The covariate order method for nonparametric exponential regression and some applications in other lifetime models.

Jan Terje Kvaløy and Bo Henry Lindqvist

Department of Mathematics and Natural Science
Stavanger University College

Department of Mathematical Sciences
Norwegian University of Science and Technology

April 1, 2003

Abstract

A new method for nonparametric censored exponential regression, called the covariate order method, is presented. It is shown that the method leads to a consistent estimator of the hazard rate as a function of the covariate. Moreover, interesting applications to more general cases of lifetime regression are presented. Possible applications include the construction of tests for covariate effect and estimation and residual plots in Cox-regression models. The key is here to perform suitable transformations to exponentiality before applying the covariate order method.

1 INTRODUCTION

A first analysis of lifetimes of a set of items, for example identical mechanical or electronic components, is often based on the assumption that the lifetimes are independent and identically exponentially distributed. This implies assuming a common and constant hazard rate λ for each item. A less restrictive and often more realistic assumption is to assume exponential lifetimes with hazard rate λ varying from item to item, for example due to differences in operating or environmental conditions. Such conditions may often be quantified by observable covariates, in which case one assumes that $\lambda = \lambda(\mathbf{x})$, where $\mathbf{x}$ is an m-dimensional vector of covariates. Exponential regression means estimating the hazard rate $\lambda(\mathbf{x})$ from observed lifetimes and covariates. If some of the observations are censored we call it censored exponential regression.

The literature contains a number of estimation methods which can be used for censored exponential regression. Parametric estimation is most conveniently done by fitting a generalized linear model (McCullagh and Nelder, 1989). Various approaches which can be used for nonparametric estimation of $\lambda(\mathbf{x})$ have furthermore been suggested. For example, Hastie and Tibshirani (1990b) consider estimation in generalized additive models as a natural nonparametric extension of generalized linear models. Other relevant references are Hastie and Tibshirani (1990a) and Gray (1992) who consider spline based methods, and Tibshirani and Hastie (1987), Staniswalis (1989), Gentleman and Crowley (1991) and Fan, Gijbels and King (1997) who consider local likelihood methods for nonparametric estimation in Cox-regression models.

In this paper we present a new nonparametric method for exponential regression, called the covariate order method. For the case of a single covariate the method can be briefly described

as follows: First, arrange the observations (both non-censored and censored) in increasing order of the covariate. Then plot the observed times successively as "interevent times" on an artificial time axis. Let the true events in the resulting point process be the ones which are endpoints corresponding to an uncensored observation. A consistent estimator of the hazard rate function $\lambda(x)$ can then be found by first estimating the intensity of the process defined above, and then combining this function with an estimated relationship between points on the artificial time axis and the covariate axis. A kernel-estimator will be used for estimating the intensity of the process. Generalizations to more than one covariate are possible, for example by assuming a generalized additive model.

It should be stressed that the covariate order method in its basic form rests heavily on the assumption of exponentially distributed lifetimes. In fact, the estimate of $\lambda(\mathbf{x})$ would have no meaning if the same procedure was tried on non-exponential lifetimes. However, often we are able to transform our data to follow, at least approximately, an exponential regression model. In these cases we can use the covariate order method on the transformed data, and this turns out to be a useful approach in applications. For example, Kvaløy (2002) used the covariate order method to suggest tests for covariate effect in general censored regression models (see Section 2.5 of the present paper), while Kvaløy and Lindqvist (2003) used the covariate order method in nonparametric estimation of covariate functions in Cox-regression (see Section 3.2).

The main purpose of the present paper is to give a formal presentation of the covariate order method and its practical implementation (Sections 2.1-2.4), and in addition to give a rigorous proof of consistency of the method in the single covariate case (Appendix). In order to illustrate the direct method we give an example with exponential data in Section 2.6. Sections 3.1 and 3.3 illustrate the use of the covariate order method to transformed data. More precisely it is shown how to make illustrative residual plots based on Cox-Snell residuals in Cox-regression models, and how the method can be used to suggest possible transformations of covariates.

2 THE COVARIATE ORDER METHOD FOR EXPONENTIAL REGRESSION

The basic formulation of the problem 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 $(T, \delta, \mathbf{X})$, 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 $\mathbf{X}$. Let Z be called the *lifetime*, C the *censoring time* and T the *observation time*. This terminology is introduced only for convenience, Z can be any kind of exponentially distributed variables.

The domain of the covariate vector $\mathbf{X}$ is a subset $\mathcal{X}$ of $\mathbb{R}^m$, and $\mathbf{X}$ is assumed to be distributed according to some density function $f_{\mathbf{X}}(\mathbf{x})$. The corresponding cumulative distribution function is denoted $F_{\mathbf{X}}(\mathbf{x})$. The covariates are assumed to remain constant over time, and $\lambda(\mathbf{x})$ is assumed to be *continuous* on $\mathcal{X}$. The method is first described for the case of a single covariate, in other words for $m = 1$. Extensions to higher dimensions are discussed in Section 2.3.

