Hazard Regression

Charles Kooperberg*
University of Washington

Charles J. Stone†
University of California at Berkeley

Young K. Truong ‡
University of North Carolina at Chapel Hill

Technical Report No. 389
April 23, 1993
Department of Statistics
University of California
Berkeley, California 94720

Abstract

Linear splines and their tensor products are used to estimate the conditional log-hazard function based on possibly censored, positive response data and one or more covariates. A fully automatic procedure involving the maximum likelihood method, stepwise addition, stepwise deletion and BIC is used to select the final model. A user interface based on S is described for obtaining estimates of the conditional hazard function, density function, distribution function and quantile function, for estimating the dependence of various quantities on the covariates, and for generating a random sample from the estimated conditional distribution.

KEY WORDS: Conditional hazard function; Covariates; Linear splines; MARS; Maximum likelihood; Model selection; Proportional hazards model; Tensor products, Time-varying coefficients.

*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.

‡This research was supported in part by a Research Council Grant from the University of North Carolina.

1 Introduction

Consider data involving a positive response variable, which may be (right-) censored, and one or more covariates. We think of the original, uncensored, response variable as having a conditional density function, given the values of the covariates, that is positive on $[0, \infty)$.

A basic assumption of the proportional hazards model (Cox (1972)) is that the conditional log-hazard function is an additive function of time and the vector of covariates or, equivalently, that the conditional hazard function is a multiplicative function of time and the vector of covariates. One of the main purposes of the present investigation is to develop a practical approach to modeling the conditional hazard function that does not depend on the validity of this assumption and which can, in fact, be used to assess departures from the assumption.

In this paper, linear splines and their tensor products are used to estimate the logarithm of the conditional 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 addition, stepwise deletion and BIC. The method is similar in spirit to MARS (Friedman (1991)). We also describe a user interface that makes our 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, various technical issues involved in the implementation of the procedure are discussed.

Traditionally, in the proportional hazards model and in in some other survival analysis models as well, the dependence of the survival time on the covariates is modeled fully parametrically, so that this regression function can be estimated independently of the baseline hazard function (see for example Cox and Oakes (1984), Kalbfleisch and Prentice (1980) or Miller (1981)). Typically the baseline hazard function is not estimated at all, but sometimes it is modeled parametrically. In particular, Etezadi-Amoli and Ciampi (1987) use polynomial splines to model this function.

Within the framework of the proportional hazard model, there have been a number of papers in which the dependence of the survival time on the covariates has been modeled using various nonparametric techniques, ignoring the baseline hazard function. In particular, Hastie and Tibshirani (1990) and O'Sullivan (1988) use smoothing splines, Sleeper and Harrington (1990) use B-splines and LeBlanc and Crowley (1992) use a regression tree algorithm. Hastie and Tibshirani (1993) introduce varying coefficient models. In the context of survival analysis this allows them to fit an additive model with time-varying coefficients of the covariates. Gray (1992) uses smoothing splines, and he allows time-varying coefficients and some interaction terms.

2 Linear Models for the Conditional Log-Hazard Function

Let T be a positive random variable whose distribution depends on M covariates $x_1, \ldots, x_M$, which range over the subsets $\mathcal{X}_1, \ldots, \mathcal{X}_M$ respectively of $\mathbf{R}$, each of which contains at least two members. Then $\boldsymbol{x} = (x_1, \ldots, x_M)$ ranges over the subset $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$ of $\mathbf{R}^M$. Let $f(\cdot|\boldsymbol{x})$ denote the dependence on $\boldsymbol{x}$ of the density function of T, which is assumed to exist and to be positive on $[0, \infty)$. Since in typical practical applications $x_1, \ldots, x_M$ are possible values of random variables, we refer to $f(\cdot|\boldsymbol{x})$ as the conditional density function of T given $\boldsymbol{x}$. Let $F(\cdot|\boldsymbol{x})$, $Q(\cdot|\boldsymbol{x})$, $\varphi(\cdot|\boldsymbol{x})$, $h(\cdot|\boldsymbol{x})$ and $\lambda(\cdot|\boldsymbol{x})$ denote the corresponding conditional distribution function, quantile function, log-density function, hazard function and log-hazard function, respectively.

Observe that $F(t|\boldsymbol{x}) = \int_0^T f(u|\boldsymbol{x})du$ for $t \geq 0$, $Q(F(t|\boldsymbol{x})|\boldsymbol{x}) = t$ for $t > 0$, and $F(Q(p|\boldsymbol{x})|\boldsymbol{x}) = p$ for $0 < p < 1$. Observe also that

$$\varphi(t|\boldsymbol{x}) = \log f(t|\boldsymbol{x}), \quad h(t|\boldsymbol{x}) = \frac{f(t|\boldsymbol{x})}{1 - F(t|\boldsymbol{x})} \text{ and } \lambda(t|\boldsymbol{x}) = \log h(t|\boldsymbol{x}), \qquad t \geq 0.$$

Moreover,

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

Since $F(t|\boldsymbol{x}) < 1$ for $0 \leq t < \infty$ and $\lim_{t\to\infty} F(t|\boldsymbol{x}) = 1$, we conclude that $\int_0^T \exp(\lambda(u|\boldsymbol{x}))du < \infty$ for $0 \leq t < \infty$ and $\int_0^\infty \exp(\lambda(t|\boldsymbol{x}))dt = \infty$. Furthermore, $h(t|\boldsymbol{x}) = \exp\lambda(t|\boldsymbol{x})$ for $t \geq 0$,

$$f(t|\boldsymbol{x}) = \exp(\lambda(t|\boldsymbol{x}))\exp\left(-\int_0^T \exp(\lambda(u|\boldsymbol{x}))du\right), \qquad t \geq 0,$$

and

$$\varphi(t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x}) - \int_0^T \exp(\lambda(u|\boldsymbol{x}))du, \qquad t \geq 0.$$

Let $1 \leq p < \infty$, let G be a p-dimensional linear space of functions on $[0, \infty) \times \mathcal{X}$ such that $g(\cdot|\boldsymbol{x})$ is bounded on $[0, \infty)$ for $g \in G$ and $\boldsymbol{x} \in \mathcal{X}$, and let $B_1, \ldots, B_p$ be a basis of this space. Motivated in part by Kooperberg and Stone (1993b), we consider the model

$$\lambda(t|\boldsymbol{x}; \boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j B_j(t|\boldsymbol{x}), \qquad t \geq 0, \tag{1}$$

for the conditional log-hazard function, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, and we refer to G as the corresponding model space. If none of the functions B_j, $1 \leq j \leq p$, in (1) depend on both t and $\boldsymbol{x}$, then (1) is a proportional hazards model (Cox (1972)).

Consider the special case of (1) in which $p = 1$ and $B_1 = 1$. Here

$$\lambda(t|\boldsymbol{\beta}) = \lambda(t|\boldsymbol{x};\boldsymbol{\beta}) = \beta_1, \qquad t \geq 0, \tag{2}$$

which does not depend on t or the vector of covariates. The corresponding conditional distribution of T given $\boldsymbol{x}$ is exponential with mean $\exp(-\beta_1)$. We refer to (2) as the minimal one-parameter model.

Given $\boldsymbol{\beta} \in \mathbf{R}^p$, set

$$h(t|\boldsymbol{x};\boldsymbol{\beta}) = \exp(\lambda(t|\boldsymbol{x};\boldsymbol{\beta})), \qquad t \geq 0,$$

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

and

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

Observe that $f(\cdot|\boldsymbol{x};\boldsymbol{\beta})$ is a positive density function on $[0,\infty)$. The corresponding distribution function and quantile function are given, respectively, by

$$F(t|\boldsymbol{x};\boldsymbol{\beta}) = \int_0^T f(u|\boldsymbol{x};\boldsymbol{\beta})du, \qquad t \geq 0,$$

and $Q(\cdot|\boldsymbol{x};\boldsymbol{\beta}) = F^{-1}(\cdot|\boldsymbol{x};\boldsymbol{\beta})$. Set $\varphi(y,1|\boldsymbol{x};\boldsymbol{\beta}) = \varphi(y|\boldsymbol{x};\boldsymbol{\beta})$ and $\varphi(y,0|\boldsymbol{x};\boldsymbol{\beta}) = \log(1 - F(y|\boldsymbol{x};\boldsymbol{\beta}))$ for $y \geq 0$. Then

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

$$\frac{\partial}{\partial\beta_j}\varphi(y,\delta|\boldsymbol{x};\boldsymbol{\beta}) = \delta B_j(y|\boldsymbol{x}) - \int_0^y B_j(u|\boldsymbol{x})\exp(\lambda(u|\boldsymbol{x};\boldsymbol{\beta}))du,$$
$$1 \leq j \leq p,\ y \geq 0 \text{ and } \delta \in \{0,1\},$$

and

$$\frac{\partial^2}{\partial\beta_j\partial\beta_k}\varphi(y,\delta|\boldsymbol{x};\boldsymbol{\beta}) = -\int_0^y B_j(u|\boldsymbol{x})B_k(u|\boldsymbol{x})\exp(\lambda(u|\boldsymbol{x};\boldsymbol{\beta}))du,$$
$$1 \leq j,k \leq p,\ y \geq 0 \text{ and } \delta \in \{0,1\}.$$

It follows from the last result that $\varphi(y,\delta|\boldsymbol{x};\cdot)$ is a concave function on $\mathbf{R}^p$ for $y \geq 0$, $\delta \in \{0,1\}$ and $\boldsymbol{x} \in \mathcal{X}$.

