The covariate order method for censored exponential regression.

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

October 5, 1998

Abstract

In reliability engineering and biomedical applications, a first analysis of life time data is often based on the assumption of independent and identically exponentially distributed life times. In many practical cases, however, the distribution of the life times depends on a set of covariates. The assumption of identically distributed data is then not valid, but the simple model of exponentially distributed data can still be used by allowing the hazard rate to be a function of these covariates. Estimating this relationship between the hazard rate and the covariates is called exponential regression. The purpose of the present paper is to present a new nonparametric method, called the covariate order method, for exponential regression with right censored data and continuous covariates. In simulations this new method shows better performance than competing nonparametric methods.

For the case of one covariate the idea of the covariate order method is to arrange the observations in increasing order of the covariate, and then create a process where the observation times are subsequent intervals on an artificial "time" axis. In this process, points being the endpoint of an interval corresponding to an uncensored observations is considered an event. A consistent kernel based estimator of the hazard rate is found by first estimating the intensity of this process and then combining this with an estimated relationship between points on the artificial "time" axis and the covariate axis. Generalizations to more than one covariate can be done in several ways. The new method has been compared to a local likelihood and a local regression method, and has shown better performance than these methods in simulations.

An interesting by-product from the derivation of the covariate order method for exponential regression is that, based on the same idea, a test for covariate effect which has good power properties both against linear and nonlinear covariate effects can be established. Other ways of doing inference and model check are also discussed.

As a practical example, the method is used to analyze a dataset concerning occurrence of cardiac arrest, and a significant nonlinear relationship between the intensity of cardiac arrest and air temperature is found.

KEY WORDS: Nonparametric estimation; Hazard rate; Kernel estimation.

1 INTRODUCTION

Various kinds of data, for example survival times, are often assumed to be independent and identically exponentially distributed, which means assuming a common and constant hazard rate λ for all observations. A less restrictive and often more realistic assumption is to assume that the data are exponentially distributed but with the hazard rate λ varying from observation to observation according to differences in certain *covariates*, for example $\lambda = \lambda(\mathbf{x})$ where $\mathbf{x}$ is an m-dimensional vector of covariates. Estimating the hazard rate $\lambda(\mathbf{x})$ from observed data is called *exponential regression.* If some of the observations are right censored we have *censored* exponential regression.

There exist a number of possible estimation methods for censored exponential regression models. If a parametric approach is taken, for instance by fitting a generalized linear model (McCullagh and Nelder, 1989), the hazard rate can be estimated by using maximum likelihood estimation. A semi-parametric approach may be taken by adapting spline based methods like for instance in Hastie and Tibshirani (1990) or Gray (1992). The methods presented in these papers are developed for estimation in Cox-regression models, and an exponential regression model is simply the special case of a Cox-regression model with a hazard function which is constant in time. Lindqvist and Tjelmeland (1989) suggested a local likelihood method for nonparametric censored exponential regression, and a number of other authors, like for instance Tibshirani and Hastie (1987), Staniswalis (1989), Gentleman and Crowley (1991) and Fan, Gijbels and King (1997), have presented similar local likelihood methods for estimation in Cox-regression models. Another nonparametric approach that may be taken is to transform the problem into an ordinary regression problem and apply a method, like for instance Fan and Gijbels (1994), for nonparametric regression with censored data.

The purpose of this paper is to present a new nonparametric method for exponential regression which for the case of one covariate is based on arranging the observations in increasing order of the covariate, and then using kernel estimation on a certain process defined by the ordered observations. This method is compared to the above mentioned local likelihood and regression methods. The new method is presented in Section 2, while the two other methods are briefly reviewed in Section 3. The methods are compared in a simulation study in Section 5.

An interesting by-product from the new method for exponential regression presented is that, in the case of a single covariate, simple tests for the assumption of no covariate effect (in other words constant hazard rate) versus the alternative of a non constant hazard rate $\lambda(x)$, can be derived. This approach for testing the covariate effect is studied in Kvaløy (1998b), and is also briefly presented in Section 6.1. Likelihood based inference is discussed in Section 6.2, and a simple graphical method for model checking is mentioned in Section 6.3. Assessing the variability of the estimated curve by bootstrapping is discussed in Section 6.4.

Methods for Cox-regression can of course be applied to exponential regression problems, but it is also possible to do the opposite. Methods for nonparametric exponential regression can be used in an iterative algorithm for estimating nonparametric covariate functions in Cox-regression models. This application is studied in Kvaløy (1998a). A special case of this approach is the local full likelihood method which among the above cited authors is used in Staniswalis (1989), Gentleman and Crowley (1991) and Fan et al. (1997).

The basic formulation of the problem considered in the present paper is as follows. 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

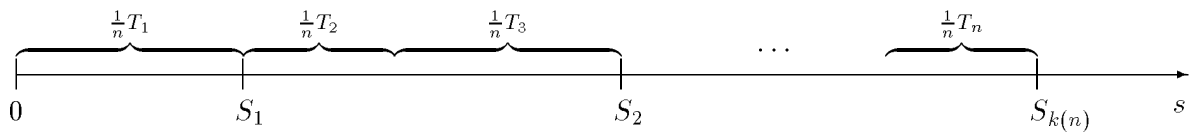

Figure 1: Construction of artificial process.

$(T, \delta, \mathbf{X})$, where $T = \min(Z, C)$, $\delta = I(Z \leq C)$ and $\mathbf{X}$ is a vector of covariates. For given $\mathbf{X} = \mathbf{x}$, Z is assumed to be exponentially distributed with an *unknown* hazard rate $\lambda(\mathbf{x})$, that is $f_Z(t|\mathbf{x}) = \lambda(\mathbf{x})\exp(-\lambda(\mathbf{x})t)$. Further C is distributed according to some unknown censoring distribution $f_C(t|\mathbf{x})$ which may depend on $\mathbf{x}$, and C is assumed to be independent of Z given $\mathbf{X}$. We call Z the *life time*, C the *censoring time* and T the *observation time*. This terminology is introduced only for convenience, Z can be any kind of exponentially distributed data.

The domain of the covariates is a subset $\mathcal{X}$ of $\mathcal{R}^m$, and the covariate vector $\mathbf{X}$ is assumed to be distributed according to some density function $f_{\mathcal{X}}(\mathbf{x})$. The corresponding cumulative distribution function is denoted $F_{\mathcal{X}}(t|\mathbf{x})$. The covariates are assumed to remain constant over time, and $\lambda(\mathbf{x})$ is assumed to be an *continuous* function for all $\mathbf{x} \in \mathcal{X}$. All methods are presented for the case $m = 1$. Extensions to higher dimensions are discussed in Section 4.

2 COVARIATE ORDER METHOD

First arrange the observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$ such that $X_1 \leq X_2 \leq \cdots \leq X_n$. Notice that ties in $X_1, \ldots, X_n$ by the above assumptions have zero probability of occurring. In practice, however, continuous data are only recorded to a finite number of digits, and ties may occur. If there is a small number of ties in the observed covariate values this can in practice be handled by arranging the observations with equal covariate values in random order. Next, for convenience, divide the observation times by n, the number of observations. Then let the scaled observation times $T_1/n, \ldots, T_n/n$, 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 points which are endpoints of an interval corresponding to an *uncensored* observation be considered as events, occurring at times denoted $S_1, S_2, \ldots$. This is visualized in Figure 1, for an example where the ordered observations are $(T_1, \delta_1 = 1), (T_2, \delta_2 = 0), (T_3, \delta_3 = 1), \ldots, (T_{n-1}, \delta_{n-1} = 0), (T_n, \delta_n = 1)$. Generally $S_i = \sum_{j=1}^{k(i)} T_j/n$ where $k(i) = \min\{r | \sum_{j=1}^{r} \delta_j = i\}$. Now the conditional intensity of the process $S_1, \ldots, S_{k(n)}$ at a point w on the s-axis equals $n\lambda(X_I)$ where I is defined from $\sum_{i=1}^{I-1} T_i/n < w \leq \sum_{i=1}^{I} T_i/n$. The basic idea is to estimate this intensity from the process $S_1, \ldots, S_{k(n)}$, yielding the estimator $\hat{\rho}_n(w)$ (a kernel estimator will be used), and then invert the relation $n\hat{\lambda}(X_I) = \hat{\rho}_n(w)$ 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 possible way of estimating a general relationship between points on the two axes is by plotting points $(X_1, T_1/n), \ldots, (X_n, \sum_{i=1}^n T_i/n)$ and drawing straight lines between successive points, in other words using the function

$$\tilde{s}(x) = \frac{1}{n}T_i + \frac{x - X_{J-1}}{X_J - X_{J-1}}\frac{1}{n}T_J, \quad X_{J-1} < x \leq X_J \tag{1}$$

where $X_0 = \min\{x | x \in \mathcal{X}\}$. An illustration is given in Figure 2.

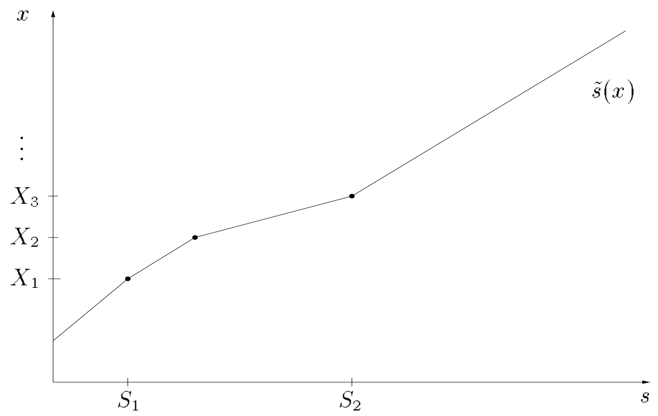

Figure 2: Illustration of estimating a monotone relationship between points on the covariate axis and the s-axis.

The motivating idea of the method is that if $\lambda(x) = \lambda$ is constant, then the process $S_1, \ldots, S_{k(n)}$ is a homogeneous Poisson process. (The test presented in Section 6 is in fact based on this observation.) Thus if $\lambda(x)$ is reasonably smooth and not varying too much, then the process $S_1, \ldots, S_{k(n)}$ could be taken to be nearly a nonhomogeneous Poisson process for which the intensity can be estimated by for instance kernel density estimation based on the points $S_1, \ldots, S_{k(n)}$. Combining this kernel estimate and (1) leads to an estimate of $\lambda(x)$. The estimator arising from this heuristic reasoning is the one presented below, but more precise arguments are needed to derive the estimator formally and to prove the consistency of the estimator. All proofs are given in the Appendix.

Let $\mathcal{F}_s^n$ be the history of the process $S_1, \ldots, S_{k(n)}$ in the interval $[0, s)$. This history is formally defined as the sub-σ-algebra $\mathcal{F}_s^n = \sigma\{X_1, \ldots, X_n\} \cup \sigma\{(T_j, \delta_j) : \sum_{i=1}^j T_i/n \leq s\}$ for $s \geq 0$. Note that $X_1, \ldots, X_n$ is contained in all the $\mathcal{F}_s^n$. Let $\rho_n(s|\mathcal{F}_s^n)$ be the conditional intensity of the process $S_1, \ldots, S_{k(n)}$ at the point s; see for example Andersen, Borgan, Gill and Keiding (1993, p. 75). Then the first step in the formal derivation of a consistent estimator for $\lambda(x)$ is Theorem 2.1 below which states that the scaled conditional intensity of the process $S_1, \ldots, S_{k(n)}$, converges in probability to a deterministic function of $\lambda(\cdot)$, and which gives an asymptotic relation between the processes running on the s-axis and the covariate axis, respectively.

Theorem 2.1 *Let the situation be as described above and in the formulation of the problem in Section 1. Further assume that* $\sup_{x\in\mathcal{X}} \lambda(x) \leq M < \infty$, $\inf_{x\in\mathcal{X}} \lambda(x) \geq a > 0$, *and that* $\sup_{x\in\mathcal{X}} \lambda'(x) \leq D < \infty$. *The conditional distribution of* C *given* x *is assumed to have finite first and second order moments and* $f_C(t|x)$ *is assumed to have bounded first derivative in*

x for all $x \in \mathcal{X}$. Then $\rho_n(s|\mathcal{F}_s^n)/n \xrightarrow{p} \lambda(\eta(s))$ as $n \to \infty$ uniformly in s, where $\eta(s)$ is a deterministic function from the s-axis to the covariate axis, the inverse of which is given by

$$s(x) = \int_{F_{\mathcal{X}}^{-1}(0)}^{x} \left[\int_0^\infty (t\lambda(v) - F_C(t|v)) \exp(-\lambda(v)t) dt \right] f_{\mathcal{X}}(v) dv.$$

The function $s(x)$ is called the correspondence function. Note that for the special case of no censoring, $s(x)$ simplifies to $s(x) = \int_{F_{\mathcal{X}}^{-1}(0)}^{x} f_{\mathcal{X}}(v)/\lambda(v)dv$.

The fact that the scaled intensity of the process $S_1, \ldots, S_{k(n)}$ converges uniformly to $\lambda(\eta(s))$ can be used to derive an estimator for $\lambda(x)$ by estimating the inverse function $s(x)$ and $\rho_n(s|\mathcal{F}_s^n)/n$. As a first step we state the following lemma.

Lemma 2.1 *Let the situation be as in Theorem 2.1. Then $\tilde{s}(x)$ in (1) is a uniformly consistent estimator of $s(x)$.*

Finally, a uniformly consistent estimator of $\lambda(x)$ is established by the following theorem.

Theorem 2.2 *Let the situation be as in Theorem 2.1. Further let $K(\cdot)$ be a positive kernel function which vanishes outside [-1,1] and has integral 1, and let h_s be a smoothing parameter which is either constant or varying along the s-axis. Assume that $h_s \to 0$ as $n \to \infty$ for all s. Further assume that there is a sequence h_n such that $h_s \geq h_n$ for all s, n where $nh_n \to \infty$ as $n \to \infty$. Then the estimator*

$$\hat{\lambda}(x) = \frac{1}{nh_s} \sum_{i=1}^{k(n)} K(\frac{\tilde{s}(x) - S_i}{h_s}) \; ; \quad x \in \mathcal{X} \tag{2}$$

is a uniformly consistent estimator of $\lambda(x)$.

In practical use the estimated correspondence function (1) may be replaced by more sophisticated estimators, improving on the smoothness of the estimator (2). We have used the super-smoother of Friedman (1984) to obtain the smooth correspondence function estimate $\hat{s}(x)$ from the points $(X_1, T_1/n), \ldots, (X_n, \sum_{i=1}^n T_i/n)$. See Figure 3 for an illustration. This estimate is calculated by using local linear regression with a variable bandwidth which is chosen locally using local cross-validation.