2.1 Method description and main theoretical results

The method proceeds conditionally on $X_1, \ldots, X_n$ and starts by first arranging 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 the number of observations, n. 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 intervals corresponding to *uncensored* observations be considered as events, occurring at times denoted $S_1, \ldots, S_r$ where $r = \sum_{j=1}^{n} \delta_j$. This is visualized in Figure 1, for an example where the ordered observations are $(T_1, \delta_1 = 1), (T_2, \delta_2 = 0), (T_3, \delta_3 = 1), \ldots, (T_{n-1}, \delta_{n-1} = 0), (T_n, \delta_n = 1)$.

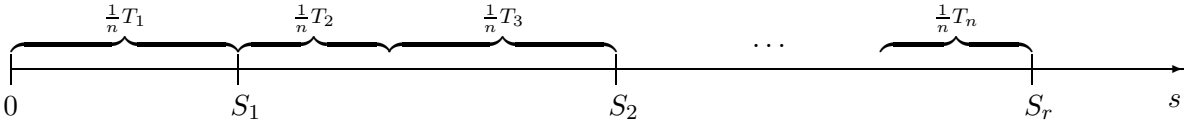

Figure 1: Construction of artificial process.

More precisely, $S_i = \sum_{j=1}^{k(i)} T_j/n$ where $k(i) = \min\{s | \sum_{j=1}^{s} \delta_j = i\}$. Now the conditional intensity of the process $S_1, \ldots, S_r$ at a point w on the s-axis, given the complete history of the T_j up to s, 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 $S_1, \ldots, S_r$, yielding the estimator $\hat{\rho}_n(w)$, 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 $X_1, \ldots, X_n$ on the "covariate-axis" and the process $S_1, \ldots, S_r$ on the "s-axis". A possible way of estimating such a relationship is to use the step-function

$$\tilde{s}(x) = \frac{1}{n} \sum_{i=1}^{j} T_i, \quad X_j \leq x < X_{j+1}, \tag{1}$$

see Figure 2 for an illustration, and then define $\hat{\lambda}(x) = \hat{\rho}(\tilde{s}(x))$.

The motivating idea of the method is that if $\lambda(x) = \lambda$ is constant, then the process $S_1, \ldots, S_r$ is a homogeneous Poisson process. (The test presented in Section 2.5 is in fact based on this observation.) Thus if $\lambda(x)$ is reasonably smooth and not varying too much, then the process $S_1, \ldots, S_r$ could be imagined to be nearly a nonhomogeneous Poisson process for which the intensity can be estimated by for instance kernel density estimation based on the points $S_1, \ldots, S_r$. 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 its consistency. All proofs are given in the Appendix.

Let $\mathcal{F}_s^n$ be the history of the process $S_1, \ldots, S_r$ in the interval $[0, s)$. This history is formally defined as the sub-σ-algebra $\mathcal{F}_s^n = \sigma\{X_1, \ldots, X_n\} \cup \sigma\{S_j : S_j \leq s\}$ 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 $S_1, \ldots, S_r$ at the point s (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. This theorem states that the scaled conditional intensity of the process $S_1, \ldots, S_r$ converges in probability to a deterministic function of $\lambda(\cdot)$, and 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 at the beginning of the section. 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*

$$s(x) = E(TI(X \leq x)).$$

The function $s(x)$ is called the correspondence function. Note that for the special case of no censoring, $s(x)$ can be written $s(x) = \int_{-\infty}^{x} f_{\mathcal{X}}(v)/\lambda(v)dv$.

The fact that the scaled conditional intensity of the process $S_1, \ldots, S_r$ 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*

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

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

2.2 Smoothing details

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 a smooth correspondence function estimate $\hat{s}(x)$ from the points $(X_1, T_1/n), \ldots, (X_n, \sum_{i=1}^{n} T_i/n)$. 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

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

where $S = \sum_{j=1}^{n} T_j/n$.

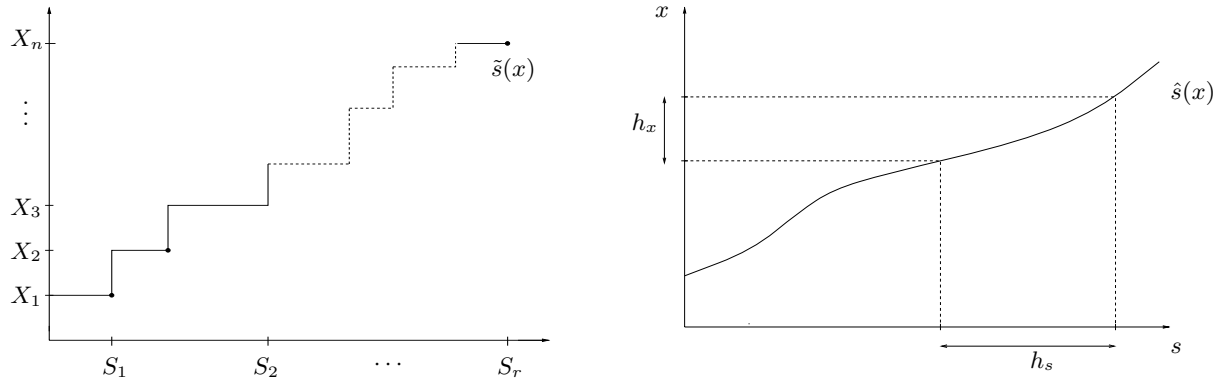

Figure 2: The left shows an example of what the estimated correspondence function $\tilde{s}(x)$ (1) might look like. The right plot illustrates a smoothed correspondence function estimate $\hat{s}(x)$ and the relationship between the smoothing parameter on the covariate axis and the s-axis.

