Hazard Regression

Charles Kooperberg Charles J. Stone and Young K. Truong *

Technical Report No. 389
May 1, 1994
Revised version
Also replaces Technical Report No. 388
University of California
Berkeley, California 94720

Abstract

Polynomial 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. An automatic procedure involving the maximum likelihood method, stepwise addition, stepwise deletion and BIC is used to select the final model. The possible models contain proportional hazards models as a subclass, which makes it possible to diagnose departures from proportionality. Two additional log terms are incorporated into a similar model for the unconditional log-hazard function to allow for greater flexibility in the extreme tails. A user interface based on S is described.

KEY WORDS: Conditional hazard function; Interactions; Model selection; Proportional hazards model; Splines; Time-varying coefficients.

*Charles Kooperberg is Assistant Professor, Department of Statistics, University of Washington, Seattle, WA 98195. Charles J. Stone is Professor, Department of Statistics, University of California, Berkeley, CA 94720. Young K. Truong is Associate Professor, School of Public Health, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7400. Charles Kooperberg's work was supported in part by a grant from the Graduate School Fund of the University of Washington. Charles J. Stone's work was supported in part by National Science Foundation Grant DMS-9204247. Young K. Truong's work was supported in part by a Research Council Grant from the University of North Carolina. The data for the example in Section 9.3 was kindly provided by the Eastern Cooperative Oncology Group.

1. INTRODUCTION

Consider data involving a positive response variable that 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 $[0, \infty)$. The hazard function and its logarithm corresponding to this density function are referred to as the conditional hazard function and the conditional log-hazard function, respectively.

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.

In this paper we describe a general framework for modeling the logarithm of the conditional hazard function with linear models. The maximum likelihood method is used to estimate the unknown parameters of the model. We describe a fully automatic method involving stepwise addition, stepwise deletion and BIC for selecting the final model in a family of allowable spaces.

We then describe particular families of allowable spaces corresponding to HARE (hazard regression) and HEFT (hazard estimation with flexible tails). In HARE, linear splines and selected tensor products are used to estimate the logarithm of the conditional hazard function. The method is similar in spirit to MARS (Friedman 1991). One of the advantages of HARE models is that they include proportional hazards models as a subclass. The presence or absence of interaction terms between covariates and time in the final model can in fact be regarded as a check on the proportionality of the underlying hazard model.

In HEFT, the unconditional log-hazard function is estimated using cubic splines. 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. With these log terms, HEFT can fit Weibull and Pareto distributions exactly. In the analysis of survival data with covariates, HEFT is useful as a preprocessor for HARE.

In order to evaluate this combination of HEFT and HARE, we apply it to various datasets that have been studied in the literature. The combined procedure appears to be a promising tool in survival analysis.

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 HEFT and HARE and, in particular, to the use of polynomial splines and their tensor products in defining the allowable spaces used in these procedures.

Traditionally, in the proportional hazards model and in some other survival analysis models, 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, and 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 hazards model, there have been a number of papers in which the dependence of the survival time on the covariates has been modeled by means of various nonparametric techniques, with the baseline hazard function being ignored. In particular, Hastie and Tibshirani (1990) and O'Sullivan (1988a) 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.

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 and hazard functions in the presence of censored data. These papers typically fall into two groups: those using smoothing splines or similar procedures include Anderson and Senthilselvan (1980), Whittemore and Keller (1986), Senthilselvan (1987), and O'Sullivan (1988b); those using polynomial splines include Etezadi-Amoli and Ciampi (1987), Efron (1988), Abrahamowicz, Ciampi and Ramsay (1992), and Kooperberg and Stone (1992). O'Sullivan (1988b) is the only one of these papers that directly models the log-hazard function. Gu (1991) contains an asymptotic analysis of the hazard estimate in O'Sullivan (1988b) 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.

2. LINEAR MODELS FOR THE CONDITIONAL LOG-HAZARD FUNCTION

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

Observe that

$$F(t|\mathbf{x}) = \int_0^t f(u|\mathbf{x})du, \qquad \lambda(t|\mathbf{x}) = \frac{f(t|\mathbf{x})}{1 - F(t|\mathbf{x})} \quad \text{and} \quad \alpha(t|\mathbf{x}) = \log \lambda(t|\mathbf{x}), \qquad t \geq 0. \tag{2.1}$$

Moreover,

$$1 - F(t|\mathbf{x}) = \exp\left(-\int_0^t \lambda(u|\mathbf{x})du\right) = \exp\left(-\int_0^t \exp(\alpha(u|\mathbf{x}))du\right), \qquad t \geq 0. \tag{2.2}$$

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

$$f(t|\mathbf{x}) = \exp(\alpha(t|\mathbf{x})) \exp\left(-\int_0^t \exp(\alpha(u|\mathbf{x}))du\right), \qquad t \geq 0. \tag{2.3}$$

In this paper we will use polynomial splines and selected tensor products to obtain a linear model for $\alpha(t|\mathbf{x})$. By modeling $\alpha(t|\mathbf{x})$, as opposed to $\lambda(t|\mathbf{x})$, $f(t|\mathbf{x})$ or $F(t|\mathbf{x})$, we do not have to worry about positivity constraints. Also, as pointed out by O'Sullivan (1988b), a model for the log-hazard function leads to a concave likelihood function, even in the context of right censoring. This is not true, for example, for a model for the log-density function. Finally, a proportional hazards model can be written as a linear model for the log-hazard function; thus by modeling the log-hazard function, we include proportional hazard models. We refer to Sections 4–6 for more details about the models for $\alpha(t|\mathbf{x})$.

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

$$\alpha(t|\mathbf{x};\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j B_j(t|\mathbf{x}), \qquad t \geq 0, \tag{2.4}$$

for the conditional log-hazard function, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. Given $\boldsymbol{\beta} \in \mathbf{R}^p$ we define $\lambda(t|\mathbf{x};\boldsymbol{\beta})$, $F(t|\mathbf{x};\boldsymbol{\beta})$ and $f(t|\mathbf{x};\boldsymbol{\beta})$ by imitating equations (2.1)–(2.3).

3. MAXIMUM LIKELIHOOD ESTIMATION

Consider n randomly selected individuals. For $1 \leq i \leq n$, let T_i be the survival time, C_i the censoring time, and $\mathbf{x}_i$ 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_i and C_i are conditionally independent and that T has conditional density function $f(\cdot|\mathbf{x})$ given $\mathbf{x}$. The random variable Y_i is said to be uncensored or censored according as $\delta_i = 1$ or $\delta_i = 0$. Note that the partial likelihood corresponding to $Y_i = y_i$, δ_i, $\mathbf{x}_i$ and $\boldsymbol{\beta}$ equals $[f(y_i|\mathbf{x}_i;\boldsymbol{\beta})]^{\delta_i}[1 - F(y_i|\mathbf{x}_i;\boldsymbol{\beta})]^{1-\delta_i}$ (see page 16 of Miller 1981), so the log-likelihood equals

$$\phi(y_i, \delta_i|\mathbf{x}_i;\boldsymbol{\beta}) = \delta_i \alpha(y_i|\mathbf{x}_i;\boldsymbol{\beta}) - \int_0^{y_i} \exp(\alpha(u|\mathbf{x}_i;\boldsymbol{\beta}))du, \qquad y_i \geq 0 \text{ and } \delta_i \in \{0,1\}.$$

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

$$l(\boldsymbol{\beta}) = \sum_i \phi(Y_i, \delta_i | \mathbf{x}_i; \boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathbf{R}^p. \tag{3.1}$$

Moreover,

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

and

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

where

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

and

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

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

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 and distribution function are given by $\hat{\alpha}(t|\mathbf{x}) = \alpha(t|\mathbf{x}; \widehat{\boldsymbol{\beta}})$, $\hat{\lambda}(t|\mathbf{x}) = \lambda(t|\mathbf{x}; \widehat{\boldsymbol{\beta}})$ 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. MODEL SELECTION

When modeling the log-hazard function with a linear model (2.4), the remaining issue to be resolved is the choice of G. In this section we describe an algorithm for determining G in an adaptive fashion. This algorithm is a hybrid of well known stepwise addition and stepwise deletion procedures in multiple regression and generalized linear models, where each regression model can be viewed as an element of a linear space spanned by the potential predictors. From this viewpoint, we will define a family of allowable spaces $\mathcal{G}$ that may be considered during model selection. This family $\mathcal{G}$ of allowable spaces is assumed to have the following properties:

- each $G \in \mathcal{G}$ is a linear space having dimension $p \geq p_{\min}$;
- there is only one $G \in \mathcal{G}$ with dimension $p_{\min}$;
- if $G \in \mathcal{G}$ has dimension $p > p_{\min}$, there is at least one subspace $G_0 \in \mathcal{G}$ of G with dimension $p-1$;
- if $G_0 \in \mathcal{G}$ has dimension p, there is at least one space $G \in \mathcal{G}$ with dimension $p+1$ and containing G_0 as a subspace.

