Nonparametric estimation in Cox-models: Time transformation methods versus partial likelihood methods.

Jan Terje Kvaløy
Department of Mathematical Sciences
Norwegian University of Science and Technology
N-7034 Trondheim, Norway
E-mail: jtk@stat.ntnu.no

January 18, 1999

Abstract

A very convenient and widely used tool for modeling the relationship between a possibly censored survival time and covariates is the Cox proportional hazard regression model. This model is usually semi-parametric, with an unspecified baseline hazard and a parametric log-linear covariate function. In this paper generalizations of the Cox-model to a completely nonparametric model with an unspecified smooth covariate function is studied.

The paper offer a review of existing methods for nonparametric Cox-regression. Moreover, a class of methods for Cox-regression called time transformation methods is proposed. Time transformation methods are iterative methods, and the idea of these methods is to transform estimation of the covariate function into an exponential regression problem by a time scale transformation. Several existing methods for Cox-regression, both for semi- and nonparametric models, are time transformation methods. A new method suggested and studied in this paper is a method called the covariate order time transformation method. In this method the exponential regression part of the problem is solved by using a nonparametric method for exponential regression based on arranging the observations in increasing order of the covariate. The covariate order time transformation method enjoys a number of useful properties. Numerical convergence is fast, implementation is easy, it is easy to fit stratified models, it is easy to fit models with a covariate function with both nonparametric and parametric terms, and ties cause no problems. A simulation study show that this method, compared to other nonparametric methods, has low variability, favorable properties in cases with few observations or heavy censoring, and no problems with numerical convergence has been encountered.

Ways of doing inference and model check in generalized nonparametric Cox-models are surveyed and demonstrated in practice by some real data examples.

1 INTRODUCTION

A primary goal in the analysis of survival data is often to investigate the relationship between a survival time response variable and one or more covariates. In practice the observed survival times will often be censored. A very popular and widely used tool for investigating the relationship between possibly censored survival times and covariates is the Cox (1972) proportional hazard model. The Cox-model assumes that the hazard function can be written in the form $\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(\beta\mathbf{x})$, where $\alpha_0(t)$ is an unknown baseline hazard function, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\mathbf{x}$ is a vector of covariates. In this paper the generalization of the Cox-model to the model $\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(g(\mathbf{x}))$ will be studied. The $g(\mathbf{x})$-function is assumed to be a general function which will be called the *covariate function*. This is an important and interesting generalization of the Cox-model.

Several authors have addressed the problem of finding methods for estimating a general smooth covariate function in a generalized Cox-model. Tibshirani (1984), Tibshirani and Hastie (1987), Staniswalis (1989), Gentleman and Crowley (1991) and Fan, Gijbels and King (1997) are all using various local likelihood approaches. Spline based methods are among others studied by O'Sullivan (1988), Hastie and Tibshirani (1990), Sleeper and Harrington (1990), Gray (1992) and Kooperberger, Stone and Truong (1995a). Section 5 offers a short review and comparison of some of these methods seen in relation to the method presented in this paper.

The approach taken in this paper is to transform the estimation of the covariate function into an exponential regression problem in an iterative algorithm. This very general approach generates a class of methods which will be called *time transformation methods*. Examples of time transformation methods are the local full likelihood method (see for example Gentleman and Crowley, 1991), and the EM algorithm suggested by Clayton and Cuzick (1985) for estimation of $\boldsymbol{\beta}$ and the cumulative baseline hazard function in the classical Cox-model. A new method for nonparametric exponential regression on right censored data called the covariate order method, were suggested and studied in Kvaløy and Lindqvist (1998a). A simulation study showed that this method had favorable properties when compared to other methods and therefore a time transformation method using the covariate order method in the exponential regression step is suggested. This method is called the covariate order time transformation method.

In principle the time transformation approach allows unrestricted estimation of the covariate function, but in the covariate order time transformation method suggested in this paper attention is restricted to models that can be written in the form of a generalized additive model, $\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(g_1(x_1) + \ldots + g_m(x_m))$, where $\mathbf{x} = (x_1, \ldots, x_m)$ and the $g_j(x_j)$-functions are unspecified smooth covariate functions. The covariate order time transformation method is very flexible, combinations of nonparametric and parametric (for example for discrete covariates) covariate functions can be modeled, stratified models can easily be fit, ties cause no problems and the baseline hazard function can be estimated either nonparametrically or parametrically. The method is easily implemented in packages that offer standard Cox-regression, and numerical convergence is fast.

The time transformation approach and the covariate order time transformation method are presented in Section 2, methods for nonparametric exponential regression are reviewed in Section 3, ways of doing inference in generalized Cox-models are discussed in Section 4 and relations to other methods are discussed in Section 5. A simulation study comparing the covariate order time transformation method to a local full likelihood method and a local partial likelihood method is presented in Section 6, and some real data examples are given in Section 7.

2 THE TIME TRANSFORMATION APPROACH FOR COX REGRESSION

Assume that we have n independent observations $(T_1, \delta_1, \mathbf{X}_1), \ldots, (T_n, \delta_n, \mathbf{X}_n)$ of the random triple $(T, \delta, \mathbf{X})$, where $T = min(Z, C)$, $\delta = I(Z \leq C)$, and $\mathbf{X}$ is a m-dimensional vector of covariates. We call T the observation time, Z the life time and C the censoring time. Let Z have an unspecified life time distribution, and assume that the censoring is independent and noninformative (see for example Kalbfleisch and Prentice, 1980, chap. 5).

The Cox (1972) proportional hazards model for the hazard function of Z is $\alpha(t|\mathbf{X} = \mathbf{x}) = \alpha_0(t)\exp(\beta\mathbf{x})$, where $\alpha_0(t)$ is a baseline hazard function and $\boldsymbol{\beta}$ is a vector of regression coefficients. In this classical formulation of the Cox-model the effect of the covariates on the hazard is modeled to be log-linear. A natural way of generalizing the model is to introduce the model

$$\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(g(\mathbf{x})), \tag{1}$$

where $g(\mathbf{x})$ in principle is any function of the covariates. If $g(\mathbf{x})$ is some parametric function, $g(\mathbf{x})$ can be estimated independently of $\alpha_0(t)$ by maximizing what is called the *Cox partial likelihood* (Cox, 1975)

$$PL(g(\cdot)) = \prod_{i=1}^{n} \frac{\delta_i \exp(g(\mathbf{x}_i))}{\sum_{j \in R_i} \exp(g(\mathbf{x}_i))} \tag{2}$$

where $R_i = \{j|T_j \geq T_i\}$ is the risk set just before time T_i. If $g(\mathbf{x})$ is a general smooth function, then a nonparametric estimate can be found by using localized versions of the partial likelihood. See Section 5.1 for details. Here a different approach than using the partial likelihood will be presented

Define the cumulative baseline hazard function by $A_0(t) = \int_0^t \alpha_0(u)du$. Then $g(\mathbf{x})$ can be estimated by using the following general algorithm:

General time transformation algorithm:

1. Find an initial estimate $\hat{A}_0(t)$ of the cumulative baseline hazard $A_0(t)$
2. Transform $T_1, \ldots, T_n$ to $\hat{A}_0(T_1), \ldots, \hat{A}_0(T_n)$.
3. Estimate $g(\mathbf{x})$ from $(\hat{A}_0(T_1), \delta_1, \mathbf{X}_1), \ldots, (\hat{A}_0(T_n), \delta_n, \mathbf{X}_n)$ using some method for exponential regression.
4. Find a new estimate of $A_0(t)$ from $(T_1, \delta_1, \hat{g}(\mathbf{X}_1)), \ldots, (T_n, \delta_n, \hat{g}(\mathbf{X}_n))$.
5. Repeat 2-4 until convergence

If we want to fit a stratified model, $\alpha_{oq}(t)\exp(g(x))$, $q = 1, \ldots, s$, where s is the number of strata, the only change needed in the algorithm is to transform each T_i in step 2 according to the estimated $\hat{A}_{oq}(\cdot)$ for the stratum which T_i belongs to.

The key to why this algorithm works is the observation that if a distribution has hazard function $\alpha(t)$, and Z is an uncensored observation from this distribution, then $A(Z) = \int_0^Z \alpha(u)du$ is exponentially distributed with parameter 1. This implies that if $\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(g(\mathbf{x}))$, $A_0(Z)$ will be exponentially distributed with parameter $\exp(g(\mathbf{x}))$, and this motivates why $g(\mathbf{x})$ in step 3 can be estimated using a method for exponential regression. By fitting a new model based on the transformed covariates in step 4 we get a better estimate of $\alpha_0(t)$ which then can be given as new input in 2 and so on. The motivating idea is that if $\alpha_0(t)$ is known, a consistent estimate of $g(\mathbf{x})$ can be found, and if $g(\mathbf{x})$ is known a consistent estimate of $\alpha_0(t)$ can be found. It is of future research to prove formally that this algorithm in general give consistent estimators of $\alpha_0(t)$ and $g(\mathbf{x})$. Proofs in special cases are given by Clayton and Cuzick (1985) and Fan et al. (1997). The algorithm is called a time transformation algorithm since a key idea in the algorithm is to transform the observation times to a new time scale before the covariate function is estimated.

So far the algorithm has only been given in general terms. There is a number of different ways to implement the algorithm in practice. Both parametric and nonparametric methods can be used to estimate $A_0(t)$ and $g(\mathbf{x})$. For $m > 1$ covariates, several variants of the algorithm can be thought of, depending on which assumptions are made. Any method for Cox-regression which represents an implementation of the general time transformation algorithm will be called a time transformation method for Cox-regression.

Examples of time transformation methods are the local full likelihood method (see for example Gentleman and Crowley, 1991) for nonparametric Cox-regression, and the EM algorithm suggested by Clayton and Cuzick (1985) for the classical Cox-model. These methods will be further discussed in subsequent sections. Kooperberger, Stone and Truong (1995a, p. 84) suggest using a preliminary time transformation in order to obtain data with a hazard function less depending on time.

Two suggested implementations of the general time transformation algorithm are presented below, one for the case of only one *continuous* covariate and a smooth $g(x)$-function, and one for the case of $m > 1$ covariates and the assumption of a generalized additive model. The covariate order method for exponential regression which is suggested for the exponential regression part of the algorithm, is presented in Section 3.1.

First notice that $g(\mathbf{x})$ is only identifiable up to a constant as we can move a constant from the baseline to the covariate function, $\alpha_0(t)\exp(g(\mathbf{x})) = c\alpha_0(t)\exp(g(\mathbf{x}) - \ln c)$. This can easily be handled, however, by imposing some simple restriction on the model. For example several of the authors cited in the introduction assume that $g(0) = 0$. In the present paper it is instead assumed that $\mathrm{E}(g(\mathbf{X})) = 0$, in which case the implementation of the algorithm for the case of one covariate is:

Algorithm 1 (a single covariate):

1. Calculate estimates $\hat{A}_0(t)$ and $\hat{\beta}$ by fitting an ordinary Cox-model $\alpha_0(t)\exp(\beta x)$ to the original data.

2. Transform $T_1, \ldots, T_n$ to $\hat{A}_0(T_1), \ldots, \hat{A}_0(T_n)$.

3. Calculate the estimate $\tilde{g}(x)$ from $(\hat{A}_0(T_1), \delta_1, X_1), \ldots, (\hat{A}_0(T_n), \delta_n, X_n)$ using the covariate order method for exponential regression, and let $\hat{g}(x) = \tilde{g}(x) - \sum_{i=1}^n \tilde{g}(X_i)/n$

4. Calculate a new estimate $\hat{A}_0(t)$ from the sample $(T_1, \delta_1, \hat{g}(X_1)), \ldots, (T_n, \delta_n, \hat{g}(X_n))$, by fitting the Cox-model $\alpha_0(t)\exp(\beta\hat{g}(x))$

5. Repeat 2-4 until convergence

Numerical convergence of the algorithm is very fast, usually only 3-4 iterations are needed if the convergence criterion is that the maximum difference in $\hat{g}(x)$ from one iteration to the other, taken over $X_1, \ldots, X_n$, should be less than 10^{-5}.

It may seem strange to fit a new Cox-model in step 4 when the only goal is to get a new estimate of the cumulative baseline hazard function. Instead of doing this, step 4 can for instance be replaced by a direct application of the Breslow estimator (Breslow, 1972). Then $A_0(t)$ is estimated by $\hat{A}_0(t) = \sum_{i=1}^n (\delta_i / \sum_{j \in R_i} \exp(\hat{g}(X_j))) I(T_i \leq t)$, and in Algorithm 1 we can indeed just as well do this. One motivation for fitting a Cox-model is that this leads to a desirable way of estimating the parametric terms in larger models with combinations of parametric and nonparametric terms as presented below, another reason is easy implementation in standard statistical software packages, and finally, fitting a Cox-model in step 4 seems to speed up the convergence of the algorithm. This last point is further discussed in Section 5.2. The estimated value of β in step 4 is of course expected to be near one if the model is correctly specified.

For the case of $m > 1$ continuous covariates we will reduce the dimensionality of the problem by considering generalized additive models of the form $g(\mathbf{x}) = g_1(x_1) + \cdots + g_m(x_m)$, that is

$$\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(g_1(x_1) + \cdots + g_m(x_m)) \tag{3}$$

where the $g_j(x_j)$ are unspecified smooth covariate functions. The general algorithm can be modified to estimating such generalized additive models by making the observation that if $\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(g_1(x_1) + \cdots + g_m(x_m))$, then $A_0(Z)\exp(g_1(x_1) + \cdots + g_{j-1}(x_{j-1}) + g_{j+1}(x_{j+1}) + \cdots + g_m(x_m))$ is exponentially distributed with parameter $\exp(g_j(x_j))$. The notation $\hat{g}_j^{(d)}(x_j)$ is introduced to indicate the dth estimate of $g_j(x_j)$, and we define $\hat{g}_j^{(0)}(x_j) = x_j$, $j = 1, \ldots, m$. With this notation the algorithm can be written as the following steps.