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 the right plot in Figure 2 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 commonly used in nonparametric regression methods. If a constant h_x is used, then (3) becomes

$$\hat{\lambda}(x) = \frac{1}{nh_s(\hat{s}(x))} \sum_{i=1}^{r} \left[K(\frac{\hat{s}(x) - S_i}{h_s(\hat{s}(x))}) + K(\frac{\hat{s}(x) + S_i}{h_s(\hat{s}(x))}) + K(\frac{\hat{s}(x) + S_i - 2S}{h_s(\hat{s}(x))}) \right] \tag{4}$$

where $h_s(\hat{s}(x)) = \hat{s}(x + h_x/2) - \hat{s}(x - h_x/2)$. For instance likelihood cross-validation can be used as criterion for choosing the "best" value of the smoothing parameter.

2.3 Several covariates

The covariate order method is not directly generalizable to higher dimensions, mainly because R^m is not linearly ordered for $m > 1$. Thus instead we suggest 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 a generalized additive model

$$\lambda(\mathbf{x}) = \exp(\alpha + g_1(x_1) +, \ldots, + g_m(x_m)), \tag{5}$$

where $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X} \subseteq \mathbb{R}^m$, and where $g_1(\cdot), \ldots, g_m(\cdot)$ are unspecified smooth functions. These functions can be estimated by the covariate order method using an iterative backfitting algorithm. The key point is that if Z is exponentially distributed with parameter $\exp(\alpha + g_1(x_1) + \ldots + g_m(x_m))$, then $Z \exp(\alpha + g_1(x_1) + \ldots + g_{j-1}(x_{j-1}) + g_{j+1}(x_{j+1}) + \ldots + g_m(x_m))$ will be exponentially distributed with parameter $\exp(g_j(x_j))$. Also note that it is possible to let some of the g-functions be parametric, for instance for discrete covariates.

2.4 Comments on the basic method

Notice the flexibility of the covariate order method. 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, and different smoothers can be used to estimate $s(x)$.

Further, the method is numerically very robust. For the $m = 1$ case numerical problems can not occur. For $m > 1$ numerical problems have never been encountered either, and convergence is fast. A useful plot related to the method is to plot the points $(X_1, T_1/n), \ldots, (X_n, \sum_{i=1}^n T_i/n)$ which corresponds to the jumps of the estimated correspondence function $\tilde{s}(x)$. This plot gives a rough unsmoothed display of the data, and is also a convenient way of simultaneously identifying outliers in observation times and covariate values.

2.5 Testing for covariate effect

Recall from Section 2.1 that if there is no covariate effect, that is $\lambda(x) \equiv \lambda$, then the process $S_1, \ldots, S_r$ is a homogeneous Poisson process (HPP). This observation suggests that in principle any statistical test for the null hypothesis of an HPP versus various non-HPP alternatives can be applied to test for covariate effect in exponential regression models. Moreover, such an approach can be extended to non-exponentially distributed lifetimes by transforming the observation times to approximately exponentially distributed data.

A detailed account of this approach for testing for covariate effect in lifetime data is given by Kvaløy (2002), who presents a number of different tests constructed based on the covariate order method. The recommendation is to use an Anderson-Darling type test which turns out to have very good power properties against both monotonic and non-monotonic alternatives to constant $\lambda(x)$. As before, let $S = \sum_{i=1}^n T_i/n$ and define

$$\hat{r} = \begin{cases} r & \text{if } S_r < S \\ r-1 & \text{if } S_r = S \end{cases}$$

Then the test statistic for the Anderson-Darling test is

$$AD = -\frac{1}{\hat{r}} \left[\sum_{i=1}^{\hat{r}} (2i-1)(\ln \frac{S_i}{S} + \ln(1 - \frac{S_{\hat{r}+1-i}}{S})) \right] - \hat{r} \tag{6}$$

The asymptotic null distribution of (6) was derived by Anderson and Darling (1952). For example, given a 5% significance level, the null hypothesis of no covariate effect is rejected if $AD \geq 2.492$. For small sample sizes the level properties of the test can be improved by use of resampling techniques (Kvaløy, 2002).