We refer to $G_{\min} \in \mathcal{G}$ with minimal dimension $p_{\min}$ as the minimal allowable space. In the context of multiple regression and generalized linear models, the space G is a linear space spanned by candidate predictors with $G_{\min}$ being the space of constants and it is typically obvious which variables can be added and deleted. However, it is more complicated in our situation. In the next two sections, we give two specific constructions of allowable spaces, which subsequently provide a complete description of HARE (Section 5) and HEFT (Section 6).

Initially, we use the minimal allowable space to model $\alpha(t|\mathbf{x})$. 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. Where for multiple regression it is possible to evaluate candidates for a new basis function by recomputing the fit, in our context this would be computationally too demanding to be practically useful. Therefore, we choose among the various candidates for a new basis function by a heuristic search that is designed approximately to maximize the absolute value of the corresponding Rao statistic. This is similar to what is sometimes done for generalized linear models; see, for example, the function `step.glm` in S and the discussion in Chambers and Hastie (1992, page 235).

Specifically, let $\widehat{\boldsymbol{\beta}}^{(o)}$ 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 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 be in G_0 is given by $R = [\mathbf{S}(\widehat{\boldsymbol{\beta}}^{(o)})]_p / \sqrt{[\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}}^{(o)})]_{pp}}$, where $\mathbf{I}(\widehat{\boldsymbol{\beta}}^{(o)}) = -\mathbf{H}(\widehat{\boldsymbol{\beta}}^{(o)})$ 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 described in Section 11.4, 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 allowable space, at each step choosing the candidate space G_0 so that the Wald statistic for a basis function that is in G but not in G_0 is smallest in magnitude. As was the case during the stepwise addition stage, we do not refit the model for each basis function that is a candidate to be dropped, since this would be computationally infeasible.

Specifically, let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ 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 would be 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 $[\mathbf{I}(\hat{\boldsymbol{\beta}})]^{-1} = -[\mathbf{H}(\hat{\boldsymbol{\beta}})]^{-1}$, 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 $\hat{\beta}_p/\mathrm{SE}(\hat{\beta}_p)$. Note that the Wald statistic used during the stepwise deletion stage and the Rao statistic used during the stepwise addition stage test the same hypothesis, but the Wald statistic is based on the maximum likelihood estimate corresponding to G, while the Rao is based on the maximum likelihood estimate corresponding to G_0.

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

$$\mathrm{AIC}_{a,\nu} = -2\hat{l}_\nu + a p_\nu \tag{4.1}$$

be the Akaike Information Criterion with penalty parameter a for this model. We select the model corresponding to the value $\hat{\nu}$ of ν that minimizes $\mathrm{AIC}_{a,\nu}$. In light of Kooperberg and Stone (1992) and our experience in the present investigation, we recommend choosing $a = \log n$ as in the Bayesian information criterion (BIC) due to Schwarz (1978). (The interface described in Section 7 allows the user to specify the penalty parameter.)

5. THE HARE MODEL

In this section we describe the family of allowable spaces for HARE and the corresponding basis functions. The HARE model involves splines and selected tensor products. In order to avoid numerous numerical integrations with respect to t and added complications in the 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 an allowable space by listing its basis functions.

Let K_0 be a nonnegative integer; if $K_0 = 0$ there are no basis functions depending on t; if $K_0 \geq 1$, let t_k for $1 \leq k \leq K_0$ be distinct positive numbers, and consider the basis functions $B_{0k}(t) = (t_k - t)_+$ for $1 \leq k \leq K_0$, where $t_+ = \max(t, 0)$. Next, for $1 \leq m \leq M$, let K_m be an integer with $K_m \geq -1$; if $K_m = -1$ there are no basis functions depending

on x_m; if $K_m = 0$, consider the basis function $B_{m0}(x_m) = x_m$; if $K_m \geq 1$, consider the basis function $B_{m0}(x_m) = x_m$, let x_{mk} for $1 \leq k \leq K_m$ be distinct real numbers, and consider the additional basis functions $B_{mk}(x_m) = (x_m - x_{mk})_+$ for $1 \leq k \leq K_m$.

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.

It is easy to check that the collection $\mathcal{G}$ of such spaces satisfies the properties listed in Section 4. In particular, the minimal allowable space $G_{\min}$ for the HARE model is the space of constant functions. Thus the minimal model for (2.4) has $p = 1$, $B_1 = 1$ and

$$\alpha(t|\boldsymbol{\beta}) = \alpha(t|\mathbf{x};\boldsymbol{\beta}) = \beta_1, \qquad t \geq 0,$$

so that α does not depend on t or the vector $\mathbf{x}$ of covariates. The corresponding conditional distribution of T given $\mathbf{x}$ is exponential with mean $\exp(-\beta_1)$, which does not depend on $\mathbf{x}$.

If none of the basis functions of G depend on both t and $\mathbf{x}$, then (2.4) is a proportional hazards model (Cox 1972). It is a particular interesting feature of the HARE model that the model selection procedure described in Section 4 may or may not result in a proportional hazards model. If any of the basis functions in the final model depend on both time and a covariate, then a proportional hazards model might not be appropriate.

Given the basis of an allowable space G as defined above, it is obvious whether any given basis function can be deleted in one step.

Example. Let $M = 3$. Then the following six basis functions span an allowable space G: $B_1 = 1$, $B_2 = (1 - t)_+$, $B_3 = x_1$, $B_4 = (x_1 - 6)_+$, $B_5 = x_2$, and $B_6 = x_1(1 - t)_+$. In this example, B_4, B_5 or B_6 could be removed and the remaining space would still be allowable. If one of the basis functions B_2 or B_3 were removed, however, the remaining space would not be allowable since it would still contain $B_6 = B_2 B_3$. The constant basis function B_1 can never be removed.

Let G_0 be the allowable 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 4

- **(i)** for all spaces that can be obtained from G_0 by adding a basis function $B_{l0}(x_l) = x_l$ to G_0;
- **(ii)** 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 $B_{lj}(x_l)$ and $B_{mk}(x_m)$, $l \neq m$, that are in G_0;
- **(iii)** 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 in Section 11.3; and

(iv) 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 in Section 11.3.

As the 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.

Example (continued). Corresponding to (i), we can add the basis function x_3 to the space in the above example. Corresponding to (ii), we can add $B_2B_4 = (1-t)_+(x_1-6)_+$, $B_2B_5 = (1-t)_+x_2$ or $B_3B_5 = x_1x_2$ to the space. The basis function $B_4B_5 = (x_1-6)_+x_2$ cannot be added, since the resulting space would not contain $B_3B_5 = x_1x_2$ so it would not be allowable. Corresponding to (iii) and (iv), a basis function $(t_k - t)_+$ with $t_k > 0$ and $t_k \neq 1$, $(x_1 - x_{1k})_+$ with $x_{1k} \neq 6$, or $(x_2 - x_{2k})_+$ could be added. No basis function of the form $(x_3 - x_{3k})_+$ could be added before x_3 is added.

6. THE HEFT MODEL

In the absence of covariates, (2.4) reduces to

$$\alpha(t|\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j B_j(t), \qquad t \geq 0.$$

Since estimation and model selection are considerably easier for this model, it is now feasible to use cubic splines.

A linear space G of functions on $[0,\infty)$ that are piecewise cubic, have two continuous derivatives and are constant near 0 and in the right tail is defined by a sequence of knots at which the third derivative may be discontinuous. In particular, given the integer $K \geq 3$ and the sequence $t_1, \ldots, t_K$ with $0 < t_1 < \cdots < t_K < \infty$, let G 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 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 $[t_K,\infty)$. The collection of such spaces G forms a family $\mathcal{G}$ of allowable spaces for the basic HEFT model. In particular, if $K = 3$, then G is the minimal allowable space $G_{\min}$, which is the one-dimensional space of constant functions on $[0,\infty)$, and $B_1 = 1$ on $[0,\infty)$.

Model selection for the basic version of HEFT is straightforward. The first three knots are placed at the quartiles of the uncensored data. During the stepwise addition stage as described in Section 4, new knots are successively added. The selection of these knots is similar to the selection of a new knot for HARE, which is described in Section 11.3. Stepwise deletion for the basic HEFT model is equivalent to stepwise deletion of knots. The main numerical task of the HEFT algorithm, the computation of the log-likelihood $l(\boldsymbol{\beta})$, the score $\mathbf{S}(\boldsymbol{\beta})$, and the Hessian $\mathbf{H}(\boldsymbol{\beta})$ for various models and values of $\boldsymbol{\beta}$ is described in Section 11.2.

