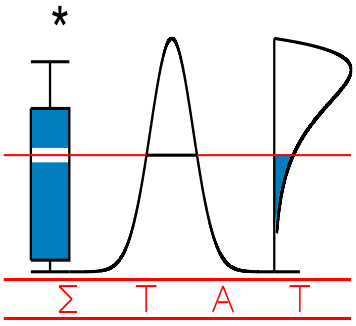
T E C H N I C A L

R E P O R T

0451

ACCELERATED FAILURE TIME MODEL FOR ARBITRARILY CENSORED DATA WITH SMOOTHED ERROR DISTRIBUTION

A. KOMÁREK, E. LESAFFRE, J. F. HILTON



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Accelerated Failure Time Model for Arbitrarily Censored Data with Smoothed Error Distribution

Arnošt Komárek[1], Emmanuel Lesaffre[1], Joan F. Hilton[2]

[1] Katholieke Universiteit Leuven, Biostatistical Centre, Kapucijnenvoer 35, B–3000, Leuven, Belgium

[2] University of California San Francisco, Dept. of Epidemiology and Biostatistics, 500 Parnassus Avenue, CA 94143–0560, San Francisco, California

Corresponding author: Arnošt Komárek

E-mail: arnost.komarek@med.kuleuven.ac.be

Tel.: +32 - 16 - 336886

Fax: +32 - 16 - 336900

ABSTRACT: We developed a semi-parametric procedure to estimate parameters of an accelerated failure time model. To express the density of the error distribution, we use the P-spline (B-splines with penalties) smoothing technique of Eilers and Marx (1996). To accommodate error densities with infinite support and for other reasons, we replace the B-splines with their limits as the degree of the B-spline goes to infinity; namely, with normal densities. The spline coefficients as well as any number of regression parameters are quickly and accurately estimated via penalized maximum likelihood. The method directly provides predictive survival distributions for fixed values of covariates while allowing for left-, right-, and interval-censored data. The approach has been implemented as an R library and is applied here to the problem of predicting AIDS-free survival in the presence of interval censoring.

KEY WORDS: Linear regression; Penalized maximum likelihood; Spline; Survival analysis.

---

1 Introduction

The aim of this article is to present the *penalized Gaussian mixture method* for analysing survival data using an accelerated failure time (AFT) model with the following characteristics: (1) the baseline survival distribution does not have to be specified; (2) it is directly estimated, thus allowing for prediction; and (3) not only right-censored but also left- and especially interval-censored data can be handled.

The accelerated failure time (AFT) model has a respected role in survival analysis today even though it is used far less broadly than Cox's proportional hazards (PH) model (Cox, 1972). Whereas Cox's model relates the hazard function to covariates, the AFT model postulates a direct relationship between the time to event and covariates. It specifies that the effect of a vector of fixed covariates $\boldsymbol{x}$ acts multiplicatively on the time to event T, or additively on $Y = \log(T)$ as

$$\log(T) = \alpha + \boldsymbol{\beta}'\boldsymbol{x} + \sigma\varepsilon, \tag{1}$$

where α and $\boldsymbol{\beta}$ are regression parameters, σ is a scale parameter, and ε is the random error with density $f(e)$.

Classical semi-parametric approaches to the AFT model that emphasize estimation of the regression parameters $\boldsymbol{\beta}$ include the method of Buckley and James (1979) and linear-rank-test-based estimators (see Chapter 7 of Kalbfleisch and Prentice (2002) for a comprehensive exposition and references). A drawback of the Buckley-James method is that it may fail to converge or may oscillate among several solutions. A drawback of linear-rank-test-based estimators is that only with considerable difficulties can they be extended to handle interval-censored data. Rabinowitz, Tsiatis, and Aragon (1995) and Betensky, Rabinowitz, and Tsiatis (2001) consider linear-rank-test-based estimation in the AFT model under interval censoring. Their method is computationally tractable only with a low dimensional covariate vector $\boldsymbol{x}$. A closely related approach is semi-parametric median regression with censored data in which the median rather than the mean of the (log)-event times is expressed as a linear function of covariates (Ying, Jung, and Wei, 1995; Yang, 1999; and McKeague, Subramanian, and Sun, 2001). However, all of the above-mentioned semi-parametric

methods become computationally intractable when the dimension of the covariate vector increases and only handle interval censoring with great difficulty. Furthermore, these methods do not provide an estimate of the baseline error distribution, which rules out their use for prediction purposes.

In contrast to frequentist approaches, Bayesian methods easily handle all types of censoring by incorporating true event times as parameters in the model and treating them in the same way as unknown covariate parameters. Bayesian semi-parametric AFT models are described by Christensen and Johnson (1988), Johnson and Christensen (1989), Kuo and Mallick (1997), Walker and Mallick (1999), Kottas and Gelfand (2001), and Hanson and Johnson (2002). Recently, Hanson and Johnson (2004) presented a Bayesian AFT model with an explicit treatment of interval censoring. However, here we concentrate on classical maximum-likelihood-based estimation, which some statisticians prefer over Bayesian methods.

Regardless of whether the baseline distribution is either not specified (in the case of the Cox's PH model) or is modelled flexibly (in the case of the AFT model, as in this paper), both the AFT and the PH models make specific assumptions regarding the effects of covariates on the respective baseline hazard and survival distributions. To weaken reliance on such assumptions, a general extended hazard regression model, encompassing both PH and AFT models, was suggested by Etezadi-Amoli and Ciampi (1987) and was studied further by Shyur et al (1999) and Chen and Jewell (2001). The first two papers use quadratic splines to express the baseline hazard function and then maximum-likelihood to estimate regression parameters. Prediction can be carried out easily and interval censoring poses no difficulties. The last paper presents a method which is in the mood of linear-rank-test-based estimators for AFT models, with all the drawbacks described earlier.

Apart from the fact that we accommodate only the AFT model, our approach could be considered a competitor to the approaches of Etezadi-Amoli and Ciampi (1987) and Shyur et al (1999). Whereas they use splines to express the baseline hazard function, we exploit them to flexibly model the density of the baseline log-event times. Additionally, instead of estimating the positions of the knots of the splines, we use penalized splines which fix the locations of the knots.

The second section describes the penalized Gaussian mixture method in detail. Section 3 gives

the inference based on our method. Results of a simulation study that evaluates performance of the method is given in Section 4. The fifth section continues with an illustration of the use of the method using a real data example. Section 6 finalizes the paper with further discussion.

2 Penalized Gaussian mixture method

2.1 Density of the error term

The motivation for our method stems from exploiting penalized B-splines (P-splines) to smooth the error density $f(e)$ (see de Boor, 1978; and Dierckx, 1993 for more details on B-splines; and Eilers and Marx, 1996 for the concept of P-splines). To model a density, it is advantageous to replace the B-spline basis by a set of Gaussian densities for the following reasons. Firstly, the error density $f(e)$ is usually viewed as having an infinite support; however, this is not provided by a B-spline basis, which is equal to zero below and above the boundary knots. Secondly, a basis formed by Gaussian densities covers the standard parametric log-normal AFT model as a special case. Further, such an approach can be viewed as the limiting case of B-spline smoothing, since an appropriately normalized B-spline basis converges uniformly on $\mathbb{R}$ (as its degree tends to infinity) to a Gaussian density (see Unser et al (1992) for details).