To avoid the estimate $\hat{\lambda}(x)$ to be seriously downward biased near the endpoints special care must be taken at the boundaries. Viewed only as a problem on the s-axis the estimator (2) is simply density estimation on the s-axis, and techniques for handling boundary problems in density estimation can be adopted. A common technique is to reflect the data points around both endpoints, see for example Silverman (1986), corresponding to using the estimator

$$\hat{\lambda}(x) = \frac{1}{nh_s} \sum_{i=1}^{k(n)} \left[K(\frac{\hat{s}(x) - S_i}{h_s}) + K(\frac{\hat{s}(x) + S_i}{h_s}) \; + K(\frac{\hat{s}(x) + S_i - 2S}{h_s}) \right] \tag{3}$$

where $S = \sum_{j=1}^n T_j/n$. In simulations not reported the pseudo-data method of Cowling and Hall (1996) for edge correction was also tried. This method is supposed to outperform the reflection method, but the experience from simulations was that the reflection method in

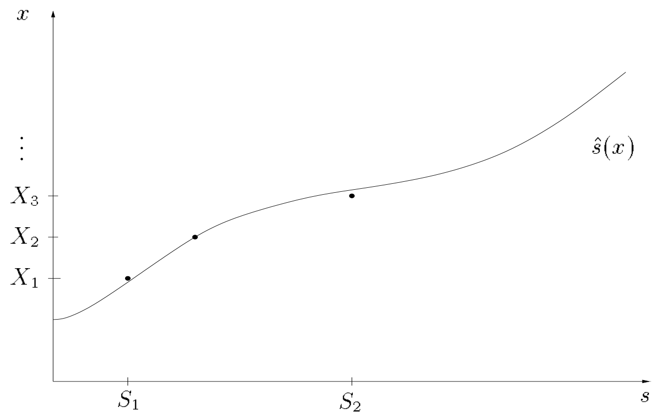

Figure 3: Illustration of estimating a smooth correspondence function from the covariate axis to the s-axis.

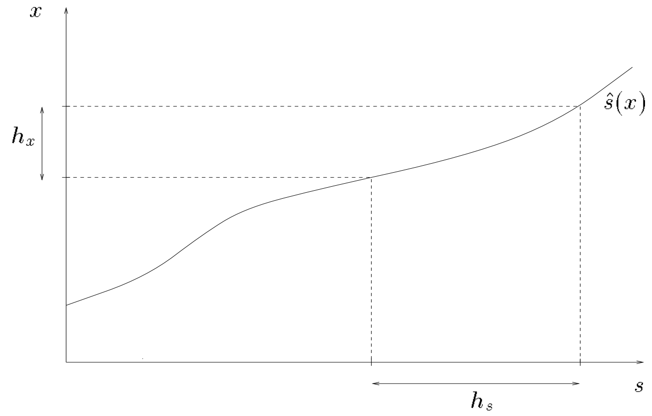

Figure 4: Correspondence between the smoothing parameter on the covariate axis and the s-axis

many cases gave better results. In addition there were often problems with the pseudo-data series not being monotone, calling for a somewhat ad-hoc adjustment of the series. Therefore the reflection method (3) is preferred.

The smoothing parameter h_s corresponds to smoothing over a certain amount of the data on the s-axis. On the covariate axis, a corresponding smoothing parameter h_x which cover approximately the same amount of the data can be defined via the relation between the points on the s-axis and the covariate axis. See Figure 4 for a rough description of the idea. If one of the smoothing parameters, h_s or h_x, is held constant the other will in general be varying (or both can be varying). Whereas a constant h_s corresponds to ordinary density estimation on the s-axis, a constant h_x corresponds to what is typically used in the other methods presented subsequently. In both cases a simple and easily applied criterion for choosing the best value of the smoothing parameter is a likelihood cross-validation criterion, see for example Simonoff (1996). The conditional likelihood for the observation times is given by

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

leading to the log-likelihood function

$$l(\lambda(\cdot)) = \sum_{i=1}^{n} [\delta_i \log \lambda(X_i) - \lambda(X_i) T_i] \tag{4}$$

The idea of the likelihood cross-validation criterion is to replace $\lambda(x)$ by $\hat{\lambda}^{(-i)}(x|h)$ which is the estimate of $\lambda(x)$ based on all data except the ith observation, and then optimize this function with respect to h, where h is a constant smoothing parameter on either the s-axis or the covariate axis. In other words, the value of h which maximizes the criterion

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

is used.

For consistency with the competing methods for nonparametric exponential regression, a constant h_x will be used in most parts of this paper. A constant smoothing parameter on the covariate axis is also easier to interpret than a constant smoothing parameter on the s-axis. Then (3) becomes

$$\hat{\lambda}(x) = \frac{1}{n h_s(\hat{s}(x))} \sum_{i=1}^{k(n)} \left[K(\frac{\hat{s}(x) - S_i}{h_s(\hat{s}(x))}) + K(\frac{\hat{s}(x) + S_i}{h_s(\hat{s}(x))}) + K(\frac{\hat{s}(x) + S_i - 2S}{h_s(\hat{s}(x))}) \right] \tag{6}$$

where $h_s(\hat{s}(x)) = \hat{s}(x + h_x/2) - \hat{s}(x - h_x/2)$.

It is difficult to say in general whether it is best to chose h_x or h_s to be constant. In the case that the covariates are approximately equally spaced the following argument seems to be in favor of letting h_x be constant. If h_x is constant, the corresponding varying smoothing parameter on the s-axis, $h_s(\hat{s}(x))$, will always be covering approximately the same amount of data as h_x in each corresponding point. In regions with long intervals on the s-axis (corresponding to a low intensity) $h_s(\hat{s}(x))$ will be large, whereas in regions with short intervals (high intensity) $h_s(\hat{s}(x))$ will adapt naturally to the data and become small. If h_s is constant there seems to be a risk of undersmoothing in the low intensity regions and oversmoothing in the high intensity regions. This will respectively increase variance in the low intensity regions and bias in the high intensity regions. On the other hand, this will reduce the bias in the low intensity region and reduce the variance in the high intensity region, so it is not clear if a constant h_s will be better or worse than a constant h_x. The nature of the censoring distribution will also play a role here. In the case of unequally spaced covariates it is even less clear whether it is preferable to held one or the other (or any at all) of the smoothing parameters constant. In general, a constant h_s will tend to increase bias and reduce variance in regions of the covariate space with few observations, and opposite in regions with many observations.

3 OTHER METHODS

3.1 Local Likelihood Method

Recall from (4) the log-likelihood function

$$l(\lambda(\cdot)) = \sum_{i=1}^{n} [\delta_i \log \lambda(X_i) - \lambda(X_i) T_i]$$

Let $\lambda(x) = \exp(g(x))$ where the form of $g(x)$ is not specified and suppose that the p^{th} order derivative of $g(x)$ at the point x exists (for example assume $\lambda(x)$ to be p times differentiable). By a Taylor expansion, if X_i is close to x,

$$g(X_i) \approx g(x) + g'(x)(X_i - x) + \ldots + g^{(p)}(x)/p!(X_i - x)^p \tag{7}$$

and we can write

$$\lambda(X_i) \approx \exp(\boldsymbol{\beta} \underset{\sim}{X}_i) \tag{8}$$

where $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\} = \{g(x), g'(x), \ldots, g^{(p)}(x)/p!\}$ and $\underset{\sim}{X}_i = \{1, X_i - x, \ldots, (X_i - x)^p\}^T$.

For a certain covariate value x, $\lambda(x)$ can be estimated *locally* using data in a weighted neighborhood of x by applying the *local likelihood* estimation method proposed by Tibshirani and Hastie (1987). Substituting $\exp(\boldsymbol{\beta} \underset{\sim}{X}_i)$ for $\lambda(X_i)$ the *local* log-likelihood function becomes

$$l(\boldsymbol{\beta}, x) = \sum_{i=1}^{n} [\delta_i \boldsymbol{\beta} \underset{\sim}{X}_i - \exp(\boldsymbol{\beta} \underset{\sim}{X}_i) T_i] \frac{1}{h_x} K(\frac{x - X_i}{h_x}) \tag{9}$$

where $K(\cdot)$ is a kernel function and h_x is a constant smoothing parameter on the x-axis. Let $\hat{\boldsymbol{\beta}}$ be the maximizer of (9). Then $\hat{g}^{(q)}(x) = q!\hat{\beta}_q$ and $\hat{\lambda}(x) = \exp(\hat{\beta}_0)$. The smoothing parameter h_x can be chosen according to the likelihood cross-validation criterion (5). Notice that as $h_x \to \infty$ this local likelihood modeling approaches the ordinary global likelihood modeling. Fan et al. (1997) have proved that point-wise for each x there exists a solution to (9) which, under mild conditions, is consistent and asymptotically normal. Further details and references on local polynomial modeling can be found in Fan and Gijbels (1996).

3.2 Local Regression Method

Estimation of $\lambda(x)$ can also be formulated as a regression problem. The idea is as follows. If Z is exponentially distributed with parameter $\lambda(x)$, then $\log Z$ is Gumbel distributed with expectation $-\log \lambda(x) - \gamma$ and variance $\pi^2/6$. Here $\gamma = 0.5772\ldots$ is Euler's constant. Let $\lambda(x) = \exp(-m(x))$ and let $Y = \log Z + \gamma$. Then $Y = m(x) + W$, where W is Gumbel distributed with expectation 0 and variance $\pi^2/6$. The function $m(x)$ can now be estimated using local polynomial regression. Let

$$m(X_i) \approx \beta_0 + \beta_1(X_i - x) + \ldots + \beta_p(X_i - x)^p = \boldsymbol{\beta} \underset{\sim}{X}_i$$

be the p^{th} order polynomial approximation to $m(X_i)$ (where, as in (7), β_i can be interpreted as $m^{(i)}(x)/i!$). Then the *local* polynomial regression estimator is the minimizer of

$$\sum_{i=1}^{n} [Y_i - \boldsymbol{\beta} \underset{\sim}{X}_i]^2 \frac{1}{h_x} K(\frac{x - X_i}{h_x}). \tag{10}$$

Notice that $h_x = \infty$ corresponds to ordinary least squares regression. Let the minimizer of (10) be denoted $\tilde{\boldsymbol{\beta}} = \{\tilde{\beta}_0, \ldots, \tilde{\beta}_p\}$, hence $\hat{m}(x) = \tilde{\beta}_0$. These estimates are easily found if no data are censored, see for example Fan and Gijbels (1996) or Simonoff (1996), but the censoring causes some difficulties. If both censored and uncensored observation times are transformed as $Y =$

$\log T + \gamma$ we get a censored sample of observations from the regression model $Y = m(x) + W$ where $\mathbf{E}W = 0$. Fan and Gijbels (1994) proposed a general method of replacing the censored observation in such regression models by an estimate of their expected value conditional on their censoring time. If Y is censored, Y is replaced by an estimate of $\mathbf{E}(Y|Y > y, X)$. With heavily censored data, however, this replacement method tends to underestimate the unobserved true values of the censored observations. In the present case, we can instead make use of the information that the censored observations are exponentially distributed. If T is censored we know that $\mathbf{E}(Z|Z > T, X) = T + 1/\lambda(X)$ due to the memoryless property of the exponential distribution. This motivates replacing a censored Y by $\log(T + 1/\hat{\lambda}(X) + \gamma)$. Using the replaced observations a new estimate of $\lambda(x)$ is found, and the process is iterated until convergence. The algorithm becomes:

1. Find an initial estimate $\hat{m}^{(0)}(x)$ from the original observations $(Y_1, X_1, \delta_1), \ldots, (Y_n, X_n, \delta_n)$ by using the method of Fan and Gijbels (1994) for local censored regression.
2. Repeat 3 for $d = 1, \ldots, D$ where $\max_{i\in\{1,\ldots,n\}} |\tilde{Y}_i^{(D-1)} - \tilde{Y}_i^{(D)}|/\tilde{Y}_i^{(D-1)} < \epsilon$.
3. Replace the observations $(Y_1, X_1, \delta_1), \ldots, (Y_n, X_n, \delta_n)$ by $(\tilde{Y}_1^{(d)}, X_1), \ldots, (\tilde{Y}_n^{(d)}, X_n)$ where
$$\tilde{Y}_i^{(d)} = \delta_i Y_i + (1 - \delta_i)\log(T_i + \exp(\hat{m}^{(d-1)}(X_i)) + \gamma)$$
and calculate the estimate $\hat{m}^{(d)}(x)$ using this replaced sample.

For $\epsilon \sim 10^{-5}$, D is typically around 10. The smoothing parameter h_x can be selected by cross-validation. Let $\hat{m}_{-i}(x|h_x)$ denote the estimator calculated with the ith data point disregarded. The optimal smoothing parameter $\hat{h}_x$ is the parameter that minimizes

$$CV(h_x) = \sum_{i=1}^{n} (\tilde{Y}_i^{(D)} - \hat{m}_{-i}(X_i|h_x))^2. \tag{11}$$

See Fan and Gijbels (1994) for further details.

3.3 The Local Methods Versus the Covariate Order Method

The ideas behind the covariate order method, the local likelihood method and the local regression method are completely different, and in general they have different properties and give different estimates. It can be verified, however, that if a rectangular kernel, ($K(u) = 1/2$ for $|u| < 1$, =0 otherwise) is used, the local constant ($p = 0$) likelihood method and the covariate order method with no boundary corrections are equal. If other kernels are used they will differ.

An advantage of the covariate order method is its flexibility. Any density estimation method should possibly be usable in the estimation of the scaled intensity of the process on the s-axis, and the smoothing parameter can be chosen either according to the values of the covariates or according to the values of the observation times. Boundary problems can be handled by adapting any edge correction technique invented for density estimation. In local polynomial modeling with $p \geq 1$ edge correction is, however, automatically incorporated (Fan and Gijbels, 1996). Simulations can indicate whether the automatically incorporated edge corrections in the local polynomial models or the specific edge corrections in the covariate order method give the best results in practice.

There are no obvious big differences in the robustness of the methods against outliers in observation time or covariate value. All methods will be affected by such outliers, but outliers will at least only be a problem locally in the neighborhood around the outlying observations. A possible way to reduce the effect of outliers in covariate value can be to apply ideas like in Tibshirani (1984) of weighting down observations that are far away from the center. A convenient way of simultaneously identifying outliers in observation times and covariate values is to plot the points $(X_1, T_1/n), \ldots, (X_n, T_n/n)$, see Figure 2 and Section 5.2.1.

The three different methods have at least conceptually quite different ways of handling censoring. In the likelihood method the censoring is taken care of in the likelihood, in the regression method censored observations are replaced, while in the covariate order method the censored observations contribute to the length of the intervals on the s-axis. Simulations can indicate which of these methods best handles censored data.

