Selection between proportional and stratified hazards models based on expected log-likelihood

Benoit Liquet, Jérôme Saracco, Daniel Commenges

Selection between proportional and stratified hazards models based on expected log-likelihood

Benoit Liquet[1], Jérôme Saracco[2] and Daniel Commenges[3]

[1] LabSAD, BSHM,
1251 avenue centrale, BP 47, 38040 Grenoble Cedex 09, France
e-mail: benoit.liquet@upmf-grenoble.fr

[2] Institut de Mathématiques de Bourgogne, UMR CNRS 5584,
Université de Bourgogne,
20 avenue Alain Savary, 21 078 Dijon Cedex, France
e-mail: Jerome.Saracco@u-bourgogne.fr

[3] INSERM E03 38, ISPED, Université Bordeaux 2,
146 rue Léo Saignat, 3076 Bordeaux cedex, France
e-mail: daniel.commenges@isped.u-bordeaux2.fr

Summary

The problem of selecting between semi-parametric and proportional hazards models is considered. We propose to make this choice based on the expectation of the log-likelihood (ELL) which can be estimated by the likelihood cross-validation (LCV) criterion. The criterion is used to choose an estimator in families of semi-parametric estimators defined by the penalized likelihood. A simulation study shows that the ELL criterion performs nearly as well in this problem as the optimal Kullback-Leibler criterion in term of Kullback-Leibler distance and that LCV performs reasonably well. The approach is applied to a model of age-specific risk of dementia as a function of sex and educational level from the data of a large cohort study.

Keywords: Kullback-Leibler information, likelihood cross-validation, model selection, proportional hazards model, smoothing, stratified hazards model.

1 Introduction

The proportional hazards model has been widely used in epidemiology. The proportional hazards assumption has several advantages: *i*) the effect of a factor can be easily summarized by the relative risk, and *ii*) a mathematical simplicity which has been exploited by Cox (1972) to produce a simple semi-parametric approach via the partial likelihood. This is however a strong assumption which may not be true: it can be relaxed by allowing strata. Nevertheless making strata incurs some loss of information. In practice, graphical methods may be used to see whether the proportional hazards assumption is violated but there is no recognized method for choosing between a proportional hazards model and a stratified model.

Consider now the important and complex problem of the choice between these two models. Gray (1994) has proposed a testing approach to this problem. In this paper, we propose to use a model selection point of view: here the problem is not to know whether the proportional hazards model is the true one but rather to choose the best model for inference, taking into account the available amount of information. In a parametric context, various model selection methods have been proposed from different perspectives including the minimization of the Kullback-Leibler information criteria such as AIC (Akaike, 1973), EIC (Ishiguro et al., 1997), and a Bayesian's point of view such as the BIC (Schwarz, 1978). In a non-parametric framework, smoothing procedures (see Hastie and Tibshirani, 1990 and Silverman, 1986) such as kernel smoothing and penalized likelihood, define families of estimators indexed by the smoothing parameter. The smoothing parameter may be chosen by cross-validation (CV), generalized cross-validation (GCV) (see Craven and Wahba, 1979) or likelihood cross-validation (LCV) (O'Sullivan, 1988). Using a criterion which is an approximation of the Kullback-Leibler information, Liquet et al. (2003) introduced a general point of view which allows to choose an estimator among parametric or semi-parametric families of estimators. Note that this approach is not exactly a model selection approach but an estimator selection approach, which is a little more general.

In the context of incomplete data, in particular for survival data, Liquet and Commenges (2004) proposed the expectation of the log likelihood (ELL) as a theoretical criterion. They have considered both families of kernel and penalized likelihood estimators of the hazard function (indexed on a smoothing parameter) and they have shown on some simulations good results for LCV and a bootstrap estimator of the ELL. In this paper, we propose to use the ELL criterion and its LCV estimator to approach the best smooth estimator in the proportional and in the stratified hazards model. In addition it is possible to choose between the proportional and the stratified hazards models

by comparing the best LCV (estimated ELL) obtained by each model.
The paper is organized as follows. In Section 2, we present the proportional and the stratified hazards models and their estimation via penalized likelihood. The theoretical selection criterion and its estimator, LCV, are described in Section 3. In order to demonstrate the quality of the proposed approach, results of a rather intensive simulation are given in Section 4. In Section 5, we apply our approach to the data of the PAQUID study, a large cohort study on dementia (Letenneur et al., 1994) in order to model onset of dementia as function of sex and educational level. We conclude in Section 6.

2 Penalized likelihood estimation for proportional and stratified hazards models

Let T be the time of the events of interest. Let f and F be the density function and the cumulative distribution function of T. The hazard function is defined by $\lambda(t) = \frac{f(t)}{S(t)}$ where $S = 1 - F$ is the survival function of T. However, we do not observe the realizations of T but only a sample $\mathcal{W} = \{W_1, \ldots, W_n\}$ of independent and identically distributed (i.i.d.) variables which bring information on the variable T. For instance, in the case of right-censored observations, the W_i's are copies of the random variable $W = (\tilde{T}, \delta)$ where $\tilde{T} = \min(T, C)$ and $\delta = \mathrm{I}_{[T \leq C]}$. Other cases of censoring are left and interval censoring. When a vector of explanatory variables X is also available, we observe an i.i.d. sample $\mathcal{W}$, where $W_i = (\tilde{T}_i, \delta_i, X_i)$ for $i = 1, \ldots, n$. In this context, the most popular model is the proportional hazards model in which the hazard function is:

