A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia.

Pierre Joly, Daniel Commenges, Luc Letenneur

A penalized likelihood approach for arbitrarily censored and truncated data : application to age-specific incidence of dementia

Pierre JOLY, Daniel COMMENGES, Luc LETENNEUR

INSERM U. 330

Université de Bordeaux II

146, rue Léo Saignat

33076 Bordeaux Cedex, France

Tel : (33) 5 57 57 11 82

Fax : (33) 5 56 99 13 60

E-mail: Daniel.Commenges@bordeaux.inserm.fr

Corresponding author : Daniel COMMENGES

October 25, 2007

Summary

Cox model is the model of choice when analyzing survival data presenting only right censoring and left truncation. There is a need for methods which can accommodate more complex observation schemes, involving general censoring and truncation. In addition it is important in many epidemiological applications to have a smooth estimate of the hazard function. We show that the penalized likelihood approach gives a solution to these problems. The solution of the maximum of the penalized likelihood is approximated on a basis of splines. The smoothing parameter is estimated using approximate cross-validation; confidence bands can be given. A simulation study shows that this approach gives better results than the smoothed Nelson-Aalen estimator. We apply this method to the analysis of data from a large cohort study on cerebral ageing. The age-specific incidence of dementia is estimated and risk factors of dementia studied.

Key words: Hazard function, penalized likelihood, spline, truncation, interval-censoring, proportional hazards.

1 Introduction

The analysis of survival data takes a prominent role in epidemiology. The model of choice for many applications is Cox model (1972) which allows estimation of the relative risks without making parametric hypotheses on the baseline hazard function. With this model, both right censoring and left truncation can easily be handled. However there are two limitations: more complex observation schemes cannot be treated; this approach does not yield a smooth estimate of the hazard function.

There are more and more instances where complex observation schemes have to be treated. In registers, the observations are right truncated because cases are registered at a certain date if and only if they have developed the disease before that date (Lagakos, Barraj and De Gruttola, 1988). In cohort studies, it often happens that there are successive visits and the time the event of interest occurred is only known to lie between two visits: this produces interval censored data. Prevalent cases of a disease can be viewed as left censored observations. For all these situations we need more general tools than the Cox model.

Peto (1973) proposed the non-parametric maximum likelihood estimator (NPMLE) of the survival function for interval censored data and Turnbull (1976) extended it to arbitrarily censored and truncated data. Finkelstein (1986), Finkelstein, Moore and Schoenfeld (1993), Tu, Meng and Pagano (1993) developed discrete proportional hazards models for such cases using the full likelihood maximized by the EM algorithm and Alioum and Commenges (1996) proposed an extension of the proportional hazards model in continuous time.

These methods are useful for estimating the survival function but not the hazard function. An estimated hazard function is often an important result in epidemiology. In particular, if age has been chosen as the time scale, the hazard function is the age-specific incidence of the disease. Choosing age as the time scale generally creates left truncated data and estimation of the age-specific incidence is most often done using the person-years method (Breslow and Day, 1987). In this method, the smoothing comes from the assumption of constant hazard rate during 5-year periods. However the person-years method (and its extension to regression model via Poisson regression) is limited because the estimate of the hazard is not really smooth but presents jumps. An interesting solution proposed by Ramlau-hansen (1983) and Andersen et al (1993) is to smooth the Nelson-Aalen estimator by kernel methods. However neither method can accomodate more than right censoring and left truncation.

A complete solution to the problem seems to be easy with a parametric approach but this approach is not satisfactory because an age-specific incidence may have any shape and we do not know it before analyzing the data. The approach of Kooperberg, Stone and Truong (1995) can be viewed as a very flexible parametric approach. The family of parametric functions is defined as those which can be constructed using a basis of splines.

Another approach is based on penalized likelihood. Here we explicitly introduce an a priori knowledge of smoothness of the hazard function, by penalizing the likelihood by a norm of the second derivative of the hazard function. The estimator is defined non-parametrically as the function which maximizes the penalized likelihood. The solution is then approximated on a

basis of splines. Such an approach has already been proposed by Senthilselvan (1987) and O'Sullivan (1988) in the case of right censored data. Senthilselvan (1992) treated the case where left truncation is also present.