3 Maximum Likelihood Estimation

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 $\delta = \text{ind}(T \leq C)$; the random variable Y is said to be uncensored or censored according as $\delta = 1$ or $\delta = 0$.

Consider n such individuals. For $1 \leq i \leq n$ 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 $\delta_i = \text{ind}(T_i \leq C_i)$. It is assumed that $T_1, \ldots, T_n, (C_1, \ldots, C_n)$ are conditionally independent given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

The log-likelihood function corresponding to the observed data $(Y_i, \delta_i, \boldsymbol{x}_i)$, $1 \leq i \leq n$, and the linear model for the conditional log-hazard function that was discussed in the previous section is given by

$$l(\boldsymbol{\beta}) = \sum_i \varphi(Y_i, \delta_i|\boldsymbol{x}_i; \boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathbf{R}^p,$$

which is a concave function on $\mathbf{R}^p$. Moreover,

$$\frac{\partial}{\partial \beta_j} l(\boldsymbol{\beta}) = \sum_i \frac{\partial}{\partial \beta_j} \varphi(Y_i, \delta_i|\boldsymbol{x}_i; \boldsymbol{\beta}), \qquad 1 \leq j \leq p \text{ and } \boldsymbol{\beta} \in \mathbf{R}^p,$$

and

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k} l(\boldsymbol{\beta}) = \sum_i \frac{\partial^2}{\partial \beta_j \partial \beta_k} \varphi(Y_i, \delta_i|\boldsymbol{x}_i; \boldsymbol{\beta}), \qquad 1 \leq j, k \leq p \text{ and } \boldsymbol{\beta} \in \mathbf{R}^p.$$

The maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ is given as usual by $l(\widehat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}} l(\boldsymbol{\beta})$ and the log-likelihood of the model is given by $\hat{l} = l(\widehat{\boldsymbol{\beta}})$. The corresponding maximum likelihood estimates of the conditional log-hazard function, hazard function, density function, distribution function and quantile function are given by $\hat{\lambda}(t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x}; \widehat{\boldsymbol{\beta}})$, $\hat{h}(t|\boldsymbol{x}) = h(t|\boldsymbol{x}; \widehat{\boldsymbol{\beta}})$ and so forth.

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

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

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

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} - 2^{-\mu} [\mathbf{H}(\widehat{\boldsymbol{\beta}}^{(m)})]^{-1} \mathbf{S}(\widehat{\boldsymbol{\beta}}^{(m)}),$$

where μ is the smallest nonnegative integer such that

$$l(\widehat{\boldsymbol{\beta}}^{(m)} - 2^{-\mu} [\mathbf{H}(\widehat{\boldsymbol{\beta}}^{(m)})]^{-1} \mathbf{S}(\widehat{\boldsymbol{\beta}}^{(m)})) \geq l(\widehat{\boldsymbol{\beta}}^{(m)} - 2^{-\mu-1} [\mathbf{H}(\widehat{\boldsymbol{\beta}}^{(m)})]^{-1} \mathbf{S}(\widehat{\boldsymbol{\beta}}^{(m)})).$$

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

4 Linear Splines and their Tensor Products

The specific model spaces that are considered in this paper involve splines and their tensor products. In order to avoid numerous numerical integrations with respect to t and added complications in context of stepwise knot addition (see (34) and (35) of Friedman, 1991), we confine our attention to linear (rather than quadratic or cubic) splines. In the present context, it is convenient to define a model space by listing its basis functions.

Let K_0 be a nonnegative integer; if $K_0 \geq 1$ let t_k, $1 \leq k \leq K_0$ be distinct positive numbers, consider the basis functions $B_{0k}(t) = (t_k - t)_+$, $1 \leq k \leq K_0$, where $t_+ = \max(t, 0)$. Next, for $1 \leq m \leq M$, let K_m be an integer with $K_m \geq -1$; if $K_m \geq 0$, consider the basis function $B_{m0}(x_m) = x_m$; if $K_m \geq 1$, let $x_{mk}, 1 \leq k \leq K_m$ be distinct real numbers and consider the basis functions $B_{mk}(x_m) = (x_m - x_{mk})_+$, $1 \leq k \leq K_m$, where $x_+ = \max(x, 0)$.

Let G be the linear space having basis functions 1, $B_{0k}(t)$ for $1 \leq k \leq K_0$, $B_{mk}(x_m)$ for $1 \leq m \leq M$ and $0 \leq k \leq K_m$, and perhaps certain tensor products of two such basis functions. It is required that if $B_{mj}(x_m)B_{0k}(t)$ be among the basis functions for some $j \geq 1$, then $B_{m0}(x_m)B_{0k}(t) = x_m B_{0k}(t)$ be among the basis functions. Similarly, it is required that if $B_{lj}(x_l)B_{mk}(x_m)$ be among the basis functions for some $j \geq 1$, then $B_{l0}(x_l)B_{mk}(x_m) = x_l B_{mk}(x_m)$ and hence $x_l x_m$ be among the basis functions. Such a linear space G is said to be *allowable*.

5 Model Selection

Initially, we fit the minimal one-parameter model (see Section 9.1). Then we proceed with stepwise addition. Here we successively replace the $(p-1)$-dimensional allowable space G_0 by a p-dimensional allowable space G containing G_0 as a subspace, choosing among the various candidates for a new basis function by a heuristic search (described in Section 9.2) that is designed approximately to maximize the absolute value of the corresponding Rao statistic.

Specifically, let $\hat{\boldsymbol{\beta}}^{(o)}$ be the maximum likelihood estimate of the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ corresponding to G, but subject to the constraint that the corresponding estimate of the conditional log-hazard function be in G_0, and let β_p be the coefficient of the basis function that is added in going from G_0 to G. Then the Rao statistic for testing the hypothesis that the conditional log-hazard function is in G_0 is given by $R = [\mathbf{S}(\hat{\boldsymbol{\beta}}^{(o)})]_p / \sqrt{[\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}^{(o)})]_{pp}}$, where $\mathbf{I}(\hat{\boldsymbol{\beta}}^{(o)}) = -\mathbf{H}(\hat{\boldsymbol{\beta}}^{(o)})$ with $\mathbf{S}(\cdot)$ and $\mathbf{H}(\cdot)$ corresponding to G. (Here R is the signed square root of the Rao statistic as usually defined; see (6e.3.6) of Rao (1973).)

Upon stopping the stepwise addition stage (according to a rule that is described in Section 9.2), we proceed to stepwise deletion. Here we successively replace the p-dimensional allowable space G by a $(p-1)$-dimensional allowable subspace G_0 until we arrive at the minimal one-parameter model, at each step choosing that candidate basis function to delete

whose Wald statistic is smallest in magnitude.

Specifically, let $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ be the maximum likelihood estimate of the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ corresponding to G, where β_p is the coefficient of the basis function that is deleted in going from G to G_0. Then the standard error $\mathrm{SE}(\hat{\beta}_p)$ of $\hat{\beta}_p$ is the positive square root of the pth diagonal entry of $[\mathbf{I}(\widehat{\boldsymbol{\beta}})]^{-1} = -[\mathbf{H}(\widehat{\boldsymbol{\beta}})]^{-1}$ with $\mathbf{H}(\cdot)$ corresponding to G, and the Wald statistic for testing the hypothesis that the conditional log-hazard function is in G_0 equals $\hat{\beta}_p/\mathrm{SE}(\hat{\beta}_p)$.

During the combination of stepwise addition and stepwise 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 $\mathrm{AIC}_{\alpha,\nu} = -2\hat{l}_\nu + \alpha p_\nu$ 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 $\mathrm{AIC}_{\alpha,\nu}$. In light of Kooperberg and Stone (1992, 1993b) and our experience in the present investigation, we recommend choosing $\alpha = \log(n)$ as in the Bayesian information criterion (BIC) due to Schwarz (1978).

6 User Interface

A program for implementing hazard regression (HARE) 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: `dhare`, `hhare`, `phare`, `qhare`, `rhare`, `hare.fit`, `hare.summary` and `hare.plot`. (Detailed documentation of each of these functions is included in the Appendix to this paper.) The functions `dhare`, `phare`, `qhare`, `rhare` 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 `dhare` gives the (estimated) conditional density function, `phare` gives the conditional distribution function, `qhare` gives the conditional quantile function, and `rhare` gives a random sample from the conditional distribution. The function `hhare` gives the conditional hazard function, `hare.fit` performs the model fitting and model selection tasks and supplies the modest output that is used as input to `dhare`, `hhare` and so forth. The function `hare.summary`, uses the output of `hare.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, `hare.plot` uses the output of `hare.fit` directly to produce a plot of the conditional density, distribution, survival or hazard function.

[1]HARE software is available from statlib. Send an email with the body `send hare from S` to statlib@stat.cmu.edu

7 Examples

In this section, we illustrate various ways of using HARE by analyzing three datasets. These analyses are not meant to be definitive.