We express the density of the error term in the AFT model (1) as

$$f(e|\boldsymbol{c}) = \sum_{j=1}^{g} c_j \varphi_{\mu_j,\sigma_0^2}(e), \tag{2}$$

where $\varphi_{\mu_j,\sigma_0^2}(e)$ is the Gaussian density with mean μ_j and variance σ_0^2, and $\boldsymbol{c} = (c_1, \ldots, c_g)^T$ are mixture coefficients that have to be estimated. Values of $\mu_1, \ldots, \mu_g$ and σ_0^2 are fixed by design, as explained below. When used in this context, we call the basis functions $\varphi_{\mu_j,\sigma_0^2}$ *basis Gaussian densities*, or briefly *BG-densities*.

The role of fixed knots in spline smoothing is played by the means $\mu_1 < \cdots < \mu_g$ in our approach. Choosing the optimal number and positions of knots are generally complex tasks. Too many knots leads to overfitting the data; too few leads to underfitting and inaccuracy. In our method, we build on the proposals of O'Sullivan (1986, 1988) and of Eilers and Marx (1996). O'Sullivan suggested

taking relatively many knots and restricting the flexibility of the fitted curve by putting a penalty on the second derivative of the spline function. Eilers and Marx extended this approach in the context of B-splines, using penalty terms based on squared finite higher-order differences between adjacent mixture coefficients c_j. They then maximized the penalized log-likelihood instead of the ordinary log-likelihood function when estimating parameters. We also adopted Eilers and Marx's suggestion to use equidistant knots.

It is important to stress that a mixture of BG-densities is different from a classical Gaussian mixture. With our BG approach, invariably a relatively large but fixed number of mixture components is needed and the smoothness of the resulting error distribution is optimized via a penalty term on the log-likelihood. Our fine grid of knots prevents inaccuracy in the estimate of the error density, while our penalization of the log-likelihood inhibits overfitting. In contrast, in the case of a classical Gaussian mixture the number of mixture components must be estimated, along with the means and the standard deviations of the Gaussian components. Although our model still requires estimation of a relatively large number of parameters, maximization of the (penalized) log-likelihood remains fairly straightforward and does not require an EM-type algorithm or a numerical search for the optimal number of mixture components.

With respect to the actual values of the knots and of the basis standard deviation σ_0 we adopted the following procedure. Since $f(e|\boldsymbol{c})$ is a standardized density, taking the range of knots from -6 to 6 is broad enough even for distributions with heavy tails such as the extreme value distribution. A distance of 0.3 between two consecutive knots is small enough to approximate $f(e|\boldsymbol{c})$ with satisfactory precision, as will be illustrated in the next paragraph. Furthermore, with a choice of $\sigma_0 = 2/3\,(\mu_{j+1} - \mu_j)$, each BG-density overlaps with its 6 neighbors practically as the cubic basis B-spline does (with choice of σ_0 as above, a BG-density is practically zero outside (μ_{j-2}, μ_{j+2})) just as a normal density is practically zero outside $\mu \pm 3\sigma_0$).

As an illustration, we computed the L_2-distance between the standard Gaussian density and its best approximation using a mixture of BG-densities with $\mu_1 = -6$, $\mu_g = 6$, different choices of $\delta = \mu_{j+1} - \mu_j$, and $\sigma_0 = 2/3\,\delta$. This distance is equal to 0.00570 for $\delta = 1$ ($g = 13$), and drops to

0.00104 for $\delta = 0.75$ ($g = 17$). When plotted, the mixture of BG-densities is indistinguishable from the Gaussian density at $\delta = 0.75$. Further, for δ equal to 0.5 ($g = 25$), 0.4 ($g = 31$), 0.3 ($g = 41$), 0.2 ($g = 61$), and 0.1 ($g = 121$) we obtain distances of 0.00031, 0.00022, 0.00017, 0.00014, and 0.00012, respectively. Clearly, the choice of $\delta = 0.3$ yields very precise correspondence between the mixture of BG-densities and the normal density.

To ensure that $f(e|\boldsymbol{c})$ is a density function, some constraints must be imposed on the mixture coefficients $\boldsymbol{c}$, i.e.,

$$\sum_{j=1}^{g} c_j = 1, \qquad c_j > 0 \quad (j = 1, \ldots, g).$$

To avoid constrained maximization, one can use an alternative parametrization based on coefficients $\boldsymbol{a}$,

$$c_j(\boldsymbol{a}) = \frac{\exp(a_j)}{\sum_{l=1}^{g} \exp(a_l)} \qquad (j = 1, \ldots, g),$$

with one of the a_j's fixed to a particular value, say $a_g = 0$.

Further, rendering the intercept α and the scale σ identifiable requires that the first two moments of the density (2) be fixed, i.e.,

$$\mathrm{E}(\varepsilon|\boldsymbol{a}) = \sum_{j=1}^{g} c_j(\boldsymbol{a})\mu_j = 0, \qquad \mathrm{var}(\varepsilon|\boldsymbol{a}) = \sum_{j=1}^{g} c_j(\boldsymbol{a})(\mu_j^2 + \sigma_0^2) = 1. \tag{3}$$

It is easily seen that the basis standard deviation σ_0 must be smaller than one to be able to satisfy the variance constraint. Finally, the two equality constraints (3) can be avoided if two coefficients, say, a_{g-2} and a_{g-1}, are expressed as functions of the remaining coefficients, denoted together as a vector $\boldsymbol{d} = (a_1, \ldots, a_{g-3})'$:

$$a_k(\boldsymbol{d}) = \log\Big\{\omega_{g,k} + \sum_{j=1}^{g-3} \omega_{j,k} \exp(a_j)\Big\} \qquad (k = g-2,\ g-1), \tag{4}$$

with

$$\begin{aligned}
\omega_{j,g-2} &= -\frac{\mu_j - \mu_{g-1}}{\mu_{g-2} - \mu_{g-1}} \cdot \frac{1 - \sigma_0^2 + \mu_{g-1}\mu_j}{1 - \sigma_0^2 + \mu_{g-1}\mu_{g-2}}, \\
\omega_{j,g-1} &= -\omega_{j,g-2} \cdot \frac{\mu_{g-2}}{\mu_{g-1}} - \frac{\mu_j}{\mu_{g-1}} \qquad\qquad (j = 1, \ldots, g-3,\ g).
\end{aligned}$$

To reflect implementation of these three constraints, the density of the error distribution, a mixture of BG-densities, subsequently will be denoted as $f(e|\boldsymbol{d}) = \sum_{j=1}^{g} c_j(\boldsymbol{d})\varphi_{\mu_j,\sigma_0^2}(e)$ rather than $f(e|\boldsymbol{c})$.

All parameters in the model (transformed mixture coefficients $\boldsymbol{d}$; regression parameters α, $\boldsymbol{\beta}$; and log-scale $\log(\sigma)$) are estimated by means of a penalized maximum likelihood method. In the next section, we construct the penalized log-likelihood function which consists of an ordinary log-likelihood and a difference penalty for the transformed spline coefficients. The penalized log-likelihood is subsequently maximized to obtain the estimates. Hence, we call our approach the *penalized Gaussian mixture method* (PGM method).