A more sophisticated version of the HEFT model is obtained by the inclusion of two extra basis functions, which adds considerable flexibility to the tails of the fitted distribution. In particular, as shown below, the inclusion of these two basis functions makes it possible for HEFT to fit Weibull and Pareto distributions exactly.

Given a positive number c (which will be defined below in terms of the observed data in a simple manner), set $B_{-1}(t) = \log(t/(t+c))$ and $B_0(t) = \log(t+c)$ for $t > 0$. Let G_0 be the space G defined above and set $p = K - 2$. Then $B_{-1}, B_0, B_1, \ldots, B_p$ is a basis of the linear space spanned by $G_0 \cup \{B_{-1}, B_0\}$. The collection $\mathcal{G}$ of such spaces G forms the family of allowable spaces for the extended HEFT model.

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 α 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 $\alpha(t) \approx \gamma \log t$ for $t \approx 0$. This motivates the inclusion of the term $\beta_{-1} B_{-1}(t)$ with $\beta_{-1} > -1$ 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

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

and hence $\alpha(t) \approx (\gamma - 1) \log t$ for $t \gg 1$. This motivates the inclusion of the term $\beta_0 B_0(t)$ with $\beta_0 > -1$ 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 $\lambda(t) \approx bt^{-1}$ for $t \gg 1$ and hence $\alpha(t) \approx (-1) \log t$ for $t \gg 1$. This motivates allowing the possibility that $\beta_0 = -1$ in the model for the log-hazard function.

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

$$\alpha(t; \boldsymbol{\beta}) = \beta_{-1} \log \frac{t}{t+c} + \beta_0 \log(t+c) + \beta_1, \qquad t > 0.$$

This three-parameter model, which is the minimal allowable space for the extended version of HEFT, includes Weibull and Pareto distributions as special cases. As the default we chose the shift parameter c to be the upper quartile of the uncensored data.

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

$$F(t) = 1 - \exp(-bt^\gamma), \qquad t > 0. \tag{6.1}$$

The corresponding log-hazard function is given by $\alpha(t) = (\gamma - 1)\log t + \log b\gamma$ for $t > 0$. Thus $\alpha(\cdot) = \alpha(\cdot; \boldsymbol{\beta})$, where $\beta_{-1} = \beta_0 = \gamma - 1$ and $\beta_1 = \log b\gamma$. (Alternatively, we can get the Weibull model by setting $c = 0$, $\beta_{-1} = 0$, $\beta_0 = \gamma - 1$ and $\beta_1 = \log b\gamma$.)

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

$$F(t) = 1 - \left(\frac{c}{t+c}\right)^b, \qquad t > 0. \tag{6.2}$$

The corresponding log-hazard function is given by $\alpha(t) = \log b - \log(t + c)$ for $t > 0$. Thus $\alpha(\cdot) = \alpha(\cdot; \boldsymbol{\beta})$, where $\beta_{-1} = 0$, $\beta_0 = -1$ and $\beta_1 = \log b$. (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 provides only an approximation to the Pareto distribution.)

7. USER INTERFACE

Programs for implementing hazard HARE and HEFT as described in this paper have been written in C, and interfaces based on S (see Becker, Chambers and Wilks 1988, and Chambers and Hastie 1992) have also been developed. The software is available from statlib by sending an email with the body `send hare from S` or `send heft from S` to statlib@stat.cmu.edu.

The current interface to HARE consists of eight S functions: `dhare`, `hhare`, `phare`, `qhare`, `rhare`, `hare.fit`, `hare.summary` and `hare.plot`. 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.

The interface to HEFT is similar to that for HARE.

8. APPLICATIONS OF HEFT

8.1. Approximating the Pareto and Weibull distribution with HEFT models

In Section 6 we discussed how Pareto and Weibull distributions can be modeled using HEFT. To illustrate the use of HEFT in estimating these distributions based on sample data, we generated a sample of size 200 from a Pareto distribution with parameters $b = 4$ and $c = 1$ in (6.2). In the left side of Figure 1, we show the true density function (solid) corresponding to this distribution together with various estimates of the density function based on the sample. The line with long dashes corresponds to the estimated density function that was obtained from HEFT using the default parameters. As we noted in the Section 6, HEFT can exactly fit a Pareto distribution if the shift parameter c in HEFT equals the parameter c in the Pareto distribution. The default value for c in HEFT is the 75th percentile of the data, which was 0.4 for this sample. The function `heft.fit` has an option `shift`, which allows the user to specify c. In particular, we 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 corresponds to the value for the shift parameter c that minimizes BIC. We determined that this was 2.7.

In the right side of Figure 1 we show the results of a similar set of computations based on a random sample of size 1000 from a Pareto distribution with parameters $b = 1$ and $c = 1$ in (6.2). 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 in the left hand side, the remaining curve is the density function corresponding to the the true density function.

From these two examples (and many more that we have 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 6. 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 6.

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 $\gamma = 0.25$ in (6.1). 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. In the right side of Figure 2 we show the results of similar calculations based upon a sample of size 1000 from a Weibull distribution with parameters $b = 1$ and $\gamma = 4$ in (6.1).

The HEFT fits to Weibull distributions that are illustrated in Figure 2 have the form of the three parameter model described in Section 6. The parameters $\hat{\beta}_{-1}$, $\hat{\beta}_0$ and $\hat{\beta}_1$ for the

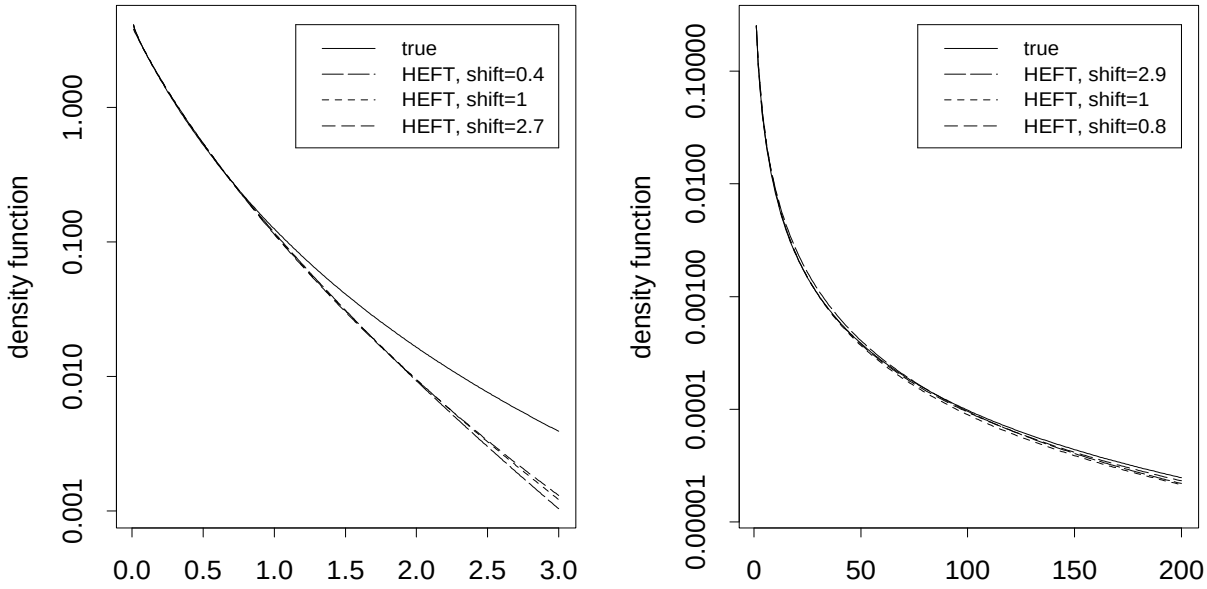

Fig. 1. Estimated density functions for Pareto distributions;
left side: n=200, b=4, c=1; right side: n=1000, b=1, c=1.