$$\lambda(t|X = x) = \lambda^0(t) \exp(x\beta), \qquad (1)$$

where $\lambda^0(t)$ is the baseline hazard function, x is a row vector $(x_1, \ldots, x_m)$ and β a column vector of regression parameters; the completely proportional hazards model will be denoted $\mathcal{M}_0$. We consider the case where a subset of X is constituted of binary variables (it is sufficient to consider binary variables rather than categorical variables with more than two categories because it is not reasonable to make a proportionality assumption on the effects of several categories; a categorical variable with p categories is generally coded by $p-1$ binary variables). Let us denote by k $(k \leq m)$ the number of binary variables, and without loss of generality we denote by $\beta_1, \ldots, \beta_k$ the regression parameters of the binary variables. The proportionality assumptions may be relaxed for some (or all) of the binary variables. If we relax the proportionality assumptions for l $(l \leq k)$ binary variables, this constitutes 2^l strata. The stratified model based on relaxing the proportionality assumptions for variables $j_1, \ldots, j_l$, denoted by $\mathcal{M}_{J_l}$ (where J_l is the subset $\{j_1, \ldots, j_l\}$ of $\{1, \ldots, k\}$) can be defined as follows:

$$\lambda^{J_l}(t|X = x) = \lambda^{J_l, x_{j_1} \cdots x_{j_l}}(t) \exp x\beta^{J_l} \qquad (2)$$

where $\beta^{J_l}_{j_1} = \ldots = \beta^{J_l}_{j_l} = 0$ and where $\lambda^{J_l, x_{j_1} \ldots x_{j_l}}(t)$ represents the baseline hazard function in stratum $x_{j_1} \ldots x_{j_l}$. Note that β^{J_l} is the same in the 2^l strata. A more radical model relaxes the latter assumption and can be written:

$$\lambda^{J_l}(t|X = x) = \lambda^{J_l, x_{j_1} \ldots x_{j_l}}(t) \exp x\beta^{J_l, x_{j_1} \ldots x_{j_l}}. \quad (3)$$

Here, $\beta^{J_l, x_{j_1} \ldots x_{j_l}}$ is specific of the stratum $x_{j_1} \ldots x_{j_l}$; this boils down to make completely separate analyses for the different strata, but can still be considered as a model for the whole data. Note that there are 2^k stratified models (including the unstratified one) since each binary variable can participate or not to define strata. The number of models is larger if we consider in addition the possibility of completely separate analyses as in (3); of course it is possible to mix stratification on some variables and separate analyses on other. Our main interest focuses on the choice for modelling $\lambda(t|X = x)$ between the proportional hazards model $\mathcal{M}_0$ and several possible stratified models $\mathcal{M}_{J_l}$ obtained with different choices of l and J_l, and also models making separate analyses in some groups.

We will only consider methods which yield smooth estimators of baseline hazard functions. The reason is that non-smooth estimators (such as the Breslow estimator) will yield a value of $-\infty$ for our model choice criteria. Moreover this is realistic to impose a smoothness condition for the applications we have in mind. Kooperberg et al. (1995) have proposed a flexible parametric approach based on splines. Ramlau-Hansen (1983) and Andersen et al. (1993) proposed to smooth the Nelson-Aalen (or the Breslow) estimator by kernel methods.

In this paper we use the approach based on penalized likelihood. An a priori knowledge of smoothness of the hazard function is introduced by penalizing the likelihood by a norm of the second derivative of the hazard function. The estimator is defined nonparametrically as the function that maximizes the penalized likelihood. The solution is then approximated on a basis of splines. Such an approach has been proposed by O'Sullivan (1988) and Joly et al. (1998). This approach has the advantage of dealing with complex cases of truncation and censoring, including interval-censoring whereas the smooth Nelson-Aalen estimator is limited to right-censoring and left truncation.

In the proportional hazards model $\mathcal{M}_0$, let $\widehat{\lambda}^0_h(t)$ and $\widehat{\beta}$ be the estimators of $\lambda^0(t)$ and β that maximize the penalized log-likelihood:

$$p\mathcal{L}_h(\mathcal{W}) = \log \mathcal{L}_p^{\lambda^0, \beta}(\mathcal{W}) - h \int {\lambda^{0}}''^{2}(u)du$$

where $\log \mathcal{L}_p^{\lambda^0, \beta}(\mathcal{W})$ is the log-likelihood (conditional on the $X_i, i = 1, \ldots, n$), h is the smoothing parameter and $\int {\lambda^{0}}''^{2}(u)du$ is the penality term which represents the a priori of the smoothness of $\lambda(t|x)$. For instance with right-censored data, the log-likelihood is defined by:

$$\log \mathcal{L}_p^{\lambda^0,\beta}(\mathcal{W}) = \sum_{i=1}^{n} \left[\delta_i \{\log(\lambda^0(\tilde{T}_i)) + x_i\beta\} - \int_0^{\tilde{T}_i} \lambda^0(u) \exp x_i\beta du \right];$$

see Commenges et al. (1998) for the case with interval-censoring and left-truncation. Thus, given a sample $\mathcal{W}$, penalized likelihood defines a family of estimators $\widehat{\lambda}_h^{\mathcal{W}}(t|x)$ of the proportional hazards model. This family is indexed by one hyper-parameter h. In the stratified model $\mathcal{M}_{J_l}$, let $\widehat{\lambda}_h^{J_l,x_{j_1}\dots x_{j_l}}(t)$ and $\hat{\beta}^{J_l}$ be the estimators of $\lambda_h^{J_l,x_{j_1}\dots x_{j_l}}(t)$ and β^{J_l} which maximize the corresponding penalized likelihood:

$$\begin{aligned} p\mathcal{L}_h(\mathcal{W}) &= \sum_{x_{j_1}=0}^{1} \dots \sum_{x_{j_l}=0}^{1} \log \mathcal{L}_p^{\lambda^{J_l,x_{j_1}\dots x_{j_l}}}(\mathcal{W}^{J_l,x_{j_1}\dots x_{j_l}}) \\ &\quad -h \int \lambda^{J_l,x_{j_1}\dots x_{j_l}}{}''^{2}(u)du \end{aligned}$$

where $\mathcal{W}^{J_l,x_{j_1}\dots x_{j_l}} = \{W_i, i = 1,\dots,n : X_{j_1} = x_{j_1},\dots,X_{j_l} = x_{j_l}\}$. Note that the baseline hazard functions of the strata are estimated using the same smoothing parameter h. Thus the family of estimators of the proportional hazards model and of the different stratified model have both only one hyper-parameter h. Since the number of functions to be estimated varies according to the models, the problem is better formulated in terms of choice in a family of estimators of conditional probabilities: given $\mathcal{W}$ for each J_l and h, the penalized likelihood gives an estimator $\hat{P}_{J_l,h}^{\mathcal{W}}$ (specified by the 2^l functions $\hat{\lambda}^{J_l,x_{j_1}\dots x_{j_l}}$); the problem is the choice in the family $(\hat{P}_{J_l,h}^{\mathcal{W}})$.