2.2 Penalized maximum-likelihood

2.2.1 Penalized log-likelihood

Let $\boldsymbol{\theta}$ be the vector of all unknown parameters to be estimated, i.e., $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', \log(\sigma), a_1, \ldots, a_{g-3})'$. Let $\ell_i(\boldsymbol{\theta}) = \ell_i(y_i|\boldsymbol{\theta})$ $(i = 1, \ldots, n)$ denote the ordinary log-likelihood contribution of the i-th observation based on model (1) with error density (2), and $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})$. With censored observations an integral of the error density has to be evaluated to get an individual $\ell_i(\boldsymbol{\theta})$. With our model, this does not cause any considerable difficulties irrespective of the type of censoring (left-, right-, interval-). Indeed, all integrals involved in the computation of the likelihood are Gaussian cumulative distribution functions which can be easily and efficiently evaluated.

To construct the penalized log-likelihood function $\ell_P(\boldsymbol{\theta}; \lambda)$, we subtract a penalty term $q\big\{\boldsymbol{a}(\boldsymbol{d}); \lambda\big\}$ based on the transformed mixture coefficients $\boldsymbol{a}(\boldsymbol{d})$ from $\ell(\boldsymbol{\theta})$, i.e.,

$$\ell_P(\boldsymbol{\theta}; \lambda) = \ell_P(\boldsymbol{y}|\boldsymbol{\theta}; \lambda) = \ell(\boldsymbol{\theta}) - q\big\{\boldsymbol{a}(\boldsymbol{d}); \lambda\big\}, \tag{5}$$

where λ is a fixed tuning parameter that controls the smoothness of the fitted error distribution and inhibits identifiability problems due to overparametrization. For a given (reasonable) λ, Eilers and Marx (1996) proposed to base the penalty on squared (higher-order) finite differences of the coefficients of adjacent B-splines, and they used second-order difference in their examples. We base our penalty on squared finite differences of order m of the transformed coefficients of adjacent

BG-densities:

$$
\begin{aligned}
q\{\boldsymbol{a}(\boldsymbol{d});\lambda\} &= \frac{\lambda}{2}\sum_{j=m+1}^{g}\{\Delta^m a_j(\boldsymbol{d})\}^2 \\
&= \frac{\lambda}{2}\boldsymbol{a}(\boldsymbol{d})'\boldsymbol{D}_m'\boldsymbol{D}_m\,\boldsymbol{a}(\boldsymbol{d}),
\end{aligned}
\tag{6}
$$

where $\Delta^1 a_j = a_j - a_{j-1}$, $\Delta^m a_j = \Delta^{m-1}a_j - \Delta^{m-1}a_{j-1}$, $m = 1, \ldots,$ and $\boldsymbol{D}_m$ is a $(g-m)\times g$ difference operator matrix. According to our experience, $m = 2$ or $m = 3$ is sufficient obtain a smooth estimate of the density. However, in our context the choice $m = 3$ has another interesting justification, as explained in Section 2.2.2.

In practice, we maximize the penalized log-likelihood first as a function of an extended parameter vector $(\alpha, \boldsymbol{\beta}', \log(\sigma),\ a_1, \ldots, a_{g-3}, a_{g-2}, a_{g-1})'$ under the constraints (3) using the sequential quadratic programming algorithm of Han (1977) to avoid negative values in the logarithmic expression (4). Upon convergence, we perform additional Newton-Raphson steps for the penalized log-likelihood as a function of parameters $\boldsymbol{\theta}$ in order to draw inferences as described in Section 3.

The estimation procedure has been implemented as a set of functions in R environment and can be downloaded from The Comprehensive R Archive Network (CRAN) on `http://www.R-project.org` as a contributed package `smoothSurv`.

2.2.2 Remarks on the penalty function

There are two reasons why we penalize the transformed mixture coefficients $\boldsymbol{a}$ instead of the original coefficients $\boldsymbol{c}$ and why we prefer the penalty of order $m = 3$.

First, the penalty based on $\boldsymbol{a}$ distinguishes between areas of the density where there are few datapoints (e.,g., where the coefficients $\boldsymbol{c}$ are close to zero) and areas where there are many datapoints (e.g., where the coefficients $\boldsymbol{c}$ are well above zero); the penalty based on $\boldsymbol{c}$ cannot distinguish between these areas. For example,

$$
\begin{array}{lll}
\text{for} & \breve{\boldsymbol{c}} = (0.001, 0.002, 0.001, 0.996)', & \tilde{\boldsymbol{c}} = (0.201, 0.202, 0.201, 0.396)' \\
\text{we have} & \breve{\boldsymbol{a}} = (-6.904, -6.211, -6.904, 0)', & \tilde{\boldsymbol{a}} = (-0.678, -0.673, -0.678, 0)'
\end{array}
$$

$$
\begin{array}{lll}
\text{and} & (\Delta^2\breve{c}_3)^2 = 0.000004 & = (\Delta^2\tilde{c}_3)^2. \\
\text{while} & (\Delta^2\breve{a}_3)^2 = 1.92 \gg 0.000099 & = (\Delta^2\tilde{a}_3)^2
\end{array}
$$

Indeed, in the areas with a sufficient amount of data, the estimated shape of the error distribution is mostly driven by the data themselves, whereas in the data-poor areas the shape of the fitted error distribution is inter- or extrapolated from the data-rich areas according to the flexibility allowed by the penalty term.

Second, the penalty of the third order ($m = 3$) based on transformed mixture coefficients $\boldsymbol{a}$ has the following interesting property which can serve as a basis for an empirical test of normality (see Section 2.2.3). Suppose that for fixed K and given set of knots $-K$, $-K+1/K, \ldots,\ -1/K,\ 0,\ 1/K,\ \ldots,\ K-1/K,\ K$, we maximize the penalized log-likelihood (5) for $\lambda \to \infty$. This is equivalent (in the limit) to minimizing the penalty term (6) under the constraints (3). For fixed K, let f_K be the fitted error density arising from the above-mentioned optimization problem. It can be shown that $\lim_{K\to\infty} f_K(e) = \varphi_{0,1}(e)$, the standard normal density. In practice, the set of knots and the basis standard deviation recommended in Section 2.1 (i.e., knots from -6 to 6 by 0.3 and $\sigma_0 = 0.2$) give already rise to a fitted error density f_K practically indistinguishable from the normal density, $\varphi_{0,1}(e)$, when only the penalty term is minimized. This property does not hold for the order $m \neq 3$ of the penalty or when the penalty is based on the original mixture coefficients $\boldsymbol{c}$.

2.2.3 Selecting the smoothing parameter

In the area of density estimation, methods for selecting the smoothing parameter, λ, that rely on cross–validation are often used. The standard modified maximum-likelihood cross–validation score that we are attempting to minimize is

$$\mathrm{CV}(\lambda) = -\sum_{i=1}^{n} \ell_i(\hat{\boldsymbol{\theta}}^{(-i)}),$$

where $\hat{\boldsymbol{\theta}}$ is the penalized maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}^{(-i)}$ the penalized MLE based on the sample excluding the ith observation. However, computation and optimization of the cross–validation score are extremely computationally intensive in our case. In a similar context, O'Sullivan (1988) suggested a one-step Newton-Raphson approximation combined with a first-order Taylor series approximation. Applying his method in our setting results in an approximate cross-

validation score given by

$$\overline{\mathrm{CV}}(\lambda) = -\Big\{\sum_{i=1}^{n} \ell_i(\hat{\boldsymbol{\theta}}) - \mathrm{trace}\big(\hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{I}}\big)\Big\}, \tag{7}$$

where $\hat{\boldsymbol{H}} = -\partial^2 \ell_P(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ and $\hat{\boldsymbol{I}} = -\partial^2 \ell(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$.