4 SEVERAL COVARIATES

All the methods can, at least in principle, be directly generalized to multidimensional covariates with, say, m components. For the local methods the generalization is in principle immediate, by considering polynomials in m variables and using m-dimensional kernels. For the covariate order method the generalization could be done by appropriately defining a point process in $\mathcal{R}^m$ with marginal processes of the type $S_1, S_2, \ldots$ as given for the case $m = 1$. This will be studied further elsewhere.

Direct generalizations will, however, soon be beaten by the curse of dimensionality. A better approach in most cases is to reduce the dimension of the problem by assuming some structure on the covariate space.

One way to proceed is to assume that the hazard rate can be written in the form of the generalized additive model

$$\lambda(\mathbf{x}) = \exp(\alpha + g_1(x_1)+, \ldots, +g_m(x_m)), \tag{12}$$

where $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X} \subseteq \mathcal{R}^m$, and where $g_1(\cdot), \ldots, g_m(\cdot)$ are unspecified continuous functions. These functions can be estimated by using the iterative algorithm presented below. Let $\hat{g}_i^{(d)}(x_i)$ be the estimate of $g_i(x_i)$ after the dth iteration of the algorithm, and define $\hat{g}_i^{(0)}(x_i) = 0$. Then the algorithm can be written as follows:

1. Repeat 2-3 for $d = 1, \ldots, D$ where $|\hat{\alpha}^{(D-1)} - \hat{\alpha}^{(D)}|/\hat{\alpha}^{(D-1)} < \epsilon$.
2. Calculate $\hat{\alpha}^{(d)} = \ln\left(\frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n T_i \exp(\hat{g}_1^{(d-1)}(X_{i1})+\ldots+\hat{g}_m^{(d-1)}(X_{im}))}\right)$.
3. Repeat 4-5 for $j = 1, \ldots, m$.
4. Transform $T_1, \ldots, T_n$ to $L_{1j}, \ldots, L_{nj}$ where $L_{ij} = T_i \exp(\hat{\alpha}^{(d)} + \hat{g}_1^{(d)}(X_{i1}) + \cdots + \hat{g}_{j-1}^{(d)}(X_{i,j-1}) + \hat{g}_{j+1}^{(d-1)}(X_{i,j+1}) + \cdots + \hat{g}_m^{(d-1)}(X_{im}))$, $i = 1, \ldots, n$.

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 some method for exponential regression, and let $\hat{g}_j^{(d)}(x_j) = \tilde{g}_j^{(d)}(x_j) - \sum_{i=1}^n \tilde{g}_j^{(d)}(X_{ij})/n$.

The speed of convergence depends on m, and for small m the convergence is fast. For $m = 2$ usually only 3 or 4 iterations are required if $\epsilon \sim 10^{-5}$. The motivation for the algorithm is the observation that if Z is exponentially distributed with parameter $\exp(\alpha + g_1(x_1) + \ldots + g_m(x_m))$, then $Z \exp(\alpha + g_1(x_1) + \ldots + g_{j-1}(x_{j-1}) + g_{j+1}(x_{j+1}) + \ldots + g_m(x_m))$ is exponentially distributed with parameter $\exp(g_j(x_j))$. This explains step 4 and 5 in the algorithm. In step 5 all the estimated functions are normalized to have mean 0 to avoid possible problems caused by constant factors jumping back and forth between the estimated functions. Finally the estimate $\hat{\alpha}^{(d)}$ in step 2 is the maximum likelihood estimate based on the sample $(L_1, \delta_1), \ldots, (L_n, \delta_n)$ where $L_i = T_i \exp(\hat{g}_1^{(d-1)}(X_{i1}) + \ldots + \hat{g}_m^{(d-1)}(X_{im}))$

Notice that the algorithm can easily be extended to allow some of the g-functions to be parametric (for example for discrete covariates) and then estimate the parameters of these functions in step 2 of the algorithm using maximum likelihood.

5 SIMULATION STUDY

The properties of the various methods discussed in the previous sections are explored and compared in this simulation study. In Section 5.1 the three methods for exponential regression are compared, while some interesting issues concerning the covariate order method are discussed in Section 5.2. All simulations are done i S-PLUS with parts of the code written in C.

5.1 Comparison of Methods

5.1.1 Simulation Models

In the comparison of the methods, data are simulated from models with three different hazard rates. These are,

$$\lambda_1(x) = \begin{cases} 1.5\exp(-\frac{x}{100}) & ,0 \le x < 100 \\ 1.5\exp(-1)[(2 - \frac{x}{100}) + 1.5(\frac{x}{100} - 1)^2] & ,100 \le x \le 250 \\ 0 & ,\text{elsewhere} \end{cases} \tag{13}$$

$$\lambda_2(x) = \begin{cases} 0.5 + 0.0025x & ,0 \le x < 100 \\ 0.25 + 0.005x & ,100 \le x \le 200 \\ 0 & ,\text{elsewhere} \end{cases} \tag{14}$$

$$\lambda_3(x_1, x_2) = \exp(0.5 + \sin x_1 - x_2^2), \quad 0 \le x_1 \le 1, \ -1 \le x_2 \le 1. \tag{15}$$

In all simulations covariate values are drawn uniformly from the covariate domain, and independent censoring points are drawn from a distribution with density

$$f_C(v) = 1.25v^{0.25}/V^{1.25}, \quad 0 \le v \le V \tag{16}$$

where V is adjusted according to the degree of censoring wanted. The biweight kernel

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

is used for estimation throughout the study. For both the local likelihood method and the local regression method local linear modeling, $p = 1$, have been used

5.1.2 Bias and variance.

To get a first picture of the bias and variance properties of the methods, ten different sets of data from hazard rate (13) and (14) have been simulated. Both without censoring and with approximately 50% censoring. Since we are primarily interested in comparing the bias and variance of the methods themselves, and not the bias and variance caused by the smoothing parameter criteria, the smoothing parameters are held fixed. For the curves shown in Figures 5 and 6, the smoothing parameters equal, respectively, $h_x = 60$ and $h_x = 200$ for all methods, and each of the ten datasets consists of respectively 200 and 100 observations

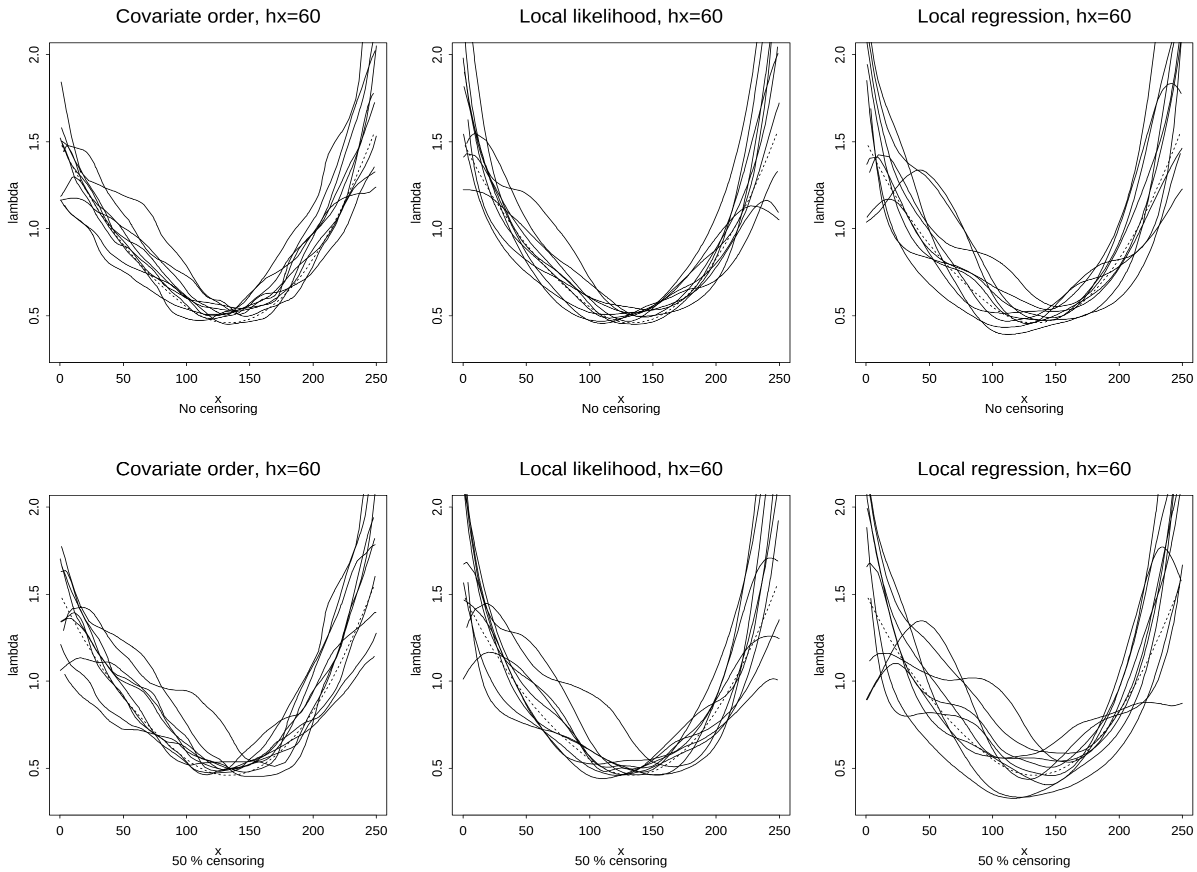

Figure 5: Ten estimated curves for each of the three estimation methods. Each estimated curve is based on 200 observations simulated using (13) and either no or 50% censoring. The true underlying curve is dashed.

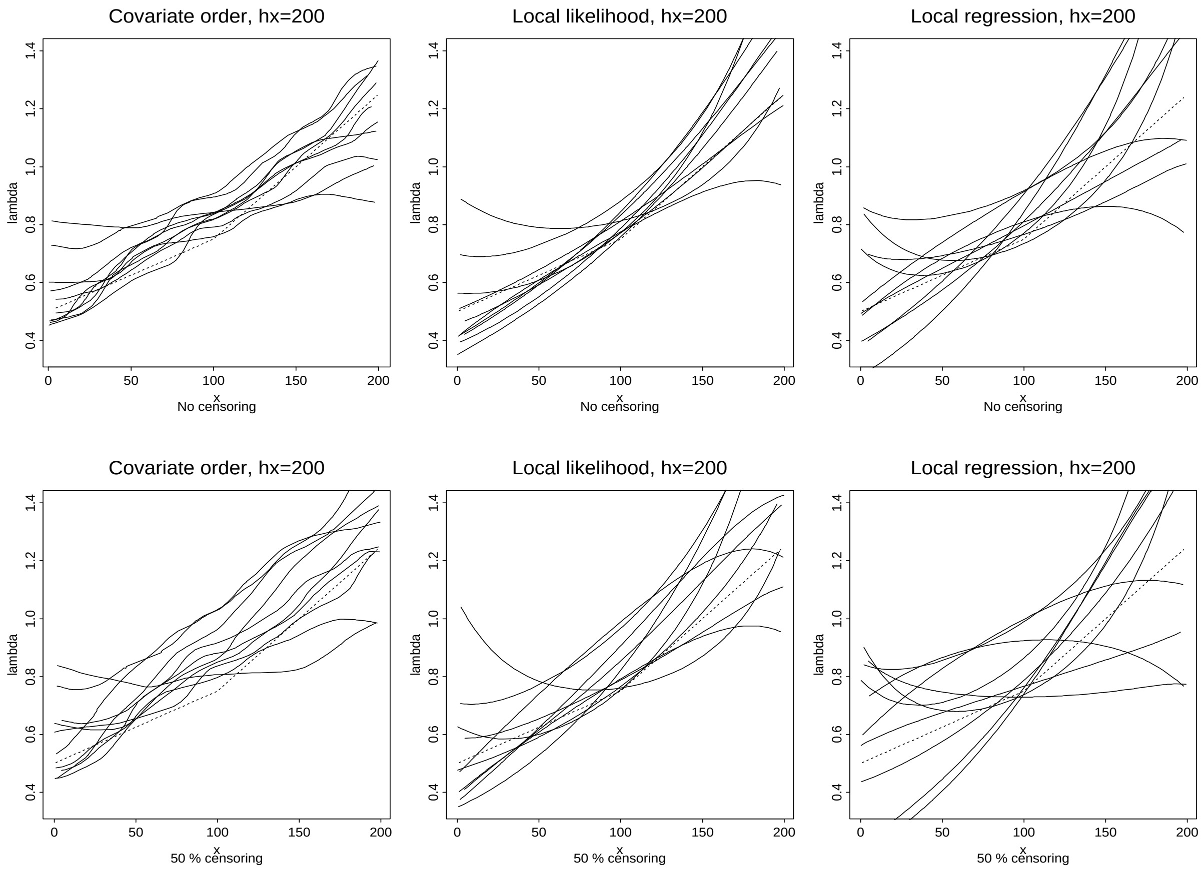

Figure 6: Ten estimated curves for each of the three estimation methods. Each estimated curve is based on 100 observations simulated using (14) and either no or 50% censoring. The true underlying curve is dashed.

At first glance we recognize in general a large variability in the estimates. This is due to the nature of exponentially distributed data. If we consider the first row of Figure 5, we see that the covariate order method seems to be the least variable and least biased method at the boundaries, especially at the left boundary . The local likelihood method is quite similar in the interior but is more variable towards the boundaries, while the regression based method also tends to be more variable in the interior. In the second row the same tendencies are seen, and as expected all methods are more variable since the data now are 50% censored. In Figure 6 the covariate order method is clearly the least variable and least biased of the three methods.

The performance of the various methods can be compared by calculating an approximate mean integrated squared error (MISE) criterion for each method. An approximate integrated squared error (ISE) for an estimate $\hat{\lambda}(x)$ based on observations with covariates $x_1 \leq x_2 \leq \cdots \leq x_n$ can be calculated as $\widehat{\mathrm{ISE}} = \sum_{i=1}^{n} \{\hat{\lambda}(x_i) - \lambda(x_i)\}^2 (x_{i+1} - x_{i-1})/2$ where $x_0 = \min_{x \in \mathcal{X}} x$ and $x_{n+1} = \max_{x \in \mathcal{X}} x$. By simulating a large number of datasets and calculating $\hat{\mathrm{ISE}}$ for the estimate based on each set, an estimate of MISE is obtained as the mean of these ISE estimates. The same simulation series as in Figures 5 and 6 have been repeated with 5000 repetitions instead of 10, and with three different choices for the smoothing parameters. Estimated

	$h_x = 40$		$h_x = 100$	
	No censoring	50% censoring	No censoring	50% censoring
Covariate order	18.42 (0.28)	28.80 (0.38)	9.08 (0.17)	15.32 (0.27)
Local likelihood	27.36 (1.15)	41.91 (1.38)	11.49 (0.25)	19.73 (0.37)
Local regression	45.52 (3.68)	78.72 (4.48)	18.74 (0.48)	39.46 (0.98)
	$h_x = 60$		$h_x = 150$	
	No censoring	50% censoring	No censoring	50% censoring
Covariate order	14.03 (0.20)	21.43 (0.28)	6.96 (0.11)	11.82 (0.19)
Local likelihood	18.96 (0.41)	29.72 (0.58)	9.19 (0.22)	16.14 (0.31)
Local regression	29.11 (0.68)	49.70 (1.29)	14.09 (0.33)	28.69 (0.58)
	$h_x = 80$		$h_x = 200$	
	No censoring	50% censoring	No censoring	50% censoring
Covariate order	12.73 (0.13)	18.12 (0.21)	6.51 (0.09)	10.75 (0.16)
Local likelihood	15.05 (0.28)	23.62 (0.41)	7.53 (0.15)	13.71 (0.26)
Local regression	23.78 (0.56)	40.40 (0.88)	12.13 (0.25)	25.01 (0.47)

Table 1: Estimated MISE-values and sample standard errors for the estimates based on 5000 different sets of data. In the two left columns, each set of data consisted of 200 observations and the data were simulated using hazard rate (13). In the two right columns, each set of data consisted of 100 observations and the data were simulated using hazard rate (14).