3 Selection Criterion

We first exhibit the different criteria for the smooth estimator $\widehat{\lambda}_h^{\mathcal{W}}(t|x)$ of the proportional hazards model and the stratified hazards model in a general censoring context. Then we expose the LCV criterion and its approximation (noted LCVa). In the sequel, we note $\widehat{\lambda}_h^{\mathcal{W}}$ instead of $\widehat{\lambda}_h^{\mathcal{W}}(t|x)$.

3.1 The expected log-likelihood

For uncensored data, Kullback-Leibler information is a useful measure of discrepancy between $f_{T|X}$, the conditional density of T given X, and a smooth estimator $\widehat{f}_{T|X,h}^{\mathcal{T}}$ indexed by h and defined on the sample $\mathcal{T} = \{(T_i, X_i), i = 1,\dots,n\}$. The useful part of this measure (called KL) is defined as the conditional expectation of the log-likelihood of a future observation (T', X') given $\mathcal{T}$:

$$\mathrm{KL}(\mathcal{T}) = \mathbf{E}\left[\log \widehat{f}_{T|X,h}^{\mathcal{T}}(T'|X')|\mathcal{T}\right], \tag{4}$$

where $(T^{'}, X')$ has the same distribution as (T, X) and is independent of the sample $\mathcal{T}$. Given a sample $\mathcal{T}$, we want to select the estimator $\widehat{f}^{\mathcal{T}}_{T|X,h}$ which has the highest KL. Based on KL, Akaike (1973) (see also DeLeeuw, 1992), in a parametric framework for complete data, defined the popular criterion AIC ($AIC = -2\log\mathcal{L} + 2p$, where $\mathcal{L}$ is the likelihood and p is the number of estimated parameters) as an estimator of the expectation of the Kullback-Leibler information, EKL=$\mathbf{E}\,[\mathrm{KL}(\mathcal{T})]$. In principle there should be an advantage in using the "adaptative" KL($\mathcal{T}$) rather than the "non-adaptative" EKL; however it seems illusory to try to estimate KL($\mathcal{T}$), while AIC (as well as some other criteria) can be derived as estimators of EKL.

In presence of incomplete data, even EKL is difficult to estimate. In particular, because $\mathcal{T}$ is not observed, it is not possible to directly estimate the different expectations by bootstrap. We consider the case where we observe an i.i.d. sample $\mathcal{W}$ with $W_i = (\tilde{T}_i, \delta_i, X_i)$ for $i = 1, \ldots, n$. In this context, Liquet and Commenges (2004) have proposed a new criterion, called ELL, which is the expectation of the observed log-likelihood of a new sample which is a copy of the original sample:

$$\mathrm{ELL}(\widehat{\lambda}^{\mathcal{W}}_h) = \mathbf{E}\left[\log\mathcal{L}^{\widehat{\lambda}^{\mathcal{W}}_h}(\mathcal{W}^{'})\right]. \tag{5}$$

where $\mathcal{W}^{'}$ has the same distribution as $\mathcal{W}$ and $\mathcal{L}^{\widehat{\lambda}^{\mathcal{W}}_h}(\mathcal{W}')$ is the likelihood function of the estimator $\widehat{\lambda}^{\mathcal{W}}_h$ for the sample $\mathcal{W}^{'}$. In a sense, this is an observed information criterion. It is easy to show that for uncensored observation $\mathrm{ELL}\{\widehat{\lambda}^{\mathcal{W}}_h\} = n\mathrm{EKL}$. (It would also be possible to define an "adaptative" or conditional version of ELL, but as it is illusory to estimate it, we skip this stage and proceed directly to ELL).

3.2 The LCVa: an estimator of ELL

Throughout this subsection, we index the sample $\mathcal{W}$ by its size n and thus use the notation $\mathcal{W}_n$. We recall that the likelihood cross-validation is defined as:

$$\mathrm{LCV}(\mathcal{W}_n) = \sum_{i=1}^{n}\log\mathcal{L}^{\widehat{\lambda}^{\mathcal{W}^{-i}}_h}(W_i)$$

where $\mathcal{L}^{\widehat{\lambda}^{\mathcal{W}^{-i}}_h}(W_i)$ is the likelihood contribution of W_i for the estimator defined on the sample $\mathcal{W}^{-i}$ in which W_i is removed. The LCV choice for $\widehat{\lambda}^{\mathcal{W}}_h$ is the estimator which maximizes LCV. An important property of LCV is that the expectation of LCV is approximatively equal to ELL and it is shown in Liquet and Commenges (2004) that when $n \longrightarrow \infty$,

$$\frac{\mathbf{E}\,[\mathrm{LCV}(\mathcal{W}_n)]}{\mathrm{ELL}(\widehat{\lambda}_h(n))} \longrightarrow 1,$$