In this paper we show how this approach can be applied in complex cases of truncation and censoring, including for instance interval censoring and right truncation. We also show how it can be applied to regression models including the proportional hazards model. We propose to represent the hazard function (rather than the log-hazard) on a basis of splines; this avoids numerical integrations to compute the cumulative hazard function and hence the likelihood. These proposals are presented in section 2 which also contains a method based on approximate cross-validation for automatically choosing the smoothing parameter and a method for obtaining confidence bands. In section 3, we present simulations comparing the penalized likelihood estimator to the smoothed Nelson-Aalen estimator. In the fourth section, the age-specific incidence of dementia is estimated using data from a large cohort study. We also study risk factors of dementia and compare the results obtained with the penalized likelihood approach to those obtained with Cox model.

2 Problem and method

2.1 Incomplete data: censoring and truncation

Let $X_1, X_2, ..., X_n$ be a sample of n positive random variables with common survival function. In the sequel we denote by f the probability density function, S the survival function, λ the hazard function and Λ the cumulative

hazard function of X. Using the notations and definitions introduced by Turnbull (1976), we say that the observation X_i is interval-censored if the only information about it is that it lies in a known interval $A_i = [L_i, R_i]$; $L_i \leq X_i \leq R_i$ with $A_i \subset I\!R^+$. Both left- and right-censoring are just particular cases of interval-censoring. Let $L_1, ..., L_n$ and $R_1, ..., R_n$ two samples of censoring times. X_i is right-censored if $X_i > L_i$. In this case we do not observe X_i but only L_i, and we have $A_i = [L_i, +\infty)$. If X_i is left-censored we just know that $X_i < R_i$ and we have $A_i = [0, R_i]$. If $L_i = R_i$ $(= X_i)$ then X_i is uncensored. A convenient assumption which is sufficient for the censoring to be non-informative is that the L_i and the R_i are fixed or independent of X_i.

We say that X_i is truncated if it is observed conditionally on the event $X_i \in B_i$; we shall restrict to the case where B_i is an interval. Let $\mathcal{L}_1, ..., \mathcal{L}_n$ and $\mathcal{R}_1, ..., \mathcal{R}_n$ two samples of truncating times. X_i is left-truncated if $B_i = [\mathcal{L}_i, +\infty)$, right-truncated if $B_i = [0, \mathcal{R}_i]$ and interval-truncated if it is both left- and right-truncated. If $B_i = [0, +\infty)$ then X_i is not truncated. If X_i is both interval-censored and truncated we have $A_i \subseteq B_i$. Similarly to the censoring, we shall assume that $\mathcal{L}_i$ and $\mathcal{R}_i$ are fixed or independent of X_i.

While dealing with censored and truncated data it is more convenient to work with the hazard function or the cumulative hazard function because the log-likelihood can be expressed simply in terms of these functions.
In the general case the log-likelihood is:

$$l = \sum_{i=1}^{n} \log \left(\frac{\Delta(L_i, R_i)}{\Delta(\mathcal{L}_i, \mathcal{R}_i)} \right), \qquad \mathcal{L}_i \leq L_i \leq R_i \leq \mathcal{R}_i \tag{1}$$

with

$$\Delta(L_i, R_i) = \begin{cases} e^{-\Lambda(L_i)} - e^{-\Lambda(R_i)} & \text{if } L_i < R_i, \\ \lambda(L_i)e^{-\Lambda(L_i)} & \text{if } L_i = R_i \end{cases}$$

Thus the log-likelihood can be expressed as a function of λ and we shall note $l(\lambda)$.

2.2 Penalized likelihood

In many situations we expect the hazard function to be smooth. A possible means for introducing such an a priori knowledge is to penalize the likelihood by a term which takes large values for rough functions. One can find a good overview of this subject in Silverman (1985).

The smooth aspect of a function is related to the value of its second derivative, which leads to take for roughness penalty $\int \lambda''^2(u)du$. We assume that $\lambda(.)$ belongs to the class of continuous functions, twice differentiable and whose second derivative is square integrable. We define the penalized log-likelihood as:

$$pl(\lambda) \quad = \quad l(\lambda) - \kappa \int \lambda''^2(u)du \tag{2}$$

Where l is the usual log-likelihood and κ is the smoothing parameter which must be positive; κ controls the balance between the fit to the data and the smoothness of the function. Maximization of (2) over the desired class of function defines the maximum penalized likelihood estimator (MPLE) $\hat{\Lambda}$ and hence $\hat{\lambda}$ and other possibly interesting functions.

2.3 Approximation via splines

The solution of (2) can be approximated on a basis of splines. We briefly present the splines used here and give some computational aspects of this approach. For more details see Ramsay (1988) (monotone splines), Wegman and Wright (1983) (splines in statistics) and de Boor (1978) (B-splines).