Algorithm 2 (m covariates):

1. Calculate estimates $\hat{\alpha}_0(t)$ and $\hat{\boldsymbol{\beta}}$ by fitting the Cox-model $\alpha_0(t)\exp(\boldsymbol{\beta}\mathbf{x}) = \alpha_0(t)\exp(\beta_1 x_1 + \cdots + \beta_m x_m)$ to the original data.

2. Repeat 3 for $d = 1, 2, \ldots, D$ where D is the number of iterations needed to reach convergence.

3. Repeat 4-6 for $j = 1, \ldots, m$.

4. Transform $T_1, \ldots, T_n$ to $L_{1j}, \ldots, L_{nj}$ where $L_{ij} = \hat{A}_0(T_i)\exp(\hat{\beta}_1\hat{g}_1^{(d)}(X_{i1}) + \cdots + \hat{\beta}_{j-1}\hat{g}_{j-1}^{(d)}(X_{i,j-1}) + \hat{\beta}_{j+1}\hat{g}_{j+1}^{(d-1)}(X_{i,j+1}) + \cdots + \hat{\beta}_m\hat{g}_m^{(d-1)}(X_{im}))$, $i = 1, \ldots, n$.

5. Calculate the estimate $\tilde{g}_j^{(d)}(X_j)$ from $(L_{1j}, \delta_1, X_{1j}), \ldots, (L_{nj}, \delta_n, X_{nj})$ using the covariate order method for exponential regression, and let $\hat{g}_j^{(d)}(x_j) = \tilde{g}_j^{(d)}(x_j) - \sum_{i=1}^{n} \tilde{g}_j^{(d)}(X_{ij})/n$

6. Calculate a new estimate $\hat{\alpha}_0(t)$ and $\hat{\boldsymbol{\beta}}$ by fitting the model $\alpha_0(t)\exp(\beta_1\hat{g}_1^{(d)}(x_1) + \cdots + \beta_j\hat{g}_j^{(d)}(x_j) + \beta_{j+1}\hat{g}_{j+1}^{(d-1)}(x_{j+1}) + \cdots + \beta_m\hat{g}_m^{(d-1)}(x_m))$

Convergence of the algorithm is generally fast but will of course depend on how large m is. The generalization to stratified models is done in the same way as discussed previously. Variants of the algorithm can be thought of, for instance can step 3 be replaced by "Repeat 4-6 for all $j \in \{1, \ldots, m\}$ picked in random order." A mix of linear and general covariate functions can also be allowed. Models of the form $\alpha_0(t)\exp(g_1(x_1) + \cdots + g_r(x_r) + \beta_{r+1}x_{r+1} + \cdots + \beta_m x_m)$ can be fitted by including all factors in the algorithm but estimating smooth functions only for the first r factors. In other words, repeat step 4-6 for $j = 1, \ldots, r$ but include the last $m - r$ linear factors in each step. These last $m - r$ factors are updated simultaneously in step 6. This was one of the motivations mentioned previously for fitting a new Cox-model in this step of the algorithm. Particularly note that the covariates modeled by linear terms are allowed to be discrete.

In the rest of the paper, Algorithm 1 and 2 as stated above, using the covariate order method for exponential regression in the exponential regression steps of the algorithms, will be referred to as the covariate order time transformation method.

3 EXPONENTIAL REGRESSION

In this section two methods for nonparametric exponential regression are presented, the covariate order method of Kvaløy and Lindqvist (1998a) used in the exponential regression step of the covariate order time transformation method and a local likelihood method. If the local likelihood method is used in the exponential regression step of a time transformation method then the method is a local full likelihood method for Cox-regression. See Section 5.1 for details.

Assume that we have n independent observations $(L_1, \delta_1, X_1), \ldots, (L_n, \delta_n, X_n)$ of the random triple (L, δ, X), where $L = \min(U, V)$, $\delta = I(U \leq V)$, and where X is a single continuous covariate which is distributed over some domain $\mathcal{X} \subseteq \mathcal{R}$ according to a density function $f_{\mathcal{X}}(x)$. Further, for given $X = x$, U is assumed to be exponentially distributed with an *unknown* hazard function $\lambda(x)$, that is $f_{U|X}(t|x) = \lambda(x)\exp(-\lambda(x)t)$, and $\lambda(x)$ is assumed to be *continuous* but otherwise unspecified. Finally, V is distributed according to some unknown censoring density, $f_{V|X}(t|x)$, and is assumed to be independent of U given X. The goal is to estimate the unknown hazard function $\lambda(x)$. Since only exponential regression in one continuous variable at a time is used in this paper, the methods are only presented for the case of a single continuous covariate.

3.1 Covariate Order Method

First arrange the observations $(L_1, \delta_1, X_1), \ldots, (L_n, \delta_n, X_n)$ such that $X_1 \leq X_2 \leq \ldots \leq X_n$. Then let the scaled observation times $L_1/n, \ldots, L_n/n$, irrespectively if they are censored or not, be subsequent inter-arrival times of an artificial point process on a time axis s. For this process, let only points being the endpoint of an interval corresponding to an uncensored observation be considered as an event, and denote these event times by $S_1, S_2, \ldots$. For the example $(L_1, \delta_1 = 1), (L_2, \delta_2 = 0), (L_3, \delta_3 = 1), \ldots, (L_{n-1}, \delta_{n-1} = 0), (L_n, \delta_n = 1)$ this is visualized in Figure 1. Generally $S_i = \sum_{j=1}^{k(i)} L_j/n$ where $k(i) = \min\{r | \sum_{j=1}^{r} \delta_j = i\}$.

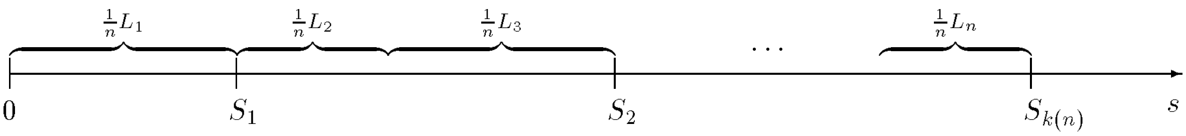

Figure 1: Construction of artificial process.

Let the conditional intensity of the process $S_1, \ldots, S_{k(n)}$ be denoted $\rho_n(s)$. We then have that $\rho_n(s) = n\lambda(X_I)$ for $\sum_{i=1}^{I-1} L_i/n < s \leq \sum_{i=1}^{I} L_i/n$. The basic idea is to estimate this intensity from the process $S_1, \ldots, S_{k(n)}$, yielding the estimator $\hat{\rho}_n(s)$, and then invert the relation $n\hat{\lambda}(X_I) = \hat{\rho}_n(s)$ to obtain an estimate of $\hat{\lambda}(x)$ at given points x. The key here is the relationship between the process $X_1, \ldots, X_n$ on the covariate axis and the process $S_1, \ldots, S_{k(n)}$ on the s-axis. A simple way of estimating the relationship between points on the two axes, also called the *correspondence function*, is by drawing straight lines between the points $(X_1, L_1/n), \ldots, (X_n, \sum_{i=1}^{n} L_i/n)$. This estimated correspondence function is denoted $\tilde{s}(x)$. Then it is proved in Kvaløy and Lindqvist (1998a) that

$$\hat{\lambda}(x) = \frac{1}{nh_s(x)} \sum_{i=1}^{k(n)} K(\frac{\tilde{s}(x) - S_i}{h_s(x)}) \tag{4}$$

is a uniformly consistent estimator of $\lambda(x)$, under relatively mild assumptions. Here $K(\cdot)$ is a symmetric density function on [-1,1] and $h_s(x)$ a smoothing parameter.

In practice the estimate (4) will be downward biased near the endpoints. To avoid this the reflection method (see for example Silverman (1986)) for handling boundary problems in density estimation is used. In practical implementation Kvaløy and Lindqvist (1998a) also suggested replacing the estimated correspondence function $\tilde{s}(x)$ by a smooth curve estimate obtained by applying the super-smoother of Friedman (1984) to the points $(X_1, L_1/n), \ldots, (X_n, \sum_{i=1}^{n} L_i/n)$, improving the smoothness of the estimate (4). See Figure 2 for an illustration.

The smoothing parameter $h_s(x)$ in (4) may or may not be a function of x. In this paper we will only use the smoothing parameter defined by $h_s(x) = \hat{s}(x + h/2) - \hat{s}(x - h/2)$ for a fixed h. This choice of $h_s(x)$ is done for consistency with the competing methods, and h can here be viewed as a constant smoothing parameter on the covariate axis. Which value of h to be used can be decided by a likelihood cross-validation criterion. Choose the value of h which maximize

$$lCV(h) = \sum_{i=1}^{n} [\delta_i \ln \hat{\lambda}^{(-i)}(X_i|h) - \hat{\lambda}^{(-i)}(X_i|h) L_i], \tag{5}$$

where $\hat{\lambda}^{(-i)}(x)$ is the estimate of $\lambda(x)$ based on all data except the ith observation.

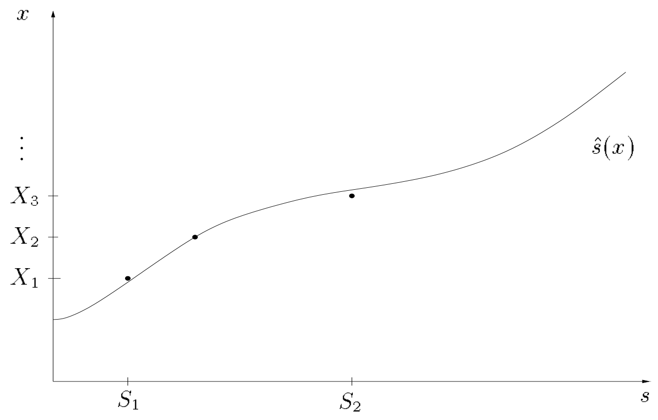

Figure 2: Illustration of estimating a smooth correspondence function from the covariate axis to the s-axis.

3.2 Local Likelihood Method for Exponential Regression

The conditional likelihood given $X_1, \ldots, X_n$, is

$$L(\lambda(\cdot)) = \prod_{i=1}^{n} [\lambda(X_i) \exp(-\lambda(X_i) L_i)]^{\delta_i} [\exp(-\lambda(X_i) L_i)]^{1-\delta_i} \tag{6}$$

leading to the log-likelihood function

$$l(\lambda(\cdot)) = \sum_{i=1}^{n} [\delta_i \ln \lambda(X_i) - \lambda(X_i) L_i] \tag{7}$$

Let $\lambda(x) = \exp(g(x))$ and approximate $g(u)$ in a neighborhood of x by a polynomial $p(u)$ where $p(u) = \beta_0 + \beta_1(u-x) + \cdots + \beta_p(u-x)^p$. Then $\lambda(x)$ is estimated locally in a weighted neighborhood of x by maximizing the local likelihood

$$l(\beta_0, \ldots, \beta_p, x) = \sum_{i=1}^{n} [\delta_i p(X_i) - \exp(p(X_i)) L_i] \, \frac{1}{h} K(\frac{x - X_i}{h}) \tag{8}$$

The estimate of $\lambda(x)$ is then $\hat{\lambda}(x) = \exp(\hat{\beta}_0)$. The smoothing parameter h can be chosen according to the likelihood cross-validation criterion (5).

4 INFERENCE AND MODEL CHECKING

Several ways of doing inference and model checking in nonparametric Cox-models are discussed in this section. These methods are presented in relation to the covariate order time transformation method, but can, as discussed in Section 5.4, also be applied with other non- or semiparametric methods for Cox-regression.

4.1 Tests for Covariate Effect

A general test for covariate effect suggested in Kvaløy (1999) can be used to make approximate inference on the effect of each covariate in Cox-models of the form (3). The test is constructed for the alternative of any non-constant covariate function in exponential regression models, and is based on the same idea as the covariate order method for exponential regression. By resampling techniques the test can be generalized to a test for covariate effect for life times having any distribution. A short review of the test is given below.

Recall the definition of the process $S_1, S_2, \ldots$ displayed in Figure 1. If there is no covariate effect, that is if the hazard of the exponential distribution is a constant function of the covariate, $\lambda(x) \equiv \lambda$ for all $x \in \mathcal{X}$, then the process $S_1, S_2, \ldots$ is a homogeneous Poisson process (HPP). If on the other hand there is a covariate effect, which means that the hazard $\lambda(x)$ is a varying function of x, then $S_1, S_2, \ldots$ is no longer a HPP. This observation suggests that statistical tests for the null hypothesis of an HPP versus various non-HPP alternatives can be applied as tests for covariate effect in this situation. In particular there exist in the literature a number of tests for testing HPP versus nonhomogeneous Poisson process (NHPP), see for instance Ascher and Feingold (1984) or Kvaløy and Lindqvist (1998b) and references therein. Moreover, it was observed in Kvaløy and Lindqvist (1998a) that the process $S_1, S_2, \ldots$ can be viewed as being approximately an NHPP if $\lambda(x)$ is reasonably smooth. This observation motivates adopting statistical tests constructed for testing the assumption of a HPP versus the alternative of an NHPP as tests for covariate effect in exponential regression models.

Among the numerous tests constructed for testing HPP versus NHPP, the Anderson-Darling test for trend studied in Kvaløy and Lindqvist (1998b) is recommended here since this test has good power properties both against NHPPs with monotonic and nonmonotonic intensity functions. In the application as a test for covariate effect it is essential that the test is able to detect both monotonic and nonmonotonic effects.