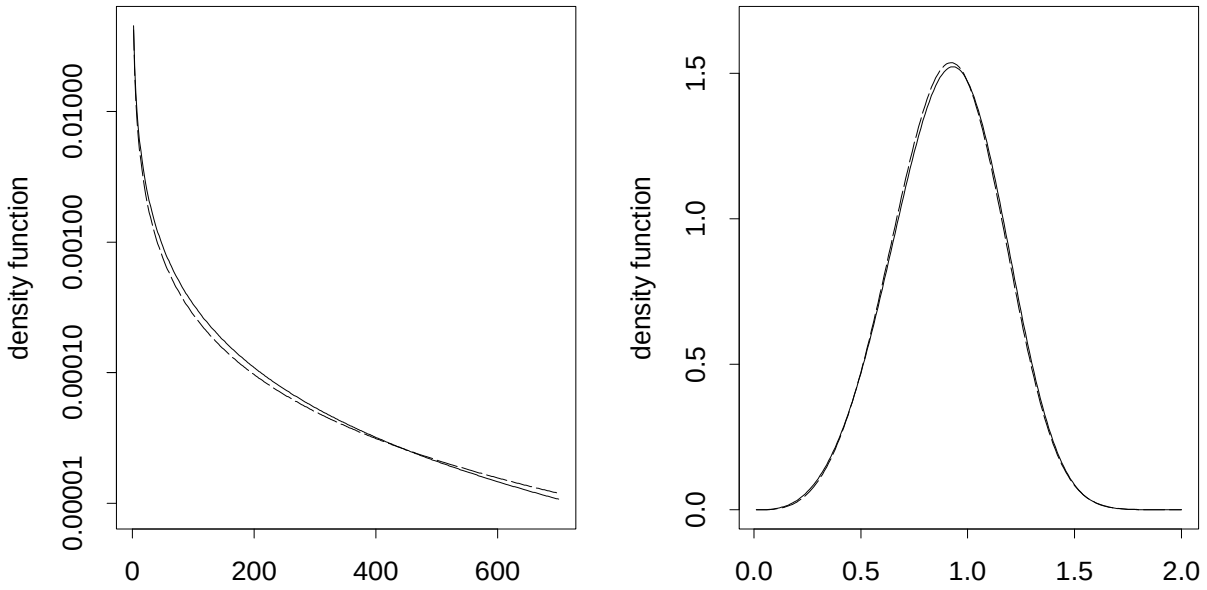

Fig. 2. Estimated density functions for Weibull distributions; left side: n=200,
gamma=0.25; right side: n=1000, gamma=4; solid = truth, dashed = HEFT.

HEFT fit in the left side of Figure 2 are -0.755, -0.819 and -1.232, 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.

We carried out a small simulation study to determine how well HEFT estimates Weibull distributions. 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 $\gamma = 4$. Between 76% and 92% of the HEFT fits have the form of the three parameter model described in Section 6. The estimated values for β_{-1}, β_0 and β_1 are typically very close to their theoretical values.

Based on experience with other examples, not reported here, we have found that HEFT yields a reasonable estimate for the hazard function even when there is a substantial percentage of right censoring.

8.2. HEFT as Preprocessor for HARE

Before applying HARE, we can use HEFT to transform time so that the transformed unconditional hazard function will be approximately equal to one. Recall the notation of hazard regression that was introduced in Section 3. Let the HEFT methodology be applied to (Y_i, δ_i), $1 \le i \le n$, to yield an estimate $\hat{\lambda}_0$ of the unconditional hazard function. Let the HARE methodology then be applied to $(\hat{q}_0(Y_i), \delta_i, \mathbf{x}_i)$, yielding an estimate $\hat{\lambda}_1$ of the conditional hazard function for the transformed data and the estimate $\hat{\lambda}(t|\mathbf{x}) = \hat{\lambda}_0(t)\hat{\lambda}_1(\hat{q}_0(t)|\mathbf{x})$ of the conditional hazard function for the untransformed data; here $\hat{q}_0 = -\log(1 - \widehat{F}_0)$ with $\widehat{F}_0$ being the distribution function corresponding to $\hat{\lambda}_0$.

The unconditional hazard function of the transformation should be approximately constant on $[0, \infty)$. To see this, let T be a continuous random variable having distribution function F. Then $U = F(T)$ is uniformly distributed on $(0, \infty)$, so $-\log(1-U) = -\log(1-F(T))$ has the exponential distribution with mean 1, whose hazard function equals one on $[0, \infty)$.

There are two advantages of such a transformation. First, because of the piecewise linear nature of HARE, the (baseline) hazard function may have big jumps in its first derivative at the various knots in time. However, the hare model for the transformed data typically has fewer knots in time, while the jumps in the first derivative of the baseline hazard function at these knots tend to be smaller. Secondly, because of the allowable spaces used for the HARE model, the fitted conditional hazard functions beyond the last knot in time are necessarily constant. This is no longer true if the transformation based on HEFT is made before applying HARE.

We refer to the examples in Sections 9.1 and 9.3 below for the practical use of HEFT as a preprocessor for HARE.

9. EXAMPLES

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

9.1. Lung Cancer Data

Our first example concerns data from a Veteran's Administration lung cancer trial, which have 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 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 the model with nine basis functions that is summarized in Table 1. Note that two of the basis functions in the 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
(Performance status)$\times(156 - t)_+$	−0.000433	0.000095
(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 roots of the diagonal entries of the inverse

of the estimated information matrix. 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 we typically do is transform time as described in Section 8.2 so that the unconditional log-hazard function of the transformed time approximately be constant. The present example will show the advantage of such a transformation.

When HEFT is applied with the default options, the estimate for the hazard rate is the solid line in the left side of Figure 3. The corresponding transformation $\hat{q}_0$ is shown

in the right side of Figure 3. The estimated hazard function has no knots remaining and the coefficient of $\log(t/(t+145.75))$ 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`, which allows us to force the coefficient of $\log(t/(t+145.75))$ to equal zero. As expected, this hazard estimate again has 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 this model corresponds to

$$\hat{\lambda}(t) \approx e^{-1.643}(t+145.75)^{-0.583}. \tag{9.1}$$

The BIC value for the model is 1503.82, which is considerably smaller than that for the previous model since the present model has one less parameter. The estimate of the unconditional hazard function and the corresponding transformation are shown as the dotted curves in Figure 3. These curves are hard to distinguish from the solid ones corresponding to the previous fit.

Finally we applied `heft.fit` with the options `leftlog=0` and `rightlog=0`, which forces the coefficients of both log-based basis functions to equal zero. This HEFT estimate has the form of a two-parameter model involving four knots, and its BIC value equals 1504.65. The estimate of the unconditional hazard function and the corresponding transformation for this fit are shown as the dashed curves in Figure 3. Observe that this estimate differs considerably from the other two estimates. All in all, we like the dotted curve corresponding to (9.1) best.

After transforming the data as described in Section 8.2 using the model (9.1) above, we applied HARE. The results of are summarized in Table 2. In Figure 4 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.

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
(Performance status $-$ 20)$_+$	-0.230	0.108
(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)$\times$(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

The fit in Table 2 is fairly similar to that in Table 1 with respect to the basis functions. However, as can be seen in Figure 4, the fits are quite different as far as the estimated condi-

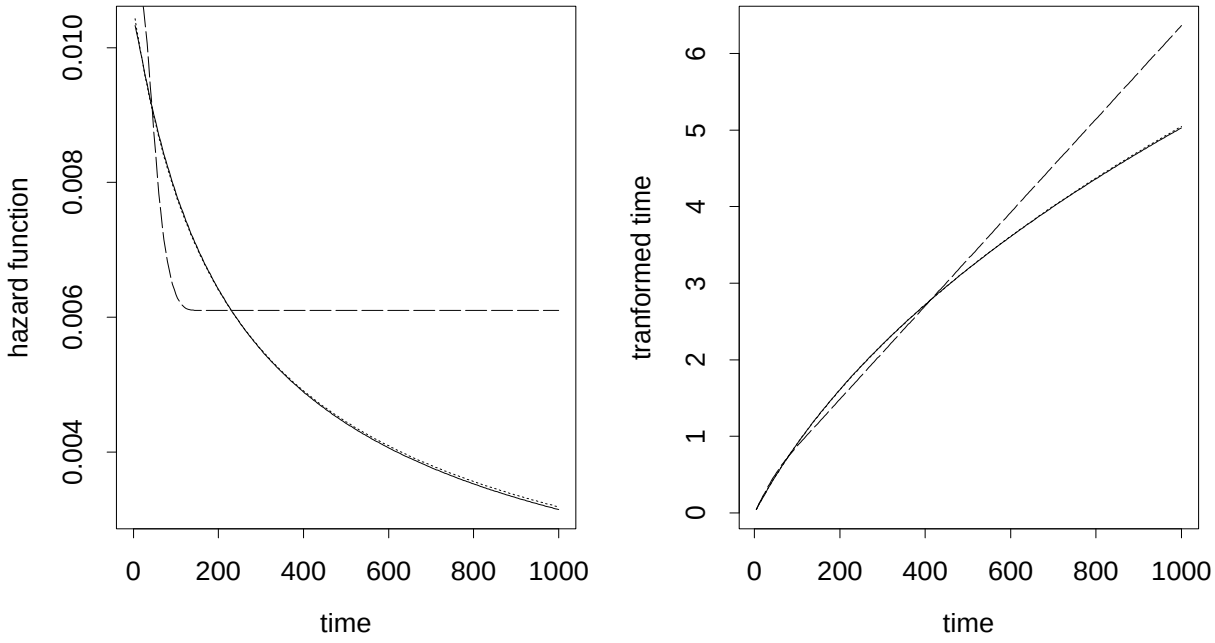

Fig. 3. Three estimates of the unconditional hazard function and the corresponding transformation of time using HEFT with various options for the lung cancer data.

tional hazard rate is concerned, the reason being that HARE, when applied to untransformed data, necessarily gives an estimate for the conditional hazard rate that is constant beyond the last knot in time.