2.6 Example: Cardiac arrest versus air temperature

We give an example of direct application of the covariate order method to data for times of out-of-hospital cardiac arrests reported to a Norwegian hospital over a 5 years period. A previous analysis of the same data by Skogvoll and Lindqvist (1999) concluded that the occurrence of cardiac arrest is reasonably well modeled by an HPP, though with some minor deviations from homogeneity. In the present example the relationship between outdoor air temperature and the occurrence of cardiac arrest is investigated. This is done by regarding inter-event times to be independent and exponentially distributed with a hazard rate $\lambda(x)$ depending on the temperature x. Assuming temperature to be varying relatively slowly this seems to be a reasonable model, although an NHPP model may seem more direct. The main finding of a more sophisticated analysis of the data (Kvaløy, 1999) coincides with the result of the simple analysis presented here. The average temperature on the day of a cardiac arrest is used as covariate for the next period between cardiac arrests. A total of 449 cardiac arrests where reported during the five years period.

Testing the significance of the covariate effect of temperature by using the Anderson-Darling test for covariate effect (6) yielded a p-value of 0.002. Several plots of the estimated

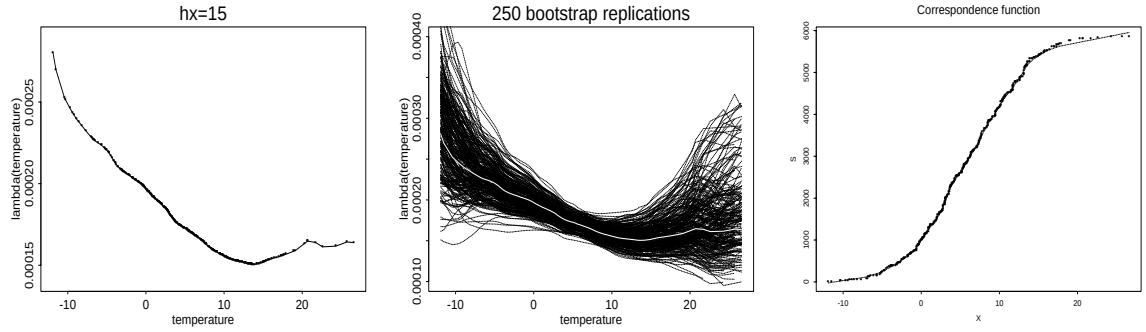

Figure 3: Analysis of cardiac arrest occurrence versus air temperature. The left plot shows the estimated hazard rate function obtained using a constant smoothing parameter on the x-axis, with the location of the observations along the curve displayed by the dots. The middle plot shows 250 bootstrap curves obtained by resampling observations (original estimate shown as white curve). The right plot shows the estimated correspondence function $\tilde{s}(x)$.

model are displayed in Figure 3. The estimated hazard rate function clearly indicates a decreasing hazard for increasing temperature. The smoothing parameter $h_x = 15$ was chosen by a likelihood cross validation criterion. The bootstrap curves indicate little variability in the estimated hazard rate in the middle temperature range where most of the observations are located, while there is large variability at the boundaries as expected. The correspondence function plot reflects the temperature distribution and the hazard function, and also displays the small number of observations near the boundaries. No particular outliers are identified by the plot.

3 APPLICATIONS IN COX REGRESSION

3.1 Model checking and model fitting in classical Cox-regression.

An interesting application of the covariate order method is in model checking for the classical Cox (1972) proportional hazards model, $\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(\boldsymbol{\beta}\mathbf{x})$, where $\alpha_0(t)$ is a baseline hazard function, $\mathbf{x}$ is a covariate vector and $\boldsymbol{\beta}$ is a vector of regression coefficients. Under the model assumptions it is well known that $A_0(T_i)\exp(\boldsymbol{\beta}\mathbf{X}_i)$, $i = 1, \ldots, n$, where $A_0(t) = \int_0^t \alpha_0(u)du$, is a censored sample from the exponential distribution with parameter 1. The Cox-Snell residuals (Cox and Snell, 1968) are defined as $r_i = \hat{A}_0(T_i)\exp(\hat{\boldsymbol{\beta}}\mathbf{X}_i)$, $i = 1, \ldots, n$, so if the model is correct, then the sample $(r_1, \delta_1), \ldots, (r_n, \delta_n)$ is expected to behave approximately as a censored sample from the exponential distribution with parameter 1. The Cox-Snell residuals are mainly used to assess an overall fit by checking whether $(r_1, \delta_1), \ldots, (r_n, \delta_n)$ is compatible with a (censored) sample from an exponential distribution. However, we shall see that the covariate order method allows one in an easy way to do further analyses of the Cox-Snell residuals, for example to check for (unexpected) relationships between residuals and variables like individual covariates, risk score and observation number.

For instance, for each single covariate X_k, say, one may fit an exponential regression model to the data $(r_1, \delta_1, X_{1k}), \ldots, (r_n, \delta_n, X_{nk})$, where X_{ik} is the kth covariate for the ith observation unit. The covariate order method gives an estimated hazard rate as a function of X_k which, if the model is correct, is expected to be approximately constant at 1. Deviations from a constant hazard rate can be tested more formally by the Anderson-Darling test described in Section 2.5. This approach is an alternative to the common plotting of residuals against covariates etc. which is routinely done in ordinary linear regression models. A similar plotting of Cox-Snell residuals is of course possible and is sometimes done, but may be misleading due to censored observations.