Recall from Section 3.1 the definitions $S = \sum_{i=1}^{n} L_i/n$ and $k(n) = \sum_{i=1}^{n} \delta_i$, and let

$$\hat{k} = \begin{cases} k(n) & \text{if } S_{k(n)} < S \\ k(n) - 1 & \text{if } S_{k(n)} = S \end{cases}$$

Then the test statistic for the Anderson-Darling test for trend is

$$AD = -\frac{1}{\hat{k}} \left[\sum_{i=1}^{\hat{k}} (2i-1)(\ln \frac{S_i}{S} + \ln(1 - \frac{S_{\hat{k}+1-i}}{S})) \right] - \hat{k} \tag{9}$$

The asymptotic null distribution of (9) was derived by Anderson and Darling (1952). On a 5% level the null hypothesis of no covariate effect is rejected if $AD \geq 2.492$. In the rest of the paper, the Anderson-Darling test for trend used as a test for covariate effect as explained above, will be called the Anderson-Darling test for covariate effect.

By adding the extra assumption that the censoring distribution is not depending on the covariate (random censorship), the test can be generalized to allow the life times to have other distributions than the exponential by using bootstrapping or permutation techniques. See Kvaløy (1999) for details. Simulations in Kvaløy (1999) showed that there is very little loss in power by using the bootstrap or permutation versions of the Anderson-Darling test for covariate effect instead of the original test. Because of this it is recommended to use such a resample test unless the assumption of exponentially distributed life times is very indisputable.

Approximate inference on each covariate in the Cox-model can be done by using the Anderson-Darling test for covariate effect in step 5 of Algorithm 2 (step 3 of Algorithm 1) for each j in the Dth iteration. In case of random censorship, a resampled Anderson-Darling test for covariat effect should be used since the transformed observations $L_{1j}, \ldots, L_{nj}$ is only approximately a sample of censored exponentially distributed variables. If there are more than one covariate the test is a test of the effect of each covariate separately and conditional on the other covariates and the covariate functions estimated for these.

An alternative approach for making inference on the effect of, say, the covariate X_m in the Cox-model $\alpha_0(t)\exp(g_1(x_1) + \cdots + g_m(x_m))$ is the following. First remove this covariate from the model and estimate the model $\alpha_0(t)\exp(g_1(x_1) + \cdots g_{m-1}(x_{m-1}))$, and then transform the observation times $T_1, \ldots, T_n$ to $L_1, \ldots, L_n$ where $L_i = \hat{A}_0(T_i)\exp(g_1(X_{i1}) + \cdots + g_{m-1}(X_{i,m-1}))$ and test if there is a covariate effect in the sample $(L_1, \delta_1, X_{1m}), \ldots, (L_n, \delta_n, X_{nm})$ by using the (resampled) Anderson-Darling test for covariate effect.

4.2 Model Selection

It would be very helpful to have an inferential tool to help choose between different models. Testing the effect of covariates included or not included in the model as suggested in the previous section may give some information on which model to chose. There is, however, several reasons why model selection should not be based solely on such tests. For instance may association between covariates cause problems, significance of the model compared to a null model can generally not be checked, and it is not possible to properly compare purely nonparametric models to models with parametric terms in the covariate function. One way of adressing such issues in model selection is by using the partial likelihood as a measure of "goodness of fit". Inference on different models can be based on differences in the partial likelihood. In parametric models this is well established, and inference is based on an asymptotic χ^2-distribution. In nonparametric models however, issues like "number of parameters" fit by a nonparametric fit and corresponding degrees of freedom in the asymptotic χ^2-distribution need to be accounted for. Moreover, there are indications (Tibshirani, 1984) that the χ^2-approximation is unconservative for inference based on the partial likelihood. These problems can be overcome by using the idea in Tibshirani (1984) of simulating the distribution of differences in the likelihood.

The idea is as follows. Let $\alpha_1(t|\mathbf{x}')$ be one Cox-model and let $\alpha_2(t|\mathbf{x}'')$ be a larger model which contains the smaller model, in other words $\mathbf{x}' \subset \mathbf{x}''$. Let $\hat{\alpha}_1(t|\mathbf{x}')$ be an estimate of the small model (which typically can be the null model) and $\hat{\alpha}_2(t|\mathbf{x}'')$ an estimate of the larger model. Then twice the observed difference in the partial log-likelihood is

$$R = 2\{\ln PL(\hat{g}_2(\mathbf{x}'')) - \ln PL(\hat{g}_1(\mathbf{x}'))\}$$

where $\hat{g}_1(\mathbf{x}')$ and $\hat{g}_2(\mathbf{x}'')$ are the estimated covariate functions from respectively the smaller and the larger model. The degrees of freedom of the larger model versus the smaller is defined as the expectation of R under the null hypothesis that the small model is correct. This expectation and in fact the distribution of R can be estimated by simulation as follows. Simulate new failure times from the estimated model $\hat{\alpha}_1(t|\mathbf{x}')$ using the same covariates as originally observed. Also estimate the censoring distribution, for instance by using the Kaplan-Meier estimator or by fitting a Cox-model to $(T_1, 1-\delta_1, \mathbf{X}_1), \ldots, (T_n, 1-\delta_n, \mathbf{X}_n)$. Simulate new censoring times from the estimated censoring distribution. From the new set of observations $(T_1^*, \delta_1^*, \mathbf{X}_1), \ldots, (T_n^*, \delta_n^*, \mathbf{X}_n)$ estimate new models $\hat{\alpha}_1^*(t|\mathbf{x}')$ and $\hat{\alpha}_2^*(t|\mathbf{x}'')$ and calculate

$$R^* = 2\{\ln PL(\hat{g}_2^*(\mathbf{x}'')) - \ln PL(\hat{g}_1^*(\mathbf{x}'))\}$$

Repeat this N times. Exactly the same method and same smoothing parameter must be used for calculating $\hat{\alpha}_1^*(t|\mathbf{x}')$ and $\hat{\alpha}_2^*(t|\mathbf{x}'')$ as was used in the calculation of $\hat{\alpha}_1(t|\mathbf{x}')$ and $\hat{\alpha}_2(t|\mathbf{x}'')$. An estimate of the density of R is found by estimating the density of $R_1^*, \ldots, R_N^*$, and the estimated degrees of freedom used by fitting the larger model is found as the mean of $R_1^*, \ldots, R_N^*$.

This way of doing inference is perhaps most easily justified if the local likelihood method is used instead of the covariate order method in the exponential regression step of the general time transformation algorithm, since the algorithm then has an interpretation as a maximization of the partial likelihood (see Section 5.2). But this method is still a reasonable way to measure the quality of different fits if another method than the local likelihood is used in the exponential regression step since inference is based on the simulated actual distribution rather than on asymptotic approximations.

4.3 Bootstrapping the Model

Bootstrapping can be used to estimate the variability of the estimated covariate functions and the baseline hazard function, and to provide information on whether certain features of the estimated functions are true features of the data or just random noise. Bootstrapping of Cox-models is for example discussed in Tibshirani (1984) and Efron (1988).

One obvious way to bootstrap the model is to resample the observed data triples. That is, draw n new observations $(T_1^*, \delta_1^*, \mathbf{X}_1^*), \ldots, (T_n^*, \delta_n^*, \mathbf{X}_n^*)$ with replacement from the original observations and fit a new model based on this sample.

Another way is based on resampling the estimated generalized residuals $\hat{e}_i = \hat{A}_0(T_i)\exp(\hat{g}(\mathbf{X}_i))$, $i = 1, \ldots, n$, and the idea is to resample with replacement from $(\hat{e}_1, \delta_1), \ldots, (\hat{e}_n, \delta_n)$ and add the new residual sample $(\hat{e}_1^*, \delta_1^*), \ldots, (\hat{e}_n^*, \delta_n^*)$ back to the model. Then the bootstrap sample is $(T_1^*, \delta_1^*, \mathbf{X}_1), \ldots, (T_n^*, \delta_n^*, \mathbf{X}_n)$ where $T_i^* = \hat{A}_0^{-1}(\hat{e}_i^* \exp(-\hat{g}(\mathbf{X}_i)))$. This last way to bootstrap is less robust against departures from the model assumptions.

4.4 Graphical Model Check

The generalized residuals $e_i = A_0(T_i)\exp(g(\mathbf{X}_i))$, $i = 1, \ldots, n$ are a censored sample from the exponential distribution with parameter 1. This implies that if the model assumptions are valid and the fitted model is good, the sample $(\hat{e}_1, \delta_1), \ldots, (\hat{e}_n, \delta_n)$, where $\hat{e}_i = \hat{A}_0(T_i)\exp(\hat{g}(\mathbf{X}_i))$, is approximately a censored sample from the exponential distribution with parameter 1. A simple graphical check of the overall adequacy of the model is then as follows. Estimate the cumulative distribution function, $\hat{F}_e(v)$, from the sample $(\hat{e}_1, \delta_1), \ldots, (\hat{e}_n, \delta_n)$, and plot $\log(1 - \hat{F}_e(v))$ against v. This should yield approximately a straight line with slope -1 if the model is good (see for example Kalbfleisch and Prentice, 1980, p. 96).

When using the covariate order time transformation method it follows naturally to plot the estimated correspondence function, $\tilde{s}(\cdot)$ (see Section 3.1) for each covariate in the exponential regression steps of the final iteration of the algorithm. The correspondence function plot can provide information on different kinds of outliers and how they affect the estimated covariate functions, and at the same time show how the observations are located in the covariate space. For covariate j the correspondence function plot can be used to determine both outliers in that certain covariate, and outliers in the transformed observations $L_{1j}, \ldots, L_{nj}$ (recall step

4 of Algorithm 2) which can be caused by either outliers in time or outliers in one of the other covariates or extreme combinations of observations times and/or covariate values. The information provided by the correspondence function plot can be obtained from other kinds of plot, but the correspondence function plot combines several kinds of information in a useful way. It was proved in Kvaløy and Lindqvist (1998a) that $\tilde{s}(x)$ converges uniformly to a function of the hazard function, the censoring distribution and the covariate distribution so that the correspondence function plot can give information on these. See Kvaløy and Lindqvist (1998a) for further details on interpretation of the correspondence function plot. Estimating the correspondence function is an incorporated part of the covariate order method for exponential regression, but the correspondence function plot will be a helpful graphical tool used in connection with any time transformation method estimating generalized additive models on the form (3).

Also notice that plots commonly used to check the assumption of proportional hazards in the classical Cox-model (for instance the so called "log-minus-log" plot) can be used to check this assumption in the generalized Cox-model as well, see for instance Andersen, Borgan, Gill and Keiding (1993, chap. 7.3.1) for details on such plots in the classical model. The generalization to Cox-models with nonparametric covariate functions is immediate.

5 RELATION TO AND SUMMARY OF OTHER METHODS

As mentioned in the introduction, several authors have studied Cox-models with general covariate functions. Most of these approaches can be divided into two groups, spline based methods and local likelihood based methods. A third method is the local scoring technique used in Hastie and Tibshirani (1986), but there are some problems by using this technique with non-exponential distributions. These problems are overcome by the spline based approach in Hastie and Tibshirani (1990). Spline based methods are also among others studied by O'Sullivan (1988), Sleeper and Harrington (1990), Gray (1992) and Kooperberger et al. (1995a). A nice discussion and overview of different ways of doing spline based modeling of the covariate function in Cox-models is given in Gray (1992). The spline based methods are generally based on the partial likelihood function or a penalized version of the partial likelihood function. One exception is the method in Kooperberger et al. (1995a) which is based on the full likelihood. This method can also estimate more general models than the Cox-model by allowing cross terms between time and the covariates in the hazard function.

The literature on spline based methods for estimation of non-linear covariate functions in Cox-models are in general more complete than for the local likelihood methods. Asymptotic convergence properties are established for many of the methods (see for example O'Sullivan, 1993, Kooperberger, Stone and Truong, 1995b and Gray, 1992), methods for doing inference are developed (all above cited articles) and extensions to time varying covariates can, at least in principle, be done (Hastie and Tibshirani, 1990). Moreover, interactions between covariate terms can be modeled (Gray, 1992, Kooperberger et al., 1995a) and models with combinations of discrete and continuous covariates can easily be handled (for example Sleeper and Harrington, 1990, Gray, 1992, Kooperberger et al., 1995a).

In the present paper, however, attention is restricted to nonparametric methods. The rest of this section is devoted to discussing properties of and relations between the local likelihood methods and the covariate order time transformation method.

5.1 Local Likelihood Methods for Nonparametric Cox-regression

The local likelihood approaches can be divided into two subclasses, local full likelihood methods as studied in Staniswalis (1989), Gentleman and Crowley (1991) and Fan et al. (1997), and local partial likelihood methods as studied in Tibshirani (1984), Tibshirani and Hastie (1987) and Fan et al. (1997).

The local partial likelihood method is based on using localized versions of the logarithm of the partial likelihood (2). For the case of one covariate, $g(x)$ is approximated by a polynomial $p(u) = \beta_0 + \beta_1(u-x) + \cdots + \beta_p(u-x)^p$. Then maximizing the local partial log-likelihood

$$lpl(\beta_0, \ldots, \beta_p, x) = \sum_{i=1}^{n} \delta_i \left[p(X_i) - \log\Big(\sum_{j \in R_i} \exp(p(X_i)) \frac{1}{h} K\big(\frac{x - X_i}{h}\big)\Big) \right] \frac{1}{h} K\big(\frac{x - X_i}{h}\big) \qquad (10)$$

leads to the estimate $\hat{\beta}_0$ for $g(x)$. It is easily realized that the constant term β_0 vanishes in (10), implying that only an estimate of the derivative curve is found in the first place. An estimate of $g(x)$ can then be found by numerical integration. Tibshirani and Hastie (1987) suggested using the trapezoidal rule for numerical integration. The method can be generalized to $m > 1$ covariates by considering polynomials in m variables and using m-dimensional kernels.