MISE-values and the sample standard errors of these estimates for each of the methods are displayed in Table 1. This table shows that the covariate order method is significantly better than the other two methods in all cases considered, while the local likelihood method is the second best method. It is also interesting to note that the sample standard errors of the MISE estimates are smallest for the covariate order method, bigger for the local likelihood method, and biggest for the local regression method.

5.1.3 Heavy Censoring

Ten samples of 100 observations drawn from hazard rate (14) with approximately 80% censoring has been simulated. The estimated hazard rates are displayed in Figure 7. Notice the difference in scale on the vertical axis compared to Figure 6. We see that the regression method gives very poor estimates in this extremely heavily censored case with only around 20 uncensored observations, while the local likelihood and the covariate order method give better estimates.

5.1.4 Two Covariates

An example with two covariates is given to demonstrate the performance of the algorithm presented in Section 4. Ten different sets of data, each consisting of 200 observations, are simulated using hazard rate (15). In the censoring distribution (16), $V = 4$ was used, corresponding to around 20% censoring in each of the datasets. The algorithm presented in Section 4 was used for all methods. For each iteration of step 5 in the algorithm the optimal

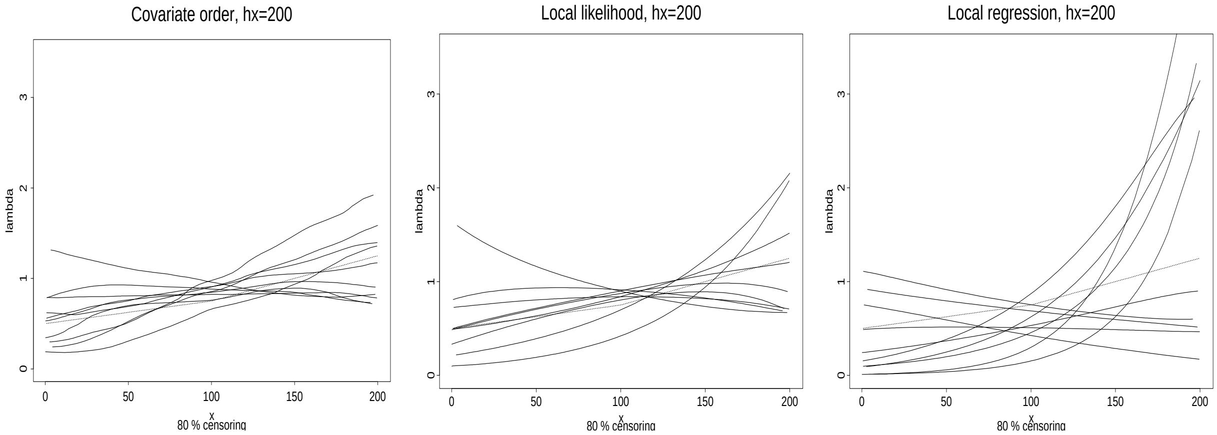

Figure 7: Ten estimated curves for each of the three estimation methods. Each estimated curve is based on 100 observations simulated using (14) and 80% censoring. The true underlying curve is dashed.

smoothing parameter according to the appropriate cross-validation criterion was used. For the covariate order method and the local likelihood method the smoothing parameter for x_1 was typically chosen to be around 0.25, and the smoothing parameter for x_2 to be around 0.85. For the local regression method the smoothing parameter chosen for x_1 varied between 0.35 and 0.9, while the smoothing parameter for x_2 typically was around 1.5.

All curve estimates are displayed in Figure 8. We see that the covariate order method and the local likelihood method in most cases capture the main features of the data fairly satisfactorily, while the regression based method again perform worse than the other two methods, especially in estimating the function of the first covariate. The main problem here is that the cross-validation criterion (11) tends to choose a far too large smoothing parameter. The covariate order method and the local likelihood method seem to behave rather similarly.

Repeating the above simulations 5000 times for some fixed values of the smoothing parameters and calculating MISE-values gave the results presented in Table 2. It is seen that the covariate order method also in this example, with two covariates, performs significantly better than the two other methods in terms of mean integrated squared error.

	$h_{x1} = 0.2$, $h_{x2} = 0.7$	$h_{x1} = 0.25$, $h_{x2} = 0.9$	$h_{x1} = 0.3$, $h_{x2} = 1.1$
Covariate order	16.92 (0.03)	15.78 (0.03)	15.02 (0.02)
Local likelihood	17.78 (0.04)	16.42 (0.03)	15.48 (0.03)
Local regression	19.10 (0.06)	17.45 (0.05)	16.29 (0.04)

Table 2: Estimated MISE-values and sample standard errors for the estimates based on 5000 different sets of data. Each set of data consisted of 200 observations, and the data were simulated using hazard rate (15) and 20% censoring.

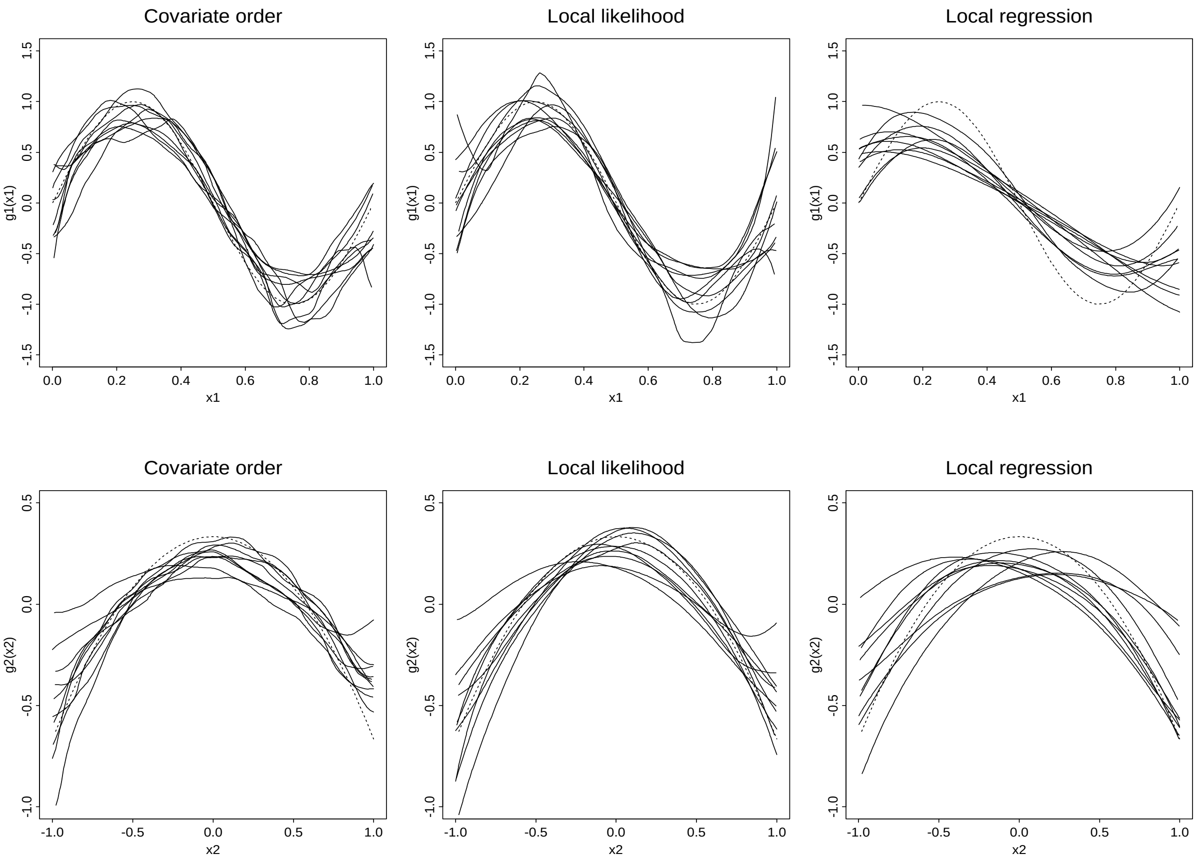

Figure 8: Ten sets of data, each consisting of 200 observation with approximately 20% censoring is simulated using a two-dimensional hazard rate. The estimates of each part of this hazard rate is displayed for each method. The estimates are for all methods obtained by using the algorithm in Section 4. The true underlying curves are dashed. Notice the difference in scale on the vertical axis compared to Figure 6

5.2 More on the Covariate Order Method

5.2.1 The Correspondence Function

Recall from Section 2 the definitions of $\tilde{s}(x)$ and $\hat{s}(x)$. The motivation for introducing the correspondence function estimate $\hat{s}(x)$ was to improve the smoothness of the estimated hazard rates compared to the estimates obtained using $\tilde{s}(x)$. An illustration of the difference is given in Figure 9 where the left and the middle plot show the estimates obtained by using respectively $\tilde{s}(x)$ and $\hat{s}(x)$ in (2). The main difference between these two curves is as expected that the left curve is much rougher than the middle curve. If we calculate the estimated MISE-values presented in Table 1 using $\tilde{s}(x)$ instead of $\hat{s}(x)$ we get the estimates displayed in Table 3. We see that the differences in MISE-values between Table 3 and Table 1 are small and not significant.

The plot on the right hand side in Figure 9 is the correspondence function plot for the data

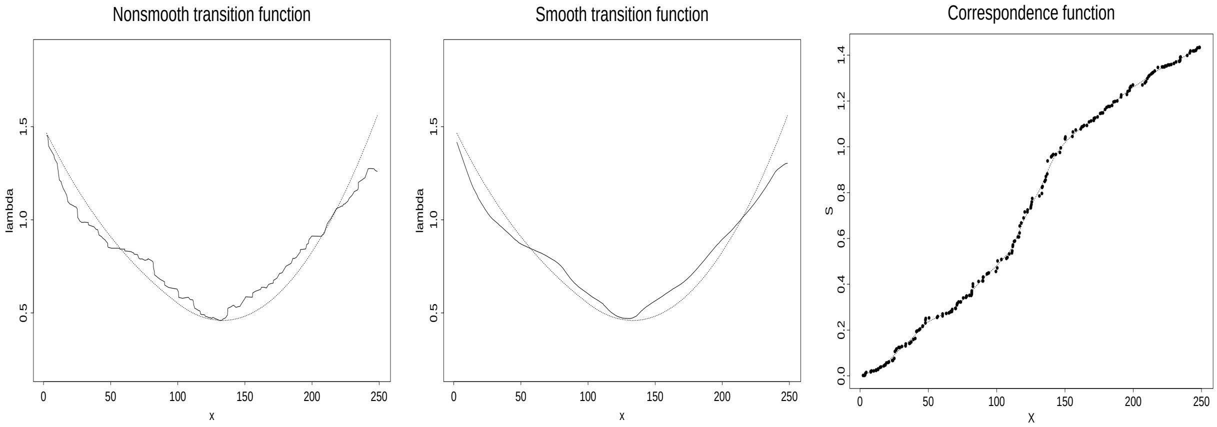

Figure 9: The left plot show a curve estimate obtained using the covariate order method and a nonsmooth correspondence function. The middle plot show the same estimate when a smooth correspondence function is used, and the correspondence function is displayed in the right plot. The estimates are based on 200 observations simulated using (13) and no censoring.

h_x	No censoring	50% censoring	h_x	No censoring	50% censoring
40	18.72 (0.29)	28.78 (0.40)	100	8.78 (0.16)	14.88 (0.25)
60	14.00 (0.18)	21.19 (0.27)	150	6.84 (0.10)	11.64 (0.18)
80	12.72 (0.13)	17.96 (0.20)	200	6.49 (0.08)	10.68 (0.16)

Table 3: Estimated MISE-values and sample standard errors for the estimates based on 5000 different sets of data. The covariate order method with no smoothing of the correspondence function is used. In the left columns each set of data consisted of 200 observations and the data were simulated using hazard rate (13). In the right columns each set of data consisted of 100 observations and the data were simulated using hazard rate (14).

which the curve estimates in the right and middle plots are based on. The dots are the points $(X_1, T_1/n), \ldots, (X_n, \sum_{i=1}^n T_i/n)$, while the dashed line is the smoothed curve $\hat{s}(x)$. We know from Theorem 2.1 that since there is no censoring in this example, and the covariates are uniformly distributed on [0,250], then asymptotically $s'(x) = 1/(250\lambda(x))$. We see from the estimated curves that in fact $\hat{s}'(x) \sim 1/(250\hat{\lambda}(x))$ in accordance with this.

The plot of the points $(X_1, T_1/n), \ldots, (X_n, \sum_{i=1}^n T_i/n)$ can also be used as a plot for detecting outliers. Especially outliers in the covariate, but also very large T-values are easily seen on this plot. Because of the relationship between the $\hat{s}(x)$-curve and the $\hat{\lambda}(x)$-curve, we can from such a plot also see in which direction an outlying observation drag the $\lambda(x)$-curve. Moreover we see in which regions the $\lambda(x)$-curve possibly depend on few and/or outlying observations.

5.2.2 Constant Smoothing Parameter on the s-axis

This far in the simulation study the smoothing parameter on the covariate axis, h_x, has been held constant for consistency with the other methods. Now some simulations which instead uses a constant h_s are reported. To be able to do a fair comparison between results from simulations with constant h_x versus simulations with constant h_s these two parameters must in some sense be "equivalent". A reasonable way to accomplish this is to choose smoothing parameters that cover approximately the same amount of the data.

