0.5$ for any $3 \le k \le K - 3$, where $\hat{l}_k$ is the log-likelihood for the model with k knots;
- the search algorithm, as described above, yields no possible position for a new knot.

References

Abrahamowicz, M., Ciampi, A. and Ramsay, J. O. (1992), "Nonparametric density estimation for censored survival data: Regression-spline approach," *The Candian Journal of Statistics*, 20, 171-185.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer.

Anderson, J. A. and Senthilselvan, A. (1980), "Smooth estimates for the hazard function," *Journal of the Royal Statistical Society*, Ser. B, 42, 322-327.

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

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

Cox, D. D. and O'Sullivan, F. (1990), "Asymptotic analysis of penalized likelihood and related estimators," *The Annals of Statistics*, 18, 1676-1695.

Cox, D. R. and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman and Hall.

Efron, B. (1988), "Logistic regression, survival analysis, and the Kaplan-Meier curve," *Journal of the American Statistical Association*, 83, 414-425.

Etezadi-Amoli, J. and Ciampi, A. (1987), "Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function," *Biometrics*, 43, 181-192.

Family Expenditure Survey (1968-1983), *Annual base tapes and reports (1968-1983)*, Department of Employment, Statistics Division - Her Majesty's Stationary Office, London. (The data utilized in this paper were made available by the ESRC Data Archive at the University of Essex.)

Friedman, J. H. (1991), "Multivariate regression splines (with discussion)," *The Annals of Statistics*, 19, 1-141.

Gray, R. J. (1992), "Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis,"

Gu, C (1991), "Penalized likelihood hazard estimation," Technical Report No. 91-58, Dept. of Statistics, Purdue University.

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

Kooperberg, C. and Stone, C. J. (1991), "A study of logspline density estimation," *Computational Statistics and Data Analysis*, 12, 327-347.

Kooperberg, C. and Stone, C. J. (1992), "Logspline density estimation for censored data," *Journal of Computational and Graphical Statistics*, 1, 301-328.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1993), "Hazard regression," Technical Report No. 389, Department of Statistics, University of California, Berkeley, California.

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.

Miller, R. G. (1981), *Survival Analysis.* Wiley, New York.

O'Sullivan, F. (1988), "Fast computation of fully automated log-density and log-hazard estimators," *SIAM Journal of Scientific and Statistical Computing*, 9, (1988) 363-379.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, second edition, New York: Wiley.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461-464.

Senthilselvan, A. (1987), "Penalized likelihood estimation of hazard and intensity functions," *Journal of the Royal Statistical Society*, Ser. B, 49, 180-174.

Stone, C. J. (1990), "Large-sample inference for log-spline models," *The Annals of Statistics*, 18, 717-741.

Tanner, M. A. and Wong, W. H. (1983), "The estimation of the hazard function from randomly censored data by the kernel method," *The Annals of Statistics*, 11, 989-993.

Tanner, M. A. and Wong, W. H. (1984),"Data-based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis," *Journal of the American Statistical Association*, 79, 174-182.

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.

Whittemore, A. S. and Keller, J. B. (1986), "Survival estimation using splines," *Biometrics*, 42, 495-506.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

Appendix: Documentation of S Functions

Heft Hazard Estimation with Flexible Tails **Heft**

```
dheft(q, fit)
hheft(q, fit)
pheft(q, fit)
qheft(p, fit)
rheft(n, fit)
```

ARGUMENTS

`q:` vector of quantiles. Missing values (NAs) are allowed.

`p:` vector of probabilities. Missing values (NAs) are allowed.

`n:` sample size. If length(n) is larger than 1, then length(n) random values are returned.

`fit:` a list like the output from heft.fit.

VALUE

densities (dheft), hazard rates (hheft), probabilities (pheft), quantiles (qheft), or a random sample (rheft) from a fit obtained by heft.fit.

heft.fit Hazard Estimation with Flexible Tails **heft.fit**

heft.fit(data, delta, penalty, knots, leftlin=F, shift, leftlog, rightlog, maxknots, silent=T)

ARGUMENTS

`data:` vector of observations. Observations may or may not be right censored. All observations should be nonnegative.