where $\widehat{\lambda}_h(n)$ is an estimator applied to a sample of size n. However we certainly can obtain more: for instance a law of large numbers should apply to $n^{-1}\mathrm{LCV}(\mathcal{W}_n)$ showing that this tends towards $\mathbf{E}[\mathrm{LCV}(\mathcal{W}_n)]$, even though the terms in the sum defining LCV are correlated. A detailed analysis of this topic is beyond the scope of this paper; however we conjecture that LCV is an "estimator" of ELL, in the sense that $\mathrm{LCV}(\mathcal{W}_n)$ will take values close to ELL when n is large, and thus will give results close to ELL when applied to model selection. In a particular case of our simulation study of Section 4, we have displayed on Figure 1 the values of the log-likelihood, ELL and LCVa (an approximation of LCV defined below); as expected the log-likelihood overestimates ELL (because for this value the maximum of the penalized likelihood is obtained) while LCVa achieves a good correction of this bias. It appears that LCVa has rather a negative bias but the shape of the curve as a function of h is similar to that of ELL. Note that when $h \to 0$ the likelihood tends towards ∞ while LCVa like ELL tends towards -∞: the likelihood choice would be to put masses at observed event times while this is strongly rejected by both LCVa and ELL.

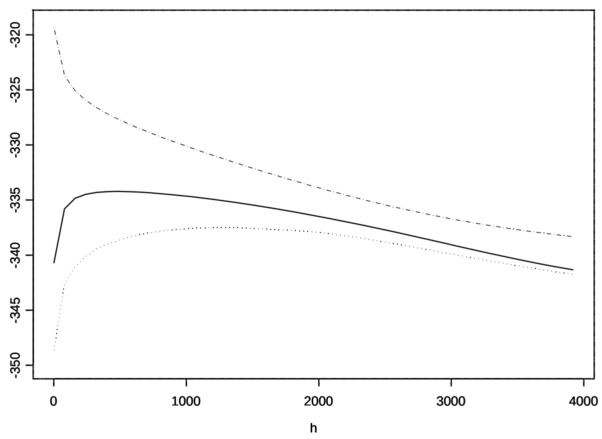


Figure 1: *Plot of ELL (solid line), mean of LCVa (dotted line) and mean of the log likelihood (dashed line) versus the smoothing parameter h for a stratified hazard model with $n = 100$ and 10% of censoring. The mean of LCVa and log-likelihood curves has been calculated over 100 replications.*

If n is large, the computation of LCV is intensive. An approximation based on a first-order expansion of $\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}_{-i}}}(W_i)$ around $\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i)$ can be used. This leads to an expression of the form

$$\mathrm{LCVa}(\mathcal{W}_n) = \sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i) - mdf,$$

where the term mdf can be interpreted as the model degrees of freedom, and this expression is analogous to an AIC criterion. For instance, in the spline approximation of the penalized likelihood, we have $mdf = \mathrm{trace}([\widehat{H} - 2h\Omega]^{-1}\widehat{H})$ where $\widehat{H}$ is the converged Hessian matrix of the log-likelihood, and Ω is the penalized part of the converged Hessian matrix, see Joly et al. (1998) for more details.

4 Simulation

To investigate the behavior of our model selection approach, we did some simulation studies. In the first simulation, the data were generated from a proportional hazards model with one binary variable X coded 0/1 (corresponding to $m = k = 1$). In the sequel, we will denote by $\mathcal{M}_0$ this model. In the second simulation, data were generated from a stratified hazards model. In accordance with the notation of Section 2, the subset of variables used for stratification is $J_1 = \{1\}$, so this model is denoted by $\mathcal{M}_{\{1\}}$ and the hazard functions are:

$$\lambda(t|X = x) = \begin{cases} \lambda^{\{1\},0}(t) & \text{if } x = 0 \\ \lambda^{\{1\},1}(t) & \text{if } x = 1 \end{cases}$$

4.1 Generation of the data

The data were generated from a Weibull distribution and we considered right-censored observations. Samples of size 50, 100 or 500 were generated. The percentages of censoring used were respectively around 10%, 25% or 50%. Each simulation involved 1000 replications.

- **True model is $\mathcal{M}_0$**
  Let $f(t; p, \gamma) = p\gamma^p t^{p-1} \mathrm{e}^{-(\gamma t)^p}$ be the probability density function of the Weibull(p,γ) distribution and let $\lambda(t; p, \gamma) = p\gamma^p t^{p-1}$ be its corresponding hazard function. For $X = 0$, we generated a random sample $\{T_1, ..., T_{n/2}\}$ of i.i.d. failure times from the Weibull(3,0.02) distribution. We generated a random sample $\{C_1, ..., C_{n/2}\}$ of i.i.d. censoring times in the following way: in order to obtain 10% (respectively 25% and 50%), we used the following hazard function for the censoring variable $\lambda_C(t|X = 0) = 0.11\lambda(t; 3, 0.02)$ (respectively $\lambda_C(t|X = 0) = 0.33\lambda(t; 3, 0.02)$ and $\lambda_C(t|X = 0) = \lambda(t; 3, 0.02)$); the C_i's were independent of the T_i's. So the observed sample was $\{(\tilde{T}_1, \delta_1, X_1 = 0), ..., (\tilde{T}_{n/2}, \delta_{n/2}, X_{n/2} = 0)\}$ where $\tilde{T}_i = \min(T_i, C_i)$

and $\delta_i = \mathrm{I}_{[T_i \leq C_i]}$. For $X = 1$, the two samples $\{T_{n/2+1}, \ldots, T_n\}$ and $\{C_{n/2+1}, \ldots, C_n\}$ were generated according to $\lambda_T(t|X = 1) = \lambda_T(t|X = 0)\exp(1)$ and $\lambda_C(t|X = 1) = \lambda_C(t|X = 0)\exp(1)$. The simulated model $\mathcal{M}_0$ is represented in Figure 2(a).