A related application is to make plots of log hazard rates against covariates not included in the model. Such plots can reveal whether these covariates should be included in the model, and in this case indicate the appropriate functional form of the covariate. This is a simple and intuitive alternative to the plotting of martingale residuals (Therneau, Grambsch and Fleming, 1990) commonly used for this purpose. A somewhat related approach, but using nonparametric Poisson regression instead of exponential regression, was used by Grambsch, Therneau and Fleming (1995), see also Therneau and Grambsch (2000, chap. 5).

An alternative approach for suggesting functional form of the covariates is to fit a complete nonparametric model including all the covariates. This is discussed below.

3.2 Nonparametric Cox-regression

The covariate order method for exponential regression can fairly easily be extended to estimation of $g(\mathbf{x})$ in the generalized Cox-model $\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(g(\mathbf{x}))$ where $g(\mathbf{x})$ in principle is any function of the covariates. The basic idea is that for an uncensored observation Z, $A_0(Z)\exp(g(\mathbf{X}))$ is exponentially distributed with parameter 1, so $A_0(Z)$ is exponentially distributed with parameter $\exp(g(\mathbf{X}))$. This motivates the following algorithm:

1. Find an initial estimate $\hat{A}_0(t)$ of the cumulative baseline hazard function $A_0(t)$.

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

3. Estimate $g(\mathbf{x})$ from $(\hat{A}_0(T_1), \delta_1, \mathbf{X}_1), \ldots, (\hat{A}_0(T_n), \delta_n, \mathbf{X}_n)$ using the covariate order method for exponential regression.

4. Find a new estimate of $A_0(t)$ from $(T_1, \delta_1, \hat{g}(\mathbf{X}_1)), \ldots, (T_n, \delta_n, \hat{g}(\mathbf{X}_n))$.

5. Repeat 2-4 until convergence

The integrated baseline hazard function $A_0(t)$ can for instance be estimated by the Breslow estimator (Breslow, 1972). Numerical convergence of the algorithm is very fast. A similar iterative algorithm but with a different approach for estimating $g(\mathbf{x})$ was used by Gentleman and Crowley (1991). Further details on this application of the covariate order method are given in Kvaløy and Lindqvist (2003).

3.3 Example: PBC data

We illustrate the use of covariate ordering in the classical Cox-model by considering model fitting and model checking for the PBC data from the Mayo Clinic. PBC (primary biliary cirrhosis) is a fatal chronic liver disease, and out of the 418 patients followed in the study, 161 died before study closure. A listing of the data can be found in Fleming and Harrington (1991). The final model proposed by Fleming and Harrington (1991) includes the five covariates age, edema, log(bilirubin), log(protime) and log(albumin).

For a demonstration of residual plotting we will look closer at the covariate bilirubin. First

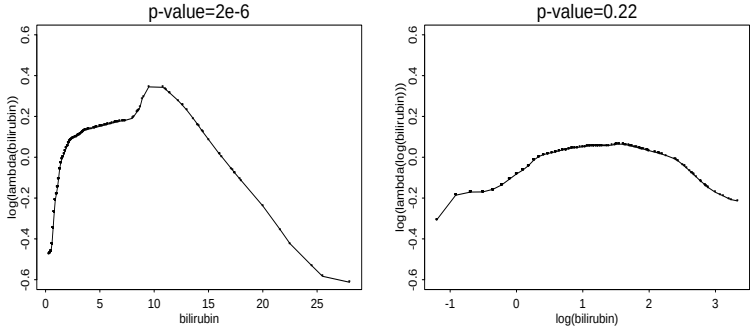

Figure 4: Residual analysis of PBC data. Plot of the log of the estimated hazard rate of the Cox-Snell residuals against bilirubin in a model using bilirubin on its original scale (left) and the same plot against log(bilirubin) in a model using log(bilirubin) (right).

we fitted a Cox-model including the five covariates mentioned above, but where the covariate bilirubin was included without making the log transformation. The left plot in Figure 4 shows, for this model, the log of the estimated hazard rate of the Cox-Snell residuals against bilirubin. The p-value reported in the plot was calculated using the Anderson-Darling test presented in Section 2.5. The low p-value certainly shows a significant deviation from constancy, which is also clear from the plot. Thus the covariate is not well modeled. The right plot shows the log of the estimated hazard rate of the Cox-Snell residuals against log(bilirubin) in a model where the bilirubin covariate was added as log(bilirubin). We see that the bilirubin covariate now seems to be much better modeled.