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 (1993) show a similar model for the data using HARE after a transformation of time by HEFT. This fit, summarized in Table 3, was obtained by using the option `linear` for performance status, which prevents HARE from entering any knots for this covariate, and the option `include` for the time $\times$ performance interaction, which makes basis functions that depend on time and performance status the only allowable interactions in the model. The function $\hat{q}_0$ is as above.

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 that we investigate in more detail for our third example, the breast cancer data). Then we applied HEFT to 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, fits well. (Since there were only 9 censored observations, it is not surprising that we obtained a very simple estimate for the unconditional hazard

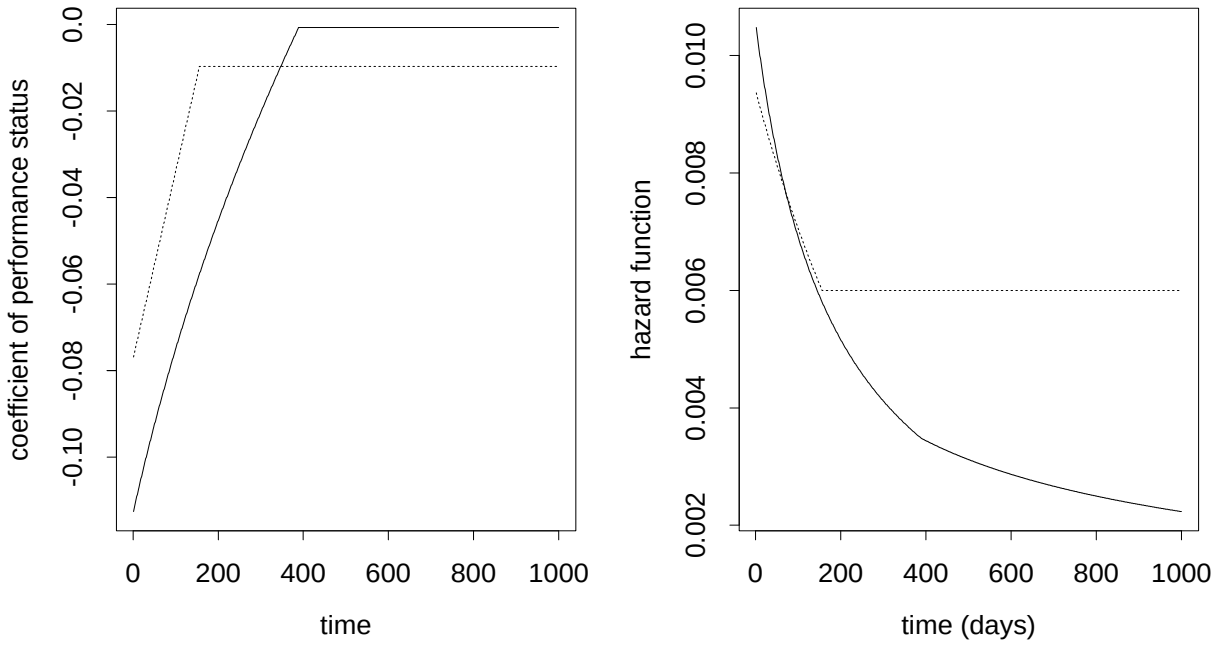

Fig. 4. Fitted coefficient of performance status as a function of time and fitted hazard function for a person with cell type squamous and performance status 40. Solid=transformed, dashed=untransformed.

function.)

For each simulation we generated a new set of survival times $T_i^* = \hat{q}_0^{-1}(t_i)$ for $1 \le i \le 137$, where t_i for $1 \le i \le 137$ is an independent sample generated using `rhare`, the fit summarized in Table 3, and the same covariates as in the original data; here $\hat{q}_0$ is obtained from the HEFT fit as described in Section 8.2. Then we generated the censoring times C_i^*, $1 \le i \le 137$, as a random sample from the exponential distribution with mean 1851. For each i we set $Y_i^* = \min(T_i^*, C_i^*)$ and $\delta_i^* = \text{ind}(T_i^* \le C_i^*)$. Using `heft.fit` with the default options, we transformed Y_i^* for $1 \le i \le 137$, after which we used `hare.fit`, also with the default options,

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

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 2 with respect to the variables involved.	

to fit a model to the conditional log-hazard function of $(\hat{q}_0^*(Y_i^*), \delta_i^*, \mathbf{x}_i)$, $1 \leq i \leq 137$.

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.

From Table 4 we see that in only 2 out of 100 simulations did the model fit by HARE have three 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. On the other hand, this simulation does not indicate whether the models in Tables 1 and 2 omit some practically important interactions.

9.2. PBC Data

Our second example illustrates many of the features of `hare.fit` that facilitate the search for the best model to fit the data. This example involves data from a double-blind, randomized trial involving primary biliary cirrhosis of the liver (PBC), which are 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 $\hat{\lambda}_0(t) \approx \exp(-8.498)$. Thus no transformation was needed to make the unconditional hazard function approximately constant.

We continued our analysis by applying HARE with the default options to the 274 cases with 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 missing values. Whichever

TABLE 5. HARE analysis of the PBC data.

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

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. Since the model selection algorithm described in Section 4 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.

It is possible to specify the maximum number of basis functions in a model, overriding the default value for $P_{\max}$ (Section 11.4). For the PBC data, changing $P_{\max}$ in `hare.fit` consistently resulted in the same fitted model as described above. It is also possible to use a model obtained from a previous `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 fit during the addition stage or 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 differs from the starting model, the new model has a lower AIC. If the original model was fit during the addition stage, HARE would necessarily return the same model. The latter was the case for the PBC data.

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

The `hare.summary` command also provides information about the influence of the choice of the penalty parameter a in (4.1). Table 6 consists of a part of the output of `hare.summary` when applied to the model from Table 5.

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

The model that was obtained with HARE using 6 as the penalty parameter turned out to be additive. The other eight nonconstant 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 use the option `additive` to fit an additive model, using the default value

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

$\log 310$ for the penalty parameter. The resulting fit is summarized in Table 7.

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 did 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. There is a discrepancy in that we include log(alkaline phosphatase) but not the indicator of edema.

9.3. Breast Cancer Data

We now consider a dataset that is much larger than those in the two previous examples. The data, discussed in Gray (1992), come from six breast cancer studies conducted by the Eastern Cooperative Oncology Group. There were 2404 patients in these studies. All patients had disease involvement in their axillary lymph nodes at diagnosis indicating some likelihood that the cancer had spread through the lympatic system to other parts of the body; however, no patients had evidence of disease at the time of entry into the study, which was following surgical removal of the primary tumor and axillary metasteses. The response is survival time (years) from entry into the study. There are six covariates, estrogen receptor status (ER: 0 is 'negative', 1 is 'positive'), the number of positive axillary 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.