- **True model is $\mathcal{M}_{\{1\}}$**
  To generate the model $\mathcal{M}_{\{1\}}$, we generated different hazard functions for the two strata. For the stratum $X = 0$, the sample $\{T_1, ..., T_{n/2}\}$ was generated from the Weibull$(2, 0.04)$ distribution. The random sample $\{C_1, ..., C_{n/2}\}$ of censoring times was generated from the Weibull(2,0.0133), Weibull(2,0.0231), Weibull(2,0.04) distributions, corresponding to a percentage of censoring around 10%, 25% and 50% respectively. For the stratum $X = 1$, the sample $\{T_{n/2+1}, ..., T_n\}$ was generated from a Weibull(4,0.03) distribution. The sample $\{C_{n/2+1}, ..., C_n\}$ of censoring times was generated from the Weibull(4,0.0173), Weibull(4, 0.0228), Weibull(4,0.03) distributions corresponding to a percentage of censoring around 10%, 25% and 50% respectively. The simulated model $\mathcal{M}_{\{1\}}$ is represented in Figure 2(b).

4.2 Description

We compared the "theoretical" criteria KL and ELL (which can be only obtained in simulation) and the proposed "practical" criterion LCVa. For each replication $\mathcal{W}$, we computed the useful part of the Kullback-Leibler information between the true density function $f_{T|X}$ and an estimator chosen by each "theoretical" or "practical" criterion

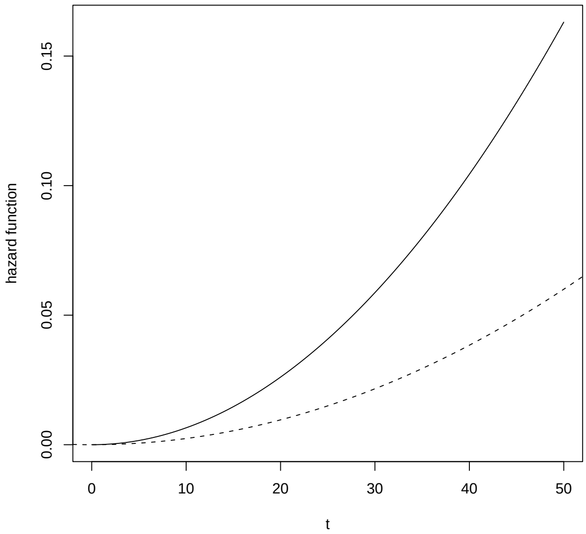
$$\mathrm{KL}(\mathcal{W}) = \mathbf{E}[\log \widehat{f}^{\widehat{\lambda}_h^{\mathcal{W}}}_{T|X}(T'|X')|\mathcal{W}] \quad \text{with} \quad (T', X') \sim F_{T,X}$$

where $\widehat{f}^{\widehat{\lambda}_h^{\mathcal{W}}}_{T|X}$ is defined by the estimator $\widehat{\lambda}_h^{\mathcal{W}}$. We calculated, for each simulation, the average (over the 1000 replications) of KL and its corresponding empirical standard error. The "theoretical" criteria (KL, ELL) were computed by a Monte Carlo method. For instance KL was evaluated as $M^{-1}\sum_{j=1}^{M} \log \widehat{f}^{\widehat{\lambda}_h^{\mathcal{W}}}_{T|X}(T^j|X^j)$, where (T^j, X^j) where independent pseudo-random numbers with the distribution $F_{T,X}$; we used $M = 50000$. For simplicity, since KL generally takes negative values, we give in Table 2 the average values -$\overline{KL}$ of -KL. Hence low values of this criterion correspond to estimators close to the true distribution.

4.3 Results

Identifying the true model is not necessarily the right thing to do. When estimation of regression function or prediction is the goal, the true model, even if assumed reasonably simple and known, may not perform the best.

(a) Proportional hazards model $\mathcal{M}_0$



(b) Stratified hazards model $\mathcal{M}_{\{1\}}$

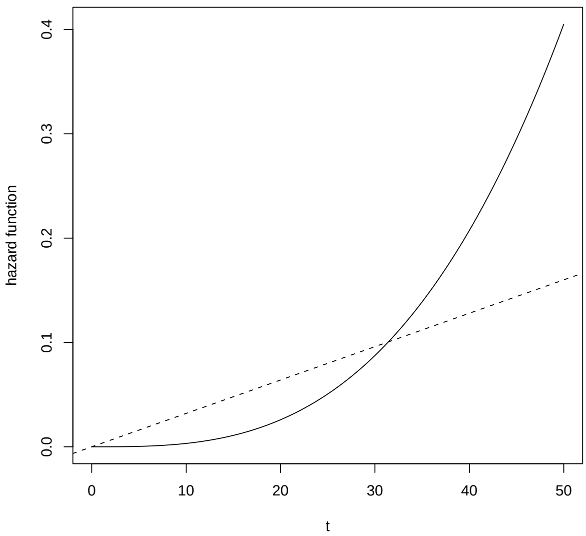


Figure 2: *The two simulated hazards models: solid line for the stratum $X = 0$ and dotted line for the stratum $X = 1$.*

The well-known trade-off between bias and variance may prefer an incorrect but simpler model (see Chapter 1 of Miller (1990) for an illustration in the context of prediction).

For a given $\mathcal{W}$, we consider that the best model between $\mathcal{M}_0$ and $\mathcal{M}_{\{1\}}$ is the one which has the lower -KL value. Therefore our "optimal" (and theoretical) choice is given by the KL criterion. We are interested to have a "practical" criterion which gives most often the same choice between $\mathcal{M}_0$ and $\mathcal{M}_{\{1\}}$ than the optimal criterion KL.

Table 1 presents the frequency of the model selected by the KL and LCVa criteria. In the two simulation cases (the true model is respectively $\mathcal{M}_0$ and $\mathcal{M}_{\{1\}}$), the LCVa criterion tends to select models that agree with those chosen by KL. Nevertheless we can observe several differences in the choice between $\mathcal{M}_0$ and $\mathcal{M}_{\{1\}}$ for small sample size ($n = 50$). We can note that, in all simulations, the "non-adaptative criterion" ELL choses the simulated model; ELL for $\mathcal{M}_0$ (noted $\mathrm{ELL}_{\mathcal{M}_0}$) is lower than $\mathrm{ELL}_{\mathcal{M}_{\{1\}}}$ when the simulated model is $\mathcal{M}_0$ and $\mathrm{ELL}_{\mathcal{M}_{\{1\}}} < \mathrm{ELL}_{\mathcal{M}_0}$ when the simulated model is $\mathcal{M}_{\{1\}}$ (these values are not presented here).