7.1 Lung Cancer Data

Our first example concerns data from a Veteran's Administration lung cancer trial. The data has been examined in Kalbfleisch and Prentice (1980) and various other publications. The response is survival time in days; the predictors are treatment (1=standard, 2=test), cell type (squamous, small, adeno and large), a performance index (between 0 and 100, higher scores are considered better), age and prior therapy (0=no, 1=yes). There are 137 cases, of which 9 are censored.

When we applied the HARE algorithm to this data, we got a model with nine basis functions, which is summarized in Table 1 below. Note that two of the nine basis functions in this model involve both time and a covariate (for one of these functions the covariate is performance status, for the other it is the indicator of cell type adeno), suggesting that a proportional hazards model might not be appropriate.

TABLE 1. First HARE analysis of the lung cancer data.

Basis function	Coefficient	Standard error
1	−9.830	2.26
Performance status	0.250	0.108
$(\text{Performance status} - 20)_+$	−0.260	0.108
Cell type: small cell	−1.39	0.634
Cell type: adeno	2.43	0.47
$(156 - t)_+$	0.0245	0.0058
(Performance status)×(Cell type: small cell)	0.0387	0.0112
$(\text{Performance status})\times(156 - t)_+$	−0.000433	0.000095
$(\text{Cell type: adeno})\times(156 - t)_+$	−0.0125	0.0045

The standard errors in the above table are obtained in the usual parametric manner as the square root of the diagonal entries of the inverse of the estimated information matrix. Thus, since they do not take the highly adaptive nature of HARE into account, they should be regarded as merely suggestive.

The default HARE analysis should not automatically be accepted as definitive. In particular, when we apply this procedure one of the first things that we typically do is to transform time such that the unconditional log-hazard function of the transformed time approximately be constant. To achieve this we use the Hazard Estimation with Flexible Tails program

(HEFT, Kooperberg and Stone (1993b)). In this manner we get a fitted model having the form

$$\hat{h}(t|\boldsymbol{x}) = \hat{h}_0(t) \exp\left(\sum_j \hat{\beta}_j B_j(\hat{q}_0(t)|\boldsymbol{x})\right),$$

for the conditional hazard function, where $\hat{h}_0$ is the HEFT fit to the unconditional hazard function and $\hat{q}_0 = -\log(1 - \hat{F}_0)$, with $\hat{F}_0$ being the distribution function corresponding to $\hat{h}_0$.

When we applied HEFT to the lung cancer data, as described in Kooperberg and Stone (1993b), we obtained

$$\hat{h}_0(t) \approx e^{-1.643}(t + 145.75)^{-0.583}$$

(145.75 is the upper quartile of the uncensored survival times). The results of the application of HARE to the transformed data are summarized in Table 2 below. In Figure 1 we show the coefficient of performance status and the hazard function for a person with specified values of the relevant variables for the fits with and without the transformation of time using HEFT.

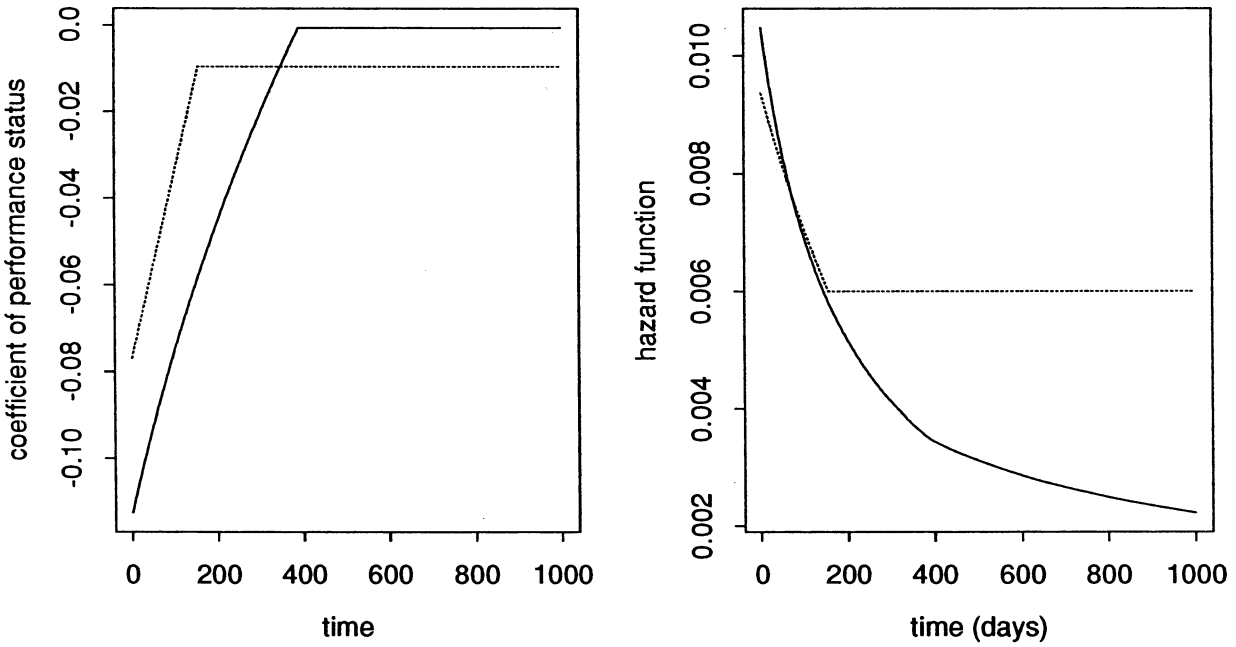

Fig. 1. Fitted coefficient of performance status as a function of time and fitted hazard function for a person with cell type squamous and performance status 40. Solid=transformed, dashed=untransformed.

The fit in Table 2 is fairly similar to the fit in Table 1 above with respect to the basis functions. However, as can be seen in Figure 1 below, the fits are quite different as far as the estimated conditional hazard rate is concerned. This difference is caused by the fact that HARE, when applied to untransformed data, will always give an estimate for the conditional hazard rate that is constant beyond the last knot in time.

TABLE 2. HARE analysis of the transformed lung cancer data.

Basis function	Coefficient	Standard error
1	−7.06	2.60
Performance status	0.272	0.110
$(\text{Performance status} - 20)_+$	−0.230	0.108
$(\text{Performance status} - 85)_+$	−0.273	0.117
Cell type: small cell	−1.16	0.65
Cell type: adeno	2.239	0.622
$(2.665 - \hat{q}_0(t))_+$	2.24	0.62
(Performance status)×(Cell type: small cell)	0.0339	0.0115
(Performance status)$\times(2.665 - \hat{q}_0(t))_+$	−0.0421	0.0095
(Cell type: adeno)$\times(2.665 - \hat{q}_0(t))_+$	−2.00	0.54

Hastie and Tibshirani (1993) analyze the same data. In their analysis, the coefficient of performance status varies with time, but no other interactions enter the model. Kooperberg and Stone (1993a) show a similar model for the data using HARE after a transformation of time by HEFT. This fit, summarized in Table 3 below, was obtained by using the option `linear` for performance status, which prevents HARE from entering any knots for performance status, and the option `include` for the combination (time, performance status), which makes basis functions that depend on time and performance status the only allowable interactions in the model. (The options of `hare.fit` are described in detail in the appendix.) The function $\hat{q}_0$ is as above.

TABLE 3. HARE analysis forcing a model similar to the model in Hastie and Tibshirani (1993)

Basis function	Coefficient	Standard error
1	0.229	0.617
Performance status	−0.00216	0.00085
Cell type: small cell	0.739	0.222
Cell type: adeno	0.963	0.255
$(1.032 - \hat{q}_0(t))_+$	2.25	0.77
(Performance status)$\times(1.032 - \hat{q}_0(t))_+$	−0.0518	0.0126

Are the two interaction terms in Table 2 but not in Table 3 real or spurious? In order to investigate this question, we carried out a small-scale simulation study. First we estimated the distribution of the censoring times under the assumption that the censoring was independent of the covariates (an assumption which we investigate in more detail for our third example, the breast cancer data). We used HEFT on the original survival times, but used $1 - \delta$ instead of δ as was done in the calculations leading to Figure 4 in Kooperberg and Stone (1992). HEFT yielded that a constant hazard function, corresponding to an exponential distribution with mean 1851, fitted well. (Since there were only 9 censored datapoints,

it is not surprising that we obtained a very simple estimate for the unconditional hazard function.)

For each simulation we first generated a new set of survival times $T_i^* = \hat{q}_0^{-1}(t_i)$, $1 \leq i \leq 137$, where t_i, $1 \leq i \leq 137$, is an independent sample generated using `rhare`, the fit summarized in Table 3 and the same covariates as in the original data. Then we generated the censoring times C_i^*, $1 \leq i \leq 137$, as a random sample from the exponential distribution with mean 1851. For each i we set $Y_i^* = \min(T_i^*, C_i^*)$ and $\delta_i^* = \text{ind}(T_i^* \leq C_i^*)$. Using `heft.fit` with the default options, we transformed Y_i^*, $1 \leq i \leq 137$, after which we used `hare.fit`, also with the default options, to fit a model to the conditional log-hazard function of $(\hat{q}_0^*(Y_i^*), \delta_i^*, \boldsymbol{x}_i)$, $1 \leq i \leq 137$.

We carried out 100 such simulations. In Table 4 the fitted models are summarized with respect to the variables involved in the two-dimensional (tensor product) basis functions, with differences in the coefficients and knot locations being ignored.