In this dataset some of the survival times are exactly zero, which makes it impossible to include $\log(t/(t+c))$ as a basis function in the HEFT fit. Thus we used the option in

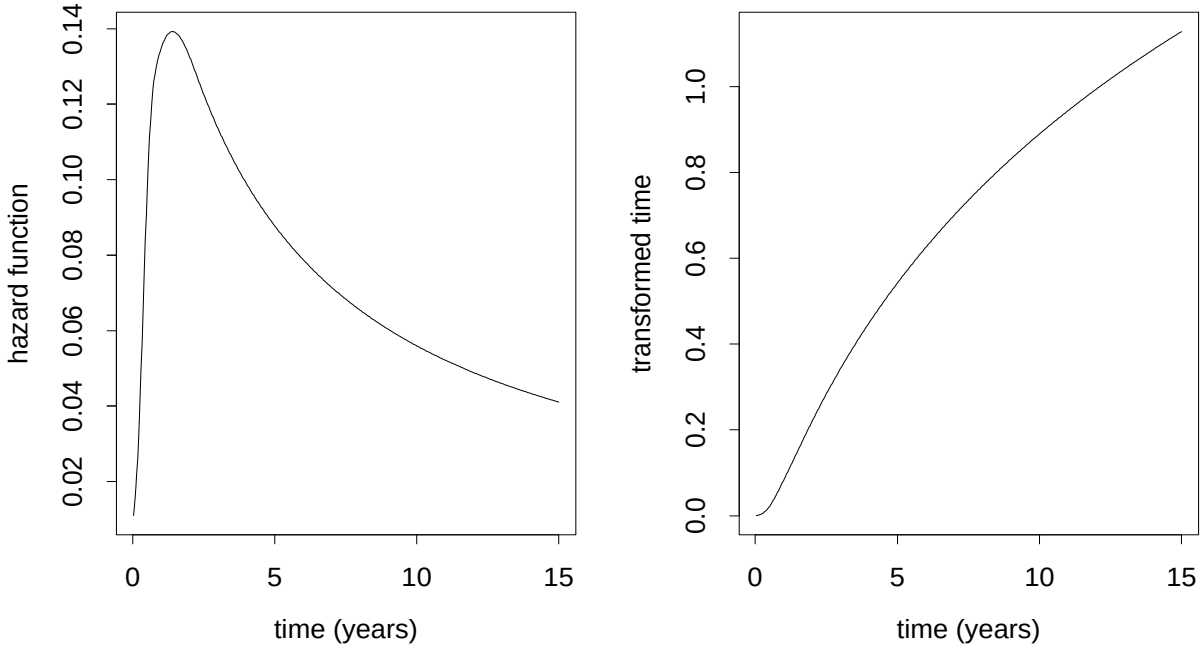

Fig. 5. Estimated unconditional hazard function and the corresponding transformation of time using HEFT for the breast cancer data.

`heft.fit` that lets 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 5. This fit has four parameters since it is based on four knots and the term $\log(t+c)$.

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 9.2 did not yield a better fit. When HARE was applied to the untransformed data, the resulting fit was similar, but it included one more knot in time. In Figure 6 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 preliminary transformation of time using HEFT. Note that the selected values of the covariates are close to the corresponding median values 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 $(Y_i, 1-\delta_i, \mathbf{x}_i)$, $1 \le i \le n$.

The estimated fit to the censoring distribution that we obtained using HEFT is multi-

TABLE 8. HARE analysis of the transformed breast cancer data.

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

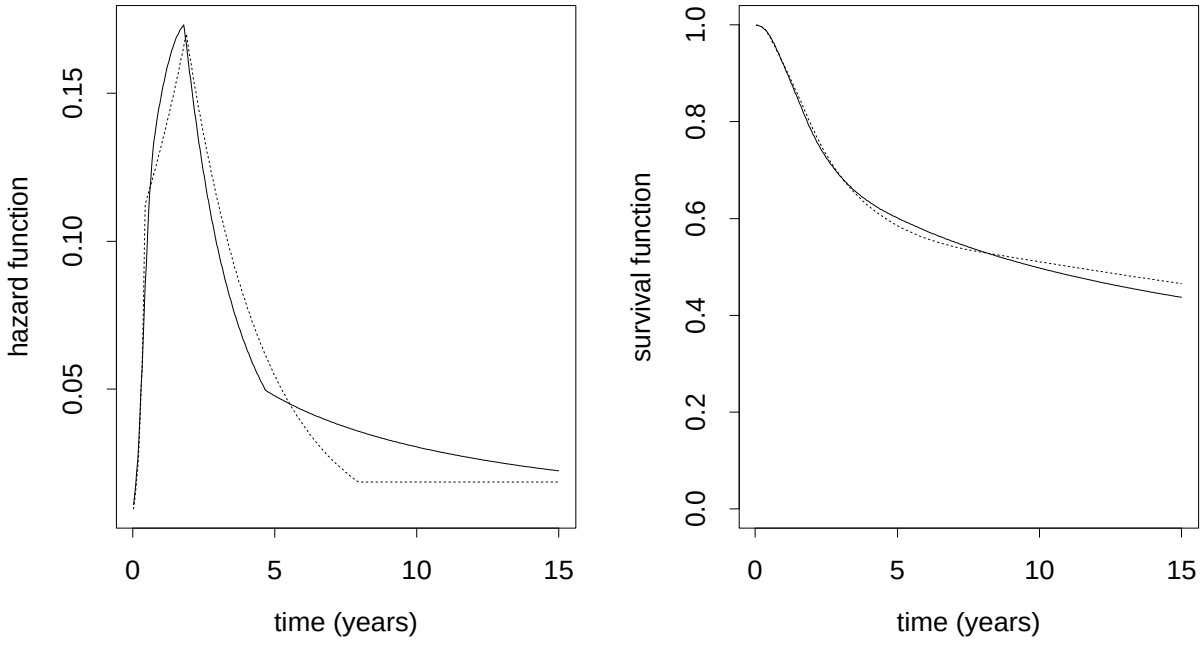

Fig. 6. Fitted hazard and survival functions for a premenopausal woman of age 50 with negative estrogen receptor status, 4 nodes, body mass index 25 and tumor size 3 cm. Solid=transformed using HEFT, dashed=untransformed.

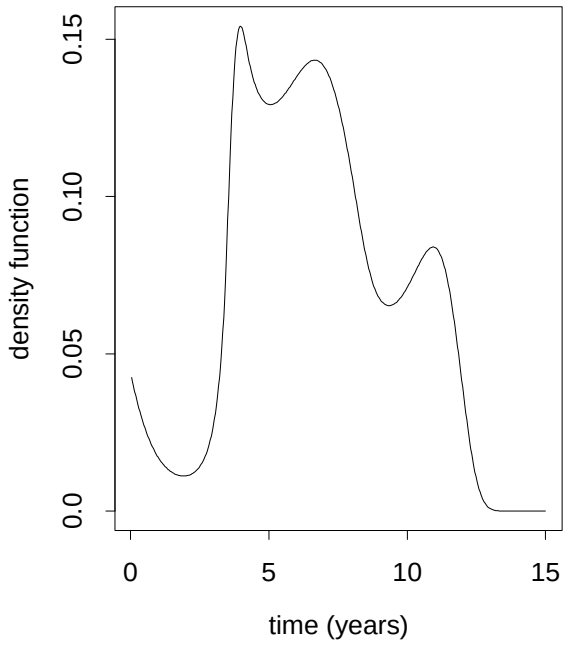

Fig. 7. Estimated unconditional density function for the censoring distribution for the breast cancer data.

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

modal, the corresponding density estimate being shown in Figure 7. Private communication with Robert Gray from the Eastern Cooperative Oncology Group suggests that the multimodality is due to different accural periods and patient populations in the six studies. After applying HARE to the transformed data, we obtained a model with two basis functions, summarized in Table 9. This analysis suggests that the conditional distribution of the censoring time depends on whether a woman is premenopausal or postmenopausal. The hazard of censoring is about 27% larger for postmenopausal women. On the other hand, three of the six studies were limited to postmenopausal women, two to premenopausal women and one study included both. Thus the apparent effect due to postmenopausality could actually be due to the different accural periods in the six studies. (At the time of this writing, we do not have the data that would allow us further to pursue this issue.)

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

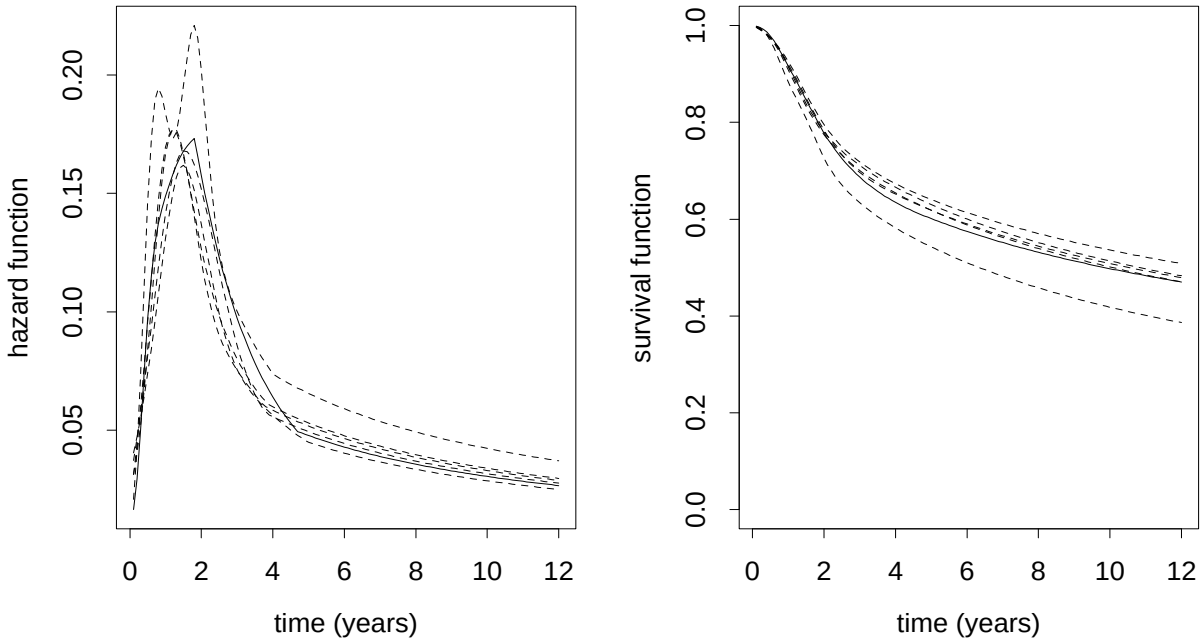

Fig. 8. Conditional hazard and survival functions for the fit of Table 8 (solid) and five random samples from this model (dashed). Same covariates as in Figure 6.