Splines are piecewise polynomial functions which are combined linearly to approximate a function on an interval. We use M-splines, which are a variant of B-splines, and I-splines.

A M-spline of order k is defined as:

$$M_j(x|k) = \begin{cases} \frac{k\left[(x-t_j)M_j(x|k-1)+(t_{j+k}-x)M_{j+1}(x|k-1)\right]}{(k-1)(t_{j+k}-t_j)}, & t_j \leq x < t_{j+k}, \\ 0 & \text{elsewhere,} \end{cases}$$

with

$$M_j(x|1) = \begin{cases} \frac{1}{(t_{j+1}-t_j)} & \text{if } t_j \leq x < t_{j+1}, \\ 0 & \text{elsewhere.} \end{cases}$$

where $t_1, ..., t_m$ is a sequence of increasing knots. Each $M_j(x|k)$ is zero outside of the interval $[t_j, t_{j+k}]$, hence is non-zero over k intervals and over each interval there are k non-zero M-splines. For our approximation we use splines of order 4.

To each M-spline we associate a I-spline:

$$I_j(x|k) = \int_0^x M_j(u|k)du.$$

Each M_j is piecewise polynomial of degree $k-1$ and each associated I_j is

piecewise polynomial of degree k defined as (if $t_j \leq x < t_{j+1}$):

$$I_h(x|k) = \begin{cases} 0 & \text{if } h > j, \\ \sum_{l=h}^{j} (t_{l+k+1} - t_l) \dfrac{M_l(x|k+1)}{k+1} & \text{if } j-k+1 \leq h \leq j, \\ 1, & \text{if } h < j-k+1. \end{cases}$$

These splines are convenient to manipulate; among other things a linear combination of splines is easy to differentiate or integrate. Note that M-splines are nonnegative and I-splines are monotonically increasing; it results that the monotonicity constraint for a function represented on a basis of I-splines can be fulfilled by constraining the coefficients to be positive. Thus the estimator $\hat{\Lambda}(.)$ can be approximated by a linear combination of m I-splines $\tilde{\Lambda}(.) = \sum_{j=1}^{m} g(\theta_j) I_j(.)$, where $g(\theta_j) \geq 0 \ \ \forall j$ (for example $g(\theta_j) = e^{\theta_j}$ or $g(\theta_j) = {\theta_j}^2$); in practice we use $g(\theta_j) = {\theta_j}^2$ to avoid convergence problems when $g(\theta_j)$ should be zero. By derivating we obtain: $\tilde{\lambda}(.) = \sum_{j=1}^{m} g(\theta_j) M_j(.)$. So with the same vector of coefficients $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)^T$ we get the cumulative hazard function with I-splines and the hazard function with M-splines. In fact the set of functions generated by the basis of splines with positive coefficients is included in the set of positive functions generated by the basis of splines. However our numerical experience shows that this set is rich enough to provide a good approximation of the hazard function.

A spline function is completely defined by a sequence of knots and the coefficients of the splines. We may put knots at every data points but the computational price would be heavy for large n. It is convenient to locate the knots at every p data points as described in O'Sullivan (1988). In any case there must be a knot before or at the first data point and after or

at the last data point. Note that the more knots we take, the better the approximation will be. The approximation $\tilde{\lambda}$ of $\hat{\lambda}$ is the function belonging to the space generated by the basis of splines which maximizes $pl(\lambda)$. The general penalized log-likelihood for interval censored and truncated data is then:

$$\sum_{i=1}^{n} \log \left(\frac{e^{-\sum_j g(\theta_j) I_j(L_i)} - e^{-\sum_j g(\theta_j) I_j(R_i)}}{e^{-\sum_j g(\theta_j) I_j(\mathcal{L}_i)} - e^{-\sum_j g(\theta_j) I_j(\mathcal{R}_i)}} \right) - \kappa \int \left(\sum_{j=1}^{m} g(\theta_j) M_j''(u) \right)^2 du,$$

$$\mathcal{L}_i \leq L_i < R_i \leq \mathcal{R}_i$$

For uncensored observation this must be modified in an obvious way as in (1). The estimated vector $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ for a fixed κ is obtained by maximizing the log-likelihood using a combination of a Newton-Raphson and a steepest descent algorithm. The Newton step involves a line search with a step reduction if the new point is not better. The steepest descent step involves a full line search and is attempted only if the Newton step has failed, due generally to a difficulty to inverse the Hessian of the log-likelihood. Few iterations are needed if the initial value is judiciously chosen because the Newton-Raphson iteration is used. In other cases the steepest descent iteration is often used because the Hessian may be singular and the convergence is slower. We stop the iterations when the difference between two consecutive log-likelihoods is small, the coefficients are stable and the gradient is small enough. We have not yet encountered local maxima that are not global maxima.