As explained in Section 3.1, one may use similar plots to suggest the functional form of covariates before they are entered into the model. Figure 5 displays plots of the log of the

estimated hazard rate of the Cox-Snell residuals from an empty model versus, respectively, age, bilirubin and log(bilirubin). Note that in this case the Cox-Snell residuals are simply $\hat{A}_0(T_i)$, where $\hat{A}_0(\cdot)$ is the Nelson-Aalen estimator of the cumulative hazard in the empty model. The (approximate) straight line seen for the plot against age in Figure 5 suggests

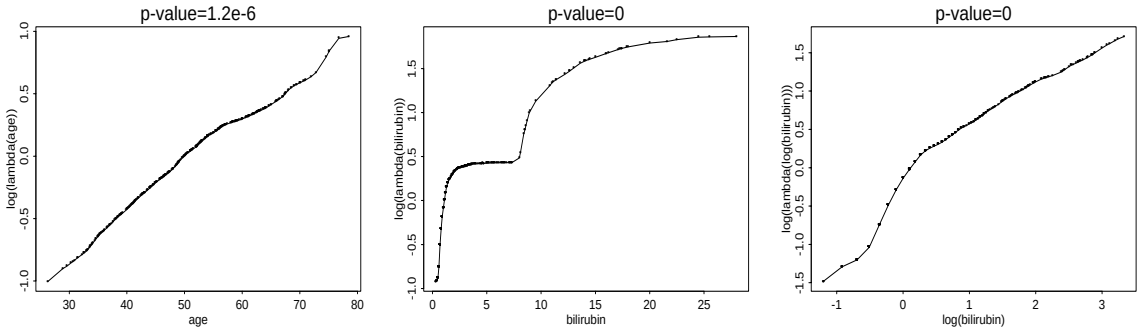

Figure 5: Functional form analysis in PBC data. Plots of the log of the estimated hazard rate of the Cox-Snell residuals from an empty model versus respectively age, bilirubin and log(bilirubin). The location of the observations along the curves are displayed by the dots.

that age can be added directly in the Cox-modeled, while the non-linear behavior of the plot against bilirubin suggests that a transformation should be made for this covariate. The plot against log(bilirubin) indicates that this covariate is much better modeled if it is transformed to log-scale.

4 CONCLUSIONS

We have presented a new method for nonparametric censored exponential regression, and shown some of its applications. While we have given emphasis to applications in Cox-regression, one may think of similar applications in any model with (approximately) exponentially distributed residuals, or in other cases where data can be transformed to (approximate) exponentiality.

The covariate order method has the advantage over some of its competitors that it is simple, flexible, intuitive and numerically very robust. Moreover, simulations (not reported here) have shown the performance in finite samples to be very similar to that of standard local linear likelihood methods.

Acknowledgements

We would like to thank Ørnulf Borgan for helpful comments to the proofs and Eirik Skogvoll for providing the cardiac arrest data. Jan Terje Kvaløy was funded by a PhD grant from the Research Council of Norway during parts of the work on this paper.

References

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

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.

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

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

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals, *Journal of the Royal Statistical Society, Series B* **30**: 248–275.

Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression, *The Annals of Statistics* **25**: 1661–1690.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.*, Wiley, New York.

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.

Grambsch, P. M., Therneau, T. M. and Fleming, T. R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models, *Biometrics* **51**: 1469–1482.

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

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

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

Karr, A. F. (1993). *Probability*, Springer-Verlag, New York.

Kvaløy, J. T. (1999). *Statistical Methods for Detecting and Modeling General Patterns and Relationships in Lifetime Data*, PhD thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Kvaløy, J. T. (2002). Covariate order tests for covariate effect, *Lifetime Data Analysis* **8**: 35–52.

Kvaløy, J. T. and Lindqvist, B. H. (2003). Estimation and inference in nonparametric Cox-models: Time transformation methods, *Computational Statistics* **18**: To appear.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman & Hall, London.

Silverman, B. W. (1986). *Density Estimation*, Chapman & Hall, London.

Skogvoll, E. and Lindqvist, B. H. (1999). Modeling the occurrence of cardiac arrest as a Poisson process, *Annals of Emergency Medicine* **33**: 409–417.

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

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models, *Biometrika* **77**: 147–160.

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

Define the process $S_1^*, \ldots, S_n^*$ by $S_j^* = \sum_{i=1}^{j} \frac{1}{n} T_i$. Let $N_n^*(s) = \sum_{i=1}^{n} I(S_i^* \leq s)$ be the counting process counting events in this process. Further, let $\mathcal{F}_s^{n^*} = \sigma\{X_1, \ldots, X_n\} \cup \sigma\{(T_j, \delta_j) : \sum_{i=1}^{j} T_i/n \leq s\}$ for $s \geq 0$. The intensity of the process $S_1, \ldots, S_r$ conditional on the history $\mathcal{F}_s^{n^*}$ is $\rho_n(s|\mathcal{F}_s^{n^*}) = n\lambda(X_{N_n^*(s)+1})$. Since $\mathcal{F}_s^n \subseteq \mathcal{F}_s^{n^*}$ it follows from the innovation theorem (Andersen et al. 1993, p. 80), that

$$\rho_n(s|\mathcal{F}_s^n)/n = \mathrm{E}[\lambda(X_{N_n^*(s)+1})|\mathcal{F}_s^n]. \tag{7}$$

Assume that it can be proved that $X_{N_n^*(s)+1} \xrightarrow{p} \eta(s)$ uniformly. Then using Markov's inequality we get