TABLE 4. Summary of the simulation study for the lung cancer data.

Interactions in model	Frequency
No interactions	16
Only a $(\hat{q}_0(t))\times$(Performance status) interaction	57
One interaction, not $(\hat{q}_0(t))\times$(Performance status)	10
A $(\hat{q}_0(t))\times$(Performance status) interaction and one other interaction	12
Two interactions, but none $(\hat{q}_0(t))\times$(Performance status)	3
Three or more interactions	2
Of the 57 simulations that yielded the correct interactions, 23 had six basis functions that coincided with those in the model in Table 3 with respect to the variables involved.	

From Table 4 we see that in only 2 out of 100 simulations did the model fit by HARE have 3 or more interactions. Since the models in Tables 1 and 2 both have three interactions, it seems reasonable to conclude that more interactions than the one in Table 3 should be included in the model.

7.2 PBC Data

Our second example illustrates many of the features of `hare.fit` that facilitate the search for the model that best fits the data. It involves data from a double-blinded randomized trial involving primary biliary cirrhosis of the liver (PBC). The data is discussed extensively in Fleming and Harrington (1991). There were 312 patients in the clinical trial. The response is survival time (days), and there are 17 covariates listed in Fleming and Harrington (1991). Of the 312 observations 187 were censored. We took the logarithm of five of the covariates, serum bilirubin, alkaline phosphatase, urine copper, SGOT and Triglycerides, since the empirical distributions of these quantities are highly skewed to the right.

As the first step in the analysis of the PBC data, we used HEFT to estimate the unconditional hazard function, getting $\hat{h}_0(t) \approx \exp(-8.498)$. Therefore no transformation was needed to make the unconditional hazard function approximately constant.

We continued our analysis by applying the HARE algorithm with the default options to the 274 cases that had no missing values for any of the covariates. This analysis yielded a model with 13 basis functions. None of these basis functions involved the covariates treatment, serum cholesterol, log(triglycerides) or platelet count, each of which had one or more of missing values. No matter which options for HARE we chose, none of these covariates entered the model. Therefore, in further analysis, we excluded these four covariates and included all 310 of the 312 cases that were complete with respect to the remaining 13 covariates. The other two cases have missing values for alkaline phosphatase. Since log(alkaline phosphatase) did appear frequently in the initial HARE fits, we excluded those two cases during the rest of the analysis. (There are other methods, such as imputation, to deal with missing data; for an overview see Little and Rubin (1987).)

Applying HARE to these 310 cases and 13 covariates, we got a fairly complicated model with 15 basis functions, which is summarized in Table 5 below. Since the model selection algorithm described in Section 5 does not guarantee an optimal model, it is reasonable to search for a model that either fits better with respect to AIC or fits about as well but is easier to interpret. HARE has several options that facilitates this search process.

TABLE 5. HARE analysis of the PBC data - 310 cases, 13 covariates.

Basis function	Coefficient	Standard error
1	−18.1	3.1
age	0.0486	0.0099
$(\text{age}-71.9)_+$	−0.503	0.230
ascites	−0.284	0.517
edema	0.149	0.410
log(serum bilirubin)	−7.56	2.61
$(\log(\text{serum bilirubin})+0.916)_+$	8.60	2.64
albumin	−0.848	0.239
log(alkaline phosphatase)	0.514	0.141
prothrombin time	0.0516	0.1293
$(1170-t)_+$	−0.00770	0.00232
$(4079-t)_+$	−0.000469	0.000140
(ascites) × (edema)	1.88	0.73
$(1170-t)_+ \times$log(serum bilirubin)	−0.000729	0.000240
$(1170-t)_+ \times$(prothrombin time)	0.000667	0.000196

It is possible to specify the maximum number of basis functions in a model, overriding the default P_{max} (Section 9.2). For the PBC data changing the option `maxdim` in `hare.fit` consistently resulted in the same fitted model as described above. It is also possible to use

a model that was previously fitted using `hare.fit` as the starting value for a new search. This is useful when combined with the output of `hare.summary`, which indicates whether the various models were fitted during the addition stage or during the deletion stage. In the latter case, a user could specify the model fit by HARE as the starting point for a new fit. If the resulting model is different from the starting model, the new model has a lower AIC. If the original model was fitted during the part of the process when basis functions were added, HARE will inevitable return the same model. The later was the case for the PBC data.

The `hare.summary` command also provides information about the influence of the choice of the penalty parameter α. Table 6 consists of a part of the output of `hare.summary`, when applied to the model from Table 5, above.

TABLE 6. Part of the output of `hare.summary`, when applied to the model from Table 5.

dim	A/D	loglik	AIC	penalty	
				min	max
1	Add	−1180.79	2367.31	113.84	inf
2	Add	−1123.87	2259.20	27.86	113.84
3	Add	−1110.50	2238.22	na	na
4	Del	−1096.00	2214.95	17.99	27.86
5	Del	−1087.01	2202.69	10.47	17.99
6	Del	−1081.77	2197.96	7.90	10.47
7	Del	−1078.54	2197.24	na	na
8	Add	−1075.81	2197.51	na	na
9	Add	−1069.92	2191.46	5.83	7.90
10	Add	−1067.78	2192.94	na	na
11	Add	−1064.42	2191.94	na	na
12	Add	−1061.70	2192.23	na	na
13	Del	−1058.29	2191.15	na	na
14	Del	−1055.61	2191.53	na	na
15	Add	−1052.42	2190.89	5.51	5.83
16	Add	−1049.97	2191.73	na	na
17	Add	−1047.38	2192.29	na	na
18	Add	−1044.15	2191.56	0.00	5.51

For each possible dimension of the model, the output shown in Table 6 indicates whether the best model of that dimension was fitted during the addition stage or the deletion stage and shows the log-likelihood and its AIC value with the choice of the penalty parameter α used in `hare.fit`. The last two columns indicate the effect of a different choice of α. For example, with $n = 310$, the default value of α is $\log 310 \approx 5.74$. As can be seen from Table 6, any α between 5.51 and 5.83 would have resulted in the same model with 15 basis functions. However, if α were to increased to 6, HARE would have fitted a model with 9 basis functions.

The model that was obtained with HARE using `penalty=6` resulted in an additive model. Besides the constant basis function, the other eight basis functions were as follows: a knot in time, age, a knot in age, ascites, log(serum bilirubin), albumin, log(alkaline phosphatase) and prothrombin time.

This led us to fit a model, using the default $\alpha = \log 310$, while forcing the model to be additive (`additive=T`). The resulting fit is summarized in Table 7 below.

TABLE 7. HARE analysis of the PBC data - forcing an additive model.

Basis function	Coefficient	Standard error
1	−18.9	3.0
age	0.0480	0.0100
$(\text{age}-71.9)_+$	−0.502	0.218
log(serum bilirubin)	−7.20	2.60
$(\log(\text{serum bilirubin})+0.916)_+$	8.06	2.62
albumin	−1.03	0.21
log(alkaline phosphatase)	0.485	0.140
prothrombin time	0.274	0.085
$(4079-t)_+$	−0.000627	0.000096

As it turned out, this model has a lower value of AIC than the model in Table 5 (2189.83 versus 2190.89). Further analysis could not improve upon this model. Note that the model in Table 7 is a proportional hazards model. As such, we can compare it with the models obtained in Fleming and Harrington (1991). In their Table 4.4.3c, they end up fitting a model that includes age, albumin, serum bilirubin, edema and prothrombin time. Thus there is a discrepancy in that we include log(alkaline phosphatase) but not the indicator of edema.

7.3 Breast Cancer Data

The dataset for our last example is considerable larger than those for the two previous examples. The data, discussed in Gray (1992), come from 6 breast cancer trials conducted by the Eastern Cooperative Oncology Group[2]. There were 2404 patients in these studies. The response is survival time (years). There are six covariates, estrogen receptor status (ER: 0 is 'negative', 1 is 'positive'), the number of positive auxiliary lymph nodes at diagnosis, size of the primary tumor (in cm), age at entry, menopause (0 is premenopause, 1 is postmenopause) and body mass index (BMI: defined as weight/height^2 in kg/m^2). Since the empirical distribution of the number of nodes is highly skewed to the right, we used log(number of nodes), instead of the number itself in our analysis. Of the 2404 cases, 1116 were uncensored and 1288 were censored. There were no missing values for any of the covariates.

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

Again, we started our analysis by estimating the unconditional log-hazard function using HEFT, getting the estimate shown in the left side of Figure 2. We then transformed time, as described in Section 7.1, so that the unconditional log-hazard function of the transformed time be approximately equal to one. Specifically, we set $\hat{q}_0(t) = -\log(1 - \widehat{F}_0(t))$, where $\widehat{F}_0$ is the distribution function corresponding to the HEFT estimate of the unconditional hazard function. The function $\hat{q}_0(t)$ is shown in the right side of Figure 2.