For a given $\mathcal{W}$, we chose by each criterion (KL, ELL, LCVa) the best estimator for the models $\mathcal{M}_0$ and $\mathcal{M}_{\{1\}}$, and then chose the best one. For the selected estimator, we computed the corresponding values of -KL, then we calculated the average of -KL (over all the 1000 $\mathcal{W}$'s replications). These averages are shown in Table 2 (the numbers in parentheses are the corresponding standard errors). The KL and ELL criteria give similar results; the larger differences occur when there is little information (small sample size and high censoring level). The LCVa criterion also yields close results. Here the larger differences also occur when the information is sparse.

To illustrate more precisely these results, we represented in Figure 3 the boxplots of -KL for the different criteria for the two models ($\mathrm{KL}_{\mathcal{M}_0}$, $\mathrm{KL}_{\mathcal{M}_{\{1\}}}$, $\mathrm{ELL}_{\mathcal{M}_0}$, $\mathrm{ELL}_{\mathcal{M}_{\{1\}}}$, $\mathrm{LCVa}_{\mathcal{M}_0}$, $\mathrm{LCVa}_{\mathcal{M}_{\{1\}}}$) when the simulated model is $\mathcal{M}_0$ and also when the simulated model is $\mathcal{M}_{\{1\}}$, with 25% of censoring and for different sample sizes ($n = 100$ and $n = 500$). For instance, the notation $\mathrm{KL}_{\mathcal{M}_0}$ represents the criterion KL applied to model $\mathcal{M}_0$. Figure 3 shows that the differences between the values of -KL for $\mathrm{KL}_{\mathcal{M}_0}$ and $\mathrm{KL}_{\mathcal{M}_{\{1\}}}$ are significative. This is particulary obvious when the true model is $\mathcal{M}_{\{1\}}$. With the $\mathrm{ELL}_{\mathcal{M}_0}$ and $\mathrm{ELL}_{\mathcal{M}_{\{1\}}}$ criteria, we observe the same significative differences. The LCVa criterion has a similar behaviour; the variability of the -KL values is larger for LCVa but this was expected since LCVa is an estimator of ELL.

5 Application

We analysed data from the PAQUID study (Letenneur et al., 1994), a prospective cohort study of mental and physical aging that evaluates social environment and healh status. The PAQUID study is based on a large cohort

(a) Simulated model $\mathcal{M}_0$

$\simeq$ 10% censoring	n= 50 $\mathcal{M}_0$	n= 50 $\mathcal{M}_{\{1\}}$	n= 100 $\mathcal{M}_0$	n= 100 $\mathcal{M}_{\{1\}}$	n= 500 $\mathcal{M}_0$	n= 500 $\mathcal{M}_{\{1\}}$
KL	837	163	897	103	962	38
LCVa	840	160	907	93	933	67
LCVa≡KL	701	24	809	5	899	4
$\simeq$ 25% censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	788	212	877	123	955	45
LCVa	828	172	892	108	932	68
LCVa≡KL	637	21	780	11	888	1
$\simeq$ 50% censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	811	189	862	138	949	51
LCVa	770	230	878	122	929	71
LCVa≡KL	616	35	748	8	878	0

(b) Simulated model $\mathcal{M}_{\{1\}}$

$\simeq$ 10% censoring	n= 50 $\mathcal{M}_0$	n= 50 $\mathcal{M}_{\{1\}}$	n= 100 $\mathcal{M}_0$	n= 100 $\mathcal{M}_{\{1\}}$	n= 500 $\mathcal{M}_0$	n= 500 $\mathcal{M}_{\{1\}}$
KL	4	996	0	1000	0	1000
LCVa	407	593	242	758	3	997
LCVa≡KL	0	589	0	758	0	997
$\simeq$ 25% censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	1	999	0	1000	0	1000
LCVa	442	558	293	707	6	994
LCVa≡KL	1	558	0	707	0	994
$\simeq$ 50% censoring	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$	$\mathcal{M}_0$	$\mathcal{M}_{\{1\}}$
KL	44	956	10	990	0	1000
LCVa	441	559	319	681	8	992
LCVa≡KL	10	525	0	671	0	992

Table 1: *Frequencies of selection of the models by different criteria using 1000 independent replications when the simulated model is (a) a proportional hazards model $\mathcal{M}_0$ or (b) a stratified hazards model $\mathcal{M}_{\{1\}}$, for different sample sizes and various levels of censoring. The notation $\equiv$ means that the two criteria agree for the choice between $\mathcal{M}_0$ and $\mathcal{M}_{\{1\}}$.*

(a) Simulated model $\mathcal{M}_0$

$\simeq$ 10% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	4.061(0.0007)	4.053(0.0004)	4.040(0.0001)
ELL	4.070(0.0011)	4.056(0.0004)	4.040(0.0001)
LCVa	4.125(0.0034)	4.072(0.0011)	4.044(0.0002)
$\simeq$ 25% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	4.074(0.0008)	4.064(0.0005)	4.036(0.0001)
ELL	4.086(0.0014)	4.067(0.0005)	4.036(0.0001)
LCVa	4.162(0.0050)	4.088(0.0013)	4.040(0.0002)
$\simeq$ 50% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	4.114(0.0018)	4.079(0.0009)	4.053(0.0003)
ELL	4.137(0.0027)	4.098(0.0011)	4.056(0.0003)
LCVa	4.297(0.0119)	4.124(0.0027)	4.062(0.0006)