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

It now easily follows by the boundedness of $\lambda(x)$ and the assumed uniform convergence of $X_{N_n^*(s)+1}$ 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 to $\eta(s)$. Since $T = \min(Z, C)$, given the covariate $X = x$, we have that

$$f_T(t|x) = f_C(t|x)\exp(-\lambda(x)t) + \lambda(x)\exp(-\lambda(x)t)(1 - F_C(t|x)). \tag{8}$$

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 (8) that there exist numbers E_{min}, E_{max} and V_{max} such that

$$\begin{aligned} 0 < E_{min} \leq \quad \mathrm{E}(T|x) \quad &\leq E_{max} < \infty, \quad \text{for all } x, \\ 0 < \quad \mathrm{Var}(T|x) \quad &\leq V_{max} < \infty, \quad \text{for all } x. \end{aligned} \tag{9}$$

We proceed by first assuming that X is 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

$$\begin{aligned} I: \quad & S_{I-1}^* \leq w < S_{I^*} \\ I_0: \quad & \textstyle\sum_{i=1}^{I_0-1} \frac{1}{n}\mathrm{E}(T|X_i) \leq w < \sum_{i=1}^{I_0} \frac{1}{n}\mathrm{E}(T|X_i) \\ I_1: \quad & \textstyle\sum_{i=1}^{I_1-1} \frac{1}{n}\mathrm{E}(T|\frac{i}{n+1}) \leq w < \sum_{i=1}^{I_1} \frac{1}{n}\mathrm{E}(T|\frac{i}{n+1}) \\ \eta(w): \quad & \textstyle\int_0^{\eta(w)} \mathrm{E}(T|v)dv = w \end{aligned} \tag{10}$$

In particular $I = N_n^*(w) + 1$. By the triangle inequality

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

What remains is to prove that each of A_1, A_2 and $A_3 \xrightarrow{p} 0$ and $A_4 \to 0$ 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 $F \equiv F_X$, then $\sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0$. Since $F(x) = x$; $0 \leq x \leq 1$, 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

$$\begin{aligned} &P(I \geq I_0 + d | X_1 = x_1, \ldots, X_n = x_n) \\ &= P\left(S^*_{I_0+d-1} \leq w | x_1, \ldots, x_n\right) \leq P\left(S^*_{I_0+d-1} \leq \sum_{i=1}^{I_0} \frac{1}{n}\mathrm{E}(T|x_i) | x_1, \ldots, x_n\right) \\ &\leq P\left(|S^*_{I_0+d-1} - \sum_{i=1}^{I_0+d-1} \frac{1}{n}\mathrm{E}(T|x_i)| \geq \sum_{i=I_0+1}^{I_0+d-1} \frac{1}{n}\mathrm{E}(T|x_i) | x_1, \ldots, x_n\right) \\ &\overset{Chebychev}{\leq} \frac{\sum_{i=1}^{I_0+d-1} \frac{1}{n^2}\mathrm{Var}(T|x_i)}{\left(\sum_{i=I_0+1}^{I_0+d-1} \frac{1}{n}\mathrm{E}(T|x_i)\right)^2} \leq \frac{V_{max}(I_0 + d - 1)/n^2}{(\frac{d-1}{n}E_{min})^2} \leq \frac{n}{(d-1)^2}\frac{V_{max}}{E^2_{min}} \end{aligned}$$

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 $\lambda'(x) \leq D$ and $f_C(t|x)$ by assumption also has finite first derivative, this implies that there exist a B such that $|\mathrm{E}(T|x_1) - \mathrm{E}(T|x_2)| \leq B|x_1 - x_2|$. Also recall that if X_i is the ith order statistic of n independent identically uniformly distributed variables on [0,1], then $\mathrm{Var}(X_i) = \frac{i(n-i+1)}{(n+1)^2(n+2)} \leq \frac{1}{4(n+2)}$. Thus for an integer d,

$$\begin{aligned} P(I_0 > I_1 + d) &= P\left(\sum_{i=1}^{I_1+d} \frac{1}{n}\mathrm{E}(T|X_i) < w) \leq P(\sum_{1}^{I_1+d} \frac{1}{n}\mathrm{E}(T|X_i) < \sum_{1}^{I_1} \frac{1}{n}\mathrm{E}(T|\frac{i}{n+1})\right) \\ &\leq P\left(|\sum_{1}^{I_1+d}(\frac{1}{n}\mathrm{E}(T|X_i) - \frac{1}{n}\mathrm{E}(T|\frac{i}{n+1}))| > \sum_{I_1+1}^{I_1+d} \frac{1}{n}\mathrm{E}(T|\frac{i}{n+1})\right) \\ &\overset{Markov}{\leq} \frac{\mathrm{E}|\sum_1^{I_1+d}(\frac{1}{n}\mathrm{E}(T|X_i) - \frac{1}{n}\mathrm{E}(T|\frac{i}{n+1}))|}{\sum_{I_1+1}^{I_1+d} \frac{1}{n}\mathrm{E}(T|\frac{i}{n+1})} \leq \frac{\frac{B}{n}\sum_1^{I_1+d}\mathrm{E}|X_i - \frac{i}{n+1}|}{\frac{d}{n}E_{min}} \end{aligned}$$

$$\overset{C.-S.}{\leq} \quad \frac{B\sum_1^{I_1+d}\sqrt{\mathrm{E}(X_i-\frac{i}{n+1})^2}}{dE_{min}} \quad = \quad \frac{Bn}{2dE_{min}\sqrt{n+2}}$$