By choosing constant smoothing parameters on the s-axis that cover approximately the same amount of data as the constant smoothing parameters on the covariate axes used in the estimates displayed in Figures 5 and 6, the estimates shown in Figure 10 are obtained. The

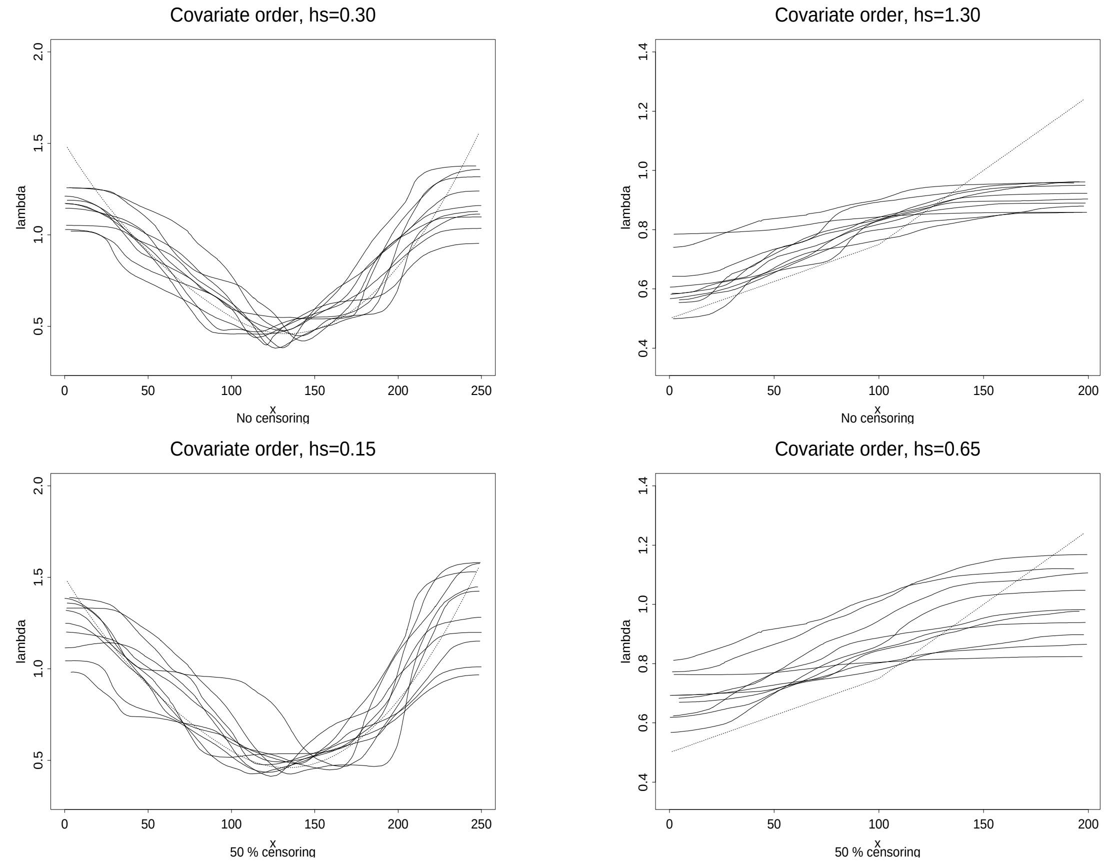

Figure 10: The covariate order method estimates in Figures 5 and 6 are repeated using a constant smoothing parameter on the s-axis instead of on the covariate axis. The true underlying curves are dashed.

estimates in Figures 5 and 6 and in Figure 10 are based on exactly the same data. If we compare the two left plots in Figure 10 with the corresponding plots for the covariate order method in Figure 5, we see that the curves estimated with a constant h_s tend to be a little bit more variable in the middle of the plots (low intensity region) and clearly less variable towards the boundaries (high intensity). This is as expected from the discussion at the end of Section 2. We would also expect a reduced bias in the middle of the plots and increased bias near the boundaries. At least there seems to be a tendency of underestimation at the boundaries. Comparing the two right plots in Figure 10 with the corresponding plots in Figure 6, we see that the curve estimates with constant h_s tend to be seriously downward biased at the right boundary.

Estimated MISE-values based on 5000 datasets and using smoothing parameters corresponding to those in Table 1 are given in Table 4. This table indicates that using a constant h_s

h_s	No censoring	h_s	50% censoring	h_s	No censoring	h_s	50% censoring
0.20	11.80 (0.09)	0.10	19.76 (0.18)	0.65	6.47 (0.08)	0.33	11.05 (0.15)
0.30	12.61 (0.08)	0.15	17.08 (0.15)	0.98	7.66 (0.07)	0.49	10.81 (0.13)
0.40	15.78 (0.07)	0.20	17.90 (0.13)	1.30	10.42 (0.07)	0.65	12.90 (0.13)

Table 4: Estimated MISE-values and sample standard errors for the estimates when the covariate order method with a constant smoothing parameter on the s-axis is used. The estimates in each column are based on 5000 sets of data. In the left columns each set of data consisted of 200 observations and the data were simulated using hazard rate (13). In the right columns each set of data consisted of 100 observations and the data were simulated using hazard rate (14).

in many cases is better than using a constant h_x, especially when the smoothing parameters are small. The impression from Table 4 and Figure 10 is that the decrease in variance in the high intensity regions when using a constant h_s is larger than the increase in bias in these regions when the smallest smoothing parameters are used. For data simulated using hazard rate (13) using at constant h_s does in fact give smaller MISE-values in all cases except when the data is uncensored and the largest smoothing parameter is used. In this case and in the same case for data simulated using hazard rate (14), the covariate order method with constant h_s have MISE-values that are higher than the values for the local likelihood method in the same cases. As opposite to the case with constant h_x, the MISE-values tend to become larger for increasing values of the smoothing parameter when a constant h_s is used, indicating that bias is the main problem when using a constant h_s while variance is the main contributer to MISE when a constant h_x is used in these examples.

The problems with bias in high intensity regions and variance in low intensity regions when using a constant h_s is the same kind of problems as often encountered when using a constant smoothing parameter in density estimation. These problems can often be overcome by using a variable smoothing parameter, and the same approach can of course also be used in our problem. If such a method had been applied, the size of h_s in a certain point would have been decided from the density of observations in the neighborhood around this point on the s-axis. Whether any such approach would improve the method further compared to having a constant h_s, or compared to the variable h_s induced by having a constant h_x, has not yet been studied.

6 INFERENCE

6.1 Nonparametric Test for Covariate Effect

Recall the definition of the process $S_1, \ldots, S_{k(n)}$ displayed in Figure 1. If there is no covariate effect, that is $\lambda(x) \equiv \lambda$, then the process $S_1, \ldots, S_{k(n)}$ is a homogeneous Poisson process (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) and references therein. The observation made in Section 2 that the process $S_1, \ldots, S_{k(n)}$ can be considered as being approximately an NHPP, further motivated by the convergence of the scaled intensity of this process to a deterministic function (Theorem 2.1), suggests adopting tests for the HPP versus NHPP setting as tests for covariate effect. A detailed presentation of this approach for testing the covariate effect is given in Kvaløy (1998b). A short review of the recommended test is given below.

Among the numerous tests constructed for testing HPP versus NHPP Kvaløy and Lindqvist (1998) recommend using a test based on the Anderson-Darling statistic and called the Anderson-Darling test for trend, 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 2 the definitions $S = \sum_{i=1}^{n} T_i/n$ 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

$$AD = -\frac{1}{\hat{k}} \left[\sum_{i=1}^{\hat{k}} (2i-1)(\ln \frac{S_i}{S} + \ln(1 - \frac{S_{\hat{k}+1-i}}{S})) \right] - \hat{k} \tag{17}$$

The asymptotic null distribution of (17) was derived by Anderson and Darling (1952). On a 5% level the null hypothesis of no covariate effect is rejected if $AD \geq 2.492$.

The test can be generalized to allow the life times to have other distributions than the exponential by using bootstrapping. See Kvaløy (1998b) for details. This generalization requires the extra assumption that the censoring distribution is *independent* of the covariate, in other words random censorship. Simulations in Kvaløy (1998b) showed that there is very little loss in power by using the bootstrapped Anderson-Darling test for trend instead of the original test in cases with exponentially distributed data, and the bootstrapped test is of course a far more general test if the assumption of random censorship can be justified. Hence, the use of the bootstrapped Anderson-Darling test for trend is recommended unless the assumption of exponentially distributed life times is very indisputable.

6.2 Model Selection

It will often be of great value to have an inferential tool to help choose between different models for the hazard rate. In cases with only one covariate it may be of interest to compare a nonparametric estimate with a certain parametric estimate, for example to check if a particular feature not captured by the parametric curve is a true feature of the data or not. In cases with several covariates it will typically be of interest to compare models with a different number of covariates included. If a model of the form (12) is assumed, the effect of each covariate can be tested by using the nonparametric test in the final pass of step 5 of the algorithm in Section 4. But to check if a certain covariate should be included in the model or not is not fully supported by the nonparametric test since this test only tests one covariate at a time and is unable to reveal correlations among the covariates. Even if two highly correlated covariates both are significant according to the covariate effect test, still probably only one of them should be included in the model. For less correlated covariates it is less clear if both should be included or not.

For parametric models, deviance tests can be used to decide which covariates should be included and which model is best. Let $\mathbf{x} = (x_1, \ldots, x_m)$ and let $\lambda(\mathbf{x}, \boldsymbol{\theta})$ be a certain parametric model for the hazard rate. If $\lambda(\mathbf{x}, \boldsymbol{\theta}')$ is another model, and these two models are nested and estimated using maximum likelihood, the two fits can be compared by comparing the values of the log likelihood functions and the number of parameters for the estimated models. Let $p = \dim(\boldsymbol{\theta})$ and $q = \dim(\boldsymbol{\theta}')$. If $q < p$ the decrease in deviance, $2(\ln L(\hat{\boldsymbol{\theta}}) - \ln L(\hat{\boldsymbol{\theta}}'))$, by fitting the larger model is asymptotically χ^2_{p-q}-distributed. The additional number of parameters in the bigger model equals the degree of freedom (df) of the corresponding deviance decrease.

When a smooth nonparametric estimate is used this does not directly apply. It is for instance not immediately clear how many degrees of freedom a nonparametric fit is using. One way to approach this problem is to use the ideas in Tibshirani (1984) of simulating the distribution of differences in the likelihood. Let $\lambda_1(\mathbf{x}')$ be one model and $\lambda_2(\mathbf{x}'')$ a bigger model which contains the smaller model. That is $\mathbf{x}' \subset \mathbf{x}''$. Typically $\lambda_1(\mathbf{x}')$ is the null model (constant hazard rate), that is $\lambda_1(\mathbf{x}') = \lambda_0$, and if so the estimate $\hat{\lambda}_1(\mathbf{x}') = \hat{\lambda}_0$ is the maximum likelihood estimate for a parametric model with constant intensity. Otherwise the estimates of $\lambda_1(\mathbf{x}')$ and $\lambda_2(\mathbf{x}'')$ are both calculated using a nonparametric method for exponential regression. The observed decrease in deviance by fitting the larger model is $R = 2(\ln L(\hat{\lambda}_2(\mathbf{x}'')) - \ln L(\hat{\lambda}_1(\mathbf{x}')))$. The degrees of freedom of the bigger model versus the smaller model is defined as the expectation of R under the null hypothesis that the smaller model is correct.

Now $\mathbf{E}R$ can be estimated by simulations. The idea presented in Tibshirani (1984) is as follows. For the observed data $(T_1, \delta_1, \mathbf{X}_1), \ldots, (T_n, \delta_n, \mathbf{X}_n)$ calculate $\hat{\lambda}_1(\mathbf{x}')$, and estimate the cumulative censoring distribution function. If the censoring can be assumed to be independent of the covariate this can for example be done by using the Kaplan-Meier estimator on $(T_1, 1-\delta_1), \ldots, (T_n, 1-\delta_n)$. From this estimated model first draw a new sample $(T_1^*, \delta_1^*, \mathbf{X}_1), \ldots, (T_n^*, \delta_n^*, \mathbf{X}_n)$. Here $T^* = \min(C^*, Z^*)$ where Z^* is exponentially distributed with parameter $\hat{\lambda}_1(\mathbf{X}')$, C^* is drawn from the estimated censoring distribution and $\delta^* = I(Z^* \leq C^*)$. Then calculate new estimates $\hat{\lambda}_1^*(\mathbf{x}')$ and $\hat{\lambda}_2^*(\mathbf{x}'')$ based on this new sample. The estimates $\hat{\lambda}_1^*(\mathbf{x}')$ and $\hat{\lambda}_2^*(\mathbf{x}'')$ must be calculated using exactly the same methods and smoothing parameters as in the calculations of $\hat{\lambda}_1(\mathbf{x}')$ and $\hat{\lambda}_2(\mathbf{x}'')$. Finally calculate $R^* = 2(\ln L(\hat{\lambda}_2^*(\mathbf{x}'')) - \ln L(\hat{\lambda}_1^*(\mathbf{x}')))$. Repeat this N times to obtain the sample $R_1^*, \ldots, R_N^*$. Then $\hat{\mathrm{df}} = \sum_{i=1}^N R_i^*/N$. Moreover, an estimate of the null distribution of R which may be better than the χ^2-approximation can be found by estimating the density of $R_1^*, \ldots, R_N^*$.

6.3 Graphical Model Check

If Z is exponentially distributed with parameter $\lambda(\mathbf{X})$, then the residual $e = Z\lambda(\mathbf{X})$ is exponentially distributed 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 $\hat{e}_i = T_i\hat{\lambda}(\mathbf{X}_i)$ should be 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, $\hat{F}_e(t)$, from the sample $(\hat{e}_1, \delta_1), \dots, (\hat{e}_n, \delta_n)$, and plot $\log(1 - \hat{F}_e(t))$ against t. This should yield approximately a straight line with slope -1 if the model is good.

6.4 Bootstrapping the model

Bootstrapping can be used to get a picture of the variability of the estimated hazard rate, and to provide information on whether certain features of the estimated function is true features of the data or just random noise.

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

Another way is based on resampling the generalized residuals $e_i = T_i\lambda(\mathbf{X}_i)$, $i = 1, \dots, n$. The estimated residuals are $\hat{e}_i = T_i\hat{\lambda}(\mathbf{X}_i)$, $i = 1, \dots, n$, and the idea is to resample with replacement from $(\hat{e}_1, \delta_1), \dots, (\hat{e}_n, \delta_n)$ and add the new residual sample $(\hat{e}_1^*, \delta_1^*), \dots, (\hat{e}_n^*, \delta_n^*)$ back to the model. That is, the bootstrap sample is $(T_1^*, \delta_1^*, \mathbf{X}_1), \dots, (T_n^*, \delta_n^*, \mathbf{X}_n)$ where $T_i^* = \hat{e}_i^*/\hat{\lambda}(\mathbf{X}_i)$. This way of bootstrapping the model assumes that the model is correct.