The local full likelihood method is based on using localized versions of the full likelihood. Again, for one covariate, approximating $g(x)$ by the polynomial $p(u)$, the function to be maximized is

$$lfl(\beta_0, \ldots, \beta_p, x) = \sum_{i=1}^{n} [\delta_i \{\ln \alpha_0(T_i) + p(X_i)\} - A_0(T_i) \exp(p(X_i))] \frac{1}{h} K\big(\frac{x - X_i}{h}\big) \qquad (11)$$

If the baseline hazard, $\alpha_0(T_i)$, is known, a direct estimate of $g(x)$ is found by optimizing (11). In practice, however, the baseline hazard is unknown and an iterative algorithm is invoked. This algorithm starts with an initial estimate and then iterates back and forth between estimation of the covariate function and the baseline hazard. The baseline hazard is estimated by using for instance the Breslow estimator (Gentleman and Crowley, 1991), a spline approximation (Staniswalis, 1989) or a parametric model (Fan et al., 1997). Generalizations to $m > 1$ covariates can be done by considering polynomials in m variables and using m-dimensional kernels. Staniswalis (1989) does this for $m = 2$ and a local constant approximation to the covariate function. Otherwise, for simulations and applications to real data, only one covariate or an additive model is considered in the other cited papers. In all these examples of practical use, either with the local full likelihood method or the local partial likelihood method, a local linear approximation to the covariate function is used.

It is easily realized that the local full likelihood method is a time transformation method. Consider the case with one covariate, and assume that the local likelihood method for exponential regression (recall Section 3.2) is used instead of the covariate order method in Algorithm 1. Then the exponential regression step of the algorithm is based on the local log-likelihood

$$l(\beta_0, \ldots, \beta_p, x) = \sum_{i=1}^{n} \left[\delta_i p(X_i) - \hat{A}_0(T_i) \exp(p(X_i)) \right] \frac{1}{h} K\big(\frac{x - X_i}{h}\big)$$

Hence, this algorithm will give exactly the same estimate of $g(x)$ as estimation based on (11) if the same estimate of $A_0(t)$ is used. This shows that the local full likelihood method is a time transformation method for the case with one covariate, and the same is true for the generalization to more than one covariate.

The covariate order method for exponential regression can be generalized to estimating general smooth functions in several variables (Kvaløy and Lindqvist, 1998a). This opens for similar non-additive versions of the covariate order time transformation algorithm in cases with several covariates as the generalizations indicated for the local likelihood methods. Such generalizations may, however, be less useful in practice due to the curse of dimensionality. In most cases it is a better idea to fit a generalized additive model of the form (3). If a local likelihood method is used in the exponential regression step of Algorithm 2 this is similar to the iterative algorithms suggested in Gentleman and Crowley (1991) and Fan et al. (1997) for estimation of additive models. They do, however, not discuss how to estimate models with a combination of smooth nonparametric and parametric (for example for discrete or categorical covariates) covariate functions.

5.2 Speed of Convergence

The local full likelihood algorithm can be viewed as an EM-algorithm, and Clayton and Cuzick (1985) is proposing the same kind of algorithm for estimation of $\alpha_0(t)$ and $\boldsymbol{\beta}$ in the classical Cox-model. They show that this in fact leads to the usual maximum partial likelihood estimate of $\boldsymbol{\beta}$ and the Breslow estimator of the baseline hazard. Hastie and Tibshirani (1990) tried to generalize the algorithm of Clayton and Cuzick (1985) to estimate spline based models for the covariate function using penalized full likelihood, but reported very slow convergence. Gentleman and Crowley (1991) also commented the slow convergence of the Clayton and Cuzick (1985) algorithm, and developed methods for speeding up the algorithm by introducing an artificial constant. The idea is to introduce some overlap between the two parts of the estimation, and that the convergence will be more rapid when there are more parameters to maximize over in the likelihood. This modification is used in their local full likelihood method.

In the time transformation methods implemented in this paper, however, speed of convergence has not appeared to be a problem, neither in the covariate order time transformation method nor the local full likelihood method. All practical experience has been that the convergence is very fast. The local full likelihood method has been implemented by using Algorithm 1 or 2 with the covariate order method replaced by the local likelihood method in the exponential regression step of the algorithm. Thus the rapid convergence could for both methods be explained by the fact that a new Cox-model is fit instead of just directly estimating the baseline hazard in the step of the algorithm where a new baseline hazard is estimated (see discussion following Algorithm 1 in Section 2). The parameter β in this Cox-model then seems to play the same role of speeding up the algorithm as the artificial parameter in the method of Gentleman and Crowley (1991).

However, for comparison, Algorithm 1 has also been implemented with direct application of the Breslow estimator in step 4, and this did not slow down the speed of convergence of the algorithm considerably. Typically 1-2 extra iterations was needed to reach the same accuracy of the estimate, but this number was varying from situation to situation. In some cases more than two extra iterations were needed and in some cases no extra iterations. Typically one more extra iteration was needed if the Breslow estimator was also used in the initial step instead of fitting an ordinary Cox-model. For analyzing real data the difference between having to run 4-7 iterations instead of 3-4 is of no practical importance. For simulations, however, it is useful to have as fast an algorithm as possible. If the additional cost of having to estimate a Cox-model instead of directly applying the Breslow estimator is small, then the suggested version of Algorithm 1 is preferable. For cases with more than one covariate the difference in number of iterations needed to reach convergence will probably be larger, and fitting a new Cox-model instead of directly applying the Breslow estimator also has the added advantage discussed in Section 2 of easy inclusion of parametric terms and easy implementation.

5.3 The Time Transformation Methods Versus the Local Partial Likelihood Method.

Gentleman and Crowley (1991) point out some benefits with the local full likelihood approach over the local partial likelihood approach which are also shared by the covariate order time transformation method. One of the benefits with the local full likelihood method is that a direct estimate of $g(\mathbf{x})$ is found. For the local partial likelihood method numerical integration of the estimated derivative curve is needed to estimate $g(\mathbf{x})$. This is a form of smoothing that needs to be accounted for. Another benefit with the local full likelihood method is that one component of the model is estimated using all of the data. This should both imply that the estimated covariate functions should be less sensitive to endpoint bias, and that the method should still perform well on heavily censored data. The covariate order time transformation method should be even better with respect to endpoint bias since the main difference between the covariate order method and the local likelihood method for exponential regression is better behavior at endpoints (Kvaløy and Lindqvist, 1998a). A benefit with the local partial likelihood method is of course that it is not iterative.

5.4 Inference

The general time transformation algorithm is a very general algorithm. Any method for exponential regression can be used in the estimation of the covariate functions. Moreover, using the (resampled) Anderson-Darling test for covariate effect for doing approximate inference on effect of each covariate follows naturally from this way of formulating the problem. This test can, however, also be used in all the other methods that estimate an additive model (including the spline based methods). Separate conditional inference for the jth covariate can be done by using the Anderson-Darling test for covariate effect on the sample $(L_{1j}, \delta_1, X_{1j}), \ldots, (L_{nj}, \delta_n, X_{nj})$ where $L_{ij} = \hat{A}_0(T_i)\exp(\hat{g}_1(X_{i1}) + \cdots + \hat{g}_{j-1}(X_{i,j-1}) + \hat{g}_{j+1}(X_{i,j+1}) + \cdots + \hat{g}_n(X_{in}))$. This implies that the baseline hazard function must be estimated. Estimation of the baseline hazard is not an issue in the papers that use the partial likelihood, but the baseline hazard can for example be estimated by using the Breslow estimator when an estimate of the covariate function is found. Inference about the estimated covariate function based on the partial likelihood as explained in Section 4 can also be used in all the above mentioned methods.

Gentleman and Crowley (1991) are using an approach similar to the likelihood based inference of Tibshirani (1984) presented in Section 4.2 to test the significance of the covariate in cases with only one covariate. They assume that the censoring is independent of the covariate and suggest a permutation test based on the difference in the likelihood between the estimated model and the null model. Under the null hypothesis of no covariate effect, the covariates are exchangeable, and the distribution of the difference in likelihood can be estimated by calculating the difference in the likelihood for a large number of permutations of the covariates. Again, this way of doing inference can be adapted in any of the methods, but is less general than the approach presented in Section 4.2. For the local likelihood method, asymptotic inference can also in principle be based on the asymptotic distributional results in Fan et al. (1997). For instance asymptotic point-wise confidence intervals for $g(x)$ (local full likelihood) or $g'(x)$ (local partial likelihood) can in principle be derived if reasonable estimates of the unknown quantities in the asymptotic distribution can be found.

5.5 Other Issues

An important problem in Cox-regression is handling of ties. Ties in the covariates is in general no problem, but ties in the observation times requires special care in the methods based on the partial likelihood. There exist, however, several ways of incorporating tied values in the partial likelihood (see for instance Kalbfleisch and Prentice, 1980, chap. 4.2) which solve the possible problem of tied values for all methods discussed in this paper.

An issue not mentioned in any of the cited articles is the possibility of fitting stratified models, but is seems like all methods can be generalized to allow stratified models with minor modifications in the likelihood or the partial likelihood.

Another issue not broadly covered is the possibility of allowing time dependent covariates. Some time transformation methods can in principle allow time dependent covariates, but the covariate order time transformation algorithm does not allow time dependent covariates. Tibshirani (1984) mention that the local partial likelihood method in principle can be used to estimate models with time dependent covariates, but that substantial computational difficulties may arise.

There are some differences in how sensitive the different methods are to outliers in time or covariate values. Outliers in observation time will affect the estimation of the covariate function worse in time transformation methods than in methods based on the partial likelihood where only the information of the ordering of the observation times is used. Outliers in covariate value is a problem in all methods. Ideas like in Tibshirani (1984) of weighting down observations that are far away from the center of the data may be used to reduce the effect of such outliers.

Different approaches for choosing the smoothing parameter are taken in the discussed papers. Gentleman and Crowley (1991) simply recommend to use several spans and compare the results visually. Fan et al. (1997) calculate the theoretical optimal bandwidth that minimizes the asymptotic mean integrated squared error, but this expression contains a lot of unknown quantities and is thus not directly applicable for practical use. Staniswalis (1989) propose to use likelihood cross-validation and Tibshirani (1984) and Tibshirani and Hastie (1987) is using a rule of thumb based on Akaike's information criterion. Likelihood cross-validation is criticized in the literature for sometimes having undesirable practical behavior, but have shown satisfying performance in all applications both in this paper and in Kvaløy and Lindqvist (1998a).

Finally some comments on asymptotic properties. In most of the papers that take the local likelihood approach (and the present paper) the authors conjecture, without giving a proof, that the estimated hazard function will be close to the true hazard function when the sample size increases. Formal proof of such a claim is, however, given in Fan et al. (1997). In the case of a parametric model for the baseline hazard they prove that the local full likelihood approach, under certain conditions, lead to point-wise consistent estimators of the covariate function and the baseline hazard. Convergence in distributions is also established. In the case of an unknown baseline hazard they prove point-wise weak convergence of the derivative curve of the covariate function estimated using local partial likelihood.

6 SIMULATION STUDY

The properties of the covariate order time transformation method (Algorithm 1), the local full likelihood and the local partial likelihood method for Cox-regression are studied by some simulated examples. All simulations are done i S-PLUS with parts of the code written in C.

The simulations focus on the primary goal to investigate the methods' ability to estimate the covariate function, $g(x)$, in the model (1). Due to the identifiability problem mentioned in Section 2, the metric used to assess and compare estimates must be invariant with respect to an additive constant in the estimated covariate function, $\hat{g}(x)$. The following metric is used. Calculate $d_i = g(X_i) - \hat{g}(X_i)$ for all simulated covariate values $X_1, \ldots, X_n$ and calculate $D = \sqrt{\sum_{i=1}^{n}(d_i - \bar{d}_i)^2/n}$. This metric is calculated for $N = 400$ different data sets and $\bar{D} = \sum_{j=1}^{N} D_j/N$ and the standard deviation of $\bar{D}$ is reported.

For the local likelihood methods, local linear modeling is used. In the local partial likelihood method, the trapezoidal rule for numerical integration is used to estimate the covariate function from the estimated derivative curve. The local full likelihood method is implemented by using the local likelihood method for exponential regression rather than the covariate order method in step 3 of Algorithm 1.

For all methods the biweight kernel

$$K(u) = \begin{cases} \frac{15}{16}(1-u^2)^2 & ,|u| \le 1 \\ 0 & ,\text{else} \end{cases}$$

is used. The abbreviations COTT for the covariate order time transformation method, LFL for the local full likelihood method and LPL for the local partial likelihood method are used in the tables reporting the simulation results.

6.1 Case 1

We first consider a model used in Gentleman and Crowley (1991) and Fan et al. (1997) to compare the local linear partial likelihood method to the local linear full likelihood method. Gentleman and Crowley (1991) and Fan et al. (1997) reach somewhat different conclusions. Gentleman and Crowley (1991) reports little difference between the local partial and local full likelihood method, while the simulations in Fan et al. (1997) are modestly in favor of the local partial likelihood method.

The model is $\alpha_0(t) = 0.007$, $g(x) = 0.54 + I(x \le 40)(x - 40)/30$ and X is sampled from a uniform(10,70) distribution. Data sets of size 210 are generated. The censoring variable C

is drawn independently of Z and X from respectively a uniform(0,400) and a uniform(0,200) distribution, corresponding to respectively on average 17% and 32% censoring. The results obtained using values of the smoothing parameter similar to the three span values used in Gentleman and Crowley (1991) and two larger span values are presented in Table 1.

Censoring	Model	$h = 15$	$h = 20$	$h = 25$	$h = 30$	$h = 35$
No	COTT	.118 (.0022)	.106 (.0020)	.113 (.0021)	.116 (.0018)	.125 (.0016)
	LFL	.129 (.0023)	.111 (.0021)	.108 (.0023)	.100 (.0021)	.100 (.0019)
	LPL	.135 (.0024)	.111 (.0021)	.108 (.0023)	.098 (.0021)	.097 (.0019)
U(0,400)	COTT	.131 (.0024)	.111 (.0021)	.111 (.0022)	.114 (.0019)	.124 (.0019)
(17%)	LFL	.143 (.0026)	.117 (.0023)	.107 (.0022)	.102 (.0021)	.103 (.0022)
	LPL	.149 (.0027)	.119 (.0024)	.107 (.0023)	.099 (.0021)	.102 (.0023)
U(0,200)	COTT	.139 (.0026)	.123 (.0024)	.118 (.0023)	.121 (.0022)	.129 (.0020)
(32%)	LFL	.151 (.0028)	.133 (.0026)	.121 (.0025)	.115 (.0024)	.113 (.0023)
	LPL	.157 (.0029)	.135 (.0027)	.121 (.0025)	.114 (.0024)	.113 (.0025)

Table 1: Comparison of the three methods for estimation of $g(x)$ for $\alpha_0(t) = 0.007$, $g(x) = 0.54 + I(x \leq 40)(x - 40)/30$, $X \sim \text{Uniform}(10, 70)$ and samples of size 210. Average distance measure, $\bar{D}$, based on 400 repetitions is reported. Standard errors are given in parentheses. Average censoring percentage is also reported.