We denote $\mathrm{trace}(\hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{I}})$ by $\mathrm{df}(\lambda)$ and call it the *effective degrees of freedom* or the *effective dimension* of the model since it necessarily plays the same role as the effective dimension of a linear smoother (see Hastie and Tibshirani, 1990). Depending on a chosen order m of the differences in the penalty, the degrees of freedom decreases in λ from $\dim(\boldsymbol{\beta}) + 2 + (g-3)$ for $\lambda = 0$ (i.e., the ordinary log-likelihood) to $\dim(\boldsymbol{\beta}) + 2 + (m-3)$ for $\lambda \to \infty$ and $m \geq 3$ (i.e., the penalized log-likelihood). For example, when $\mu_{j+1} - \mu_j = 0.3$, $\sigma_0 = 0.2$ ($g = 41$) and $m = 3$, penalized likelihood estimation as $\lambda \to \infty$ depends effectively on $g - m = 38$ fewer parameters than does ordinary likelihood estimation. Interestingly, this reduction in parameters is not just of theoretical value; in practice, values of λ that are virtually equal to ∞ do arise, as illustrated in Section (5); they result in significant computational savings.

Further, minimizing the expression (7) is essentially the same as maximizing Akaike's information criterion $\mathrm{AIC}(\lambda) = \ell(\hat{\boldsymbol{\theta}}) - \mathrm{df}(\lambda)$ (Akaike, 1974). This can be a valuable means of comparing different models and assessing the importance of covariate contributions (see an example in Section 5).

In our R programs, a grid search using user-defined values $\lambda_1^*, \ldots, \lambda_S^*$ (in our applications we used values $\lambda_1^* = e^2, \lambda_2^* = e^1, \ldots, \lambda_S^* = e^{-9}$) is used to find the optimal AIC. Since the log-likelihood is of the order $O(n)$, using a factor of $n\lambda_s^*/2$ in the penalty term (6) instead of $\lambda/2$ allows one to use approximately the same grid for datasets of different sizes while also maintaining the proportional importance of the penalty term in the penalized log-likelihood at the same level.

The result of the second paragraph of Section 2.2.2 further implies that with a sufficiently dense set of knots, we can check the normality of the error term. When the optimal value of the tuning parameter λ approaches infinity (i.e., takes a high value in practical situations) the error density of the model can be considered to be normal.

3 Inference based on the penalized Gaussian mixture method

For $\lambda > 0$, the penalized MLE $\hat{\boldsymbol{\theta}}$ is necessarily a biased estimator. For that reason, its standard errors may not be very informative if that bias is high. However, there are two possibilities for drawing accurate inferences based on penalized MLE.

3.1 Pseudo-variance

Wahba (1983) described a pseudo-Bayesian technique for generating confidence bands around the cross-validated smoothing spline. O'Sullivan (1988) used this technique in the penalized ML framework and his approach can be adopted also here. Basically, the penalized log-likelihood ℓ_P is viewed as a "posterior" log-density for the parameter $\boldsymbol{\theta}$ and the penalty term as a "prior" negative log-density of that parameter. Then, the second order Taylor series expansion of the "posterior" log-density around its mode $\hat{\boldsymbol{\theta}}$ leads to

$$\ell_P(\boldsymbol{\theta}) \approx \ell_P(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{H}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

Finally the Gaussian approximation gives "posterior" normal distribution for $\boldsymbol{\theta}$ with covariance matrix

$$\widehat{\mathrm{var}}_P(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{H}}^{-1}. \tag{8}$$

We call this estimate of the variance of the penalized MLE $\hat{\boldsymbol{\theta}}$ the "pseudo-variance estimate."

3.2 Asymptotic variance

More formal inference is possible under the following assumptions. First, we assume independent noninformative censoring (see Kalbfleisch and Prentice, 2002). Further, as the sample size n increases, we require that the knots (both number and positions) and the basis standard deviation be fixed. Let $\boldsymbol{\theta}_T$ be the true parameter value of $\boldsymbol{\theta}$, assuming initially that it exists. To be able to get asymptotically unbiased estimates we have to either keep the value of the smoothing parameter λ constant as $n \to \infty$ or let it increase at a rate lower than n (i.e., $\lambda = \lambda_n$ and $\lim_{n\to\infty} \lambda_n/n = 0$). Under these conditions, the penalty part of the penalized log-likelihood reduces its importance rela-

tive to the log-likelihood part as $n \to \infty$ (i.e., as the sample size n increases, the smoothness of the fitted error distribution is determined to greater extent by the data and to a lesser extent by the penalty). Then, in combination with standard maximum likelihood arguments, for arbitrary $\varepsilon > 0$ the penalized MLE $\hat{\boldsymbol{\theta}}$ satisfies $\mathrm{P}_{\boldsymbol{\theta}_T}\big(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T| < \varepsilon\big) \to 1$. Using the same arguments as in Gray (1992), one can further show that $\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T\big)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\lim_{n\to\infty}(n\,\boldsymbol{W})$ where the matrix $\boldsymbol{W}$ can be consistently estimated by

$$\widehat{\mathrm{var}}_A(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{I}}\,\hat{\boldsymbol{H}}^{-1}, \tag{9}$$

which we call the "asymptotic variance estimate." As pointed out by Gray (1992), the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ remains the same if the smoothing parameters λ_n are replaced by estimates satisfying $\hat{\lambda}_n/\lambda_n \xrightarrow{\mathrm{P}} 1$.

3.3 The pseudo-variance versus the asymptotic variance

In various applications, the pseudo-variance estimate (8) has been shown to be useful. When smoothing a spline curve $g(t)$, Wahba (1983) showed it yielded pointwise confidence intervals $\hat{g}(t) \pm z\sqrt{\widehat{\mathrm{var}}_P\{\hat{g}(t)\}}$, where z is a quantile of the normal distribution, that have good frequentist coverage properties. Verweij and Van Houwelingen (1994) used it in the context of penalized likelihood estimation in Cox regression; they called the square roots of its diagonal elements "pseudo-standard errors." Joly et al (1998) exploited this technique to get confidence bands on the hazard function smoothed using M-splines. In contrast, for the asymptotic variance estimate (9) there is no guarantee that for finite samples its middle matrix $\hat{\boldsymbol{I}}$ is positive semidefinite. Based on our experience, this problem is not rare. Finally, according to our simulations (results not shown), the pseudo-variance estimate (8) yields confidence intervals $\hat{\beta} \pm z\sqrt{\widehat{\mathrm{var}}_P(\hat{\beta})}$ for regression parameters with better coverage properties than the corresponding confidence intervals based on the asymptotic estimate (9).

3.4 Remarks

We have assumed in this section that the true parameter vector $\boldsymbol{\theta}_T$ exists. This does not have to be true. In particular, true $\boldsymbol{a}$ coefficients may fail to exist when the true error distribution is

not a mixture of BG-densities determined by the choice of knots and the standard deviation σ_0. However, if the distance between two consecutive knots is small enough, we argue that the mixture of BG-densities can approximate every continuous distribution sufficiently well that the assumption on the existence of the true parameter vector $\boldsymbol{\theta}_T$ is not restrictive at all. Generally, by increasing the sample size, the estimated coefficients $\boldsymbol{a}$ will yield a BG-density which is close to the true error density.