When we get the vector of coefficients, with the knots sequence we have all the functions of interest, as in a parametric method.

2.4 Selection of the smoothing parameter

A rough estimate of the smoothing parameter κ is often enough, so we may select it empirically. But an automatic choice of this parameter seems almost always necessary because it is less subjective. The method of cross-validation gives a solution to this problem.

The standard cross-validation score which must be maximized to obtain κ is:

$$V(\kappa) = \sum_{i=1}^{n} l_i(\hat{\boldsymbol{\theta}}_{-i})$$

where $\hat{\boldsymbol{\theta}}_{-i} = \hat{\boldsymbol{\theta}}_{-i}(\kappa)$ is the maximum penalized likelihood estimator of $\boldsymbol{\theta}$ for the sample in which the i^{th} individual is removed and l_i is the log-likelihood contribution of this individual. This score is equivalent, in the case of the log-density estimation, to the Kullback-Liebler cross-validation score (see Silverman, 1985 and O'Sullivan, 1988).

The maximization of this score however is computationally expensive, because it requires a maximization for each individual and for each different value of κ. So we use an approximation. O'Sullivan (1988) proposed to use a one step Newton-Raphson expansion to approximate the leave-out-one estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_i$ and derived the formulas for the estimator of a density. We extend his method to our case.

The one-step Newton-Raphson approximation is: $\overline{\boldsymbol{\theta}}_{-i} = \hat{\boldsymbol{\theta}} - \left[\hat{\boldsymbol{H}} - 2\kappa\boldsymbol{\Omega}\right]^{-1} \hat{\boldsymbol{d}}_{-i}$ when $\hat{\boldsymbol{H}}$ is the converged Hessian $\dfrac{\partial^2 l}{\partial \boldsymbol{\theta}^2}(\hat{\boldsymbol{\theta}})$, $\boldsymbol{\Omega}$ is the penalized part of the converged Hessian and $\hat{\boldsymbol{d}}_{-i} = -\hat{\boldsymbol{d}}_i = -\dfrac{\partial l_i}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})$. If $g(\theta_j) = \theta_j$ we have $\boldsymbol{\Omega} = \int M_l''(u) M_m''(u) du$.

So we must maximize the approximate score:

$$\overline{V}(\kappa) = \sum_{i=1}^{n} l_i(\overline{\boldsymbol{\theta}}_{-i}) \tag{3}$$

Then we denote $\overline{\Lambda}_{-i}(x)$ the function obtained with the vector $\overline{\boldsymbol{\theta}}_{-i}$.

In the general case we obtain:

$$\overline{V}(\kappa) = \sum_{i=1}^{n} \log\left(\frac{e^{-\overline{\Lambda}_{-i}(L_i)} - e^{-\overline{\Lambda}_{-i}(R_i)}}{e^{-\overline{\Lambda}_{-i}(\mathcal{L}_i)} - e^{-\overline{\Lambda}_{-i}(\mathcal{R}_i)}}\right), \qquad \mathcal{L}_i \leq L_i < R_i \leq \mathcal{R}_i$$

So we have after some simplifications and first order approximations

$$\overline{V}(\kappa) \simeq l(\hat{\boldsymbol{\theta}}) + trace\left(\left[\hat{\boldsymbol{H}} - 2\kappa\boldsymbol{\Omega}\right]^{-1}\boldsymbol{H}^*\right) \tag{4}$$

$$\simeq l(\hat{\boldsymbol{\theta}}) - trace\left(\left[\hat{\boldsymbol{H}} - 2\kappa\boldsymbol{\Omega}\right]^{-1}\hat{\boldsymbol{H}}\right) \tag{5}$$

with $\boldsymbol{H}^* = \sum_{i=1}^{n} \hat{\boldsymbol{d}}_{-i}\hat{\boldsymbol{d}}_{-i}^T = \sum_{i=1}^{n} \hat{\boldsymbol{d}}_i\hat{\boldsymbol{d}}_i^T$.