Table 1 basically shows that there are no large differences between the methods. The table indicates that the covariate order time transformation method is slightly better than the other methods for small smoothing parameters, while the other methods are a bit better for larger smoothing parameters. The local likelihood methods achieve a somewhat smaller best value of $\bar{D}$, while the standard errors in all cases are smallest for the covariate order time transformation method. There is, as in the same simulation in Gentleman and Crowley (1991), hardly no difference in the performance of the local full versus the local partial likelihood method. There seems to be no difference between the methods with regard to censoring.

The results in Table 1 was obtained by assuming the simulated data to be data from a Cox-model. The effect of having to deal with an unknown baseline hazard in the estimation of the covariate function in Cox-models can be illustrated by now using the fact that we know that the baseline hazard is constant, and estimating $g(x)$ directly by using the methods of Section 3 for exponential regression rather than Cox-regression methods. The results obtained by using the covariate order and the local likelihood method for exponential regression are summarized in Table 2. Comparing the results in Table 2 to the corresponding results in Table 1 we see that there are in fact no practical difference in the estimates of $g(x)$ obtained when knowing and not knowing that the baseline hazard is constant. In the next section the effect of different forms of the baseline hazard is studied.

Notice that the estimates of the covariate function obtained when using the methods for exponential regression for comparison are assessed by the same metric as the estimates obtained when using the Cox-regression methods. A probably more relevant quantity for assessing exponential regression estimates would, since there are no identifiability problem in exponential regression, be mean integrated squared error.

Censoring	Model	$h = 15$	$h = 20$	$h = 25$	$h = 30$	$h = 35$
No censoring	CO	.116 (.0020)	.104 (.0021)	.105 (.0019)	.110 (.0018)	.121 (.0016)
	LL	.126 (.0021)	.108 (.0021)	.100 (.0019)	.092 (.0019)	.094 (.0019)
Uniform(0,400)	CO	.125 (.0022)	.113 (.0022)	.110 (.0022)	.113 (.0019)	.124 (.0018)
	LL	.137 (.0023)	.119 (.0022)	.109 (.0022)	.099 (.0020)	.103 (.0021)
Uniform(0,200)	CO	.137 (.0026)	.122 (.0025)	.115 (.0022)	.118 (.0020)	.127 (.0020)
	LL	.150 (.0027)	.131 (.0025)	.117 (.0023)	.109 (.0022)	.111 (.0025)

Table 2: Estimating $g(x)$ using methods for exponential regression in the model $\alpha_0(t) = 0.007$, $g(x) = 0.54 + I(x \leq 40)(x - 40)/30$, $X \sim$ Uniform$(10, 70)$ and samples of size 210. The abbreviations CO for covariate order and LL for local likelihood are used. Average distance measure, $\bar{D}$, based on 400 repetitions is reported. Standard errors are given in parentheses.

6.2 Case 2

The second model considered is also a model used by Gentleman and Crowley (1991) to compare the local partial and the local full likelihood methods. Their conclusion in this case is that the local full likelihood is doing better for small spans, while there is little difference for longer spans.

The model is $\alpha_0(t) = 0.0036t^{0.5}$, $g(x) = (x-40)^2/900$ and X is sampled from a uniform(10,70) distribution. This is a slight modification of the model in Gentleman and Crowley (1991) where a constant baseline hazard equal to 0.012 was used. As in the previous example the censoring variable C is drawn independently of Z and X from respectively a uniform(0,400) and a uniform(0,200) distribution, again, by the choice of the constant in the baseline hazard, corresponding to on average 17% and 32% censoring. In addition a case with C drawn from the uniform(0,50), corresponding to approximately 80% censoring is included. The results obtained are presented in Table 3.

Again the covariate order time transformation method seems to be slightly better for the smallest smoothing parameter and a bit worse for the large smoothing parameters, and is, with one single exception, again the method with smallest standard errors. There is also an indication that the local full likelihood is handling heavy censoring worse than the other methods. A slight indication that the covariate order time transformation method is performing relatively better for heavily censored data may also be seen. In their similar example Gentleman and Crowley (1991) reported the local full likelihood method to be better than the local partial likelihood method for the smallest span values. In the present simulations there seems, in the three less heavily censored cases, to be hardly no difference between the two methods for the three smallest smoothing parameters (corresponding to the span values used in Gentleman and Crowley (1991)), but the local full likelihood method is a bit better for the largest smoothing parameter.

To study the effect of different baseline hazard functions on the estimated covariate function, the simulations in Table 3 with on average 32% censoring have been repeated with the alternative baseline hazard $\alpha_0(t) = 9t^{-0.5}$. The results obtained using this model are summarized in Table 4. As we see, there are no notable differences between the results in Table 4 and the corresponding results in Table 3. This illustrates that the form of the baseline hazard has marginal influence on the estimation of the covariate function. The simulations in Section 6.1

Censoring	Model	$h = 15$	$h = 20$	$h = 25$	$h = 30$	$h = 35$
No	COTT	.122 (.0022)	.112 (.0021)	.116 (.0023)	.131 (.0023)	.149 (.0020)
	LFL	.126 (.0023)	.109 (.0022)	.106 (.0024)	.113 (.0025)	.126 (.0023)
	LPL	.132 (.0024)	.113 (.0022)	.111 (.0024)	.117 (.0025)	.134 (.0022)
U(0,400)	COTT	.127 (.0022)	.124 (.0024)	.120 (.0023)	.136 (.0022)	.153 (.0023)
(17%)	LFL	.134 (.0024)	.125 (.0026)	.112 (.0026)	.118 (.0025)	.128 (.0026)
	LPL	.137 (.0024)	.125 (.0025)	.114 (.0025)	.121 (.0024)	.138 (.0024)
U(0,200)	COTT	.141 (.0025)	.130 (.0024)	.131 (.0024)	.137 (.0026)	.158 (.0024)
(32%)	LFL	.153 (.0028)	.136 (.0027)	.127 (.0027)	.126 (.0029)	.135 (.0027)
	LPL	.154 (.0027)	.133 (.0026)	.124 (.0026)	.127 (.0028)	.144 (.0026)
U(0,50)	COTT	.252 (.0050)	.232 (.0049)	.218 (.0044)	.195 (.0039)	.208 (.0039)
(80%)	LFL	.283 (.0055)	.259 (.0055)	.243 (.0051)	.204 (.0047)	.208 (.0049)
	LPL	.273 (.0051)	.242 (.0048)	.220 (.0045)	.192 (.0044)	.205 (.0044)

Table 3: Comparison of the three methods for estimation of $g(x)$ for $\alpha_0(t) = 0.036t^{0.5}$, $g(x) = (x-40)^2/900$, $X \sim \text{Uniform}(10,70)$ and samples of size 210. Average distance measure, $\bar{D}$, based on 400 repetitions is reported. Standard errors are given in parentheses. Average censoring percentage is also reported.

Censoring	Model	$h = 15$	$h = 20$	$h = 25$	$h = 30$	$h = 35$
U(0,0.02)	COTT	.143 (.0024)	.134 (.0024)	.135 (.0026)	.141 (.0024)	.157 (.0023)
(32%)	LFL	.154 (.0026)	.140 (.0027)	.132 (.0028)	.127 (.0026)	.134 (.0027)
	LPL	.151 (.0026)	.134 (.0025)	.128 (.0027)	.131 (.0026)	.145 (.0026)

Table 4: Comparison of the three methods for estimation of $g(x)$ for $\alpha_0(t) = 9t^{-0.5}$, $g(x) = (x-40)^2/900$, $X \sim \text{Uniform}(10,70)$ and samples of size 210. Average distance measure, $\bar{D}$, based on 400 repetitions is reported. Standard errors are given in parentheses.

indicated no practical difference between the estimated covariate functions when knowing (exponential regression) and not knowing (Cox-regression) that the baseline hazard is constant. The conclusion seems to be that the fact that we have to deal with an unknown baseline hazard in the Cox-model, and the actual shape of this unknown baseline hazard, have little effect on the estimation of the covariate function.

6.3 Case 3

This final simulated example is a modification of a simulation model used in Staniswalis (1989). The model is $\alpha_0(t) = 0.5t^{-0.5}$, $g(x) = \sin 2\pi x$ and X is sampled from a uniform(0,1) distribution. The censoring variable C is drawn independently of Z and X from the density $1.25t^{0.25}/V^{1.25}$, $0 \le t \le V$, where V is chosen according to the degree of censoring wanted. In Staniswalis (1989) a constant baseline hazard and 10% censoring was used.

Estimated covariate functions and cumulative baseline hazard functions using each of the methods with $h = 0.2$ on 10 simulated data sets of size 200 with approximately 40% censoring ($V = 2$) are shown in Figure 3. We see that the estimated covariate functions capture the

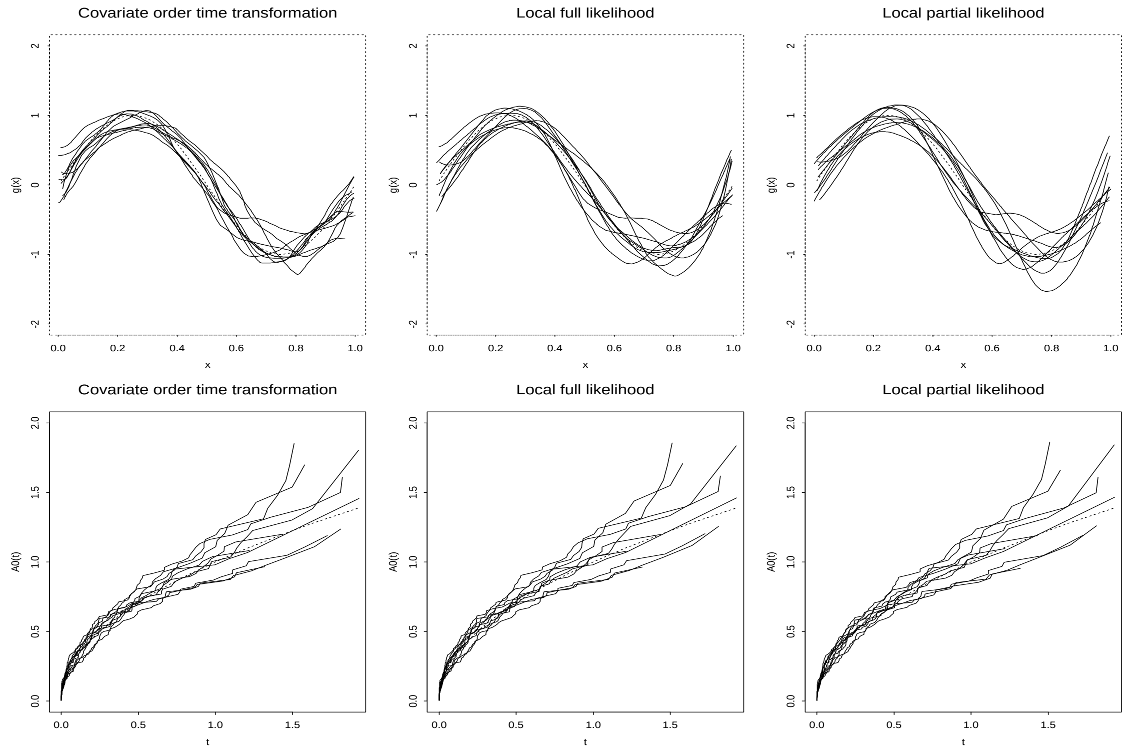

Figure 3: Estimated covariate functions and cumulative baseline hazard functions for ten data sets of size 200 and 40% censoring using each of the methods and $h = 0.2$. The true curves are dashed.

structure of the true curve fairly well, and the estimates of the cumulative baseline hazard functions are also reasonable. For all methods the cumulative baseline hazard function is estimated using the Breslow estimator, and we see that there is very little difference between the estimates produced by each method. There are a bit more differences in the estimated covariate functions. Repeating these simulations 400 times for different smoothing parameters and varying degree of censoring gave the results summarized in Table 5. We see that the covariate order time transformation method again seems to be doing relatively better as the censoring increases. In the most heavily censored case the covariate order time transformation method is in fact at least as good or better than the other methods for all smoothing parameters considered. In the uncensored case on the other hand the covariate order time transformation method is doing clearly worse than the other methods for the three largest smoothing parameters, but the difference in the smallest $\bar{D}$-values achieved by the three methods is not big. The tendency seen in Table 3 that the local full likelihood method tend to be doing worse compared to the other methods in the heavily censored case is not seen in the simulations reported in Table 5. The covariate order time transformation method is again the least variable method.