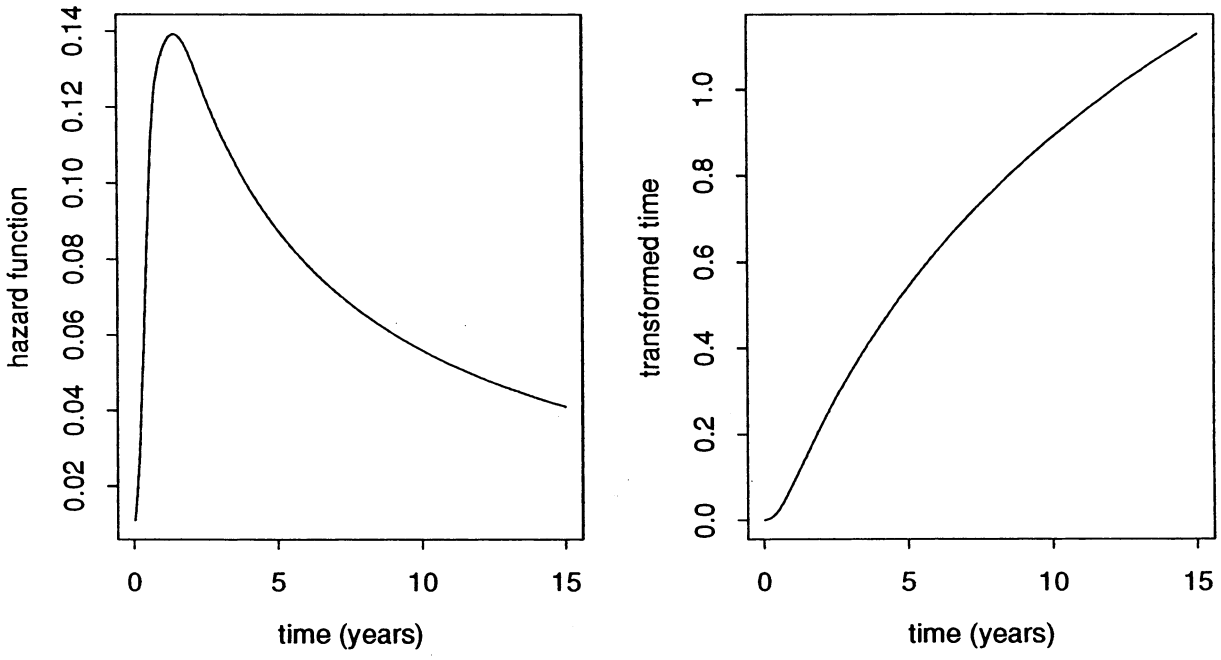

Fig. 2. Estimated unconditional hazard function and the corresponding transformation of time using HEFT for the breast cancer data.

When we applied HARE to the transformed data, we obtained the fit summarized in Table 8. Further analysis, along the lines of that in Section 7.2, did not yield a better fit. When HARE was applied to the untransformed data, the resulting fit was similar, but included one more knot in time. In Figure 3 we show the hazard function and the survival function for a person with specified values of the relevant covariates for the fits with and without the transformation of time using HEFT. Note that we chose values of the covariates that are all close to the median value as observed in the study.

A plausible assumption in survival analysis is that the censoring time is independent of the vector of covariates. This assumption can be investigated using HEFT and HARE by treating $T_1, \ldots, T_n$ as the censoring times and $C_1, \ldots, C_n$ as the survival times; that is by applying these procedures to $(Y_i, 1 - \delta_i, \boldsymbol{x}_i)$, $1 \leq i \leq n$.

The estimated fit to the censoring distribution that we obtained using HEFT is surprisingly complicated, the corresponding density estimate being shown in Figure 4. After applying HARE to the transformed data, we obtained a model with two basis functions,

TABLE 8. HARE analysis of the transformed breast cancer data 2404 cases, 6 covariates.

Basis function	Coefficient	Standard error
1	−0.0443	0.3990
ER	0.426	0.119
log(nodes)	0.686	0.070
size	0.158	0.035
age	−0.0401	0.0093
$(\text{age}-43)_+$	0.0408	0.0115
menopause	0.409	0.105
$(0.194 - \hat{q}_0(t))_+$	−6.58	1.33
$(0.514 - \hat{q}_0(t))_+$	2.66	0.41
log(nodes)×size	−0.0650	0.0181
$(0.514 - \hat{q}_0(t))_+ \times$ER	−2.91	0.39
$(0.194 - \hat{q}_0(t))_+ \times$size	0.878	0.266

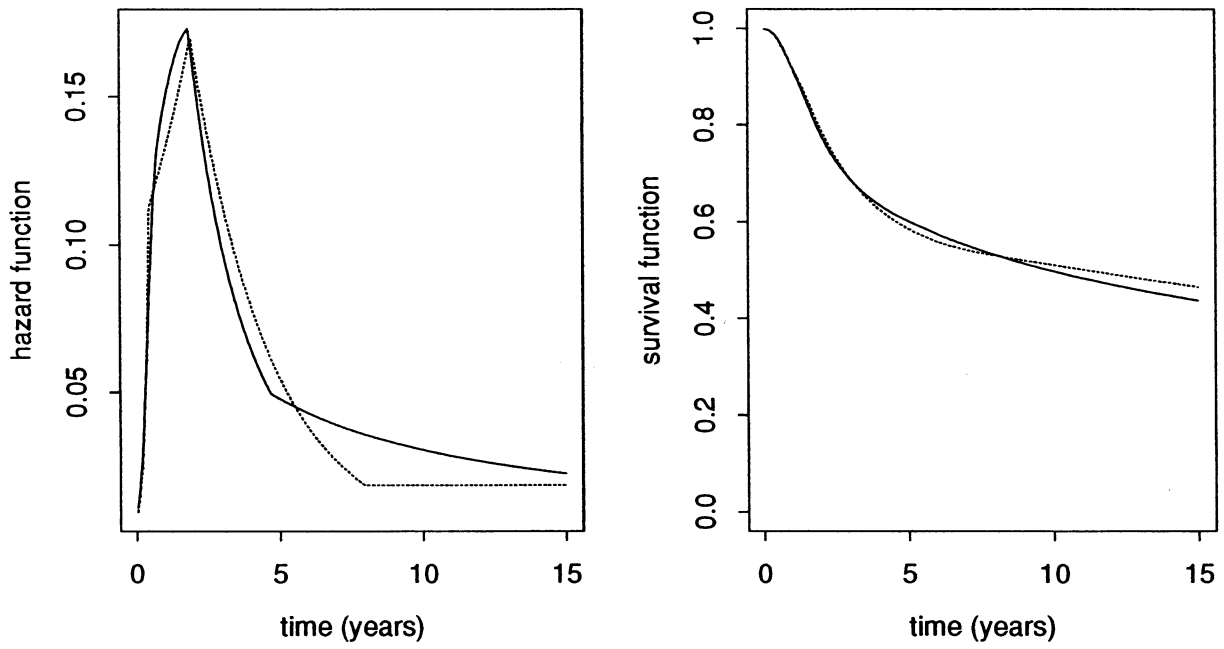

Fig. 3. Fitted hazard and survival functions for a premenopausal woman of age 50 with negative estrogen receptor status, 4 nodes, body mass index 25 and tumor size 3 cm. Solid=transformed using HEFT, dashed=untransformed.

summarized in Table 9 below. This analysis suggests that the conditional distribution of the censoring times depends on whether a woman is premenopausal or postmenopausal. The hazard of censoring is about 27% larger for postmenopausal women.

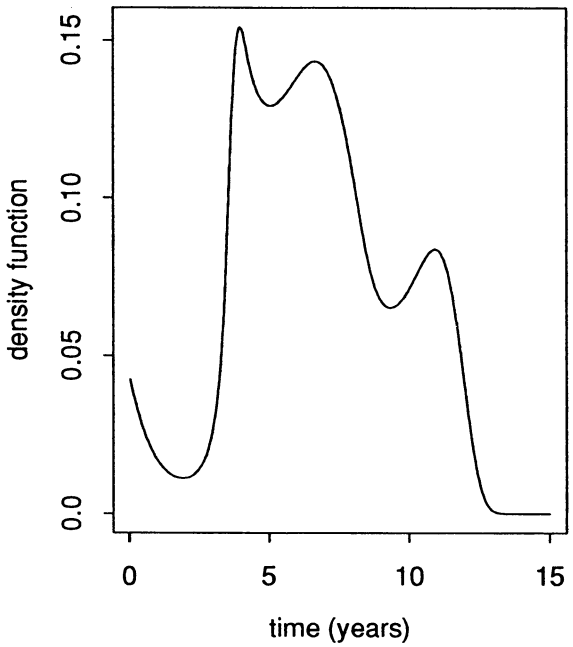

Fig. 4. Estimated unconditional density function for the censoring distribution for the breast cancer data.

TABLE 9. HARE analysis of censoring times for the breast cancer data

Basis function	Coefficient	Standard error
1	−0.109	0.039
menopause	0.236	0.056

In order to investigate the sensitivity of the fit summarized in Table 8 to random fluctuations in the data, we carried out the following simulation 200 times. First we generated a new set of survival times $T_i^* = \hat{q}_0^{-1}(t_i)$, $1 \leq i \leq 2404$, where t_i, $1 \leq i \leq 2404$, is an independent sample generated using `rhare`, the fit summarized in Table 8 and the same covariates as in the original data, while $\hat{q}_0^{-1}$ is the inverse of the transformation displayed in the right hand side of Figure 2. Then we generated the censoring times C_i^*, $1 \leq i \leq 2404$, as a random sample from the distribution corresponding to the density displayed in Figure 4. For each i we set $Y_i^* = \min(T_i^*, C_i^*)$ and $\delta_i^* = \text{ind}(T_i^* \leq C_i^*)$. Using `heft.fit` with the default options, we transformed Y_i^*, $1 \leq i \leq 2404$, after which we used `hare.fit`, also with the default options, to fit a model to the conditional log-hazard function of $(\hat{q}_0^*(Y_i^*), \delta_i^*, \boldsymbol{x}_i)$, $1 \leq i \leq 2404$.