It should be noted that the above expression essentially is an AIC criterion (Akaike 1974) if we interpret $trace\left(\left[\hat{\boldsymbol{H}} - 2\kappa\boldsymbol{\Omega}\right]^{-1}\hat{\boldsymbol{H}}\right)$ as the model degrees of freedom (mdf). Indeed mdf decreases in κ from m (if $\kappa = 0$) to 2 (if $\kappa \to +\infty$) which is the number of degrees of freedom of a straight line. The boundary at $+\infty$ is 2 instead of 0 because $\boldsymbol{\Omega}$ has two zero eigenvalues. To maximize (5) we use a golden section search.

2.5 Approximate Bayesian confidence bands

A Bayesian technique for generating confidence bands for penalized likelihood estimators was proposed by Wahba (1983), Silverman (1985) and O'Sullivan (1988). Then, $\boldsymbol{\theta}$ is regarded as a random variable. Up to a constant, the penalized log-likelihood pl is a posterior log-likelihood for $\boldsymbol{\theta}$ and the penalty

term the prior log-likelihood. After a Gaussian approximation, the covariance of $\boldsymbol{\theta}$ is $-\left[\frac{1}{2}\hat{\boldsymbol{H}} - \kappa\boldsymbol{\Omega}\right]^{-1}$. Then, an approximate 95% Bayesian confidence interval for $\tilde{\lambda}$ at point x is:

$$\tilde{\lambda}(x) \pm 1,96\tilde{\sigma}(x),$$

where the approximate standard error is:

$$\tilde{\sigma}(x) = \sqrt{\boldsymbol{M}(x)^T \left[-\frac{1}{2}\hat{\boldsymbol{H}} + \kappa\boldsymbol{\Omega}\right]^{-1} \boldsymbol{M}(x)},$$

where $\boldsymbol{M}(x) = (M_1(x), ..., M_m(x))^T$. To obtain approximate Bayesian confidence bands for $\tilde{\Lambda}$, we can use the same formula with the I-splines basis. Hence we can easily deduce an approximate Bayesian confidence bands for the survival function. Another possibility would be to determine confidence intervals using a boostrap technique.

2.6 Generalization to regression models

The penalized likelihood can be applied for estimating the hazard function in a general regression model defined by: $\lambda_i(.) = \varphi[\lambda_0(.), \boldsymbol{z}_i\boldsymbol{\beta}]$, where $\lambda_0(.)$ is the baseline hazard function, $\boldsymbol{\beta}$ is a vector of regression parameters and $\boldsymbol{z}_i$ the vector of covariates for subject i. The accelerated failure time model, the additive model and the proportional hazards model are particular cases of this general form. The penalized log-likelihood used is then:

$$l[\lambda_0(.), \boldsymbol{Z}\boldsymbol{\beta}] - \kappa \int \lambda_0^{''2}(u)du \tag{6}$$

where $\boldsymbol{Z}$ is the matrix with rows equal to $\boldsymbol{z}_i$, $i = 1, ..., n$. In our method, the selection of the smoothing parameter is a serious difficulty. We use a two

steps search: firstly we maximize (2), ignoring the explanatory variables, to obtain an estimator of κ and a good initial guess of $\tilde{\lambda}_0(.)$ and subsequently we maximize (6) with κ fixed to obtain $\hat{\boldsymbol{\beta}}$ and $\tilde{\lambda}_0(.)$. The regression parameters and the baseline functions are estimated simultaneously by the robust Newton-Raphson method described in section 2.3.

3 Simulation study

In order to see how our method performs, we compared it to the smoothed Nelson-Aalen estimator. The data were generated from a Weibull distribution or a mixture of gamma distributions. Samples of sizes 50, 100 and 500 were generated. Each simulation involved 100 replications. For each sample the distances between the true hazard function and both MPLE and smoothed Nelson-Aalen estimator were calculated. The distance used is the integrated squared error (ISE):

$$\int_J \left(\lambda(u) - \hat{\lambda}(u)\right)^2 du$$

where λ and $\hat{\lambda}$ are the true and estimated hazard function respectively and $J = [x_{(1)} + b, x_{(N)} - b]$, the interval on which the smoothed Nelson-Aalen estimator is defined with $x_{(1)}$ and $x_{(N)}$ the smaller and higher generated failure times and b the bandwidth for the kernel estimator.