It is also interesting to see how the methods perform for varying number of observations. In Table 6 some simulations with on average 40% censoring and respectively 400, 100 and 50 observations are reported. The relative performance of the different methods for varying number of observations reported in Table 6 roughly seems to be the same as for varying degree

Censoring	Model	$h = 0.15$	$h = 0.20$	$h = 0.25$	$h = 0.30$	$h = 0.35$
No	COTT	.169 (.0024)	.177 (.0023)	.206 (.0023)	.245 (.0022)	.253 (.0023)
	LFL	.176 (.0027)	.157 (.0024)	.161 (.0025)	.183 (.0026)	.190 (.0027)
	LPL	.208 (.0034)	.176 (.0028)	.169 (.0028)	.184 (.0029)	.190 (.0029)
$V = 2$	COTT	.209 (.0032)	.194 (.0030)	.206 (.0028)	.232 (.0029)	.270 (.0027)
(40%)	LFL	.224 (.0036)	.200 (.0032)	.195 (.0031)	.210 (.0033)	.232 (.0031)
	LPL	.242 (.0035)	.214 (.0035)	.200 (.0031)	.205 (.0033)	.223 (.0031)
$V = 0.05$	COTT	.407 (.0062)	.350 (.0061)	.339 (.0066)	.313 (.0058)	.331 (.0059)
(82%)	LFL	.500 (.0112)	.407 (.0088)	.386 (.0092)	.331 (.0064)	.333 (.0066)
	LPL	.503 (.0091)	.408 (.0072)	.394 (.0086)	.342 (.0078)	.347 (.0074)

Table 5: Comparison of the three methods for estimation of $g(x)$ for $\alpha_0(t) = 0.5t^{-0.5}$, $g(x) = \sin 2\pi x$, $X \sim \text{Uniform}(0,1)$ and samples of size 200. The parameter in the censoring distribution and average censoring percentage is reported. Average distance measure, $\bar{D}$, based on 400 repetitions is reported. Standard errors are given in parentheses.

of censoring reported in Table 5. There seems though to be a tendency that the local partial likelihood method is doing worse than the local full likelihood method for decreasing number of observations.

Finally an example with stratification is given. For the time transformation methods the modification needed to allow stratification, mentioned after the general time transformation algoritm in Section 2, is used. That is, for the covariate order time transformation method Algorithm 1 is used with each $\hat{A}_0(T_i)$ replaced by the $\hat{A}_{0q}(T_i)$, $q \in \{1, \ldots, s\}$ according to the strata which observation i belongs to. A similar modification is used for the local full likelihood method. For the local partial likelihood method the standard modification of the partial likelihood needed to allow stratified models, see for example Kalbfleisch and Prentice (1980, p. 88), was incorporated into the local partial likelihood estimate (10). The modification is simply to split the likelihood in one part for each stratum.

The model considered is $\alpha_{01}(t) = 2t$ and $\alpha_{02}(t) = 0.5t^{-0.5}$, with approximately 50% of the observations in each stratum. Ten datasets of size 200 with approximately 40% censoring have been generated. The estimated covariate functions, using $h = 0.2$, and the estimated baseline hazard functions for each stratum are, for the covariate order time transformation method and the local partial likelihood method, displayed in Figure 4. We see for both methods that the estimated curves are capturing the shape of the true curves quite satisfactorily. The local partial likelihood method seems to be a bit more variable in estimating the covariate function, and the estimated cumulative baseline hazard curves are, not surprisingly, a bit more variable compared to the case in Figure 3 with one common baseline hazard. Repeating the above simulations 400 times using different smoothing parameters gave the results displayed in Table 7.

Comparing the results in Table 7 to the results for the similar case with no stratification in Table 5 shows that the methods in general are able to estimate the covariate function equally good in the case with stratification. The local partial likelihood method is doing a bit worse for the smallest smoothing parameter, and is a bit more variable, otherwise the differences are negligible. The tendency that the local partial likelihood method is a bit more

n	Model	$h = 0.15$	$h = 0.20$	$h = 0.25$	$h = 0.30$	$h = 0.35$
400	COTT	.153 (.0021)	.150 (.0021)	.178 (.0022)	.209 (.0020)	.251 (.0019)
	LFL	.158 (.0023)	.144 (.0022)	.158 (.0023)	.176 (.0023)	.209 (.0023)
	LPL	.168 (.0023)	.149 (.0022)	.152 (.0023)	.163 (.0023)	.192 (.0022)
100	COTT	.296 (.0047)	.260 (.0044)	.249 (.0041)	.267 (.0041)	.291 (.0041)
	LFL	.340 (.0061)	.280 (.0047)	.253 (.0045)	.255 (.0044)	.263 (.0046)
	LPL	.379 (.0072)	.301 (.0052)	.275 (.0048)	.260 (.0045)	.266 (.0048)
50	COTT	.415 (.0072)	.359 (.0061)	.340 (.0071)	.328 (.0065)	.339 (.0053)
	LFL	.518 (.0092)	.416 (.0079)	.381 (.0075)	.347 (.0071)	.340 (.0068)
	LPL	.722 (.0320)	.479 (.0126)	.418 (.0093)	.376 (.0077)	.356 (.0072)

Table 6: Comparison of the three methods for estimation of $g(x)$ for $\alpha_0(t) = 0.5t^{-0.5}$, $g(x) = \sin 2\pi x$, $X \sim \text{Uniform}(0,1)$ and different sample sizes. 40% censoring ($V = 2$). Average distance measure, $\bar{D}$, based on 400 repetitions is reported. Standard errors are given in parentheses.

Censoring	Model	$h = 0.15$	$h = 0.20$	$h = 0.25$	$h = 0.30$	$h = 0.35$
$V = 2$	COTT	.212 (.0029)	.192 (.0028)	.211 (.0029)	.242 (.0028)	.276 (.0026)
(39%)	LFL	.225 (.0031)	.194 (.0030)	.197 (.0032)	.215 (.0031)	.240 (.0031)
	LPL	.278 (.0048)	.221 (.0036)	.207 (.0038)	.211 (.0035)	.224 (.0033)

Table 7: Comparison of the three methods for estimation of $g(x)$ for $\alpha_{01}(t) = 0.5t^{-0.5}$, $\alpha_{02}(t) = 2t$ and $g(x) = \sin 2\pi x$, $X \sim \text{Uniform}(0,1)$ and samples of sizes 200. On average 50% of the observations in each stratum. Average distance measure, $\bar{D}$, based on 400 repetitions is reported. Standard errors are given in parentheses.

variable especially for small smoothing parameters in cases with stratification compared to the similar case with no stratification has also been seen in other simulations not reported. A possible explanation is that the variability of the estimate increases since the likelihood is split in two parts and then each neighborhood in each part of the likelihood will contain fewer observations.

6.4 Comments

A main conclusion of the above simulation study is that the differences between the methods are for practical purposes in general not very large. If we look at the finer details of the simulations we notice that the covariate order time transformation method is generally the least variable of the methods, and this method also seems to have preferable properties in "difficult" cases with few observations or heavy censoring. The covariate order time transformation method is in general better, compared to the other methods, for small smoothing parameters than for large smoothing parameters. By plotting estimated curves as in Figure 3 in a number of different situations and with different smoothing parameters it can be seen that the covariate order time transformation method in general is less variable and is doing better at the endpoints than the other methods, but is more biased for large smoothing parameters. As for the local partial versus the local full likelihood method, there are more cases where

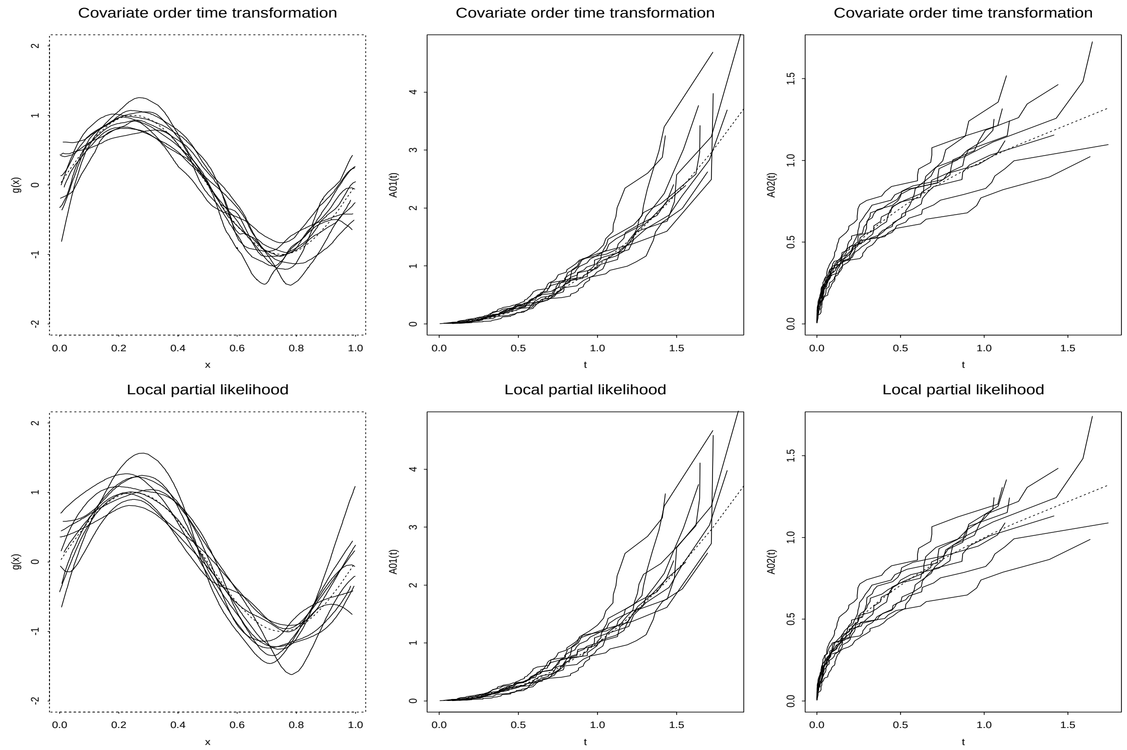

Figure 4: Estimated covariate functions and cumulative baseline hazard functions for ten data sets of size 200 with 40% censoring and two strata using the covariate order time transformation and the local partial likelihood method with $h = 0.2$. The true curves are dashed.

the local full likelihood method is better than the local partial likelihood method than the opposite way. This is in accordance with the simulation results in Gentleman and Crowley (1991), but somewhat in conflict with the results in Fan et al. (1997). The differences are, however, not big or systematic enough to conclude that the local full likelihood method is generally preferable over the local partial likelihood method.

The motivation for the covariate order time transformation method was that the covariate order method for exponential regression showed better performance than competing non-parametric methods for exponential regression , including the local likelihood method, in a simulation study in Kvaløy and Lindqvist (1998a). This does, however, not seem to fully carry over to better performance in Cox-regression since the local full likelihood method in many cases give equally good or even better results than the covariate order time transformation method. A possible explanation is that the estimated curves in the Cox-regression simulations are assessed by a metric which is invariant with respect to an additive constant, while mean integrated squared error (with no adjustment for different means) was used to assess the curves in the exponential regression simulations.

The local partial likelihood method has some possible drawbacks mentioned in Section 5.3, but seems in practice in many cases to produce equally good estimates as the other methods. This is somewhat in accordance with the theoretical result in Fan et al. (1997) that in certain cases and under some conditions the local partial likelihood and the local full likelihood with a correctly parameterized baseline hazard has the same asymptotic bias and variance for estimation of the derivative of the covariate function. The local partial likelihood method also has the benefit of being non-iterative. On the other hand it is easier to implement the iterative methods in a statistical software package that offers Cox-regression functions, as then only the covariate order or the local likelihood method for exponential regression has to be coded. Especially the implementation of the iterative methods is more easy if models with several covariates, possibly including some categorical covariates, are to be fitted.

It has been demonstrated in different ways that the baseline hazard has little influence on the estimation of the covariate function. This is not surprising when seen in relation to results for the classical Cox-model, see for instance Kalbfleisch and Prentice (1980, chap. 4.7), that knowledge of the baseline hazard is of little importance for estimation of the regression coefficients $\boldsymbol{\beta}$.

In practical use of the local likelihood methods, especially the local full likelihood method, a problem of numerical instability has been encountered. Estimating the linear term β_1 in (10) or (11) requires a numerical solution. In the implementation of the local likelihood method used for the simulations, a simple Newton-Raphson method was used to solve this problem. This is generally a fast and stable method, but in some cases, especially in cases with few observations or heavy censoring, the Newton-Raphson algorithm fails to converge even if different choices of starting-points is tried. This is a problem most frequently encountered for the local full likelihood method, but was also in certain cases a problem with the local partial likelihood method. In all cases where the Newton-Raphson method failed to converge in the calculation of one or more estimated β_1-values, the current data set was skipped for all methods and a new data set was drawn instead. In the most difficult cases (heavy censoring/few observations) this could happen in around 10% of the cases for the local full likelihood method and around 1% of the cases for the local partial likelihood method. In the less difficult cases convergence problems occurred in 0-2% of the cases for the local full likelihood and almost never for the local partial likelihood method. This numerical instability is also, which is reasonable, more frequently a problem for small smoothing parameters than for larger smoothing parameters.

It is easily realized from (10) and (11) that the local maximum likelihood estimates will be undefined if there are no uncensored observations in the neighborhood considered. Such situations probably account for most of the cases in which both the local full and local partial likelihood methods run into problems. Some of the other cases are probably numerically unstable situations close to this undefined situation. For the rest of the cases the numerical problems could possibly have been avoided by using more advanced (or less advanced but slow) numerical methods.

In practical use, the problem of no uncensored observations in the neighborhood considered can of course always be overcome by choosing a large enough smoothing parameter. This may, however, not always be a satisfactory solution since this may lead to an oversmoothed estimate in other regions. A better solution may be to use a varying smoothing parameter.

For the covariate order time transformation method numerical problems of this kind has never been encountered. Even neighborhoods with no uncensored observations cause no problems for the covariate order time transformation method. A possible source of numerical problems in the covariate order time transformation method is the estimation of a smooth correspondence function in the exponential regression step of the algorithm. The super-smoother used for this purpose is in practice local linear regression which involves some iterative numerical routines. But estimation of the correspondence function is based on all observations, censored as well as uncensored, and numerical problems has never occurred in the estimation of the smooth correspondence function. If such problem did occur, however, a simple and and satisfactory way of getting around the problem would be to use the simple estimate of the correspondence function denoted $\tilde{s}(x)$ in Section 3.1. Simulations in Kvaløy and Lindqvist (1998a) show that the only difference between using the estimate $\tilde{s}(x)$ and the smoother estimate denoted $\hat{s}(x)$ is that the latter remove some small scale roughness in the estimated hazard function. With this implementation of the covariate order method and if the Breslow estimator is used for estimating the cumulative baseline hazard, there are no "external" sources for numerical problems in the covariate order time transformation method. The only possible numerical problem would then be that the algorithm itself did not converge, but this has never happened with data from Cox-models in our simulations and computations.