A third way to bootstrap the model which also assumes that the model is correct and which in addition requires estimation of the censoring distribution is as follows. Draw n new life time observations $Z_1^*, \dots, Z_n^*$ where Z_i^* is drawn from the exponential distribution with parameter $\hat{\lambda}(\mathbf{X}_i)$. Similar draw n new censoring observations $C_1^*, \dots, C_n^*$. This gives the bootstrap sample $(T_1^*, \delta_1^*, \mathbf{X}_1), \dots, (T_n^*, \delta_n^*, \mathbf{X}_n)$. By estimating the covariate distribution this approach can also be extended to include resampling of the covariates.

7 EXAMPLE: CARDIAC ARREST

Times of out-of-hospital cardiac arrest reported to a Norwegian hospital over a 5 years period have been recorded. Previous analysis of essentially the same data in Skogvoll (1997) concludes that cardiac arrests occur as a homogeneous Poisson process. In the present example the relationship between the intensity of cardiac arrests of primary cardiac origin, and outdoor air temperature is investigated. This is done by regarding inter-event times to be independent and exponentially distributed with an intensity $\lambda(x)$ depending on the temperature x. Assuming temperature to be varying relatively slowly this seems to be a reasonable model, although a nonhomogeneous Poisson process model may seem more direct. The temperature is recorded three times a day at a local weather station, and the temperature recordings closest in time to each reported cardiac arrest is used as covariate x.

In Kvaløy (1998b) the significance of a number of weather covariates possibly related to the intensity of cardiac arrests is tested. The conclusion of that study is that the only covariate found to be significant is temperature. Using the bootstrapped Anderson-Darling test for trend (Section 6.1), the p-value of the covariate effect of temperature is 0.002.

Plots of the estimated hazard rate as a function of temperature, using each of the three methods and with smoothing parameters decided by the corresponding cross-validation criteria are displayed in Figure 11. We see that the covariate order method and the local likelihood method give very similar estimated hazard rate curves, while the curve obtained using the

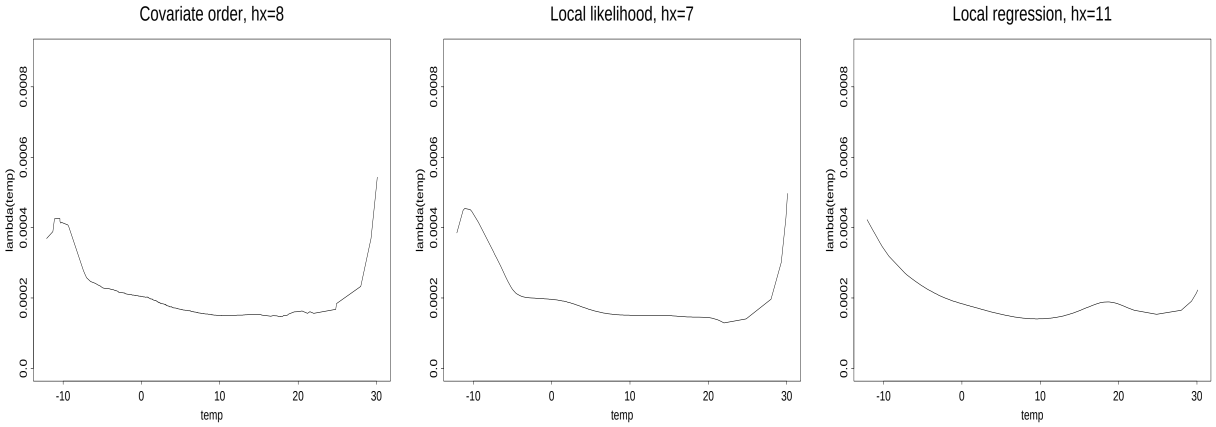

Figure 11: Estimated hazard rate functions for the cardiac arrest data using each of the three methods.

regression based method looks a bit different. In particular, the sharp increase in hazard rate for very high temperatures is not seen in the local regression estimate. Otherwise all curves indicate an increased cardiac arrest intensity at very low temperatures, and then a decreasing intensity for increasing temperatures until a certain point.

In Figure 12 some additional plots for the covariate order method are displayed. The correspondence function plot reflects the shape of the estimated hazard rate in a similar manner as in Section 5.2.1. We also see from this plot that the number of observations at the boundaries is small. In other words, the variance of the estimated hazard rate curve at the boundaries is large. This is reflected by the bootstrap curves. The bootstrap curves are obtained by resampling observations, and they indicate that there is a large variability at the boundaries, while there is very little variability in the estimated hazard rate in the middle temperature range where most of the observations are located. The bootstrap curves seem to indicate that the increased frequency in reported cardiac arrests for decreasing temperature is a true feature of the data. The increased frequency at very high temperatures also seems to be true, but the variability in this part of the curve is large, indicating that the steepness of this increase is unclear.

In the bottom left plot in Figure 12, the hazard rate is estimated using the covariate order method with a constant h_s. From the correspondence function plot we realize that the value $h_s = 3100$ chosen by (5) corresponds to smoothing over half of the data, but choosing a considerably smaller value of h_s does not change the estimate curve much. This is an example of a situation in which using a constant smoothing parameter on the s-axis is not such a good idea. The increased hazard rate at the boundaries, which according to the bootstrap curves seems to be a true feature of the data, is completely masked when using a constant h_s. As discussed at the end of Section 2, using a constant h_s tends to increase bias in regions of the covariate space with few observations and in regions with high hazard rate. In the present case both these effects add up to produce an estimated curve which is very biased, or oversmoothed, at the boundaries. Only the decreasing cardiac arrest intensity in the middle of the temperature range is reflected in the estimate with constant h_s.

The model check plot in the bottom right plot in Figure 12 indicates a very good fit in the

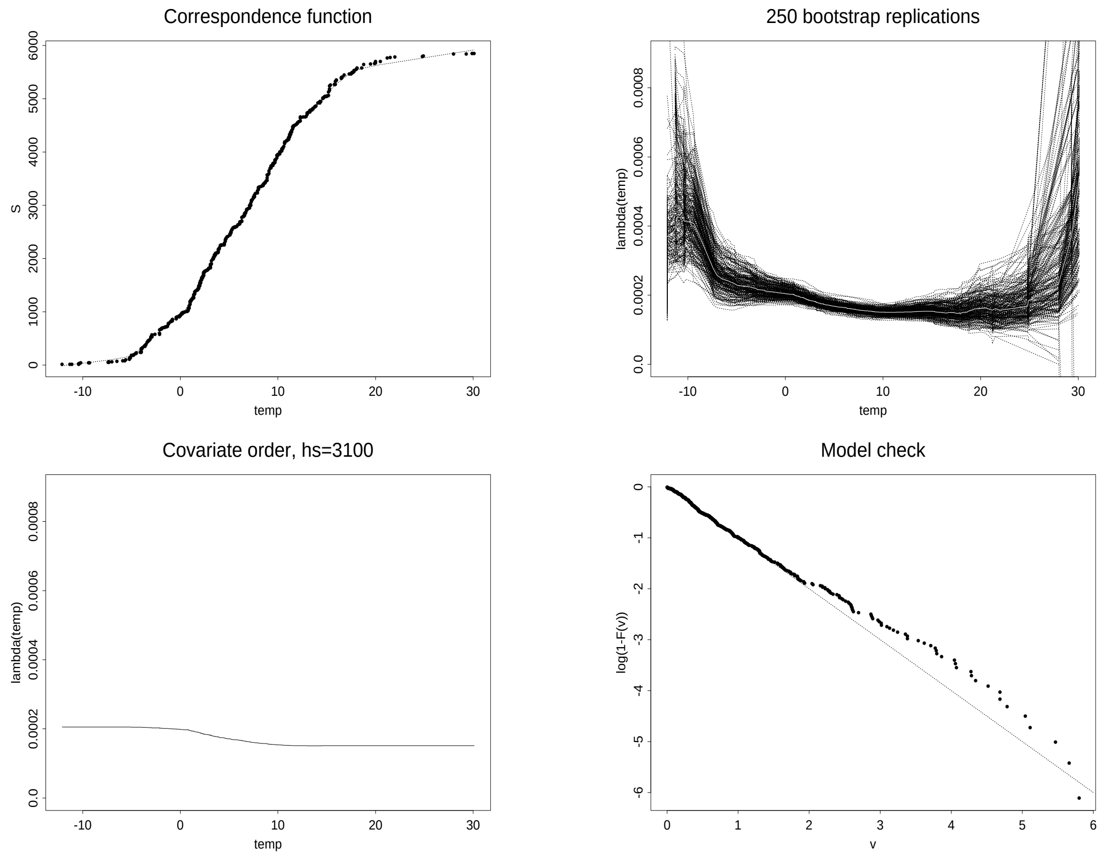

Figure 12: Various plots for the covariate order method. A correspondence function plot, 250 bootstrap curves constructed by resampling observations (original estimate shown as the white curve), estimated hazard rate for the cardiac arrest data using the covariate order method with a constant h_s and a model check plot.

lower tail where most of the observations are located, while there are some not too serious deviations from the straight line in the upper tail.

The degrees of freedom of the estimated hazard rate compared to a null model with constant hazard rate was found to be 5.40 by simulations as explained in Section 6.2, using 10000 replications. The estimated variance of the decrease in deviance was 9.40, indicating that the distribution of the decrease in deviance does not deviate too much from a χ^2-distribution. The observed decrease in deviance was 17.69, and calculating the p-value of the estimated model using the empirical distribution of the simulated decreases in deviance gave the p-value 0.0027. If instead the simple parametric model $\lambda(x) = \exp(a + bx)$ is used, the observed decrease in deviance is only 9.39 compared to the null model, but this model is only using one degree of freedom. The p-value of this model is 0.0024. Using Akaike's Information Criterion, *AIC=-2*maximized log-likelihood+2*#parameters*, as a criterion to decide which model is best gave *AIC*-values of respectively 8681.3 and 8680.8, indicating no practical differences between the parametric and the nonparametric model in this example.

8 CONCLUSIONS

The simulation study in Section 5 shows that, at least for the cases studied, the covariate order method for nonparametric censored exponential regression is better than the competing methods. These results are obtained when using a constant smoothing parameter on the covariate axis. In addition, the covariate order method has the flexibility of possibly choosing a smoothing parameter according to the values of the observation times, and simulations indicated that choosing a constant smoothing parameter on the s-axis often improved the performance of the covariate order method further. The example in Section 7 is, however, an example of a situation in which using a constant smoothing parameter on the s-axis is not suitable.

The main reason for the success of the covariate order method in the simulation study seems to be that the explicit boundary correction used in this method is better in practice than the automatically incorporated boundary corrections in the local linear methods. The differences between the covariate order method and the local likelihood method are mainly seen at the boundaries. The regression based method is more variable both at the boundaries and in the interior.

It should be noted that neither the full potential of the covariate order nor of the local methods has been exploited in the simulation study. For the covariate order method, the simple kernel density estimator could have been replaced by more sophisticated density estimation methods, and other methods than the reflection method for boundary corrections could have been tried. One such method has in fact been tried, the pseudo-data method of Cowling and Hall (1996). We did, however, as mentioned in Section 2, encounter some problems in practical use of this method, and we also observed that the reflection method in many cases gave better results. The obvious way of improving the local polynomial methods is by using a higher order polynomial approximation. This will reduce the bias of the estimated curve, but will also increase the variability by introducing more local parameters. Fan and Gijbels (1996) argue that odd order fits are generally preferable, implying that at least a third order fit should be used to possibly improve the performance of the local linear fits used in the simulation study. This does though require a considerable increase in the computational costs. Instead of exploring the possible but probable minor improvements of the methods by the various above mentioned refinements, it seems more interesting and reasonable to compare the easily programmed simple basic versions of the different methods as was done in the simulation study.

The likelihood cross-validation criterion (5) is somewhat criticized in the literature for sometimes having undesirable practical behavior, but have shown quite satisfying performance in all applications in the work with this paper. On the other hand, the cross-validation criterion (11) used in the local regression method has in several situations not been able to suggest reasonable smoothing parameters, for example in the simulated example in Section 5.1.4. It is of future research to study other, and possibly find better, ways of choosing the smoothing parameters.

Doing inference in nonparametric models is in general more cumbersome than in classical parametric models, but it is demonstrated in Sections 6 and 7 that there is a number of useful techniques at hand for doing inference in nonparametric models as well.

Acknowledgement

The authors would like to thank Ørnulf Borgan for help with the proof of Theorem 2.2, and Eirik Skogvoll for providing the cardiac arrest data. Jan Terje Kvaløy was funded by a Dr.ing. 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). *Reipairable Systems Reliability. Modeling, Inference, Misconceptions and Their Causes*, Marcel Dekker, Inc., New York.

Cowling, A. and Hall, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation, *Journal of the Royal Statistical Society, Series B* **58**(3): 551–563.

Fan, J. and Gijbels, I. (1994). Censored regression: Local linear approximations and their applications, *Journal of the American Statistical Association* **89**(426): 560–570.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Monographs on Statistics and Applied Probability, London: Chapman & Hall.

Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression, *The Annals of Statistics* **25**(4): 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**(420): 942–951.

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

Karr, A. F. (1993). *Probability*, Springer Texts in Statistics, New York: Springer-Verlag.

Kvaløy, J. T. (1998a). Nonparametric estimation in Cox-models., *Technical report*, Norwegian University of Science and Technology, Department of Mathematical Sciences. Under preparation.

Kvaløy, J. T. (1998b). Nonparametric 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. (1998). TTT-based tests for trend in repairable systems data, *Reliability Engineering and System Safety* **60**(1): 13–28.

LINDQVIST, B. AND TJELMELAND, H. (1989). An exponential regression model for censored failure data: Estimation and graphical model checking, *in* T. Aven (ed.), *Reliability Achievement: The Commercial Incentive*, Elsevier Applied Science, London, pp. 192–202.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized Linear Models*, Monographs on Statistics and Applied Probability, London: Chapman & Hall.

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

SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*, Springer Series in Statistics, New York: Springer-Verlag.

SKOGVOLL, E. (1997). Occurrence of out-of-hospital cardiac arrest modelled by the Poisson and exponential probability distributions., *Acta Anaesthesiologica Scandinavica* **41**: 194.

STANISWALIS, J. G. (1989). The kernel estimate of a regression function in likelihood-based models, *Journal of the American Statistical Association* **84**(405): 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**(398): 559–567.

A PROOFS

A.1 Proof of Theorem 2.1

In this proof and in the proof of Lemma 2.1, the Glivenko-Cantelli theorem, and the Chebychev, Markov and Cauchy-Schwarz inequalities will be used repeatedly. See Karr (1993) for a general reference.