(b) Simulated model $\mathcal{M}_{\{1\}}$

$\simeq$ 10% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	3.734(0.0008)	3.715(0.0004)	3.695(0.0001)
ELL	3.740(0.0009)	3.719(0.0004)	3.696(0.0001)
LCVa	3.827(0.0039)	3.751(0.0016)	3.697(0.0002)
$\simeq$ 25% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	3.740(0.0010)	3.729(0.0006)	3.695(0.0001)
ELL	3.746(0.0011)	3.734(0.0006)	3.697(0.0002)
LCVa	3.863(0.0103)	3.769(0.0017)	3.699(0.0003)
$\simeq$ 50% censoring	n= 50 $-\overline{KL}$	n= 100 $-\overline{KL}$	n= 500 $-\overline{KL}$
KL	3.781(0.0020)	3.750(0.0011)	3.713(0.0003)
ELL	3.795(0.0021)	3.772(0.0022)	3.717(0.0003)
LCVa	3.978(0.0092)	3.821(0.0043)	3.719(0.0006)

Table 2: *Average Kullback-Leibler information* $-\overline{KL}$ *and the corresponding standard errors (numbers in the parentheses) for each criterion when the simulated model is (a) a proportional hazards model* $\mathcal{M}_0$ *or (b) a stratified hazards model* $\mathcal{M}_{\{1\}}$*, for different sample sizes* n *and various levels of censoring.*

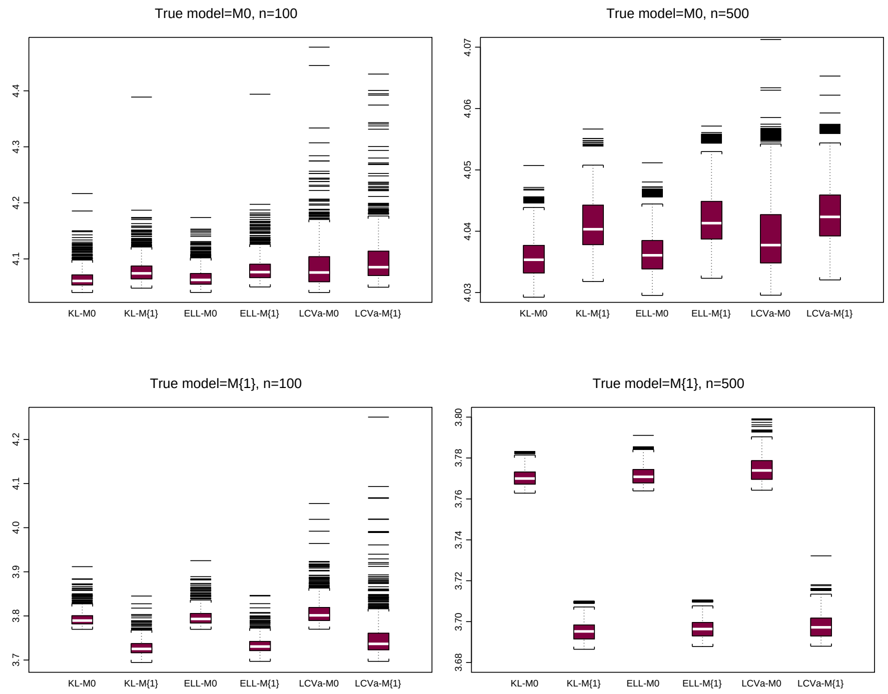


Figure 3: *Boxplots of the -KL for the different criteria when the simulated model is a proportional hazard model* $\mathcal{M}_0$ *(pictures on the top) and when the simulated model is a stratified hazard model* $\mathcal{M}_{\{1\}}$ *(pictures on the bottom) for sample sizes* $n = 100$ *and* $n = 500$ *with 25% of censoring.*

randomly selected in a population of subjects aged 65 years or more, living at home in two departments of southwest France (Gironde and Dordogne). There were 3675 non-demented subjects at entry in the cohort and each subject has been visited six times or less, between 1988 and 2000; 431 incident cases of dementia were observed during the follow up. The risk of developing dementia was modeled as a function of age. As prevalent cases of dementia were excluded, data were left-truncated and the truncation variable was the age at entry in the cohort (for more details see Commenges et al., 1998). Two explanatory variables were considered: sex (noted S) and educational level (noted E). In the sample, there were 2133 women and 1542 men. Educational level was classified into two categories: no primary school diploma and primary school diploma (Letenneur et al., 1999). The pattern of observations involved interval-censoring and left-truncation. We use the likelihood for interval-censored and left-truncated data. For sake of simplicity, we kept here the survival data framework, treating death as censoring rather than the more adapted multistate framework (Commenges et al., 2004).

There are two layers of model selection: the selection of the variables and the selection of the model given the variables; a third layer is the selection of the smoothing coefficient h given variables and model. Even with only two variables there is a total number of 11 possible models. It is common in epidemiology not to examine all possible models: we adopted a strategy involving the evaluation of only five models and giving up the notation of Section 2 which would be somewhat cumbersome here we denoted them A, B, C, D, E. We first examined the effect of sex. We used penalized likelihood method and the LCVa criterion to choose between the stratified hazards model on sex (called here model A) and the proportional hazards model on sex (called model B). The selected model is model A since Table 3 shows that model B had lower value for LCV_a than model A. We represented in Figure 4 the penalized likelihood estimate of the risk of dementia for men and women with the corresponding estimated stratified hazards model.