a new set of survival times $T_i^* = \hat{q}_0^{-1}(t_i)$ for $1 \leq i \leq 2404$, where t_i for $1 \leq i \leq 2404$ 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 5. 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 7. 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 $(\hat{q}_0^*(Y_i^*), \delta_i^*, \mathbf{x}_i)$, $1 \leq i \leq 2404$.

In Figure 8 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 6. In Figure 9 we summarize these quantities for all 200 simulations. In particular, for every time t we computed the 2.5th and 97.5th percentiles 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 10 and 11 summarize the effect of some of the covariates. The bootstrap bands in these figures are constructed as in Figure 9. In the left side of Figure 10 we show log(hazard ratio) $\hat{\alpha}(t|\mathbf{x}_1) - \hat{\alpha}(t|\mathbf{x}_2)$ as a function of time; here $\mathbf{x}_1$ and $\mathbf{x}_2$ are identical to the vector of covariates used in Figure 6, except that Estrogen Receptor status equals one (ER is positive) in $\mathbf{x}_1$ and it equals zero (ER is negative) in $\mathbf{x}_2$. The right side of Figure 10 displays the effect

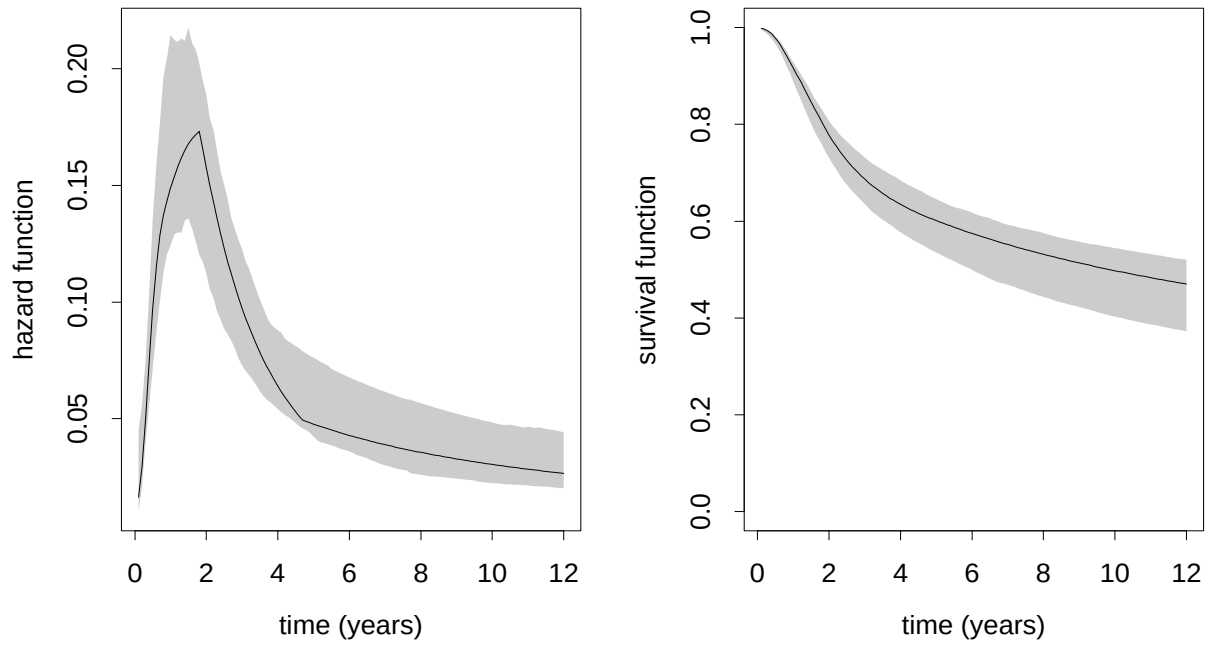

Fig. 9. Conditional hazard and survival functions for the fit of Table 8 and 95% bootstrap bands from that model. Same covariates as in Figure 6.

of the number of nodes on the log hazard. Specifically, in this figure we show

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

That is, we show log(hazard ratio) when time = 2 and all covariates are kept fixed at the same value as in Figure 6, 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 equals zero when nodes = 4 is a consequence of the fact that $g(4)$ equals zero by definition.

Similarly, the left side of Figure 11 shows log(hazard ratio) when time = 2 and all covariates are kept fixed at the same value as in Figure 6, except that the tumor size is allowed to vary and is compared with size = 3. The right side of Figure 11 shows log(hazard ratio) when time = 2 and all covariates are kept fixed at the same value as in Figure 6, except that age is allowed to vary and is compared with age = 50.

The results of our analysis of the breast cancer data are similar to those in Gray (1992). In particular, compare the left and right sides of Figure 10 and the left and right sides of Figure 11 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.

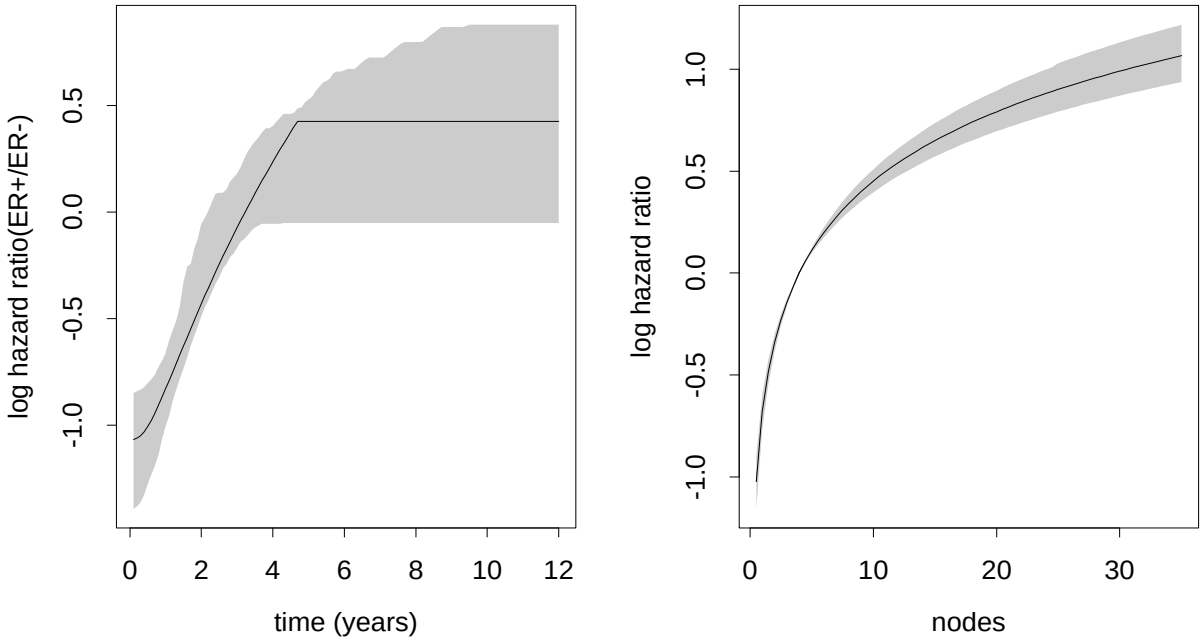

Fig. 10. Log of the hazard ratio for the fit of Table 8 and 95% bootstrap bands for that fit; left side: as a function of time for the ratio ER positive/ER negative; right side: as a function of nodes, relative to nodes = 4 after 2 years; other covariates as in Figure 6.