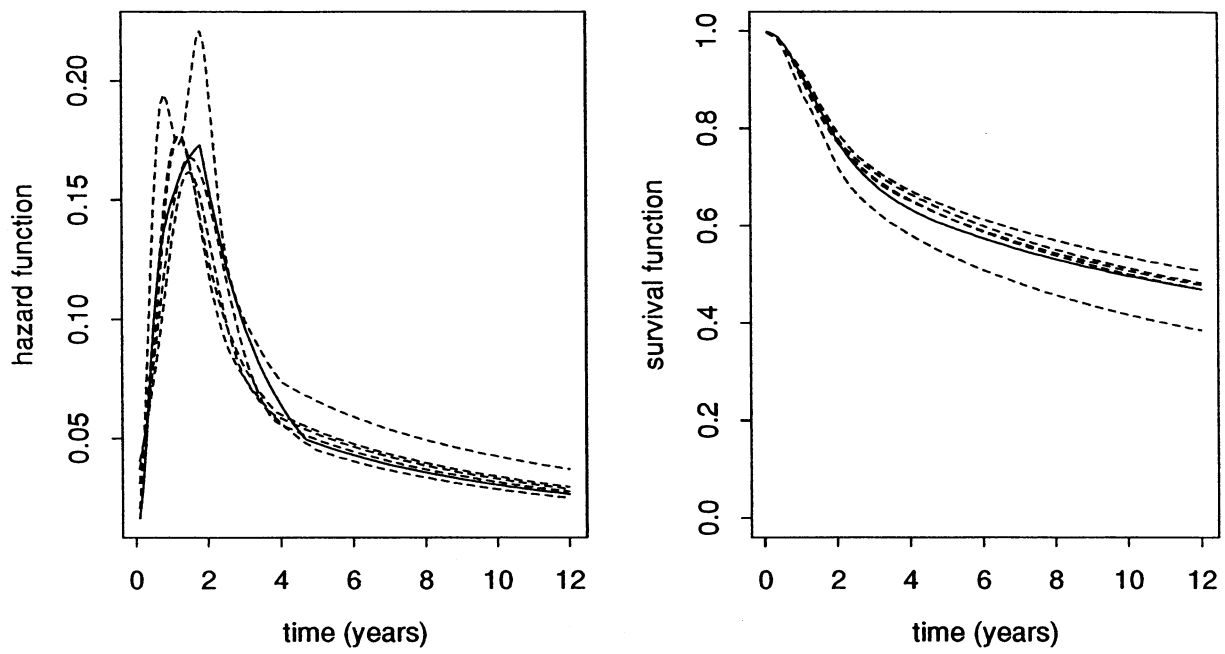

Fig. 5. Conditional hazard and survival functions for the fit of Table 8 (solid) and five random samples from this model (dashed). Same covariates as in Figure 3.

In Figure 5 we show, for the fit from Table 8 and five randomly selected simulations from that fit, the fitted conditional hazard and survival functions for the same vector of covariates as used in Figure 3. In Figure 6 we summarize these same quantities for all 200 simulations. In particular for every time t we computed the 2.5th and 97.5th percentile of the simulated fit to the conditional hazard and survival functions. At each time 95% of the simulations fell in the gray band (the solid line is again the fit from Table 8).

Figures 7 and 8 summarize the effect of some of the covariates. The bootstrap bands in these figures are constructed as in Figure 6. In the left side of Figure 7 we show log(hazard ratio) $\hat{\lambda}(t|\boldsymbol{x}_1) - \hat{\lambda}(t|\boldsymbol{x}_2)$ as a function of time; here $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are identical to the vector of covariates used in Figure 3, except that Estrogen Receptor status equals 1 (ER is positive) in $\boldsymbol{x}_1$ and it equals 0 (ER is negative) in $\boldsymbol{x}_2$. The right side of Figure 7 displays the effect of the number of nodes on the log hazard. Specifically, in this figure we show

$$\begin{aligned} g(x) &= \hat{\lambda}(2|\text{nodes} = x,\ \text{Age} = 50,\ \text{ER} = 1,\ \text{BMI} = 25,\ \text{size} = 3,\ \text{menopause} = 1) \\ &- \hat{\lambda}(2|\text{nodes} = 4,\ \text{Age} = 50,\ \text{ER} = 1,\ \text{BMI} = 25,\ \text{size} = 3,\ \text{menopause} = 1). \end{aligned}$$

That is, we show log(hazard ratio) when time is 2 years and all covariates are kept fixed at the same value as in Figure 3, except that the number of nodes is allowed to vary and is compared with nodes = 4. The fact that both the estimate corresponding to Table 8 and the width of the 95% bootstrap band are 0 when nodes = 4 is a consequence of the fact that $g(4) = 0$ by definition.

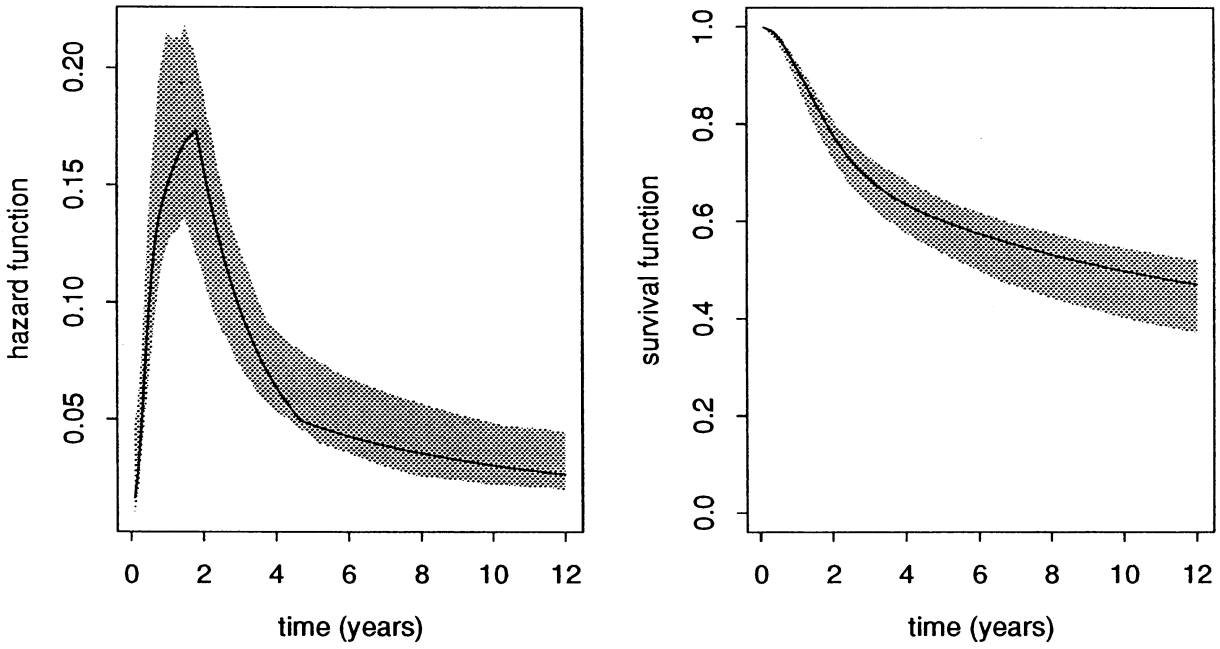

Fig. 6. Conditional hazard and survival functions for the fit of Table 8 and 95% bootstrap bands from that model. Same covariates as in Figure 3.

Similarly, the left side of Figure 8 shows log(hazard ratio) when time is 2 years and all covariates are kept fixed at the same value as in Figure 3, except that the tumor size is allowed to vary and is compared with size = 3. The right side of Figure 8 shows log(hazard ratio) when time is 2 years and all covariates are kept fixed at the same value as in Figure 3, except that age is allowed to vary and is compared with age = 50.

It is interesting to observe that our results are similar to those in Gray (1992). In particular, compare the left and right sides of Figure 7 and the left and right sides of Figure 8 with Figures 3a, 4a, 4c and 4d in Gray (1992) respectively. Furthermore, the only interaction between covariates that is significant in Table 3 of Gray (1992) is Nodal Group $\times$ Tumor Size. Similarly, in Table 8 the only interaction between covariates that ends up in the model is that between log(nodes) and size.