We generated a random sample $X_1, ..., X_n$ of i.i.d failure times and $C_1, ..., C_n$ of censoring times, the C_i were independent of the X_i. So the observed samples were $(Y_1, \delta_1), ..., (Y_n, \delta_n)$ where $Y_i = \min(X_i, C_i)$ and $\delta_i = \mathrm{I}_{[X_i \leq C_i]}$. The density of X was a simple Weibull ($f(t; 2, 0.06)$) or a mixture of Gamma ($0.4\Gamma(t; 14, 1.8) + 0.6\Gamma(t; 50, 2)$), with the probability density functions $f(t; \gamma, p) =$

$p\gamma^p t^{p-1} e^{-(\gamma t)^p}$ and $\Gamma(t;\alpha,\gamma) = \frac{\alpha^\gamma t^{\gamma-1} e^{-\alpha t}}{\Gamma(\gamma)}$. The probability density function of C_i was a simple Weibull or a simple Gamma. The percentage of censoring was around 10% and 50%.

The Epanechnikov kernel was used to smooth the Nelson-Aalen estimator. The optimal bandwidth was automatically chosen for each sample by a method of cross-validation described in Andersen et al (1993). For the MPLE the smoothing parameter κ was chosen by the approximate cross-validation method. The number of knots was 12. Figure 1 displays the smoothed Nelson-Aalen estimate and the MPLE for one simulated example from a mixture of Gamma.

The results of the simulations are summarized in Table 1 for the Weibull distribution and in Table 2 for the mixture of Gamma. We computed 5%-trimmed means of the distances to eliminate the rare cases when numerical problems occured in the automatic choice of κ or b. In these tables, we give the average of the length of the interval over which the distance was calculated, because it differs from one sample to the other. Note that the mean length of J varies much between the different simulations so we cannot compare the rows of the tables. The increase of the distance with n for the smoothed Nelson-Aalen estimator is due to the increase of the length of J. We notice that though the distance is calculated on the interval J the MPLE is closer to the true hazard function on average for each case.

4 Application

In order to illustrate the MPLE method we present an application to modeling the risk of developing dementia. Dementia is a common disease among the elderly in the developed countries. The incidence of dementia is not well known because of the lack of studies and direct estimation demands the follow-up of many people for several years. The application is based on the Paquid research program (Letenneur et al, 1994), a prospective cohort study of mental and physical aging that evaluates social environment and health status. The target population consists of subjects aged 65 years and older living at home in 75 civil parishes in southwestern France. The baseline variables registered included socio-demographic factors, medical history and psychometric tests. Subjects were re-evaluated 1, 3 and 5 years after the initial visit. Prevalent cases were removed from the sample because of their high mortality relative to non-demented people of the same age. So this produced a left-truncation problem. We have restricted our analysis to the sample of people living in the administrative area of Gironde and for which educational level was recorded. The sample consisted of 2717 subjects and during the follow-up 128 incident cases of dementia were observed. Age distribution of the study sample is presented in Table 3. Age was used as the basic time scale in order to get age-specific incidence of dementia. The age of onset of dementia of the subjects was left-truncated by the age of the subject at inclusion in the study, and right censored if the subject had not developed dementia at the time last seen. The information available on incident cases of dementia was the date of the visit last seen without dementia and the date of the visit first seen with dementia. Thus the ages of onset of dementia for

these 128 subjects were interval-censored. We used splines of order 4 with 12 knots and the smoothing parameter was automatically selectioned. The method of approximation discussed in this paper allows the direct treatment of these data which are both interval-censored and left-truncated. The approximate cross-validation method lead to $mdf = 2.9$ (which is nearly the number of degrees of freedom of a quadratic curve). Figure 2 displays the estimated hazard function of dementia in Gironde. It increases steadily with age with no evidence of a plateau in the oldest age, but approximate bayesian confidence bands are very large at that time.

Using prevalent cases, Dartigues et al (1992) found that educational level and main occupation during life-time were risk factors of cognitive impairment. Thus the hypothesis that educational level is a risk factor of dementia is an interesting one and we can test it using follow-up data. Two explanatory variables were considered: gender and educational level. In the sample there are 1623 females and 1094 males. Educational level was classified into three categories: no schooling (118 subjects), grade school level (1660 subjects) and high school or university level (939 subjects). Two methods of inference were used for a proportional hazards model. The first one was a semi-parametric model based on penalized likelihood, as described in section 2.5. The baseline hazard function was approximated by M-splines of order 4 with 12 knots. The second method was a Cox model with delayed entry (Cnaan and Ryan, 1989), inference was based on Cox partial likelihood. This method does not take into account interval-censoring. So, we pretend to know exactly the age of onset by taking the middle of the interval of censoring. This analysis was performed with BMDP software. We analyzed three

models: one model including gender, the second education level and the third both variables (Table 4). Note that we cannot compare the log-likelihoods of the two methods because one is a penalized log-likelihood and the other a partial log-likelihood. It is remarkable that the results for the regression coefficients for the Cox model and the penalized likelihood approach are very close. Subjects with no schooling have an increased risk of dementia. Incident dementia is not significantly related to gender. However, the graph of the two hazard functions estimated separately (not shown here) seem to be different. So the question remains open.