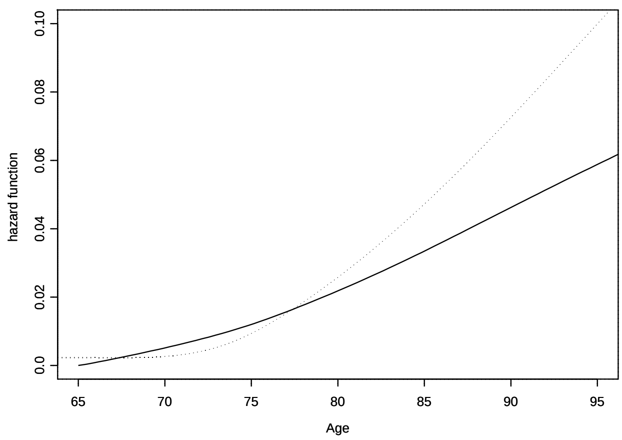


Figure 4: *Estimates of the hazard function of dementia for male (solid line) and female (dotted line) chosen by LCVa criterion for selected model A.*

It appears that women tend to have a lower risk of dementia than men before 78 years and a higher risk above that age.

Another important risk factor for dementia is educational level. As the proportional hazards assumption does not hold as we have seen previously, we performed several analyses on the educational level stratified on sex. We considered three models. The first model is a model stratified on sex and proportional for the educational level; the coefficient of proportionality of educational levels is the same for women and men. We denote it model C:

$$\lambda(t|S_i, E_i) = \begin{cases} \lambda_h^0(t) \exp \beta E_i & \text{if } \; S_i = 0 \text{ (women)}, \\ \lambda_h^1(t) \exp \beta E_i & \text{if } \; S_i = 1 \text{ (men)}. \end{cases}$$

Secondly, we considered a proportional hazards model on educational level with different coefficient of proportionality for men and women (denoted model D):

$$\lambda(t|S_i, E_i) = \begin{cases} \lambda_h^0(t) \exp \beta_0 E_i & \text{if } \; S_i = 0 \text{ (women)}, \\ \lambda_h^1(t) \exp \beta_1 E_i & \text{if } \; S_i = 1 \text{ (men)}. \end{cases}$$

Finally, we used a model stratified on both sex and educational level (denoted model E):

$$\lambda(t|S_i, E_i) = \begin{cases} \lambda_h^{0,0}(t) & \text{if } \; S_i = 0 \text{ and } E_i = 0, \\ \lambda_h^{1,0}(t) & \text{if } \; S_i = 1 \text{ and } E_i = 0, \\ \lambda_h^{0,1}(t) & \text{if } \; S_i = 0 \text{ and } E_i = 1, \\ \lambda_h^{1,1}(t) & \text{if } \; S_i = 1 \text{ and } E_i = 1. \end{cases}$$

Table 3 presents the values of the LCVa criterion for each model. The selected model is the stratified proportional hazards model (highest value for the model C). Subjects with no primary school diploma have an increased risk of dementia. For this model (model C), the estimated relative risk for educational level is equal to 1.97; the corresponding 95% confidence interval is [1.63; 2.37].

	model A	model B	model C	model D	model E
LCVa	-1517.45	-1519.92	-1496.28	-1497.18	-1498.42

Table 3: *Comparison of stratified and proportional hazards models according to the LCVa criterion; models A and B are respectively: stratified and unstratified models on sex; models C, D, E are 3 models stratified on sex with educational level as new covariate (see text).*

6 Conclusion

We have presented a method for the choice between stratified and proportional hazards models. We have shown that this could be done using LCVa which is an estimator of ELL. We observed that ELL has a very similar behaviour as the optimal Kullback-Leibler criterion and LCVa also provided good performances. We considered here a case where all the models are indexed by a single hyper-parameter. This raises a completely new problem which is how to compare families of models of different complexities, i.e. indexed by a different number of hyper-parameters. For instance this problem would arise if we compared a proportional hazards model (one hyper-parameter) to a stratified model with one hyper-parameter for each stratum. We conjecture that there is a principle of parsimony at the level of the hyper-parameter, similar to that known for the ordinary parameters. Further research would be needed to explore this field.

Acknowledgements

We thank Luc Letenneur and Jean-François Dartigues for allowing the use of the PAQUID data. We would like to thank the two anonymous referees and the Associated Editor for providing helpful comments that greatly improved the final version of this article.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *Second International Symposium on Information Theory*, 267-281, Budapest. Akademiai Kiado.

Andersen, P.K., Borgan, R., Gill, R. & Keiding, D. (1993) *Statistical models based on counting processes*. Springer-Verlag, New-York.

Commenges, D., Joly, P., Letenneur, L. & Dartigues, JF. (2004). Incidence and prevalence of Alzheimer's disease or dementia using an Illness-death model. *Statistics in Medicine*, **23**, 199-210.

Commenges, D., Letenneur, L., Joly, P., Alioum, A. & Dartigues, J. (1998). Modelling age-specific risk: application to dementia. *Statistics in Medicine*, **17**, 1973-1988.

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

Craven, P & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, **31**, 377-403.

DeLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics, volume I: Foundations and basic theory* (Edited by Kotz, S. & Johnson, N. L.), 599-609. Springer-Verlag, New York.

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

Gray, R.J. (1994). Splines-based tests in survival analysis. *Biometrics*, **50**, 640-652.

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

Ishiguro, M., Sakamoto, Y. & Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics*, **49**, 411-434.

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

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

Letenneur, L., Commenges, D., Dartigues, J. & 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.

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, J. & Dartigues, J. (1999). Are sex and educational level independent predictors of dementia and alzheimer's disease? Incidence data from the PAQUID project. *Journal of Neurology, Neurosurgery and Psychiatry*, **66**, 177-183.

Liquet, B. & Commenges, D. (2004). Estimating the expectation of the log-likelihood with censored data for estimator selection. *Lifetime Data Analysis*, **10**, 351-367.

Liquet, B., Sakarovitch, C. & Commenges, D. (2003). Bootstrap choice of estimators in non-parametric families: an extension of EIC. *Biometrics*, **59**, 172-178.

Miller, A.J. (1990). *Subset Selection in Regression.* Chapman and Hall, New York.

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.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions.

Annals of Statistics, **11**, 453-466.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464.

Silverman, B. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall, London.