The bootstrap bands in Figures 6 through 8 reflect the contribution of the variance of the corresponding point estimates but not their bias. To see this in a simple manner, consider just the HARE procedure by itself and its dependence on the penalty parameter α. If α is sufficiently large (say, $\alpha = 200$), when HARE is applied to the real data it estimates the conditional log-hazard function by a constant $\hat{\beta}_0$. Similarly, when applied to the simulated data from the initial fit, it typically estimates the conditional log-hazard function by a constant that is rather close to $\hat{\beta}_0$. Thus the corresponding bootstrap bands, obtained as in Figures 6 through 8, are very narrow. In the opposite direction, when α is extremely small (in particular, when $\alpha = 0$), the corresponding bootstrap bands are very large. Clearly,

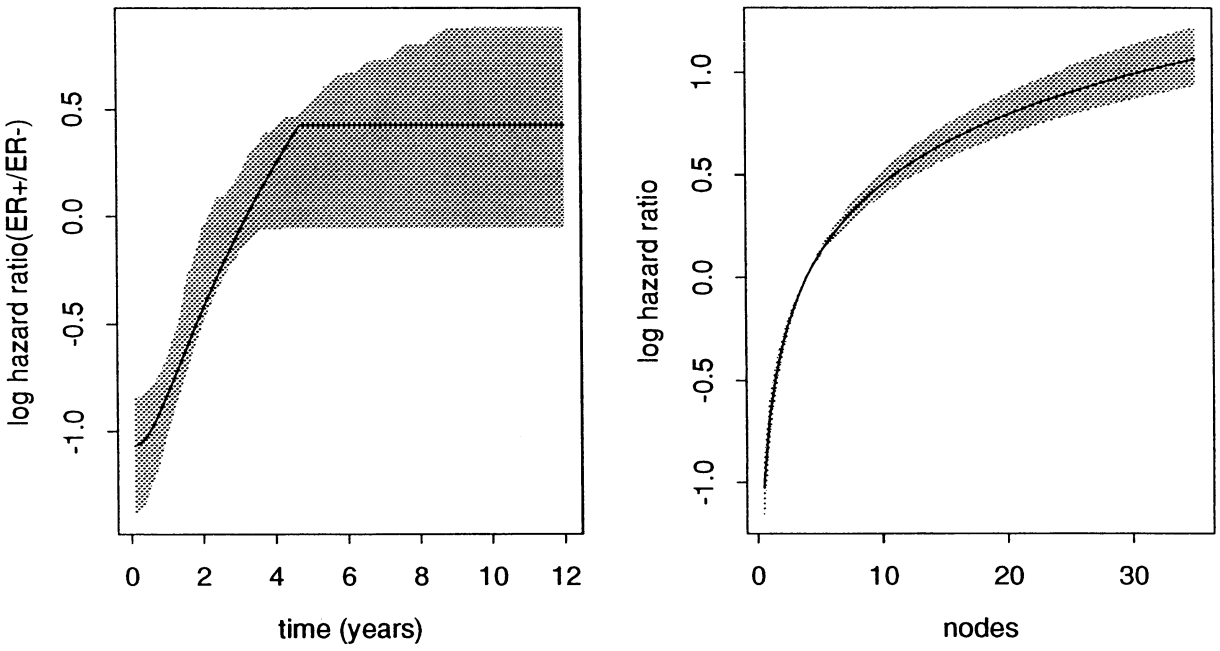

Fig. 7. Log of the hazard ratio for the fit of Table 8 and 95% bootstrap bands for that fit; left side: as a function of time for the ratio ER positive/ER negative; right side: as a function of nodes, relative to nodes = 4 after 2 years; other covariates as in Figure 3.

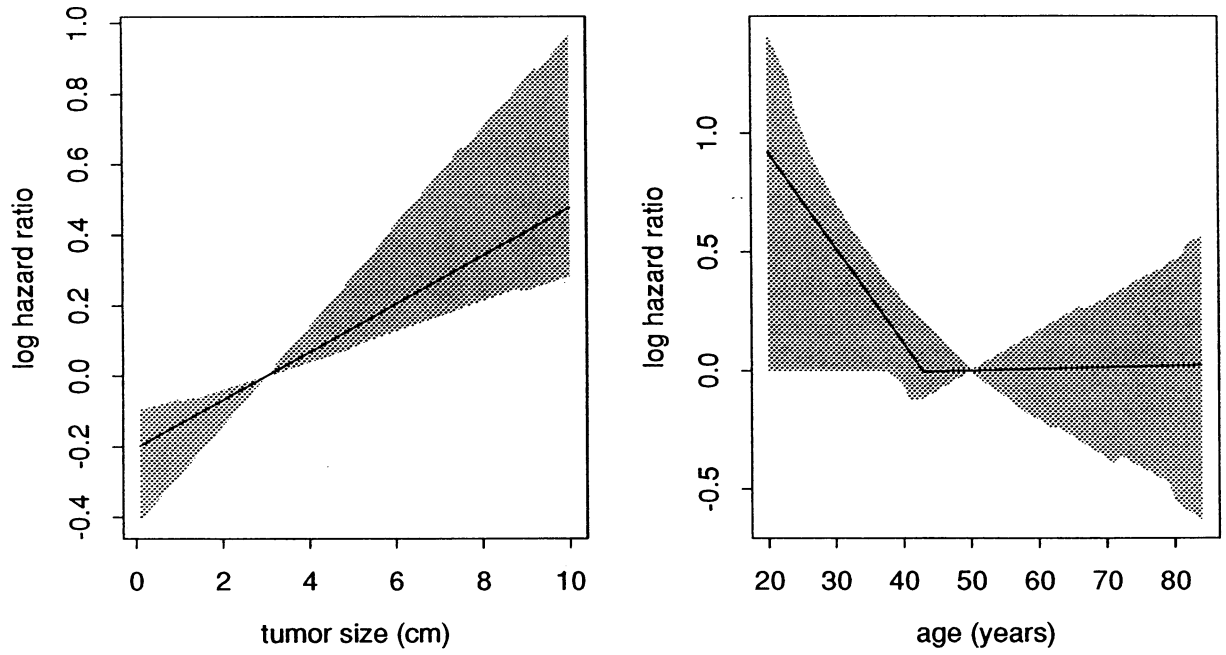

Fig. 8. Log of the hazard ratio for the fit of Table 8 and 95% bootstrap bands for that fit after 2 years; left side: as a function of size relative to size = 3; right side as a function of age, relative to age = 50; other covariates as in Figure 3.

however, we should not think of the various estimates as steadily improving in accuracy as α increases from zero to infinity.

The reservations about the bootstrap bands that we have just described and similar reservations about standard errors apply much more generally in statistics, especially in the context of highly adaptive procedures but even in the context of parametric models that are not exactly valid, and these reservations deserve much greater emphasis in the literature.

8 Concluding Remarks

In light of the examples in Section 7 and considerable additional experience with HARE and its user interface, we are convinced that the methodology is of considerable practical value. The available features make it very easy to try a variety of models on a given set of data. In particular linear proportional hazard models, additive proportional hazard models, proportional hazard models with time-varying coefficients and non-parametric proportional hazard models can conveniently be fitted and compared.

As Figure 3 illustrates, it is very easy to plot hazard and survival functions for an individual with a given vector of covariates after a model has been fitted. Thus HARE is potentially useful for a health care practitioner in coming up with a prognosis for a particular patient.

Under suitable conditions, Kooperberg, Stone and Truong (1993) obtain the L_2 rate of convergence for a nonadaptive version of the methodology treated in the present paper.This result lends theoretical support to HARE.

9 Numerical Implementation

9.1 Starting Values

As the starting value for the maximum likelihood estimate of the conditional log-hazard function having the form of the minimal one-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 space contains the previous one as a proper subspace.

In the context of stepwise knot deletion, let $\hat{\beta}_1 B_1 + \cdots + \hat{\beta}_p B_p$ be the maximum likelihood estimate of the conditional log-hazard function having the form corresponding to the p-dimensional linear space G with basis $B_1, \ldots, B_p$, and let $\tilde{B}_1, \ldots, \tilde{B}_{p-1}$ be the basis of an allowable $(p-1)$-dimensional subspace G_0 of G. Also, for $1 \leq j \leq p$, let $\sum_{k=1}^{p} a_{jk} \tilde{B}_k$ be the

orthogonal projection of B_j onto G_0 relative to the inner product

$$\langle h_1, h_2 \rangle = \sum_i h_1(Y_i|\boldsymbol{x}_i) h_2(Y_i|\boldsymbol{x}_i).$$

As the starting value for the maximum likelihood estimate of the conditional log-hazard function corresponding to G_0, we use

$$\sum_{j=1}^{p} \hat{\beta}_j \left(\sum_{k=1}^{p-1} a_{jk} \tilde{B}_k \right) = \sum_{k=1}^{p-1} \left(\sum_{j=1}^{p} a_{jk} \hat{\beta}_j \right) \tilde{B}_k.$$

9.2 Stepwise Addition of Basis Functions

Let G_0 be the linear space having basis functions 1, $B_{0k}(t)$ for $1 \leq k \leq K_0$, $B_{mk}(x_m)$ for $1 \leq m \leq M$ and $1 \leq k \leq K_m$, and perhaps certain tensor products of two such basis functions. To decide which basis function to add to this model, we compute the Rao statistic, as described in Section 5,

- for all spaces that can be obtained from G_0 by adding a basis function $B_{l0}(x_l) = x_l$ to G_0;
- for all allowable spaces that can be obtained from G_0 by adding a basis function to G_0 that is a tensor product of two basis functions B_{lj}, B_{mk}, $l \neq m$, that are in G_0;
- for a space that can be obtained from G_0 by adding a basis function based upon a potential new knot in time, located using the algorithm described below; and
- for a space that can be obtained from G_0 by adding a basis function based upon a potential new knot in covariate m, for $1 \leq m \leq M$, located using the algorithm described below.