7 REAL DATA EXAMPLES

In this section two real data sets are analysed using the covariate order time transformation method.

7.1 Childhood Leukemia Data

This data set is studied in Seidel et al. (1997) and consists of observations of 42 children with leukemia. The response variable is time in months until relapse after diagnosis, and for this example we will only consider the two covariates age at diagnosis and MCL. MCL is mean clearance of a drug called Methotrexate, and the mean is taken over up to eight treatments. All the patients were followed from they entered the study at time of diagnosis and until the end of the study or until they suffered relapse. This means that the censoring is random. The data are highly censored, only 16 of the patients suffered relapse during the observation period, in other words 62% censoring.

First the model $\alpha_0(t)\exp(g_1(\text{age}) + g_2(\text{MCL}))$ is fitted. The algorithm converged after 6 iterations, and the smoothing parameters chosen by (5) were respectively 48 for $\hat{g}_1$ and 44 for $\hat{g}_2$ in all iterations. The bootstrapped Anderson-Darling test for covariate effect with 10000 bootstrap replications gave a p-value of 0.1358 for age and 0.0158 for MCL. The estimated covariate functions are displayed in Figure 5. Taking into account the small number of observations, it seems reasonable that the effect of MCL can be modeled as a linear covariate function, while a possible (in this model non-significant) effect of age seems to be non-linear.

Hence the model $\alpha_0(t)\exp(g_1(\text{age}) + \beta_2\text{MCL})$ was tried next. The algorithm converged after 6 iterations and the smoothing parameter was chosen to be 48 in the two first iteration and 44 in all subsequent iterations. The bootstrapped Anderson-Darling test for covariate effect now gave p-values of respectively 0.0682 for age and 0.0139 for MCL.

In Figure 6 both a plot of $\hat{g}_1(\text{age})$ and the correspondence function (recall Figure 2 and Section 4.4) from the last step of the algorithm is given. The plot of $\hat{g}_1(\text{age})$ basically show the same non-linear effect as was seen previously, and the correspondence function plot both reflects the main features of the $\hat{g}_1(\text{age})$ curve and show that the covariates are fairly well spread over the domain of observed covariates. None of the covariate values are outlying, but there is one jump on the s-axis that is relatively much larger than the others. A closer

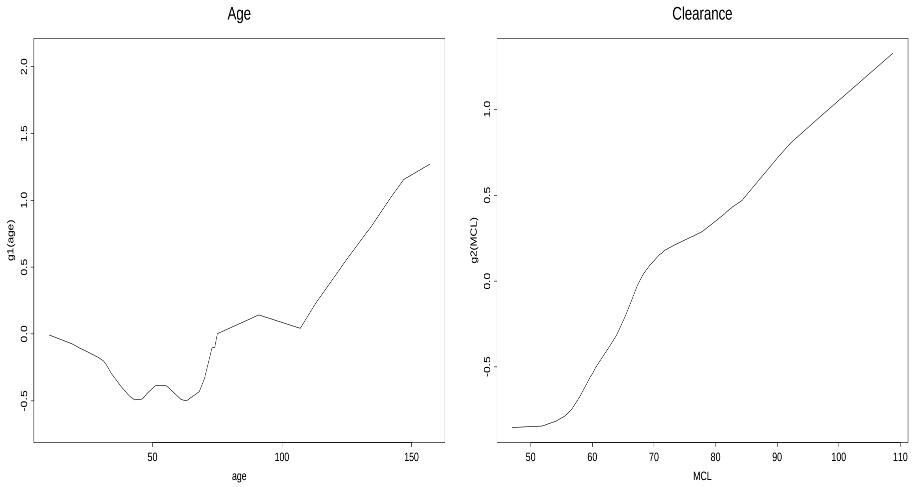

Figure 5: Estimated covariate functions for age and MCL in the model $\alpha_0(t)\exp(g_1(\text{age}) + g_2(\text{MCL}))$.

examination revealed that this was a child with a long observed life time and the highest of all recorded scores on MCL. It could be investigated whether this child's MCL score should be regarded an outlier, but we do not push this further here. The large jump occurs in a very steep region of $\hat{s}(\text{age})$ and thus does not alter the curve seriously.

The nearly U-shaped form of $\hat{g}_1(\text{age})$ suggests a quadratic function of age as a reasonable transformation to try if we want to fit a parametric model. This model and some other basic parametric fits, plus the two nonparametric models discussed, are summarized in Table 8. From this table it is clear that the two models with a linear function of MCL and either a

Model	df	p-value	$-2\log PL(\hat{g}(\cdot))$
null			112.8
$\beta_1\text{age}$	1	$p_1 = 0.24$	111.4
$\beta_1\text{age} + \beta_2\text{age}^2$	2	$p_1 = 0.160$, $p_2 = 0.093$	108.7
$\beta_1\text{MCL}$	1	$p_1 = 0.06$	109.5
$\beta_1\text{age} + \beta_2\text{MCL}$	2	$p_1 = 0.073$, $p_2 = 0.020$	106.3
$\beta_1\text{age} + \beta_2\text{age}^2 + \beta_3\text{MCL}$	3	$p_1 = 0.024$, $p_2 = 0.0076$, $p_3 = 0.0010$	98.8
$g_1(\text{age}) + g_2(\text{MCL})$	2.96(5.97)	$p_{1_B} = 0.1358$, $p_{2_B} = 0.0158$	103.3
$g_1(\text{age}) + \beta_2\text{MCL}$	2.16(5.14)	$p_{1_B} = 0.0682$, $p_{2_B} = 0.0139$	100.8

Table 8: Degrees of freedom compared to the null model, p-values and value of twice the log partial likelihood for different models. For the models with nonparametric covariate terms are the degrees of freedom estimated based on 1000 simulations, and the variance of the distribution of the difference in partial likelihood is given in parentheses. The p-values for these models are calculated using the bootstrapped Anderson-Darling test for covariate effect with 10000 replications.

nonparametric or a quadratic function of age are the two best models. The parametric model achieves the lowest value of $-2\log PL$, but is using more degrees of freedom. The amount of

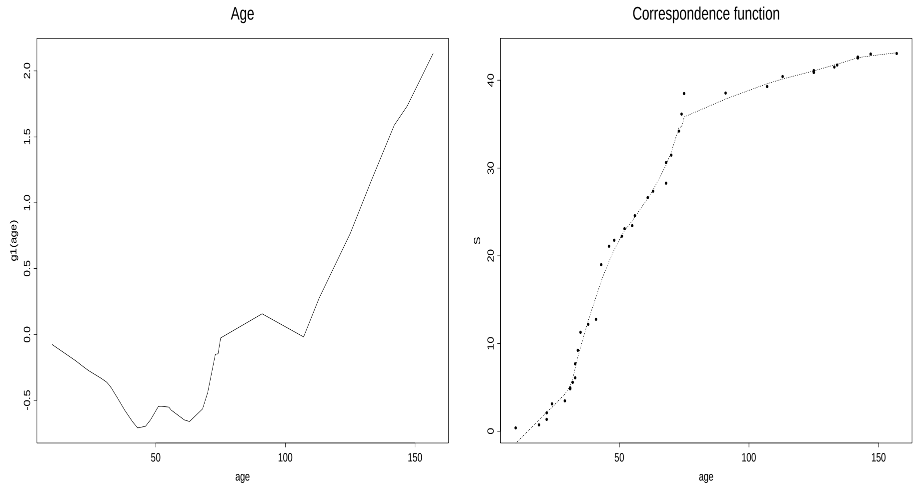

Figure 6: Estimated covariate function for age, and a plot of $(L_{11}, x_{11}), \ldots, (L_{42,1}, x_{42,1})$ and $\hat{s}(\text{age})$ (dashed line) from the last iteration of the algorithm, using the model $\alpha_0(t)\exp(g_1(\text{age}) + \beta_2\text{MCL})$.

data is not sufficient to distinguish between these two models based on differences in partial likelihood. All covariates are significant at a 2.5% level in the parametric model, while age is only significant at a 7% level in the model with a nonparametric covariate function of age. However, with only 16 uncensored observations, the bootstrapped Anderson-Darling test for covariate effect should not be expected to have the same power as a test based on a well specified parametric model.

Bootstrap curves for $g_1(\text{age})$ in the model $\alpha_0(t)\exp(g_1(age) + \beta_2\text{MCL})$ are displayed in Figure 7. These curves are obtained by respectively resampling the observations and the gener-

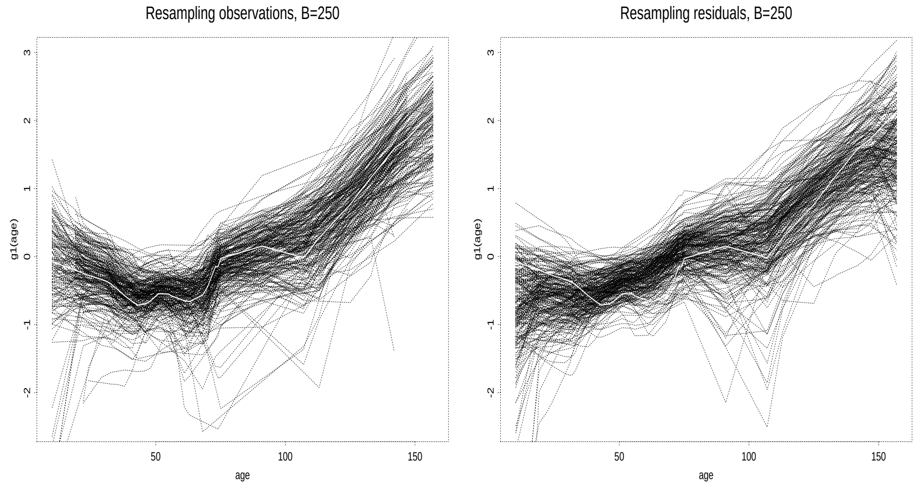

Figure 7: The plots show 250 bootstrap-replications of $\hat{g}_1(\text{age})$ in the model $\alpha_0(t)\exp(g_1(age) + \beta_2\text{ML})$ resampling respectively the observations and the residuals. The original estimate is displayed as the white curve.

alized residuals as explained in Section 4.3. The number of bootstrap curves is 250 in both cases, and the curves basically reflect the sparseness of data by showing great variability in all parts of the domain. The decline in $\hat{g}_1(\text{age})$ for ages up to around 60 months is supported by the bootstrap curves that is based on resampling observations, but not by the curves based on resampling residuals.

The estimated cumulative baseline hazard function, using the model $\alpha_0(t)\exp(g_1(age)+\beta_2\text{ML})$, and the graphical model check explained in Section 4.4 are displayed in Figure 8. The cumulative baseline hazard plot shows an approximately constant baseline hazard up to

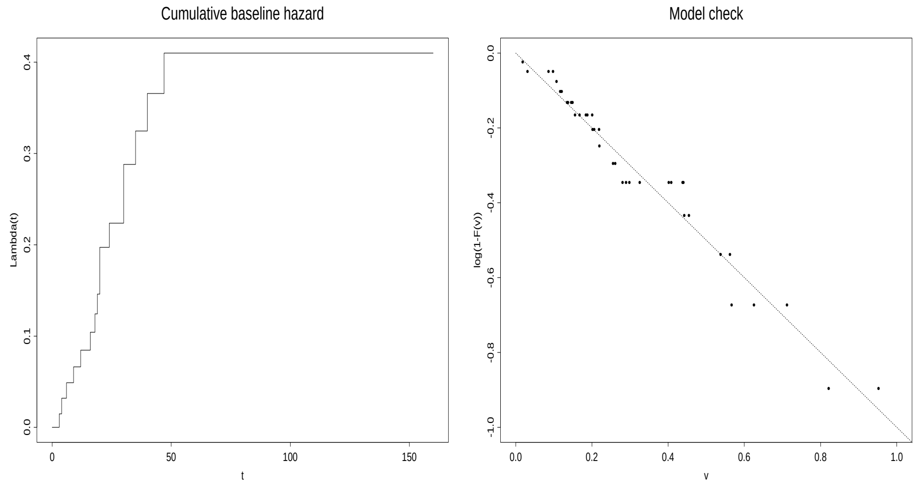

Figure 8: The left plot shows the estimated cumulative baseline hazard function using the model $\alpha_0(t)\exp(g_1(age)+\beta_2\text{ML})$ and the right plot shows the graphical model check plot for this model.

time 50, after which time no uncensored observations exist. The graphical model check plot (recall Section 4.4) indicates that the fitted model is good.

Seidel et al. (1997) fitted the parametric model $\alpha_0(t)\exp(\beta_1\text{age}+\beta_2\text{age}^2+\beta_3\text{ML})$. The example in this section shows that even if we ultimately want to fit a parametric model for the covariates in a Cox-model, a nonparametric model used in an early step of the analysis can be very useful for suggesting covariate transformations.

7.2 Stanford Heart Transplant Data

The Stanford Heart Transplant data have been studied by a number of authors. Among those cited in this paper Kalbfleisch and Prentice (1980), Tibshirani (1984), Hastie and Tibshirani (1986), Efron (1988) and O'Sullivan (1988) can be mentioned. Various versions of the data exist, the version used in this example is the one reported in Miller and Halpern (1982).

The data set consists of survival times in days after a patient had received a heart transplant, and two covariates, age in years and T5 mismatch score. The data reported in Miller and Halpern (1982) consists of totally 184 observations, but some have missing values, and only the 157 observations that are complete are used. The patients entered the study at random and where followed until the end of the study or until they died. Consequently, the censoring is random. 55 patients were still alive at the end of the study.