5 Discussion

We have shown that the penalized likelihood approach yields a method for analyzing survival data arising from complex observation schemes where the conventional methods including Cox model and Nelson-Aalen estimators were not applicable. Our simulation study shows that in the case where the smooth Nelson-Aalen estimator of the hazard function can be applied, the penalized likelihood estimator is better. In addition there is no problem of edge effects in the latter while every method based on kernel smoothing will have such problems (see however in Andersen et al (1993) an attempt to remove this problem). Thus the present method and program make it possible to analyze a wide variety of epidemiologic problems; for instance we could treat a data set on pediatric AIDS which would be a combination of data from registers, that is right truncated, and from cohorts, that is right censored. This approach can also be applied to non-proportional hazards models; it

would be just as easy to treat an additive risk model or an accelerated failure time model. Time varying variables can also be treated, however this implies more computations: in this case $\lambda_0(t)$ is approximated using a basis of M-splines and $\Lambda_i(t)$ must be computed by numerical integration, except for simple form of $z_i(t)$.

REFERENCES

Akaike, H., 1974, A new look at the statistical model identification, *IEEE Trans. Automat. Control.*, **AC-19**, 716-723.

Alioum, A. and Commenges, D. (1996). A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, **52,** 512-524.

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

Breslow, N. E. and Day, N.E. (1987), Statistical methods in cancer research, Volume II, The design and analysis of cohort studies, *IARC Scientific Publications 82.* International Agency for Research on Cancer, Lyon.

Cnaan, A. and Ryan, L. (1989). Survival analysis in natural history studies of disease. *Statistics in Medicine* **8,** 1255-1268.

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

Dartigues, J.-F., Gagnon, M., Letenneur, L., Barberger-Gateau, P., Commenges, D., Evaldre, M. and Salamon, R. (1992). Principal lifetime occupation and cognitive impairment in a French elderly cohort (Paquid). *American Journal of Epidemiology* **135,** 981-988.

de Boor, C. (1978). *A practical guide to splines.* Springer, New York.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42,** 845-854.

Finkelstein, D. M., Moore, D. F. and Schoenfeld, D. A. (1993). A proportional hazards model for truncated AIDS data. *Biometrics* **49,** 731-740.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure time data*. New-York: Wiley.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995). Hazard Regression. *Journal of the American Statistical Association* **90,** 78-94.

Lagakos, S. W., Barraj, L. M. and De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika* **75,** 515-523.

Letenneur, L., Commenges, D., Dartigues, J.-F. and Barberger-Gateau, P. (1994). Incidence of dementia and Alzheimer's disease in elderly community residents of south-western France. *International Journal of Epidemiology* **23,** 1256-1261.

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

Peto, R. (1973). Experimental survival curves for interval-censored. *Applied Statistics* **22,** 86-91.

Ramlau-Hansen, H. (1983). The choice of a kernel function in the graduation of counting process intensities. *Scandinavian Actuarial Journal*, 165-182.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3,** 425-461.

Senthilselvan, A. (1987). Penalized Likelihood Estimation of Hazard and Intensity Functions. *Journal of the Royal Statistical Society, Series B* **49,** 170-174.

Senthilselvan, A. (1992). Nonparametric estimation of hazard function from left truncated and right censored data. *Nonparametric Statistics* **2,** 29-35.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B* **47,** 1-52.

Tu, X. M., Meng, X-. L. and Pagano, M. (1993). The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association* **88,** 26-36.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **38,** 290-295.

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

Wegman, E. J. and Wright, I. W. (1983). Splines in Statistics. *Journal of the American Statistical Association* **78,** 351-365.