As new space G we choose the one corresponding to the largest absolute value of the Rao statistic among those candidates listed above that are nonvacuous.

To find a potential new knot in covariate m, let $t_1 < t_2 < \cdots < t_{K_m}$ be the corresponding knots presently in the model, to which we want to add one more knot, and let $X_{(1)}, \ldots, X_{(n)}$ be the values $X_{1m}, \ldots, X_{nm}$ of covariate m written in nondecreasing order. Define l_i and u_i by

$$\begin{aligned}
l_i &= 6 + \arg \max_{1 \leq j \leq n} X_{(j)} \leq t_i, \qquad i = 1, \ldots, k, && (3) \\
u_i &= -6 + \arg \min_{1 \leq j \leq n} X_{(j)} \geq t_{i+1}, \qquad i = 0, \ldots, k-1, && (4) \\
l_0 &= 1 \text{ and} \\
u_k &= n.
\end{aligned}$$

For those $i = 0, \ldots, K_m$ for which $u_i \geq l_i$ we compute the Rao statistic r_i for the model with $(x_m - X_{(j_i)})_+$ as new basis function, where $j_i = [(l_i + u_i)/2]$. Because of the 6 and -6 in (3) and (4) 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 for a given covariate in the model to be at least 6 order statistics apart, which improves the numerical and statistical stability. If there is no i for which $u_i \geq l_i$, then no knots can be added to the model.

We place the potential new knot in the interval $[X_{(l_{i^*})}, X_{(u_{i^*})}]$, where $i^* = \arg\max |r_i|$. We proceed by computing the Rao statistic r_l, for the model with $(x_m - X_{(l)})_+$ as new basis function, where $l = [(l_{i^*} + j_{i^*})/2]$, and r_u, for the model with $(x_m - X_{(u)})_+$ as new basis function, where $u = [(j_{i^*} + u_{i^*})/2]$. If $|r_{i^*}| \geq |r_l|$ and $|r_{i^*}| \geq |r_u|$, we place the new knot at $T_{(m_{i^*})}$; if $|r_{i^*}| < |r_l|$ and $|r_l| \geq |r_u|$, we continue searching for a knot location in the interval $[X_{(l_{i^*})}, X_{(j_{i^*})}]$; if $|r_{i^*}| < |r_u|$ and $|r_l| < |r_u|$, we continue searching for a knot location on the interval $[X_{(j_{i^*})}, X_{(u_{i^*})}]$.

To find a potential new knot in time we proceed identically, except that we select the location of the potential new knot based on the ordered statistics of just the uncensored data.

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

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

- the number P of basis functions is equal to P_{max}, where $P_{max} = \min(6n^{.2}, n/4, 50)$;
- $\hat{l}_P - \hat{l}_p < \frac{1}{2}(P - p) - 0.5$ for some p with $3 \leq p \leq P - 3$, where $\hat{l}_p$ is the log-likelihood for the model with p basis functions;
- the search algorithm, as described above, yields no possible new basis function.

References

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. R. (1972), "Regression models and life tables (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, 34, 187-220.

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

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.

Fleming, T. R. and Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, New York: Wiley.

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," *Journal of the American Statistical Association*, 87, 942-951.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Hastie, T. and Tibshirani, R. (1993), "Varying-coefficient models (with discussion)" *Journal of the Royal Statistical Society*, Ser. B, 55 to appear.

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

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

Kooperberg, C. and Stone, C. J. (1993a), Contribution to the discussion of "Varying-coefficient models" by Hastie, T. and Tibshirani, R., *Journal of the Royal Statistical Society*, Ser. B, to appear.

Kooperberg, C. and Stone, C. J. (1993b), "Hazard Estimation with Flexible Tails," Technical Report No. 388, Department of Statistics, University of California, Berkeley, California.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1993), "The L_2 rate of convergence for hazard regression," Technical Report No. 390, Department of Statistics, University of California, Berkeley, California.

LeBlanc, M. and Crowley, J. (1992), "Relative risk trees for censored data," *Biometrics*, 48, 411-425.

Little, R. J. A. and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

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

O'Sullivan, F. (1988), "Nonparametric estimation of relative risk using splines and cross-validation," *SIAM Journal on Scientific and Statistical Computing*, 9, 531–542.

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.

Sleeper, L. A. and Harrington, D. P. (1990), "Regression splines in the Cox model with Application to covariate effects in liver disease," *Journal of the American Statistical Association*, 85, 941-949.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-7400

Appendix: Documentation of S Functions

Hare	Hazard Regression	**Hare**

```
dhare(q, cov, fit)
hhare(q, cov, fit)
phare(q, cov, fit)
qhare(p, cov, fit)
rhare(n, cov, 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.

`cov:` covariates. There are several possibilities. If a vector of length fit$ncov is provided, these covariates are used for all elements of p or q or for all random numbers. If a matrix of dimension length(p) or length(q) or n by fit$ncov is provided, the rows of cov are matched with the elements of p or q or every row of cov has its own random number. If a matrix of dimension m times fit$ncov is provided, while length(p)=1 or length(q)=1 or n=1, the single element of p or q is used m times, or m random numbers with different sets of covariates are generated.

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

VALUE

Densities (dhare), hazard rates (hhare), probabilities (phare), quantiles (qhare), or a random sample (rhare) from a hare density.

hare.fit	Hazard Regression	**hare.fit**

```
hare.fit(data, delta, cov, penalty, maxdim, exclude, include,
prophaz = F, additive = F, linear = F, fit, 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.

`cov:` covariates: matrix with as many rows as the length of data. May be omitted if there are no covariates.

`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(samplesize) as in BIC. The effect of this parameter is summarized in hare.summary().

`maxdim:` maximum dimension (default is 6*length(data)$^{0.2}$).

`exclude:` combinations to be excluded - this should be a matrix with 2 columns - if for example exclude(1,1)=2 and exclude(1,2)=3 no interaction between covariate 2 and 3 is included. 0 represents time.

`include:` those combinations that can be included. Should have the same format as exclude. Only 1 of the exclude and include can be specified

`prophaz:` should a proportional hazard model be fitted?

`additive:` should an additive model be fitted?

`linear:` vector indicating for which of the variables no knots should be entered. For example, if linear=c(2,3) no knots for either covariate 2 or 3 are entered. 0 represents time.

`fit:` object created by hare.fit. If a fit is specified, hare.fit adds basis functions starting with those in the specified fit.

`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 hare.plot, hare.summary, dhare, hhare, phare, qhare and rhare.
The function returns a list with the following members:

`ncov:` number of covariates

`ndim:` number of dimensions of the fitted model

`fcts:` matrix of size ndim x 6. each row is a basis function. First element: first covariate involved (0=time);
second element: which knot (0 means: constant (time) or linear (covariate));
third element: second covariate involved (NA means: this is a function of one variable);
fourth element: knot involved (if the third element is NA, of no relevance);
fifth element: beta;
sixth element: standard error of beta.

`knots:` a matrix of size ncov $\times$? one row for each dimension. Covariate i has row i+1, time has row 1. first element - number of knots in this dimension other elements - the knots, appended with NAs to make it a matrix

`penalty:` the parameter used in the AIC criterion.

`max:` maximum element of data.

`ranges:` column i gives the range of the i-th covariate.

`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.

hare.plot Hazard Regression **hare.plot**

```
hare.plot(fit, cov, n = 100, which = 0, what = "d", time, add =
F, ...)
```

ARGUMENTS

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

`cov:` a vector of length fit$ncov, indicating for which combination of covariates the plot should be made. Can be omitted only if fit$ncov is 0.

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

`which:` for which coordinate should the plot be made. 0: time; positive value i: covariate i. Note that if which is the positive value i, then the element corresponding to this covariate must be given in cov even though its actual value is irrelevant. (See example 2 below.)

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

`time:` if which is not equal to 0, the value of time for which the plot should be made.

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

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

This function produces a plot of a hare 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 <- hare.fit(time, delta, covs)
```

hazard curve for covariates like case 1

```
hare.plot(fit, covariates[1,], what = "h")
```

survival function as a function of covariate 2, for covariates as case 1 at t=3

```
hare.plot(fit, covariates[1,], which = 2, what = "s", time = 3)
```

hare.summary Hazard Regression **hare.summary**

```
hare.summary(fit)
```

ARGUMENTS

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

VALUE

This function produces only printed output. The main body consists of two tables.

The first table has six columns: the first column is a possible number of dimensions for the fitted model;

the second column indicates whether this model was fitted during the addition or deletion stage;

the third column is the log-likelihood for the fit;

the fourth column is -2*loglikelihood + penalty*(dimension), which is the AIC criterion - hare.fit selected the model with the minimum value of AIC;

the last two columns give the endpoints of the interval of values of penalty that would yield the model with the indicated number of dimensions (NAs imply that the model is not optimal for any choice of penalty).

At the bottom of the first table the dimension of the selected model is reported, as is the value of penalty that was used.

Each row of the second table summarizes the information about a basis function in the final model. It shows the variables involved, the knot locations, the estimated coefficient and its standard error and Wald statistic (estimate/SE).