First we assume no censoring, which implies that $S_i - S_{i-1}$ is exponentially distributed with parameter $n\lambda(X_i)$ for $i = 1, \ldots, n$. Let $N_n(s)$ be the counting process that counts events on the s-axis, that is $N_n(s) = \sum_{i=1}^n I(S_i \leq s)$. The conditional intensity of this process is given by, see for example Andersen, Borgan, Gill and Keiding (1993, p. 75), the predictable process

$$\begin{aligned}
\rho_n(s|\mathcal{F}_s^n) &= \lim_{\Delta s \to 0} \frac{P(\Delta N_n(s)|\mathcal{F}_s^n)}{\Delta s} \\
&= \lim_{\Delta s \to 0} \frac{P(s - S_{N_n(s)} < \frac{1}{n} T_{N_n(s)+1} \leq s + \Delta s - S_{N_n(s)}|\mathcal{F}_s^n)}{\Delta s} \\
&= n\lambda(X_{N_n(s)+1})
\end{aligned}$$

To prove uniform convergence of $\rho_n(s|\mathcal{F}_s^n)/n$ we first consider the case when $X_1, \ldots, X_n$ are independent and identically uniformly distributed on $[0, 1]$. Let a point w on the s-axis be fixed in the following, and define I, I_0, I_1 and $\eta(w)$ by the following relations.

$$\begin{aligned}
I &: \quad S_{I-1} \leq w < S_I \\
I_0 &: \quad \textstyle\sum_{i=1}^{I_0 - 1} \frac{1}{n\lambda(X_i)} \leq w < \sum_{i=1}^{I_0} \frac{1}{n\lambda(X_i)} \\
I_1 &: \quad \textstyle\sum_{i=1}^{I_1 - 1} \frac{1}{n\lambda(\frac{i}{n+1})} \leq w < \sum_{i=1}^{I_1} \frac{1}{n\lambda(\frac{i}{n+1})} \\
\eta(w) &: \quad \textstyle\int_0^{\eta(w)} \frac{1}{\lambda(v)} dv = w
\end{aligned} \tag{18}$$

In particular $I = N_n(w) + 1$, see Figure 13 for an illustration. To prove that $\rho_n(w|\mathcal{F}_w^n)/n =$

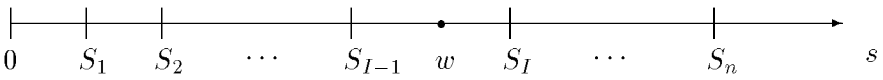

Figure 13: Definition of I.

$\lambda(X_I) \xrightarrow{p} \lambda(\eta(w))$ uniformly in w, it is by the uniform continuity of $\lambda(\cdot)$ sufficient to prove that $X_I \xrightarrow{p} \eta(w)$ uniformly. By the triangle inequality,

$$\begin{aligned}
|X_I - \eta(w)| &\leq |X_I - \frac{I}{n+1}| + |\frac{I}{n+1} - \frac{I_0}{n+1}| + |\frac{I_0}{n+1} - \frac{I_1}{n+1}| + |\frac{I_1}{n+1} - \eta(w)| \\
&\equiv A_1 + A_2 + A_3 + A_4
\end{aligned}$$

so it is enough to show that each of $A_1, \ldots, A_3 \xrightarrow{p} 0$ and $A_4 \to 0$ (A_4 is deterministic) uniformly.

($A_1 \xrightarrow{p} 0$): This follows by the Glivenko-Cantelli theorem which states that if F_n is the empirical distribution function based on n i.i.d. observations from F, then $\sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0$. Letting $F(x) = x$; $0 \leq x \leq 1$, we have $F(X_i) = X_i$, while $F_n(X_i) = \frac{i}{n}$. Thus, $|X_I - \frac{I}{n}| = |F(X_I) - F_n(X_I)| \leq \sup_x |F(x) - F_n(x)| \xrightarrow{a.s.} 0$, which implies that $A_1 \xrightarrow{p} 0$ uniformly.

($A_2 \xrightarrow{p} 0$): Let $d \geq 2$ be an integer. Then

$$\begin{aligned}
P(I \geq I_0 + d | x_1, \ldots, x_n) &= P(S_{I_0+d-1} \leq w | x_1, \ldots, x_n) \\
&\leq P(S_{I_0+d-1} \leq \sum_{i=1}^{I_0} \frac{1}{n\lambda(x_i)} | x_1, \ldots, x_n) \\
&\leq P(|S_{I_0+d-1} - \sum_{i=1}^{I_0+d-1} \frac{1}{n\lambda(x_i)}| \geq \sum_{i=I_0+1}^{I_0+d-1} \frac{1}{n\lambda(x_i)} | x_1, \ldots, x_n) \\
&\overset{Chebychev}{\leq} \frac{\sum_{i=1}^{I_0+d-1} \frac{1}{n^2\lambda(x_i)^2}}{\left(\sum_{i=I_0+1}^{I_0+d-1} \frac{1}{n\lambda(x_i)}\right)^2} \\
&\leq \frac{(I_0 + d - 1)/(n^2 a^2)}{(\frac{d-1}{nM})^2} \\
&= \frac{nM^2}{a^2(d-1)^2}
\end{aligned}$$

Since the upper bound on the conditional probability is not a function of $x_1, \ldots, x_n$ this implies that the inequality also holds for the unconditional probability $P(I \geq I_0 + d)$. By choosing $d = [n^{3/4}]$ we get $P(I \geq I_0 + [n^{3/4}]) \leq cn^{-1/2}$ for a suitable constant c. A similar calculation gives $P(I \leq I_0 - [n^{3/4}]) \leq cn^{-1/2}$.

Hence

$$P(|\frac{I}{n+1} - \frac{I_0}{n+1}| \leq \frac{[n^{3/4}]}{n+1}) \geq 1 - \frac{c}{\sqrt{n}}.$$

so $|\frac{I}{n+1} - \frac{I_0}{n+1}| \xrightarrow{p} 0$ uniformly in w.

($A_3 \xrightarrow{p} 0$): A key step in the following is the observation that since $\lambda'(x) \leq D$, $|\{\frac{1}{\lambda(x)}\}'| \leq \frac{D}{a^2} = C$, which implies $|\frac{1}{\lambda(x_1)} - \frac{1}{\lambda(x_2)}| \leq C|x_1 - x_2|$. Also recall that if X_i is the ith order statistic of n independent identically uniformly distributed variables on [0,1], then $\mathbf{V}ar(X_i) = \frac{i(n-i+1)}{(n+1)^2(n+2)} \leq \frac{1}{4(n+2)}$. Thus for an integer d,

$$\begin{aligned}
P(I_0 > I_1 + d) &= P(\sum_1^{I_1+d} \frac{1}{n\lambda(X_i)} < w) \\
&\leq P(\sum_1^{I_1+d} \frac{1}{n\lambda(X_i)} < \sum_1^{I_1} \frac{1}{n\lambda(\frac{i}{n+1})}) \\
&\leq P(|\sum_1^{I_1+d} (\frac{1}{n\lambda(X_i)} - \frac{1}{n\lambda(\frac{i}{n+1})})| > \sum_{I_1+1}^{I_1+d} \frac{1}{n\lambda(\frac{i}{n+1})}) \\
&\overset{Markov}{\leq} \frac{\mathbf{E}|\sum_1^{I_1+d}(\frac{1}{n\lambda(X_i)} - \frac{1}{n\lambda(\frac{i}{n+1})})|}{\sum_{I_1+1}^{I_1+d} \frac{1}{n\lambda(\frac{i}{n+1})}}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{\frac{C}{n}\sum_1^{I_1+d}\mathbf{E}|X_i - \frac{i}{n+1}|}{\frac{d}{nM}} \\
&\stackrel{Cauchy-Schwarz}{\leq} \frac{CM\sum_1^{I_1+d}\sqrt{\mathbf{E}(X_i - \frac{i}{n+1})^2}}{d} \\
&= \frac{CMn}{2d\sqrt{n+2}}
\end{aligned}$$

Proving the parallel inequality for $P(I_0 < I_1 - d)$ and letting $d = [n^{3/4}]$ this implies that

$$P(|\frac{I_0}{n+1} - \frac{I_1}{n+1}| \leq \frac{[n^{3/4}]}{n+1}) \geq 1 - cn^{-1/4}$$

For a suitable constant c. Hence $|\frac{I_0}{n+1} - \frac{I_1}{n+1}| \xrightarrow{p} 0$ uniformly.

$(A_4 \to 0)$: Observe that $|\sum_{i=1}^{I_1}\frac{1}{n\lambda(\frac{i}{n+1})} - w| \leq \frac{1}{n\lambda(\frac{I_1}{n+1})} \leq \frac{1}{na}$ which implies that $\sum_{i=1}^{I_1}\frac{1}{n\lambda(\frac{i}{n+1})} \to w = \int_0^{\eta(w)}\frac{1}{\lambda(w)}dv$ uniformly. Note that $\eta(w)$ is uniquely defined since $\lambda(v) > 0$ for all v, and it follows that $\frac{I_1}{n+1} \to \eta(w)$ uniformly.

This completes the proof that $\rho_n(w|\mathcal{F})/n \xrightarrow{p} \lambda(\eta(w))$ uniformly in w in the case of uniformly distributed covariates on [0,1].

For covariates $X_1, \ldots, X_n$ drawn from a distribution $F_{\mathcal{X}}(\cdot)$, let $U_i = F_{\mathcal{X}}(X_i)$ be transformed covariates which are now independent and identically uniformly distributed on [0,1]. Further let $\lambda^\star(u) = \lambda(F_{\mathcal{X}}^{-1}(u))$ be the hazard rate of the transformed covariates. Then (18) gives

$$\int_0^{\eta^\star(w)}\frac{1}{\lambda^\star(u)}du = w$$

which by substituting $u = F_{\mathcal{X}}(x)$ and letting $\eta(w) = F_{\mathcal{X}}^{-1}(\eta^\star(w))$ can be written

$$\int_{F_{\mathcal{X}}^{-1}(0)}^{\eta(w)}\frac{1}{\lambda(x)}f_{\mathcal{X}}(x)dx = w \tag{19}$$

Replace $\eta(w)$ with x and w with $s(x)$ to get the result of Theorem 2.1.

For the proof in the censored case, it is useful to consider the process $S_1^*, \ldots, S_n^*$, where $S_j^* = \sum_{i=1}^{j}\frac{1}{n}T_i$. Let $N_n^*(s) = \sum_{i=1}^{n} I(S_i^* \leq s)$ be the counting process counting events in this process. Further, let $\mathcal{F}_s^{n^*}$ denote the history of the process $S_1^*, \ldots, S_n^*$, and let this history be defined similarly as $\mathcal{F}_s^n$ is defined for $S_1, \ldots, S_{k(n)}$. Clearly $\mathcal{F}_s^n \subseteq \mathcal{F}_s^{n^*}$. The intensity of the process $S_1, \ldots, S_{k(n)}$ conditional on the history $\mathcal{F}_s^{n^*}$ is

$$\begin{aligned}
\rho_n(s|\mathcal{F}_s^{n^*}) &= \lim_{\Delta s \to 0}\frac{P(\Delta N_n(s)|\mathcal{F}_s^{n^*})}{\Delta s} \\
&= \lim_{\Delta s \to 0} P(s - S_{N_n^*(s)}^* \leq Z_{N_n^*(s)+1} < s + \Delta s - S_{N_n^*(s)}^* \cap \\
&\qquad C_{N_n^*(s)+1} > s + \Delta s - S_{N_n^*(s)}^* | \mathcal{F}_s^{n^*})/\Delta s \\
&= \lim_{\Delta s \to 0}\frac{P(s - S_{N_n^*(s)}^* \leq Z_{N_n^*(s)+1} < s + \Delta s - S_{N_n^*(s)}^*)}{\Delta s P(Z_{N_n^*(s)+1} > s - S_{N_n^*(s)}^*)}\frac{P(C_{N_n^*(s)+1} > s + \Delta s - S_{N_n^*(s)}^*)}{P(C_{N_n^*(s)+1} > s - S_{N_n^*(s)}^*)} \\
&= n\lambda(X_{N_n^*(s)+1})
\end{aligned}$$

which is intuitive. Assume that it can be proved that $X_{N_n^*(s)+1} \xrightarrow{p} \eta(s)$ uniformly, implying that $\rho_n(s|\mathcal{F}_s^{n^*})/n = \lambda(X_{N_n^*(s)+1}) \xrightarrow{p} \lambda(\eta(s))$ uniformly. Since $\mathcal{F}_s^n \subseteq \mathcal{F}_s^{n^*}$ it follows from the innovation theorem, for example Andersen et al. (1993, p. 80), that

$$\rho_n(s|\mathcal{F}_s^n)/n = \mathbf{E}[\lambda(X_{N_n^*(s)+1})|\mathcal{F}_s^n]. \tag{20}$$

Using Markov's inequality we get

$$\begin{aligned} P(\,|\rho_n(s|\mathcal{F}_s^n)/n - \lambda(\eta(s))| > \gamma) &= P(\,|\mathbf{E}[\lambda(X_{N_n^*(s)+1}) - \lambda(\eta(s))|\mathcal{F}_s^n]| > \gamma) \\ &\overset{Markov}{\leq} \frac{1}{\gamma}\mathbf{E}(\,|\mathbf{E}[\lambda(X_{N_n^*(s)+1}) - \lambda(\eta(s))|\mathcal{F}_s^n]|\,) \\ &\leq \frac{1}{\gamma}\mathbf{E}(\mathbf{E}[\,|\lambda(X_{N_n^*(s)+1}) - \lambda(\eta(s))|\,|\mathcal{F}_s^n]) \\ &\leq \frac{1}{\gamma}\mathbf{E}[\,|\lambda(X_{N_n^*(s)+1}) - \lambda(\eta(s))|\,] \end{aligned}$$

It now easily follows by the boundedness of $\lambda(x)$ and the assumed uniform convergence of $\lambda(X_{N_n^*(s)+1})$ that $|\rho_n(s|\mathcal{F}_s^n)/n - \lambda(\eta(s))| \xrightarrow{p} 0$ uniformly in s.