A number of different models have been fitted to the data, and a summary is given in Table 9. This table clearly shows that T5 is not a significant covariate. The model with a quadratic

Model	df	p-value	$-2\log PL(\hat{g}(\cdot))$
null			902.4
$\beta_1$age	1	$p_1 = 0.0085$	894.8
$\beta_1\text{age} + \beta_2\text{age}^2$	2	$p_1 = 0.0088$, $p_2 = 0.0014$	886.3
$g_1(\text{age})$	2.07(4.31)	$p_{1_B} = 0.0002$	889.9
β_1T5	1	$p_1 = 0.31$	901.4
$\beta_1\text{T5} + \beta_2\text{T5}^2$	2	$p_1 = 0.27$, $p_2 = 0.13$	901.4
$\beta_1\text{age} + \beta_2\text{T5}$	2	$p_1 = 0.0091$, $p_2 = 0.35$	894.0
$\beta_1\text{age} + \beta_2\text{age}^2 + \beta_3\text{T5}$	3	$p_1 = 0.0084$, $p_2 = 0.0014$, $p_3 = 0.34$	885.4
$g_1(\text{age}) + g_2(\text{T5})$	2.96(6.20)	$p_{1_B} = 0.0006$, $p_{2_B} = 0.70$	888.5
$g_1(\text{age}) + \beta_2\text{T5}$	3.03(6.04)	$p_{1_B} = 0.0003$, $p_{2_B} = 0.6939$	889.1

Table 9: Degrees of freedom compared to the null model, p-values and values of twice the log partial likelihood for different models. For models with nonparametric covariate terms, the degrees of freedom are estimated from 1000 simulations. The variance of the distribution of the difference in partial likelihood is given in parentheses. The p-values for these models are calculated using the bootstrapped Anderson-Darling test for covariate effect with 10000 replications.

function in age achieves the smallest value of $-2\log PL$ among the models that only include age. The nonparametric model uses roughly the same number of "parameters" as the best parametric estimate, but achieves a higher value of $-2\log PL$. The variance of the increase in likelihood also indicates a distribution not very different from the χ^2-distribution, in other words, there is no support for using the nonparametric rather than the parametric estimate based on the partial likelihood.

Some relevant plots for the nonparametric model, $\alpha_0(t)\exp(g_1(age))$, are given in Figure 9. The algorithm for estimation of the nonparametric model converged after 4 iterations and the smoothing parameter was chosen to be 22 i all iterations. The left plot in the first row shows the nonparametric and the best parametric estimate of the covariate functions. The nonparametric curve shows an approximately constant risk of dying after a heart transplantation until age 40, and then an increasing risk for age greater than 40. A similar increasing risk for age greater than 35 is shown by the parametric curve, but this curve shows a surprising declining risk from age 15 until age 35. The correspondence function plot in Figure 9 shows that the main part of the patients are in the age interval 30-55, with only three patients older than 55. Also, the data for ages less than 30 are more sparse than for the middle ages. No obvious outliers in observation time are seen. The cumulative hazard indicates that the risk of dying after a heart transplantation is very high during the first few days after the transplantation, then decreasing to an approximately constant level.

The bootstrap curves in Figure 9 show, as expected, a greater variability toward the endpoints where there are less observations than in the region with many observations. Resampling of the residuals seems to indicate that the estimated curve is upward biased around age 35-45 and downward biased toward the right endpoint. But when resampling of the observations is used, the estimated curve is everywhere in the middle of the region spanned by the bootstrap

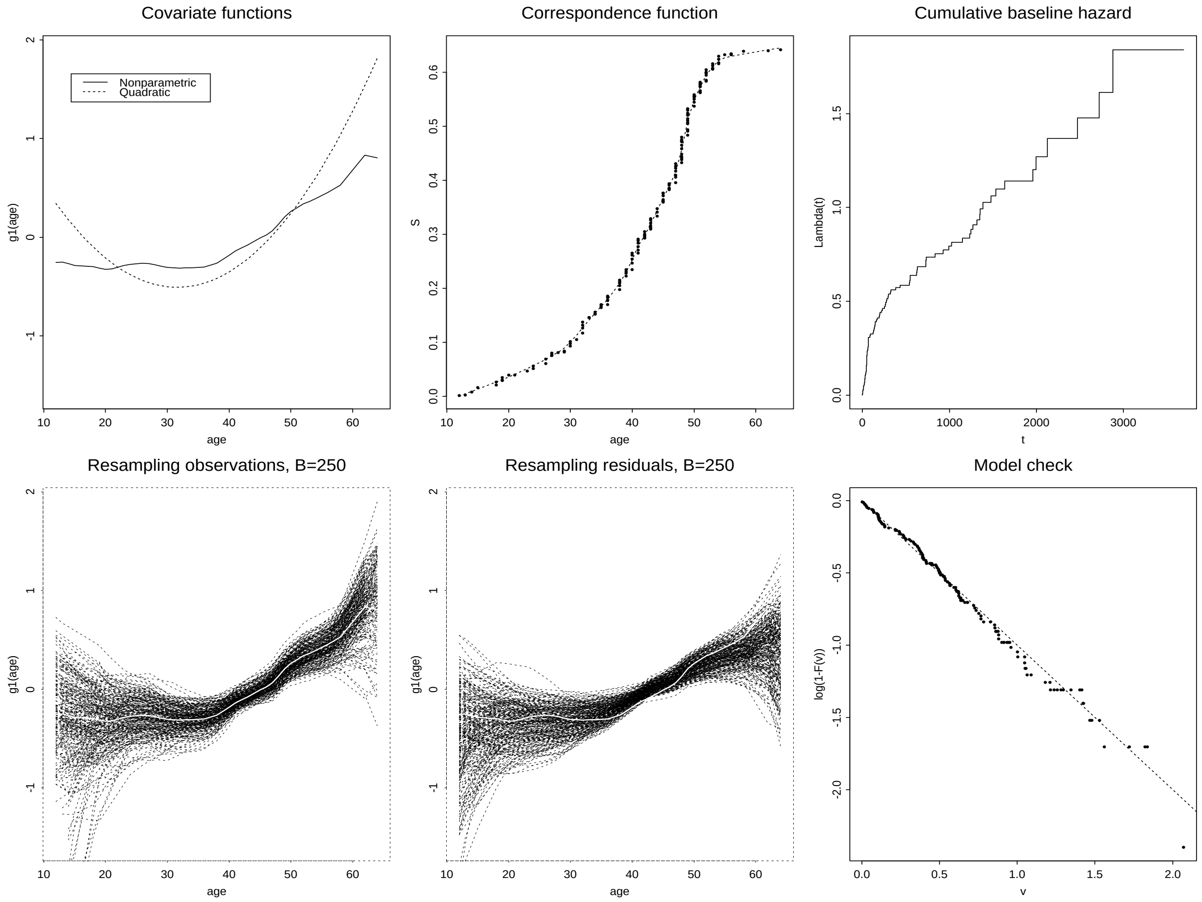

Figure 9: The plots in the first row show respectively the nonparametric and the quadratic estimate of the covariate function, the correspondence function and the cumulative baseline hazard function estimated from the nonparametric model. The two first plots in the second row show 250 bootstrap-replications of the nonparametric covariate function resampling respectively the observations and the residuals, and the original estimate is displayed as the white curve. The right plot in the second row is the graphical model check plot.

curves. The bootstrap curves indicate that the unexpected declining curve until age 35 shown by the quadratic function in age is just an artifact caused by using a parametric model. This agrees with Efron's (1988) conclusion from bootstrapping the local score estimates of Hastie and Tibshirani (1986).

Finally, the graphical model check based on the generalized residuals indicates a very good adequacy of the fitted nonparametric model.

8 CONCLUSIONS

The time transformation approach defines a very broad class of methods for Cox-regression in that any method for exponential regression can be used in the estimation of the covariate function and that the baseline hazard function can be estimated in a number of different ways. Examples of time transformation methods is the local full likelihood, the covariate order time transformation method, and, if the classical semiparametric approach is taken, the EM algorithm of Clayton and Cuzick (1985). In this paper focus has been on nonparametric methods, and the implementations of the general time transformation algorithm studied, the covariate order time transformation method and the local full likelihood method, enjoy several useful properties. Implementation is easy, especially in software packages that offer standard Cox-regression, numerical convergence is fast, stratified models can easily be fit, combinations of nonparametric and parametric models can easily be fit and ties cause no problems.

Doing inference in the generalized Cox-model is not as easy as in the classical Cox-model, but it is demonstrated in Sections 4 and 7 that there are several useful techniques at hand for doing at least approximate inference in the generalized nonparametric Cox-model as well. It follows naturally from the time transformation approach how to test the significance of each covariate separately in the exponential regression step of the algorithm. The correspondence function plot, arising from the covariate order method for exponential regression, is a helpful tool to visualize information on location of the observations in the covariate space and to show how the observation times and other covariates affect the estimation of the present covariate. Doing inference based on simulating the distribution of differences in the partial likelihood and using bootstrapping to assess the variability in the estimated covariate function is also very useful.

It is not possible to clearly distinguish between the three methods studied in the simulation study. Suppose that the criterion used to choose the smoothing parameter for each of the methods tends to choose the smoothing parameter which minimizes the metric D. Then all the three methods will, measured by the same metric, give very similar results in a broad class of situations. If, however, one of the methods studied is to be recommended over the other methods for general use, the conclusion is modestly in favor of recommending the covariate order time transformation method. The covariate order time transformation method is less variable than the other methods, it is performing well in cases with few observations or heavy censoring and numerical problems has never been encountered. Low variability and good properties at endpoints make the covariate order time transformation method better than the other methods for small smoothing parameters, but the method is more biased than the other methods for larger smoothing parameters. As for the local full likelihood method versus the local partial likelihood method, the simulation results are weakly in favor of recommending the local full likelihood method, but the local partial likelihood has the benefit of being non-iterative.

Acknowledgement

I would like to thank my advisor Bo Lindqvist for introducing me to this topic, for suggesting the time transformation approach and for a number of very helpful discussions and suggestions during the work on this report. Part of this work was done at Department of Statistics, Stanford University, and I am grateful for their kind hospitality. This work was funded by a PhD grant from the Research Council of Norway.

References

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. AND KEIDING, N. (1993). *Statistical Models Based on Counting Processes*, Springer Series in Statistics, New York: Springer-Verlag.

ANDERSON, T. W. AND DARLING, D. A. (1952). Asymptotic theory of certain goodness of fit criteria based on stochastic processes., *Annals of Mathematical Statistics* **23**: 193–212.

ASCHER, H. AND FEINGOLD, H. (1984). *Repairable Systems Reliability. Modeling, Inference, Misconceptions and Their Causes*, Marcel Dekker, Inc., New York.

BRESLOW, N. E. (1972). Contribution to the discussion of "Regression models and life-tables" by D. R. Cox., *Journal of the Royal Statistical Society, Series B* **34**: 187–220.

CLAYTON, D. AND CUZICK, J. (1985). The EM algorithm for Cox's regression model using GLIM., *Applied Statistics* **34**: 148–156.

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

COX, D. R. (1975). Partial likelihood, *Biometrika* **62**: 269–276.

EFRON, B. (1988). Computer-intensive methods in statistical regression, *SIAM Review* **30**: 421–449.

FAN, J., GIJBELS, I. AND KING, M. (1997). Local likelihood and local partial likelihood in hazard regression, *The Annals of Statistics* **25**: 1661–1690.

FRIEDMAN, J. (1984). A variable span smoother, *Technical Report 5*, Stanford University, Department of Statistics.

GENTLEMAN, R. AND CROWLEY, J. (1991). Local full likelihood estimation for the proportional hazards model, *Biometrics* **47**: 1283–1296.

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

HASTIE, H. AND TIBSHIRANI, R. (1986). Generalized additive models, *Statistical Science* **1**: 297–318.

HASTIE, T. AND TIBSHIRANI, R. (1990). Exploring the nature of covariate effects in the proportional hazards model, *Biometrics* **46**: 1005–1016.

KALBFLEISCH, J. D. AND PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley series in probability and mathematical statistics, Wiley, New York.

KOOPERBERGER, C., STONE, C. J. AND TRUONG, Y. (1995a). Hazard regression, *Journal of the American Statistical Association* **90**: 78–94.

KOOPERBERGER, C., STONE, C. J. AND TRUONG, Y. (1995b). The l_2 rate of convergence for hazard regression, *Scandinavian Journal of Statistics* **22**: 143–157.

KVALØY, J. T. (1999). Covariate order tests for covariate effect., *Technical report*, Norwegian University of Science and Technology, Department of Mathematical Sciences. Under preparation.

KVALØY, J. T. AND LINDQVIST, B. H. (1998a). The covariate order method for censored exponential regression., *Technical Report 10*, Norwegian University of Science and Technology, Department of Mathematical Sciences.

Kvaløy, J. T. and Lindqvist, B. H. (1998b). TTT-based tests for trend in repairable systems data, *Reliability Engineering and System Safety* **60**: 13–28.

Miller, R. and Halpern, J. (1982). Regression with censored data, *Biometrika* **69**: 521–531.

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

O'Sullivan, F. (1993). Nonparametric estimation in the Cox model, *The Annals of Statistics* **21**: 124–145.

Seidel, H., Nygaard, R., Moe, P. J., Jacobsen, G., Lindqvist, B. and Slørdal, L. (1997). On the prognostic value of systemic methotrexate clearance in childhood acute lymphocytic leukemia, *Leukemia Research* **21**: 429–434.

Silverman, B. W. (1986). *Density Estimation*, Monographs on Statistics and Applied Probability, London: Chapman & Hall.

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

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models, *Journal of the American Statistical Association* **84**: 276–283.

Tibshirani, R. (1984). Local likelihood estimation, *Technical Report 223*, Stanford University, Department of Statistics.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation, *Journal of the American Statistical Association* **82**: 559–567.