4 Simulation study

To see how the proposed method performs, we carried out a simulation study. 'True' uncensored data were generated according to the model

$$\log(T) = 1.6 - 0.8 \cdot z_1 + 0.4 \cdot z_2 + 1.4 \cdot \varepsilon,$$

where covariate z_1 was binary taking a value of 1 with probability 0.4 and covariate z_2 was generated according to the extreme value distribution of a minimum, with location 8.5 and scale 1. The model attempts to mimic an AFT model used for the dataset presented in the next section with z_1 playing the role of the covariate *lesion* and z_2 being distributed as $\log_2(1+\text{CD4 count})$. Time to the event T is expressed in months. The error term ε was generated from a standard normal distribution $\mathrm{N}(0,1)$, from a standardized extreme value distribution, and from a mixture of two normal distributions $0.4\ \mathrm{N}(-1.4, 0.8^2) + 0.6\ \mathrm{N}(0.93, 0.8^2)$. Samples of sizes 50, 100, 300, and 600 were generated. Each simulation involved 100 replications.

For each uncensored dataset we created four censored datasets that were then used to compute the estimates: a dataset with (1) approximately 20% right-censored and 80% uncensored observations *(light RC)*; (2) approximately 20% right and 80% interval-censored observations *(light RC + IC)*; (3) approximately 60% right and 40% uncensored observations *(heavy RC)*; (4) approximately 60% right and 40% interval-censored observations *(heavy RC + IC)*. The censoring was created by simulating consecutive 'visit times' for each subject in the dataset. Times of the first 'visits' were drawn from $\mathrm{N}(7, 1)$ distribution. Further, times between each consecutive 'visits' were simulated from $\mathrm{N}(6, 0.5^2)$.

This approach reflects the idea that subjects in our Oral Substudy were seen for the first time about 7 months after the onset of the parent study and then approximately every 6 months for several years. At each visit, subjects were withdrawn (censored) according to a prespecified percentage (between 0.4% and 0.7% for light censoring and between 4.0% and 5.0% for heavy censoring) creating right-censored observations provided that the uncensored event time T was greater than the visit time at which the subject was withdrawn. To obtain interval-censored observations, we took the 'visit' interval that contained the uncensored event time T.

For comparison, estimates for each dataset were computed using our smoothed procedure and using two parametric models: an AFT model on the log scale with a correctly specified error distribution (normal, extreme value or mixture of normals, respectively) and a log-normal AFT model For the smoothing procedure, the third order penalty, equidistant knots with a distance of 0.3 between consecutive knots, and the basis standard deviation of 0.2 were used.

Figure 1 shows average estimates of the regression parameters based on our smoothed procedure, on the AFT model with correctly specified error distribution, and on the log-normal AFT model for selected (least favorable) simulation settings. It is seen that, in most cases, our smoothed procedure performs better than the incorrectly specified log-normal AFT model and often only but slightly worse than the correctly specified parametric AFT model. Additionally, when our smoothing approach is used, the error distribution is reproduced rather satisfactory as can be seen in Figure 2. This property is quite important especially when the estimated model is to be used for prediction purposes. Further, it is seen that even for small samples the performance of our smooth procedure is quite similar to the performance of a parametric AFT model with a correctly specified error distribution. Complete results of the simulation study can be found on the journal's web page.

< Figure 1 about here.>

< Figure 2 about here.>

5 Illustration: the Women's Interagency HIV Study

To illustrate our method in real data, we present an example from AIDS research. We analyzed the subsample of the HIV-seropositive women participating in the Women's Interagency HIV Study (WIHS). The total study population (over 3000 participants) was enrolled between October 1994 and November 1995 through six clinical consortia at 23 sites throughout the United States. More information on the setup of the study can be found in Barkan et al (1998). Our subsample consisted of the 224 AIDS-free women who participated in the dental sub-study and for whom the HIV RNA viral load, the CD4 T-lymphocyte count, and lesion-marker status (described below) were available at the baseline visit.

For HIV positive people, it is of interest to describe the distribution of the time to the onset of an AIDS-related illness based on some measured quantities. Classically used predictors include the number of copies of the HIV RNA virus and the count of CD4 T-cells per ml of blood. We examined whether presence of one of the three lesion markers, oral candidiasis, hairy leukoplakia and angular cheilitis, is useful, possibly together with one or both laboratory predictors, in describing the distribution of the residual time to onset of AIDS.

As a response, we used the time in months between the baseline visit, defined as the first visit at which the lesion markers were collected by dental professionals, and the onset of an AIDS-related illness. Clinical AIDS diagnoses were self-reported in 73.5% of cases, presumptive or definitive in 17.5%, and indeterminate in 9%; the case definition did not depend on CD4 T-lymphocytes. For 66 cases the response was interval-censored, while for 158 cases it was right-censored. The average length of the interval between two examinations at which AIDS could be detected was 7 months. The average follow-up time was 41 months and the maximal follow-up time was 84 months.

The three lesion markers were summarized in one binary covariate, $lesion$, equal to one if at least one of the above mentioned three lesion markers was present. We also analyzed functions of the viral load and the CD4 count in the AFT model (i.e., $lvload = \log_{10}(1 + \text{viral load})$ and $lcd4 = \log_2(1 + \text{CD4 count})$). All three covariates are moderately to strongly associated with one another since, as AIDS progresses, viral load increases, CD4 count falls, and oral lesions occur more

frequently. In our sample, for women with $lesion = 0$ and 1, respectively, the median $lvload$ was 3.60 and 4.23 (Mann-Whitney p-value, 0.001), and the median $lcd4$ was 8.85 and 8.52 (Mann-Whitney p-value, 0.005). There was also a moderate negative correlation of -0.46 between $lcd4$ and $lvload$. These associations have to be taken into account when interpreting the results.

To obtain the results shown below, we used a sequence of 41 equidistant knots from -6 to 6 with a distance of 0.3 between each pair. The basis standard deviation was 0.2 and the third order difference were used in the penalty. Different models were compared using Akaike's information criterion and claims concerning the significance of the parameters were based on Wald's tests using the pseudo-variance estimate (8). Summary of the fitted models is shown in Table 1.

< Table 1 about here.>

< Figure 3 about here.>

If used alone (model (1) in Table 1) the effect of $lesion$ on the time to onset of AIDS is statistically significant ($p = 0.018$) and the estimated time is $\exp(-0.87) \approx 0.42$ times shorter for women with $lesion = 1$ than women with $lesion = 0$. According to the AIC values for models (2) and (3) in Table 1, the transformed CD4 count and viral load are equally good predictors of the time to onset of AIDS. Addition of the lesion marker (models (4) and (5)) improves the model with $lcd4$ considerably but improves the model with $lvload$ only slightly. Finally, some additional improvement is gained by considering the model with all three predictors (model (7)).