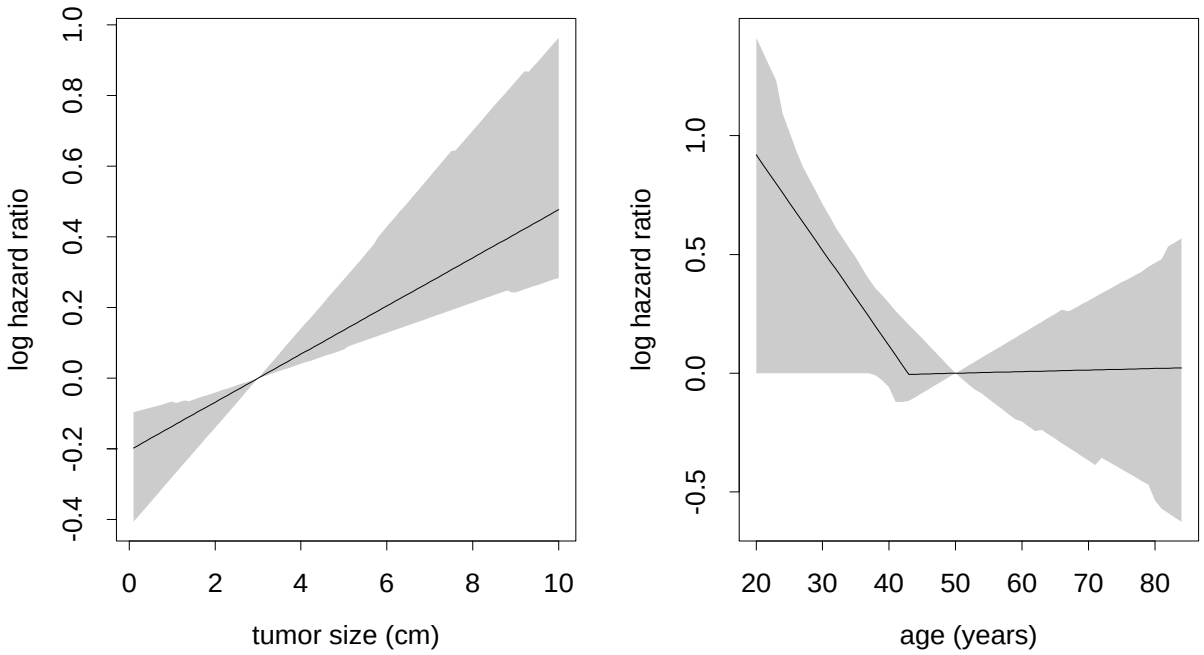

Fig. 11. Log of the hazard ratio for the fit of Table 8 and 95% bootstrap bands for that fit after 2 years; left side: as a function of size relative to size = 3; right side as a function of age, relative to age = 50; other covariates as in Figure 6.

The bootstrap bands in Figures 9 through 11 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. If this parameter is sufficiently large (say, 200 or more), 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 9 through 11, are very narrow. In the opposite direction, when the penalty parameter is extremely small (in particular, when it equals zero), the corresponding bootstrap bands are very large. Clearly, however, we should not think of the various estimates as steadily improving in accuracy as the penalty parameter increases from zero to infinity. These reservations about the bootstrap bands and similar reservations about standard errors, which apply 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, deserve much greater emphasis in the literature.

10. CONCLUDING REMARKS

In light of the examples in Section 9 and considerable additional experience with HARE and HEFT and their user interfaces, we are convinced that these methodologies are of considerable practical value. In particular, as Figure 6 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 fit using HARE. Thus this methodology is potentially useful for a health care practitioner in coming up with a prognosis for a particular patient.

HEFT, as described in this paper, combines cubic splines with stepwise addition and deletion of knots for the estimation of the log-hazard function. It has two extra log terms, which are specifically tailored to fit the tails of the underlying distribution. 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. Another important advantage of the two log terms is that it is possible to estimate a distribution with a simple parametric form in some situations (as in the lung cancer and PBC examples), while in other situations HEFT can estimate well using a spline model with a small number of knots (as is the case in the breast cancer example). These features make HEFT an ideal preprocessor for HARE.

The available features in HARE make it easy to try a variety of models on a given set of data. In particular linear proportional hazards models, additive proportional hazards models, proportional hazards models with time-varying coefficients, and nonparametric proportional hazards models can conveniently be fitted and compared. A limitation of HARE is that the present implementation, although ideally suited for the study of time-varying coefficients (sometimes called time-dependent covariate effects), does not allow for time-dependent covariates; that is, covariates whose value may change during the study. A future version of

HARE should be able to deal with such covariates.

11. NUMERICAL IMPLEMENTATION

11.1. Starting Values

As the starting value in the Newton-Raphson algorithm for obtaining the maximum likelihood estimate of the log-hazard function in the minimal allowable space, we use the maximum likelihood constant estimate $\hat{\alpha} = \log(\sum_i \delta_i / \sum_i Y_i)$ 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 $\hat{\beta}_1 B_1 + \cdots + \hat{\beta}_p B_p$ 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 $1 \le j \le p$, let $\sum_{k=1}^{p-1} a_{jk} \tilde{B}_k$ be the orthogonal projection of B_j onto G_0 relative to the inner product

$$\langle \lambda_1, \lambda_2 \rangle = \sum_i \lambda_1(Y_i|\mathbf{x}_i)\lambda_2(Y_i|\mathbf{x}_i).$$

Since

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

we use $\sum_{j=1}^{p} a_{jk} \hat{\beta}_j$, $1 \le k \le p-1$, as the starting value for the maximum likelihood estimate of the conditional log-hazard function to G_0.

11.2. Computation of the Log-Likelihood Function, Score Function and Hessian for HEFT

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

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

for many functions ψ that are 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 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 evaluation 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 are determined.

Suppose one or more of the uncensored observations equal zero. If the coefficient β_{-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.

11.3. Optimizing the Location of a New Knot

In this section we describe the algorithm for finding the location of a potential new knot in a covariate for the HARE model. The addition of a new knot in time for HEFT or HARE is similar. Since it is not possible to come up with simple updating formulas to compute score functions and Hessians in HEFT and HARE for many new potential knots with little effort as in MARS (Friedman 1991), we need to limit the number of knots for which we compute the Rao statistic.

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

$$l_i \;=\; 6 + \arg\max_{1\le j\le n} X_{(j)} \le t_i, \qquad i = 1, \ldots, K_m, \tag{11.1}$$

$$
\begin{aligned}
u_i &= -6 + \arg\min_{1 \le j \le n} X_{(j)} \ge t_{i+1}, \qquad i = 0, \ldots, K_m - 1, \qquad (11.2)\\
l_0 &= 1 \qquad \text{and}\\
u_k &= n.
\end{aligned}
$$

For $i = 0, \ldots, K_m$ such that $u_i \ge l_i$ we compute the Rao statistic r_i for the model with $(x_m - X_{(j_i)})_+$ as the new basis function, where $j_i = [(l_i + u_i)/2]$. Because of the 6 and -6 in (11.1) and (11.2) 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 $u_i \ge l_i$, 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 the new basis function with $l = [(l_{i^*} + j_{i^*})/2]$, and r_u for the model with $(x_m - X_{(u)})_+$ as the new basis function with $u = [(j_{i^*} + u_{i^*})/2]$. 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 $[X_{(l_{i^*})}, X_{(j_{i^*})}]$; and 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 in a similar manner, except that we select its location based on the ordered statistics just of the uncensored data.

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

11.4. Maximal Number of Basis Functions

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

- the number P of basis functions equals $P_{\max}$, where the default value for $P_{\max}$ is $\min(6n^{1/5}, n/4, 50)$ in HARE and $\min(4n^{1/5}, n/4, 30)$ in HEFT;
- $\hat{l}_P - \hat{l}_p < \frac{1}{2}(P - p) - 0.5$ for some p with $3 \le p \le P - 3$, where $\hat{l}_p$ is the log-likelihood for the model with p basis functions;
- the search algorithm yields no possible new basis function.

Note that the default value for $P_{\max}$ is somewhat arbitrary and mainly the result of experience. However, the power rate is somewhat motivated by the theoretical results in Kooperberg, Stone and Truong (1993), the $n/4$ bound prevents models for small datasets from being too large, while the constant upper bound prevents models for large datasets from being too large.

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.

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

——— 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 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.

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.

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

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

——— and ——— (1993), "Varying-coefficient models (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, 55, 757–800.

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.

——— and ——— (1993), Contribution to the discussion of "Varying-coefficient models" by Hastie, T. and Tibshirani, R., *Journal of the Royal Statistical Society*, Ser. B, 55, 791–793.

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

——— (1988b), "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, 170–174.

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.

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