`delta:` binary vector with the same length as data. Elements of data for which the corresponding element of delta is 0 are assumed to be right censored, elements of data for which the corresponding element of delta is 1 are assumed to be uncensored. If delta is missing, all observations are assumed to be uncensored.

`penalty:` the parameter to be used in the AIC criterion. The method chooses the number of knots that minimizes -2*loglikelihood+penalty*(dimension). The default is to use penalty=log(sample size) as in BIC. The effect of this parameter is summarized in heft.summary().

`knots:` ordered vector of values, which forces the method to start with these knots. If knots is not specified, a default knot-placement rule is employed.

`leftlin:` if leftlin is T an extra basis-function, which is linear to the left of the first knot, is included in the basis. If any of the data is exactly 0, the default of leftlin is T.

`shift:` parameter for the log terms. Default is quantile(data,.75).

`leftlog`: coefficient of log(x/(x+shift)), which must be greater than -1. (In particular, if leftlog equals zero no log(x/(x+shift)) term is included.) If leftlog is missing its maximum likelihood estimate is used. If any of the data is exactly zero, leftlog is set to zero.

`rightlog`: coefficient of log(x+shift), which must be greater than or equal to -1. (In particular, if rightlog equals zero no log(x+shift) term is included.) If rightlog is missing its maximum likelihood estimate is used.

`maxknots`: maximum number of knots allowed in the model (default is 4*length(data)$^{0.2}$).

`silent`: suppresses the printing of diagnostic output about basis functions added or deleted, Rao-statistics, Wald-statistics and log-likelihoods.

VALUE

The output is organized to serve as input for heft.plot, heft.summary, dheft, hheft, pheft, qheft and rheft.
The function returns a list with the following members:

`knots` : vector of the locations of the knots in the logspline model.

`logl`: the k-th element is the log-likelihood of the fit with k knots.

`thetak`: coefficients of the knot part of the spline. The k-th coefficient is the coefficient of $(x - t(k))_+^3$. If a coefficient is zero the corresponding knot was deleted from the model.

`thetap`: coefficients of the polynomial part of the spline. The first element is the constant term and the second element is the linear term.

`thetal`: coefficients of the logarithmic terms. The first element equals leftlog and the second element equals rightlog.

`penalty`: the penalty that was used.

`shift`: parameter used in the definition of the log terms.

`sample`: the sample size that was used.

`logse`: the standard errors of thetal.

`max`: the maximum element of data.

`ad`: vector indicating whether a model of this dimension was not fitted (2), was fitted during the addition stage (0) or during the deletion stage (1).

`logl`: matrix with two columns. The i-th element of the first column is the loglikelihood of the model of dimension i. The second column indicates whether this model was fitted during the addition stage (1) or during the deletion stage (0).

`sample`: sample size.

```
heft.plot(fit, n = 100, what = "d", add = F, ...)
```

ARGUMENTS

`fit:` a list like the output from heft.fit.

`n:` the number of equally spaced points at which to plot the fit.

`what:` what should be plotted: d (density), p (distribution function), s (survival function) or h (hazard function).

`add:` should the plot be added to an existing plot?

`...:` all regular plotting options as desired.

This function produces a plot of a heft fit at n equally spaced points roughly covering the support of the density. (Use xlim=c(from,to) to change the range of these points.)

EXAMPLES

```
fit <- heft.fit(time, delta, covs)
heft.plot(fit)
```



```
heft.summary(fit)
```

ARGUMENTS

`fit:` a list like the output from heft.fit.

VALUE

This function produces only printed output. The main body is a table with six columns:
the first column is a possible number of knots for the fitted model;
the second column is 0 if the model was fitted during the addition stage and 1 if the model was fitted during the deletion stage;
the third column is the log-likelihood for the fit;
the fourth column is -2*loglikelihood + penalty*(number of knots-1), which is the AIC criterion - heft.fit selected the model with the minimum value of AIC;
the fifth and sixth columns give the endpoints of the interval of values of penalty that would yield the model with the indicated number of knots. (NAs imply that the model is not optimal for any choice of penalty.)
At the bottom of the table the number of knots corresponding to the selected model is reported, as are the value of penalty that was used and the coefficients of the log-based terms in the fitted model and their standard errors.