Figure 3 shows predictive survivor and hazard curves and predictive densities for women with $lesion = 0$ and $lesion = 1$ based on the simplest model $lesion$ and on the most complex model considered $lesion$+$lvload$+$lcd4$. The predictive survivor curves based on the model $lesion$ are further overlaid with the nonparametric estimate of Turnbull (1976) in each group. The two estimates are quite close to each other, illustrating the semiparametric nature of our approach. However, our procedure gives smooth estimates of the survival curves and moreover enables quantification of the difference in survival between the two groups. Notice further that due to the fact that the hazard is obtained as a ratio of the density and the survivor function, which relatively slowly varies from one, only a slight difference is observed between the predictive density and the hazard.

Further, we point out that the predictive densities for models where $lcd4$ was not involved are very close to the log-normal density. This is not surprising since the optimal tuning parameter λ for these models was equal to $224 \cdot \exp(2)$, essentially a value of infinity in this practical situation and thus implying that the fitted error distributions are close to the normal distribution, as discussed in Section 2.2.3. On the other hand, models where $lcd4$ was used in combination with other covariates gave much lower optimal tuning parameters λ, implying also non-normal error densities. This is seen on the right-hand side of Figure 3. The phenomenon could indicate presence of a risk-group mixture in the data or absence of another important predictor. Indeed, a factor that could play an important role is antiretroviral therapy, which might have been used by some women in our sample. However, this factor requires modelling time–dependent covariates, which cannot be done with our model.

In conclusion, the time to AIDS onset in this study population is notably shorter in women with oral lesions. Further, this marker improves the prediction of that time based on any of the classical indicators (CD4 count and viral load). When interpreting these findings, one must bear in mind that only a limited number of WIHS women opted to participate in the Oral Substudy, the source of the dental data. Thus they may differ in unknown ways from the overall set. Nonetheless, our findings are consistent with those of others who have evaluated oral lesions as predictors of AIDS onset and they illustrate use of our method in the area of AIDS research. Our method restricts us to analysis of baseline covariates. Although this is a very widely applicable special case, extension of the method to accommodate time-dependent covariates would allow more complex relationships between outcomes and covariates.

The model (7) of Table 1 can be fitted in R using the library `smoothSurv` in the following way. We assume that the dataset is stored in a `data.frame` called `wihs` with columns `t.left` and `t.right` giving the lower and upper limit of the observed interval with `t.right` equal to `NA` for right-censored observations and `lesion, lvload, lcd4` giving the covariate values.

```
> library(smoothSurv)
> fit7 <- smoothSurvReg(Surv(t.left, t.right, type=‘‘interval2’’) ~ lesion+
```

```
            lvload+lcd4, knots=seq(-6,6,0.3), sdspline=0.2, difforder=3, data=wihs)
> fit7

Estimated Regression Coefficients:

              Value Std.Error Std.Error2      Z     Z2         p        p2
(Intercept)  2.8257    0.6148    0.59992  4.596  4.710 4.311e-06 2.475e-06
lesion      -0.6033    0.2272    0.21899 -2.655 -2.755 7.931e-03 5.872e-03
lvload      -0.3025    0.1074    0.10290 -2.816 -2.939 4.867e-03 3.291e-03
lcd4         0.3905    0.0467    0.04618  8.363  8.458 6.096e-17 2.729e-17
Log(scale)   0.3565    0.1036    0.09774  3.442  3.647 5.783e-04 2.653e-04

Scale = 1.428

                    Lambda: 0.000911882
               Log(Lambda): -7
                        df: 10.02823
AIC (higher is better):  -250.0086
```

Most of the labels in the output are self-explanatory. Columns `Std.Error, Z, p` refer to the pseudo-variance estimate (8) while columns `Std.Error2, Z2, p2` to the asymptotic variance estimate (9). Information concerning the fitted error distribution is stored in the resulting object `fit7` and can be extracted if necessary.

There exist methods to plot the fitted error distribution or compute predictive functions, e.g., predictive survivor and hazard curves and survival densities for a new subject with $lesion = 0$ and $lesion = 1$ and median values of $lvload$ (3.875) and $lcd4$ (8.735) based on the above model are drawn as follows.

```
> covar7 <- matrix(c(0, 1, rep(3.875, 2), rep(8.735, 2)), ncol=3)
> survfit(fit7, cov=covar7, plot=TRUE)
> hazard(fit7, cov=covar7, plot=TRUE)
> fdensity(fit7, cov=covar7, plot=TRUE)
```

6 Discussion

We have suggested and implemented as an R library a method useful for fitting the linear regression model for censored observations while avoiding overly restrictive parametric assumptions on the error distribution. Most classically, the logarithmic transformation of the response leads to the well known AFT model. However, other transformations of the response leading to its potential range covering the whole real line are also possible. The density of the error distribution is specified in a semi-parametric way as a mixture of basis Gaussian densities (Gaussian densities with given means – knots – and given common standard deviation). Mixture coefficients are then estimated using the penalized maximum-likelihood method. Such model specifications allow flexibility with respect to the resulting error distribution yet retain tractability such that data carrying censoring of several types, especially interval censoring, can be handled naturally.

The penalized Gaussian mixture method also has been used successfully by Ghidey et al (2005) in the context of the linear mixed model. They exploited a mixture of BG-densities to approximate a density of the distribution of the random intercept and slope in the linear mixed model while assuming standard normal distribution for the random error. They did not assume censored observations; however, they showed an additional potential of this method by using a tensor product of two univariate mixtures of BG-densities to approximate a bivariate distribution. Generally, it would be interesting to join their and our models to form an AFT model with random effects in which the random effects distribution would be approximated by one mixture of BG-densities and the error distribution by another mixture of BG-densities. A question that remains and would need additional research is the extent of the computational difficulties that could be encountered with such a complex model.

In some specific situations, it may be desirable to have a finite support for the density of the error distribution and keep the finite support also in the estimated model. The mixture of BG-densities as presented in this paper is then inappropriate, but one could use a mixture of B-splines without additional complications. However, we think that situations that require a finite support for the error distribution in survival models are rather rare.

In the literature, Kooperberg and Stone (1992) and Eilers and Marx (1996) considered spline estimation of the logarithm of a density based on a set of i.i.d. observations. Kooperberg and Stone (1992) also allowed for censoring. Although their approach could be extended to the regression context, it would require that the logarithm of the error density ($\log f(e)$) be expressed as a spline ($s(e)$). When computing the likelihood contributions for censored observations one would have to evaluate an integral of the form $\int e^{s(e)}\,de$, which would generally require numerical methods. This complication is avoided by our technique since all integrals needed to evaluate the likelihood are expressed as linear combinations of values of cumulative normal distribution functions, and quantities related to the normal distribution can be computed using fast, precise numerical methods.

Acknowledgements

This work was primarily supported by the Research Grant OE/03/29, Katholieke Universiteit Leuven. The first two authors acknowledge support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs. Funding for Dr. Hilton's research was provided by the U.S.A. National Institutes of Health (DHHS NIAID R01-AI55085 and NIDCR P01-DE07946).