It remains to prove that $X_{N_n^*(s)+1}$ really converges uniformly in probability. Since $T = \min(Z, C)$, given the covariate $X = x$, we have that

$$f_T(t|x) = f_C(t|x)\exp(-\lambda(x)t) + \lambda(x)\exp(-\lambda(x)t)(1 - F_C(t|x)). \tag{21}$$

With the assumption $0 < a \leq \lambda(x) \leq M < \infty$, for all x and the assumption that the censoring distribution for all x has finite first and second order moments, it follows from (21) that there exist numbers E_{min}, E_{max} and V_{max} such that

$$\begin{aligned} 0 < E_{min} \leq \quad \mathbf{E}(T|x) &\leq E_{max} < \infty, \quad \text{for all } x \\ 0 < \quad \mathbf{V}ar(T|x) &\leq V_{max} < \infty, \quad \text{for all } x. \end{aligned} \tag{22}$$

We proceed as in the proof for the uncensored case by assuming uniformly distributed covariates and choosing some fixed point w on the s-axis. Define I^*, I_0^*, I_1^* and $\eta(w)$ by

$$\begin{aligned} I^*: &\quad S^*_{I-1^*} \leq w < S^*_{I^*} \\ I_0^*: &\quad \textstyle\sum_{i=1}^{I_0^*-1} \frac{1}{n}\mathbf{E}(T|X_i) \leq w < \sum_{i=1}^{I_0^*} \frac{1}{n}\mathbf{E}(T|X_i) \\ I_1^*: &\quad \textstyle\sum_{i=1}^{I_1^*-1} \frac{1}{n}\mathbf{E}(T|\frac{i}{n+1}) \leq w < \sum_{i=1}^{I_1^*} \frac{1}{n}\mathbf{E}(T|\frac{i}{n+1}) \\ \eta(w): &\quad \textstyle\int_0^{\eta(w)} \mathbf{E}(T|v)dv = w \end{aligned}$$

Now similar as in the uncensored case

$$\begin{aligned} |X_I^* - \eta(w)| &\leq |X_I^* - \frac{I^*}{n+1}| + |\frac{I^*}{n+1} - \frac{I_0^*}{n+1}| + |\frac{I_0^*}{n+1} - \frac{I_1^*}{n+1}| + |\frac{I_1^*}{n+1} - \eta(w)| \\ &= A_1^* + A_2^* + A_3^* + A_4^*. \end{aligned}$$

What remains is to prove that each of $A_1^*, \ldots, A_3^* \xrightarrow{p} 0$ and $A_4^* \to 0$ uniformly.

$(A_1^* \xrightarrow{p} 0)$: Exactly the same proof as in the uncensored case.

$(A_2^* \xrightarrow{p} 0)$: Essential as in the uncensored case, but with the $\frac{1}{n\lambda(x_i)}$ replaced by $\frac{1}{n}\mathbf{E}(T|x_i)$ and using the bounds of (22).

$(A_3^* \xrightarrow{p} 0)$: Essential as before. Note that (21) and the assumption that $\lambda(x)$ and $f_C(t|x)$ have finite first derivative with respect to x, implies that $\mathbf{E}(T|x)$ also will have finite first derivative with respect to x. Then repeating the calculations from the uncensored case with $\frac{1}{\lambda(x)}$ replaced by $\mathbf{E}(T|x)$ gives the result.

$(A_4^* \to 0)$: The proof is the same as in the uncensored case. Notice that

$$\begin{aligned} w = \int_0^{\eta(w)} \mathbf{E}(T|v)dv &= \int_0^{\eta(w)} \left[\int_0^\infty t f_C(t|v)\exp(-\lambda(v)t)dt \right. \\ &\quad \left. + \int_0^\infty t\lambda(v)\exp(-\lambda(v)t)(1-F_C(t|v))dt\right] dv \\ &= \int_0^{\eta(w)} \left[\int_0^\infty (t\lambda(v) - F_C(t|v))\exp(-\lambda(v)t)dt\right] dv \qquad (23) \end{aligned}$$

This completes the proof in the censored case for the situation with uniformly distributed covariates on [0,1]. The generalization to covariates drawn from a general domain $\mathcal{X}$ can be done by the same arguments as before, and the generalization of (23) becomes

$$\int_{F_{\mathcal{X}}^{-1}(0)}^{\eta(w)} \left[\int_0^\infty (t\lambda(v) - F_C(t|v))\exp(-\lambda(v)t)dt\right] f_{\mathcal{X}}(v)dv = w. \qquad (24)$$

We see that (24) simplifies to (19) in the case of no censoring, and we get the formula in Theorem 2.1 by substituting $\eta(w)$ by x and w by $s(x)$.

A.2 Proof of Lemma 2.1

For the ease of notation consider the simpler estimator $\tilde{\tilde{s}}(x) = \sum_{i=1}^{J-1}\frac{1}{n}T_i, \quad X_{J-1} < x \leq X_J$, rather than $\tilde{s}(x)$. The proof for $\tilde{s}(x)$ is exactly the same as for $\tilde{\tilde{s}}(x)$ except that the asymptotically negligible extra term $\frac{x-X_{J-1}}{X_J-X_{J-1}}\frac{1}{n}T_J$ requires more writing. We proceed in the same manner as in the proof of Theorem 2.1 by first assuming the covariates to be uniformly distributed on [0,1]. Define J^* by $\frac{J^*-1}{n+1} < x \leq \frac{J^*}{n+1}$. We prove the uniform consistency of $\tilde{\tilde{s}}(x)$ by considering each term in the inequality

$$\begin{aligned} |\tilde{\tilde{s}}(x) - s(x)| &\leq |\tilde{\tilde{s}}(x) - \sum_{i=1}^{J-1}\frac{1}{n}\mathbf{E}(T|X_i)| + |\sum_{i=1}^{J-1}\frac{1}{n}\mathbf{E}(T|X_i) - \sum_{i=1}^{J-1}\frac{1}{n}\mathbf{E}(T|\frac{i}{n+1})| \\ &\quad + |\sum_{i=1}^{J-1}\frac{1}{n}\mathbf{E}(T|\frac{i}{n+1}) - \sum_{i=1}^{J^*-1}\frac{1}{n}\mathbf{E}(T|\frac{i}{n+1})| + |\sum_{i=1}^{J^*-1}\frac{1}{n}\mathbf{E}(T|\frac{i}{n+1}) - s(x)| \\ &= B_1 + B_2 + B_3 + B_4 \end{aligned}$$

$(B_1 \xrightarrow{p} 0)$: By Chebychev and (22)

$$P(|\tilde{\tilde{s}}(x) - \sum_{i=1}^{J-1}\frac{1}{n}\mathbf{E}(T|x_i)| > \epsilon | x_1, \ldots, x_n) \leq \frac{V_{max}}{n\epsilon^2}.$$

Notice that J is solely a function of x and $x_1,\dots,x_n$, and thus deterministic here. Since the right hand side of the inequality is independent of both x and $x_1,\dots,x_n$ it follows that $|\tilde{\tilde{s}}(x) - \sum_{i=1}^{J-1}\frac{1}{n}\mathbf{E}(T|X_i)| \xrightarrow{p} 0$ uniformly in x.

($B_2 \xrightarrow{p} 0$): Let $C \geq \max_{x\in\mathcal{X}} \frac{d\mathbf{E}(T|x)}{dx}$ which by the assumptions in Theorem 2.1 and (21) exist and is finite. Then

$$
\begin{aligned}
P(|\sum_{i=1}^{J-1}\frac{1}{n}\mathbf{E}(T|X_i) - \sum_{i=1}^{J-1}\frac{1}{n}\mathbf{E}(T|\frac{i}{n+1})| > \epsilon) &\stackrel{Markov}{\leq} \frac{\mathbf{E}|\sum_{i=1}^{J-1}\frac{1}{n}[\mathbf{E}(T|X_i) - \mathbf{E}(T|\frac{i}{n+1})]|}{\epsilon} \\
&\leq \frac{\frac{C}{n}\mathbf{E}\sum_{i=1}^{J-1}\mathbf{E}|X_i - \frac{i}{n+1}|}{\epsilon} \\
&\stackrel{Cauchy-Schwarz}{\leq} \frac{\frac{C}{n}\mathbf{E}\sum_{i=1}^{J-1}\sqrt{\mathbf{E}(X_i - \frac{i}{n+1})^2}}{\epsilon} \\
&\leq \frac{C}{2\epsilon\sqrt{n+2}}
\end{aligned}
$$

($B_3 \xrightarrow{p} 0$): It is easily realized that $J-1$ has a binomial distribution with expectation nx and variance $nx(1-x)$. It is also easy to show that $|x - \frac{J^*-1}{n}| \leq \frac{1}{n+1}$. Also recall (22). Then

$$
\begin{aligned}
P(|\sum_{i=1}^{J-1}\frac{1}{n}\mathbf{E}(T|\frac{i}{n+1}) - \sum_{i=1}^{J^*-1}\frac{1}{n}\mathbf{E}(T|\frac{i}{n+1})| > \epsilon) &\leq P(|J-J^*|\frac{E_{max}}{n} > \epsilon) \\
&\stackrel{Chebychev}{\leq} \frac{E_{max}^2\mathbf{E}(J-J^*)^2}{n^2\epsilon^2} \\
&= \frac{E_{max}^2[\mathbf{E}(J-\mathbf{E}J)^2 + (\mathbf{E}J - J^*)^2]}{n^2\epsilon^2} \\
&= \frac{E_{max}^2[nx(1-x) + (nx - (J^*-1))^2]}{n^2\epsilon^2} \\
&\leq \frac{E_{max}^2}{\epsilon^2}(\frac{1}{4n} + \frac{1}{(n+1)^2})
\end{aligned}
$$

($B_4 \to 0$): In the case of uniformly distributed covariates on [0,1], $s(x) = \int_0^x \mathbf{E}(T|v)dv$, and it follows from convergence of Riemann-sums that $\sum_{i=1}^{J^*-1}\frac{1}{n}\mathbf{E}(T|\frac{i}{n+1})$ converges uniformly to $\int_0^x \mathbf{E}(T|v)dv$.

The generalization to covariates drawn from a general domain is done by the same arguments as in the proof of Theorem 2.1, and the proof is completed.

A.3 Proof of Theorem 2.2

Let $N_n(s)$ and $k(n)$ be defined as before. It follows from counting process theory, for example Andersen et al. (1993), that $M_n(s) = N_n(s) - R_n(s)$, where $R_n(s) = \int_0^s \rho_n(u|\mathcal{F}_u^n)du$, is a local square integrable martingale. The general expression for $\rho_n(s|\mathcal{F}_s^n)$ is given in (20).

Introduce the notation $\tau_n(s) = \rho_n(s|\mathcal{F}_s^n)/n$ and $\mathcal{T}_n(s) = R_n(s)/n$. The first part of the proof is to find an estimator of $\tau_n(s)$ and to prove that this estimator is an uniformly consistent estimator of $\tau(s) = \lim_{n\to\infty}\tau_n(s) = \lambda(\eta(s))$.

The fact that $M_n(s)$ is a martingale also implies that

$$M^n(s) = M_n(s)/n = N_n(s)/n - \mathcal{T}_n(s) \tag{25}$$

is a martingale. Following the same reasoning as in the derivation of the Nelson-Aalen estimator in Andersen et al. (1993, chap. 4) it follows from (25) that a natural estimator for $\mathcal{T}_n(s)$ is

$$\hat{\mathcal{T}}_n(s) = \int_0^s \frac{dN_n(u)}{n}$$

and then a kernel estimator for $\tau_n(s)$ is

$$\hat{\tau}_n(s) = \frac{1}{h_s}\int_0^\infty K(\frac{s-u}{h_s})\frac{dN_n(u)}{n} = \frac{1}{nh_s}\sum_{i=1}^{k(n)} K(\frac{s-S_i}{h_s}) \tag{26}$$

By this an estimator of $\tau_n(s)$ is motivated, it only remains to prove its consistency as an estimator of $\tau(s)$. It follows from (25) that

$$\begin{aligned}\hat{\tau}_n(s) &= \frac{1}{h_s}\int_0^\infty K(\frac{s-u}{h_s})dM^n(u) + \frac{1}{h_s}\int_0^\infty K(\frac{s-u}{h_s})\tau_n(u)du \\ &\equiv d_n(s) + \tilde{\tau}_n(s)\end{aligned}$$

By showing

$$|\hat{\tau}_n(s) - \tilde{\tau}_n(s)| \xrightarrow{p} 0 \tag{27}$$

uniformly and

$$|\tilde{\tau}_n(s) - \tau_n(s)| \xrightarrow{p} 0 \tag{28}$$

uniformly, uniform consistency of $\hat{\tau}_n(s)$ follows from the triangle inequality since uniform convergence of $|\tau_n(s) - \tau(s)|$ was proved in Theorem 2.1. For (27), first notice that by results on stochastic integration and the fact that $< M_n >$ is defined as the compensator of M^{n^2} it follows (Andersen et al. 1993) that

$$\begin{aligned}\mathbf{E}d_n^2(s) &= \mathbf{E} < d_n > (s) \\ &= \frac{1}{h_s^2}\int_0^\infty K^2(\frac{s-u}{h_s})\mathbf{E}d < M^n > (u) \\ &= \frac{1}{h_s^2}\int_{s-h_s}^{s+h_s} K^2(\frac{s-u}{h_s})\frac{1}{n}\mathbf{E}\tau_n(u)du \\ &= \frac{1}{nh_s}\int_{-1}^1 K^2(v)\mathbf{E}\tau_n(s-h_s v)dv \\ &\leq \frac{M}{nh_n}\int_{-1}^1 K^2(v)dv\end{aligned}$$

Then Markov's inequality gives

$$\begin{aligned}P(|\hat{\tau}_n(s) - \tilde{\tau}_n(s)| > \epsilon) = P(|d_n(s)| > \epsilon) &\leq \frac{\mathbf{E}d_n(s)^2}{\epsilon^2} \\ &\leq \frac{M}{\epsilon^2 nh_n}\int_{-1}^1 K^2(v)dv \\ &\to 0\end{aligned}$$

For (28) the convergence follows from

$$\begin{aligned}
|\tilde{\tau}_n(s) - \tau_n(s)| &= |\int_{-1}^{1} K(v)(\tau_n(s - h_s v) - \tau_n(s))dv| \\
&\leq \int_{-1}^{1} |K(v)||\tau_n(s - h_s v) - \tau_n(s)|dv \\
&\xrightarrow{p} 0
\end{aligned}$$

uniformly because

$$|\tau_n(s - h_s v) - \tau_n(s)| \leq |\tau_n(s - h_s v) - \tau(s - h_s v)| + |\tau(s) - \tau_n(s)| + |\tau(s - h_s v) - \tau(s)|$$

where the two first terms converge uniformly to zero in probability by Theorem 2.1 and where the last term converges numerically uniformly to 0 by uniform continuity of $\lambda(x)$.

This completes the proof that $\hat{\tau}_n(s)$ given in (26) is a uniformly consistent estimator of $\tau(s)$. It now only remains to prove that replacing s by $\tilde{s}(x)$ in (26) yields a consistent estimator of $\lambda(x)$. By the triangle inequality

$$|\hat{\tau}_n(\tilde{s}(x)) - \tau(s(x))| \leq |\hat{\tau}_n(\tilde{s}(x)) - \tau(\tilde{s}(x))| + |\tau(\tilde{s}(x)) - \tau(s(x))|$$

where the second part converges uniformly to 0 in probability by Lemma 2.1 and the uniform continuity of $\tau(s)$. This completes the proof that $\hat{\lambda}(x) = \hat{\tau}_n(\tilde{s}(x))$ is a uniformly consistent estimator of $\lambda(x)$.