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

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

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 $|\sum_{i=1}^{I_1}\frac{1}{n}\mathrm{E}(T|\frac{i}{n+1}) - w| \leq \frac{1}{n}\mathrm{E}(T|\frac{I_1}{n+1}) \leq \frac{1}{n}E_{max}$ which implies that $\sum_{i=1}^{I_1}\frac{1}{n}\mathrm{E}(T|\frac{i}{n+1}) \to w = \int_0^{\eta(w)}\mathrm{E}(T|v)dv$ uniformly. Note that $\eta(w)$ is uniquely defined since $\mathrm{E}(T|v) > 0$ 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 general continuous distribution $F_X(\cdot)$, let $U_i = F_X(X_i)$ be transformed covariates which are now independent and identically uniformly distributed on [0,1]. Further let $\mathrm{E}^\star(T|u) = \mathrm{E}(T|F_X^{-1}(u))$. Then (10) gives $\int_0^{\eta^\star(w)}\mathrm{E}^\star(T|u)du = w$ which by substituting $u = F_X(x)$ and letting $\eta(w) = F_X^{-1}(\eta^\star(w))$ can be written

$$\int_{F_X^{-1}(0)}^{\eta(w)}\mathrm{E}(T|x)f_X(x)dx = w \tag{11}$$

Replacing $\eta(w)$ with x and w with $s(x)$ we get

$$\begin{aligned} s(x) &= \int_{F_X^{-1}(0)}^{x}\mathrm{E}(T|v)f_X(v)dv = \int_{-\infty}^{\infty}I(v \leq x)\mathrm{E}(T|v)f_X(v)dv \\ &= \mathrm{E}(I(X \leq x)\mathrm{E}(T|X)) = \mathrm{E}(\mathrm{E}(TI(X \leq x)|X)) = \mathrm{E}(TI(X \leq x)). \end{aligned}$$

A.2 Proof of Lemma 2.1

We can write

$$\tilde{s}(x) = \frac{1}{n}\sum_{i=1}^{n}T_iI(X_i \leq x)$$

Noting that $s(x) = \mathrm{E}(\tilde{s}(x))$ we have by Chebyshev's inequality, for each fixed x and any $\epsilon > 0$,

$$P(|\tilde{s}(x) - s(x)| > \epsilon) \leq \frac{\mathrm{Var}(TI(X \leq x))}{n\epsilon^2} \leq \frac{\mathrm{E}(T^2)}{n\epsilon^2} \leq \frac{\mathrm{E}(Z^2)}{n\epsilon^2}$$

which tends to 0 as $n \to \infty$ since $\mathrm{E}(Z^2) < \infty$. In fact, we have $\mathrm{E}(Z^2) = \mathrm{E}[\mathrm{E}(Z^2|X)] = \mathrm{E}[2/\lambda(X)^2] \leq 2/a^2$. This proves the result.

A.3 Proof of Theorem 2.2

Let $N_n(s)$ and m 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 (7). 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 a 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

$$M^n(s) = M_n(s)/n = N_n(s)/n - \mathcal{T}_n(s) \tag{12}$$

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 (12) that a natural estimator for $\mathcal{T}_n(s)$ is $\hat{\mathcal{T}}_n(s) = \int_0^s dN_n(u)/n$ and then a kernel estimator for $\tau_n(s)$ is

$$\hat{\tau}_n(s) = \frac{1}{h_s}\int_0^\infty K(\frac{s-u}{h_s})\frac{dN_n(u)}{n} = \frac{1}{nh_s}\sum_{i=1}^r K(\frac{s-S_i}{h_s}) \tag{13}$$

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 (12) that

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

By showing

$$|\hat{\tau}_n(s) - \tilde{\tau}_n(s)| \xrightarrow{p} 0 \tag{14}$$

uniformly and

$$|\tilde{\tau}_n(s) - \tau_n(s)| \xrightarrow{p} 0 \tag{15}$$

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 (14), 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, chap. 4) that

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

Then Markov's inequality gives

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

For (15) the convergence follows from

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

uniformly because

$$|\tau_n(s-h_sv) - \tau_n(s)| \leq |\tau_n(s-h_sv) - \tau(s-h_sv)| + |\tau(s) - \tau_n(s)| + |\tau(s-h_sv) - \tau(s)|$$

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 (13) is a uniformly consistent estimator of $\tau(s)$. It now only remains to prove that replacing s by $\tilde{s}(x)$ in (13) 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 $\lambda(x)$.