Data used in Section 5 of this manuscript were collected by the Women's Interagency HIV Study (WIHS) Collaborative Study Group and its Oral Substudy with centers (Principal Investigators) at New York City/Bronx Consortium (K. Anastos, J. A. Phelan); Brooklyn, NY (H. Minkoff); Washington DC Metropolitan Consortium (M. Young); The Connie Wofsy Study Consortium of Northern California (R. Greenblatt, D. Greenspan, J. S. Greenspan); Los Angeles County/Southern California Consortium (A. Levine, R. Mulligan, M. Navazesh); Chicago Consortium (M. Cohen, M. Alves); Data Coordinating Center (A. Muñoz). The WIHS is funded by the National Institute of Allergy and Infectious Diseases, with supplemental funding from the National Cancer Institute, the National Institute of Child Health & Human Development, the National Institute on Drug Abuse, the National Institute of Dental and Craniofacial Research, the Agency for Health Care Policy and Research, the National Center for Research Resources, and the Centers for Disease Control and

Prevention. U01-AI-35004, U01-AI-31834, U01-AI-34994, U01-AI-34989, U01-HD-32632 (NICHD), U01-AI-34993, U01-AI-42590, M01-RR00079, and M01-RR00083. The WIHS Oral Substudy is funded by the National Institute of Dental and Craniofacial Research.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* **AC–19,** 716–723.

Buckley, J., and James, I. (1979). Linear regression with censored data. *Biometrika,* **66,** 429–436.

Barkan, S. E., Melnick, S. L., Preston–Martin, S., Weber, K., Kalish, L. A., Miotti, P., Young, M., Greenblatt, R., Sacks, H., and Feldman, J. (1998). The Women's Interagency HIV Study. *Epidemiology,* **9,** 117–125.

Betensky, R. A., Rabinowitz, D., and Tsiatis, A. A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika,* **88,** 703–711.

Chen, Y. Q., and Jewell, N. P. (2001). On a general class of semiparametric hazards regression models. *Biometrika,* **88,** 687–702.

Christensen, R., and Johnson, W. O. (1988). Modeling accelerated failure time with a Dirichlet process. *Biometrika,* **75,** 693–704.

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

de Boor, C. (1978). *A Practical Guide to Splines.* Springer, Berlin.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines.* Clarendon, Oxford.

Eilers, P. H. C., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science,* **11,** 89–121.

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.

Ghidey, W., Lesaffre, E., and Eilers, P. (2005). P-spline smoothing of the random effects distribution in a linear mixed model. *To appear in Biometrics.*

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association,* **87,** 942–951.

Han, S. P. (1977). A globally convergent method for nonlinear programming. *Journal of Optimization Theory and Applications,* **22,** 297–309.

Hanson, T., and Johnson, W. O. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association,* **97,** 1020–1033.

Hanson, T., and Johnson, W. O. (2004). A Bayesian semiparametric AFT model for interval-censored data. *Journal of Computational and Graphical Statistics,* **13,** 341–361.

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

Biometrika, **90,** 341–353.

Johnson, W. O., and Christensen, R. (1989). Nonparametric Bayesian analysis of the accelerated failure time model. *Statistics and Probability Letters,* **8,** 179–184.

Joly, P., Commenges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: application to age–specific incidence of dementia. *Biometrics,* **54,** 185–194.

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, 2nd Ed.* Chichester: John Wiley & Sons.

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

Kottas, A., and Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association,* **96,** 1458–1468.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics,* **22,** 79–86.

Kuo, L., and Mallick, B. K. (1997). Bayesian semiparametric inference for the accelerated failure time model. *Canadian Journal of Statistics,* **25,** 457–472.

Leroy R., Bogaerts K., Lesaffre E., and Declerck D. (2003). Impact of caries experience in the deciduous molars on the emergence of the successors. *European Journal of Oral Sciences,* **111**, 106-110.

McKeague, I. W., Subramanian, S., and Sun, Y. (2001). Median regression and the missing information principle. *Journal of Nonparametric Statistics,* **13,** 709–727.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problem (with discussion). *Statistical Science,* **1,** 505–527.

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

R Development Core Team (2004). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL `http://www.R-project.org`.

Rabinowitz, D., Tsiatis, A., and Aragon, J. (1995). Regression with interval-censored data. *Biometrika,* **82,** 501–513.

Shyur, H.-J., Elsayed, E. A., and Luxhøj, J. T. (1999). A general model for accelerated life testing with time-dependent covariates. *Naval Research Logistics: an International Journal,* **46,** 303–321.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B,* **37,** 290–295.

Unser, M., Aldroubi, A., and Eden, M. (1992). On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Transactions on Information Theory,* **38,** 864–872.

Vanobbergen, J., Martens, L., Lesaffre, E., and Declerck, D. (2000). A longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry,* **2,** 87–96.

VERWEIJ, P. J. M., and VAN HOUWELINGEN, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine,* **13,** 2427–2436.

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross–validated smoothing spline. *Journal of the Royal Statistical Society, Series B,* **45,** 133–150.

WALKER, S. G., and MALLICK, B. K. (1999). Semiparametric accelerated life time model. *Biometrics,* **55,** 477–483.

YANG, S. (1999). Censored median regression using weighted empirical survival and hazard functions. *Journal of the American Statistical Association,* **94,** 137–145.

YING, Z., JUNG, S., and WEI, L. J. (1995). Survival analysis with median regression models. *Journal of the American Statistical Association,* **90,** 178–184.

Model	*AIC*	*df*	$\log(\lambda/n)$	*lesion*	*logvload*	*logcd4*
(1) *lesion*	−262.39	3.2	2	−0.87 (0.37; 0.018)		
(2) *lvload*	−256.16	3.4	2		−0.76 (0.19; <0.001)	
(3) *lcd4*	−256.94	3.4	2			0.44 (0.11; <0.001)
(4) *lesion* + *lvload*	−255.63	4.4	2	−0.62 (0.36; 0.080)	−0.70 (0.19; <0.001)	
(5) *lesion* + *lcd4*	−253.19	8.9	−7	−0.78 (0.26; 0.003)		0.39 (0.07; <0.001)
(6) *lvload* + *lcd4*	−253.45	8.4	−6		−0.39 (0.14; 0.004)	0.38 (0.06; <0.001)
(7) *lesion* + *lvload*+ +*lcd4*	−250.01	10.0	−7	−0.60 (0.23; 0.008)	−0.30 (0.11; 0.005)	0.39 (0.05; <0.001)

Table 1: WIHS Data. Akaike's information criterion, degrees of freedom, the optimal $\log(\lambda/n)$, estimates of the regression parameters (standard error; p-value) for the fitted models.

Figure 1: Simulation Study. Average estimate of the regression parameters $\beta_1 = -0.8$ (left column) and $\beta_2 = 0.4$ (right column) for simulation patterns involving extreme value and normal mixture error distributions and interval censoring. Dotted line: true parameter value, solid line: estimate based on our procedure, dashed line: estimate based on the AFT model with correctly chosen error distribution, dotted-dashed line: estimate based on the log-normal AFT model.