Captions for Figures

Table 1

Comparison of the performances of the MPLE and the smoothed Nelson-Aalen (sNA) estimator for estimating the Weibull hazard function based on 100 replications. The table gives: the sample sizes (1^{rst} column), the density of censoring variable (2^{nd} column), the trimmed estimates of the mean of the integrated squared error (MISE) (and standard deviation) of: distance for MPLE (3^{rd} column) and for the sNA estimator (4^{th} column); the length of $J = [x_{(1)} + b, x_{(N)} - b]$, on which the sNA estimator is defined.

Sample size	C density	MISE[MPLE]		MISE[sNA]		length of J	
50	$f(t; 4, 0.031)$	0.033	(0.037)	0.073	(0.079)	27.38	(3.44)
100	-	0.025	(0.030)	0.050	(0.041)	30.67	(3.03)
500	-	0.054	(0.101)	0.084	(0.079)	35.16	(1.98)
50	$f(t; 2, 0.06)$	0.018	(0.019)	0.044	(0.037)	19.93	(3.08)
100	-	0.021	(0.025)	0.043	(0.036)	23.45	(2.93)
500	-	0.036	(0.064)	0.044	(0.039)	28.26	(2.58)

Table 2

Comparison of the performances of the MPLE and the smoothed Nelson-Aalen (sNA) estimator for estimating the hazard function of the mixture of Gamma $(0.4\Gamma(t, 14, 1.8) + 0.6\Gamma(t, 50, 2))$ based on 100 replications. The table gives: the sample sizes (1^{rst} column), the density of censoring variable (2^{nd} column), the trimmed estimates of the mean of the integrated squared error (MISE) (and standard deviation) of: distance for MPLE (3^{rd} column) and for the sNA estimator (4^{th} column); the length of $J = [x_{(1)} + b, x_{(N)} - b]$, on which the sNA estimator is defined.

Sample size	C density	MISE[MPLE]	MISE[sNA]	length of J
50	$\Gamma(t; 50, 1.65)$	0.038 (0.030)	0.107 (0.079)	22.63 (1.45)
100	-	0.036 (0.030)	0.142 (0.108)	24.44 (1.23)
500	-	0.017 (0.015)	0.176 (0.151)	28.08 (1.14)
50	$\Gamma(t; 50, 2.4)$	0.009 (0.005)	0.010 (0.011)	15.13 (1.62)
100	-	0.009 (0.006)	0.020 (0.021)	17.86 (1.45)
500	-	0.009 (0.006)	0.030 (0.027)	21.43 (1.01)

Table 3

Number of subjects in different age groups according to age at inclusion in the study, age at censorship and age of dementia, and number of person-years. $n = 2717$ Paquid 1989-1994

Age	65-70	70-75	75-80	80-85	85-90	90+
Inclusion	833	602	630	384	206	62
Censoring	232	775	605	551	280	146
dementia	2	13	33	38	26	16
person-years	1564	2500	2129	1542	742	272

Table 4

Regression analysis of age of dementia, comparison between penalized likelihood method and Cox model with delayed entry. In the columns there are respectively, the covariates, and for the two methods the estimated regression coefficients, standard errors, relative risks, confidence intervals for the relative risk and log-likelihood. Three models involving gender and education level were fitted.

	Penalized likelihood method					Cox model with delayed entry				
	β	σ	RR	95% CI	l	β	σ	RR	95% CI	l
					-526.68					
gender	0.27	0.17	1.32	[0.93,1.85]	-525.40	0.25	0.19	1.29	[0.88,1.89]	-711.5
no schooling	0.90	0.36	2.47	[1.21,5.04]	-523.14	0.93	0.36	2.55	[1.24,5.23]	-708.8
grade school	0.40	0.20	1.49	[0.99,2.24]		0.40	0.21	1.50	[0.99,2.27]	
gender	0.22	0.18	1.24	[0.87,1.77]	-522.40	0.23	0.19	1.25	[0.86,1.84]	-708.1
no schooling	0.86	0.36	2.36	[1.15,4.85]		0.91	0.36	2.50	[1.22,5.15]	
grade school	0.36	0.20	1.44	[0.95,2.17]		0.39	0.21	1.47	[0.98,2.23]	

Figure 1: True hazard function, smoothed Nelson-Aalen estimator and approximated maximum penalized likelihood estimator for a simulated example. The sample size is 500, with 40 right-censored data. The smoothed Nelson-Aalen estimator is plotted within the interval J on which it is defined, the other curves are plotted within $[0, x_{(n)}]$.

Figure 2: Approximation of the hazard function of dementia in Gironde (solid line) and bayesian confidence bands (dotted line).