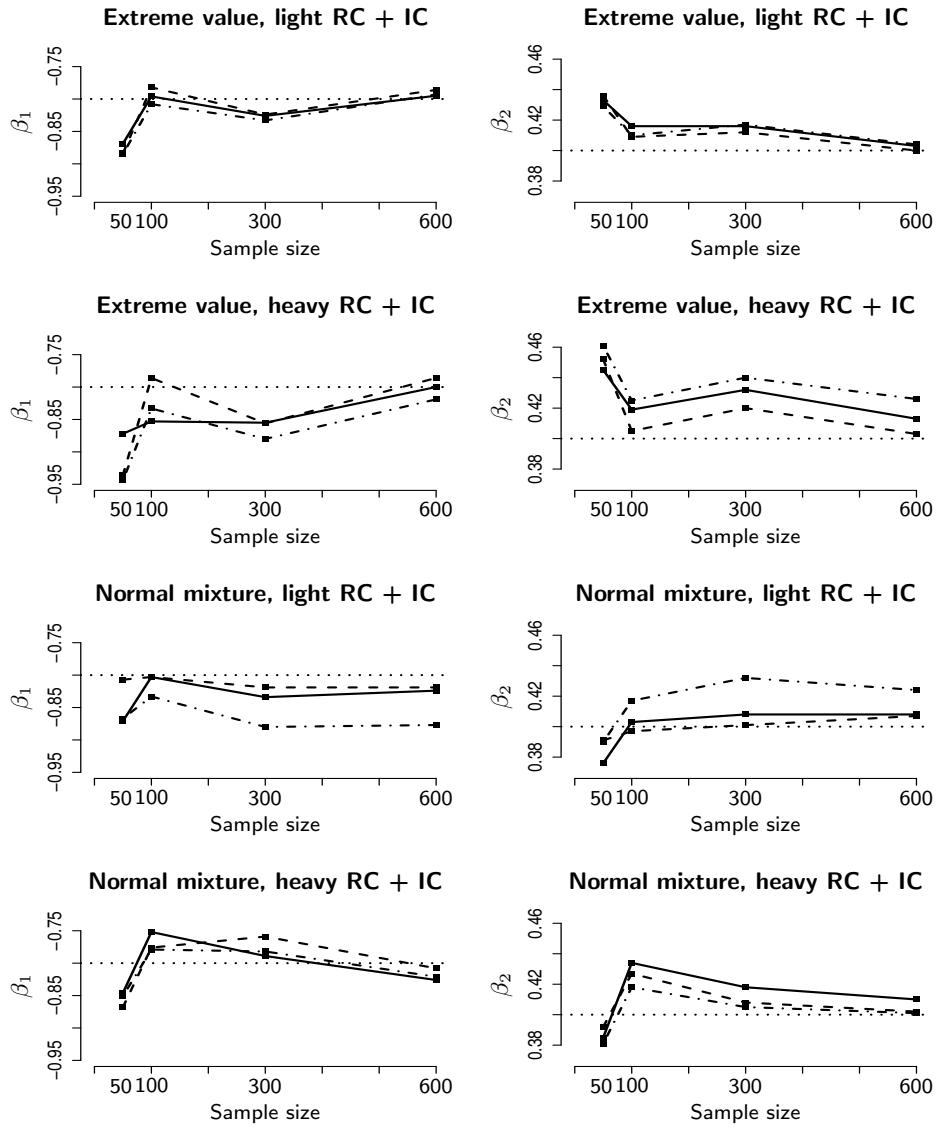


Figure 2: Simulation Study. The average of the fitted error density (solid line), 95% *pointwise* confidence band (dotted line) and the true error density (dashed line) for selected simulation patterns. Extreme value as the true distribution in the upper part, normal mixture as the error distribution in the bottom part.

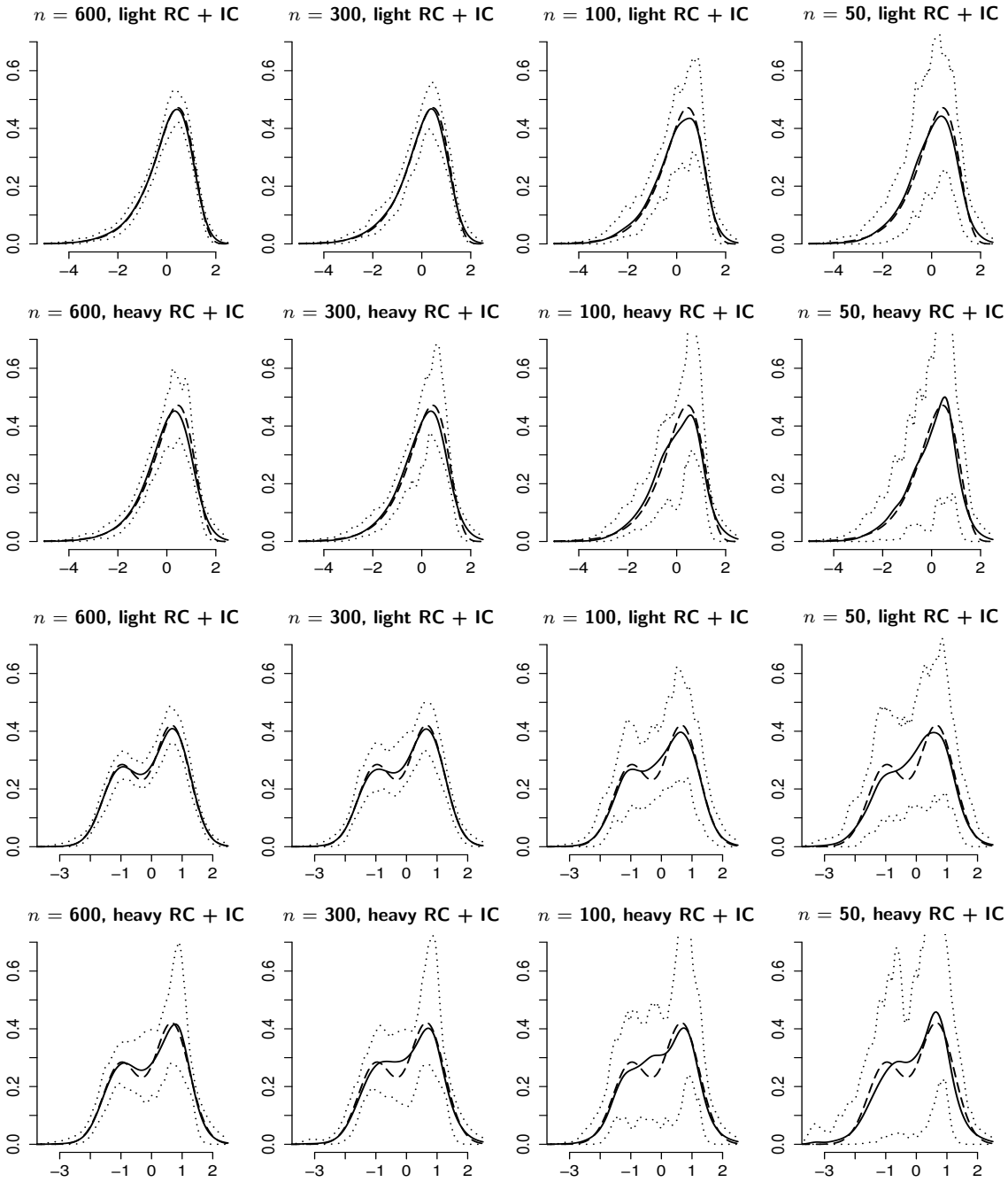


Figure 3: WIHS Data. Predicted survivor curves, hazard curves and densities for women with $lesion = 1$ (dotted-dashed line) vs. women with $lesion = 0$ (solid line) based on models $lesion$ (left part) and $lesion + lvload + lcd4$ (right part). Predictive curves for the latter model control for a median value of $lvload = 3.875$ and a median value of $lcd4 = 8.735$. Predictive survivor curves for model $lesion$ are further compared to the nonparametric estimate of Turnbull (1976) in each group.